Re-evaluating the Word Token for Bilingual Speech Processing: The Case for Intonation Units

Rebecca Pattichis[1], Dora LaCasse[2], and Rena Torres Cacoullos[3]

[1]Independent Researcher
rebeccapattichis2000@gmail.com
[2]Department of World Languages and Cultures, University of Montana
dora.lacasse@mso.umt.edu
[3]Department of Spanish, Italian and Portuguese, The Pennsylvania State University
rena@psu.edu

Natural Language Processing (NLP) metrics for bilingual code-switching (CS) have, until now, used words as the token level. However, the assumption that any two words constitute an equally likely switch point is erroneous. In spoken language, a major delimiter of CS is a prosodic chunk known as the Intonation Unit (IU). Switch points are far more likely between words at IU boundaries than between words in the same IU. The word as an elementary NLP unit is thus incommensurate with bilingual speech patterns. Here, we put forward an IU-based adaptation of a familiar metric of CS probability. We then compare the token levels on this metric for ten bilingual datasets featuring multi-word CS. Our comparison shows that the currently standard two-significant-figure precision of the word-based metric is insufficient, as the token level compresses the range of values by inflating the universe of CS. More discerning CS probability values can be obtained by normalizing word-based counts using mean IU length.

1. Introduction

Code-switching (CS), or alternating between languages in everyday conversations, is well documented in bilingual communities (e.g., Deuchar 2020; Poplack 1980). CS has gained traction in Natural Language Processing (NLP) research, being recently acknowledged as an important research direction at the Association for Computational Linguistics (Doğruöz, Sitaram, and Yong 2023). CS has attracted attention in NLP mainly due to its presence in social media data (Winata et al. 2023). A general challenge facing CS research, however, is the lack of usable and representative CS language data (Doğruöz et al. 2021). This is due to the difficulty of automatically identifying natural CS in well-defined speech communities or on the Web, the labor it requires to effectively transcribe multilingual speech data, and the task of distinguishing different

Action Editor: Preethi Jyothi. Submission received: 1 December 2024; revised version received: 7 September 2025; accepted for publication: 23 October 2025.

https://doi.org/10.1162/COLI.a.580

language mixing patterns. To this end, our work focuses on the latter point by empirically comparing the word token—up to now the accepted token level for CS metrics in NLP—with the Intonation Unit (IU), a prosodically defined speech chunk. IUs are defined by prosodic segmentation cues that are shared across human languages (Chafe 1994; Himmelmann et al. 2018).

CS is illustrated in Example (1), reproduced verbatim from a bilingual corpus, where each line of transcription represents an IU.[1] The translation appears on the right. Italics represent speech originally in English, and roman type, Spanish. The letters following each line are language tags, where "E" indicates an English IU, "S" a Spanish IU, and "SLS" a Spanish IU containing a lone other-language item "L". Between lines (a) and (b) there is a switch from Spanish ("S") to English ("E") and between (c) and (d), from English to Spanish. These are instances of multi-word CS, or alternating-language strings. Within line (a) there is an English noun embedded in Spanish (*kid*). This illustrates a lone item (L), or other-language, single-word incorporation.

(1) Multi-word Code-switching (CS)

a.	cuando tú haces una cosa por un *kid* o asina,	'when you do something for a *kid* or suchlike,	SLS
b.	... *through the years*,	... *through the years*,	E
c.	... *they appreciate*,	... *they appreciate*,	E
d.	.. cosas asina.	.. things like that.'	S

[23, 42:59–43:05]

Metrics such as CS probability enable comparisons across bilingual communities and individuals. Quantitative metrics for CS give values to qualifiers—such as "abundant", "dense", and "extensive"—invoked in debates on how CS may affect language processing and grammatical structure, due to parallel activation or mutual influence of bilinguals' two languages (e.g., Green and Wei 2014; Myers-Scotton 1997, page 211). Replicable definitions and reliable metrics of CS are sorely needed, as underscored by discrepant calculations. For example, the frequency of codeswitched utterances was given as 5.8% (2,527/43,327) in one report but 9.2% (3,923/42,854) in another, for the same bilingual corpus (Fricke, Kroll, and Dussias 2016, page 114; Wintner et al. 2023, pages 1475–1476).

Currently, CS metrics within NLP use words as elementary units. This means that in a sequence of four words, there are three eligible switch points, and thus there could be between one ($w_{L1}\, w_{L1}\, w_{L2}\, w_{L2}$) and three ($w_{L1}\, w_{L2}\, w_{L1}\, w_{L2}$) instances of CS (Gambäck and Das 2016:1851 n.3). However, there are two a priori problems with the assumption that any two words constitute an equally likely switch point.

First, switch points are structurally constrained. Multi-word CS is avoided at points where there is a word placement mismatch between the two languages, for example, where adjectives are placed before nouns in one language but postnominally in the other. Instead, CS occurs at shared syntactic boundaries of the two languages, for

1 Examples are from the New Mexico Spanish-English Bilingual (NMSEB) corpus (Torres Cacoullos and Travis 2018: Chs. 2 & 3; `https://nmcode-switching.la.psu.edu`). Within brackets is the recording number and time stamp (minutes:seconds). The corpus is transcribed in IUs (Du Bois et al. 1993; see Appendix A for transcription conventions).

example, where determiners are placed before nouns in both languages (the Equivalence constraint, Poplack 1980, page 586; cf. Deuchar 2020, page 255; Muysken 2000, page 27; Sankoff 1998). Second, CS frequency is lower than may have been assumed, far from rates of 33% ($w_{L1}w_{L1}w_{L2}w_{L2}$), let alone 100% ($w_{L1}w_{L2}w_{L1}w_{L2}$). One Spanish-English bilingual speech corpus of 7,000 words (Solorio and Liu 2008) yielded 6% and another, of 500,000 words (Bullock and Toribio 2013), just 2% (Guzmán et al. 2017, pages 69–70).

More generally, development of CS metrics has been hampered by the problem of the *possible* outcomes, or eligible CS locations (often represented by a metric's denominator). While it can be straightforward to determine the favorable outcome (i.e., the numerator) by counting occurrences of CS, determining the denominator requires circumscribing the universe of contexts where CS *could have* occurred (and not only those where it actually did occur) (cf. Labov 1972, page 72). The word placement equivalence constraint and the actual low frequency of CS indicate that it is useful to refer to strings of more than one word as a suitable scale for probability measures.

In Pattichis et al. (2023), we adapted metrics of text multilinguality (how balanced the distribution of languages is) and CS probability (the percent, or rate, of CS points) to the IU as the basic token, calculating CS probability according to prosodic position (IU boundary). Despite individual differences in CS rates, which are independent of multilinguality, there is a shared tendency to perform multi-word CS at the boundaries of IUs rather than within them. In addition, we showed that this IU-boundary constraint on CS is blurred by failure to distinguish lone (single-word) items (see Section 4.2 Distinguishing Multi-word CS and Lone Item Incorporations).

Here, we devise a unitary version of the CS probability metric (the I-Index; Guzmán et al. 2017) that combines IU-boundary and within-IU CS rates. Applying word- and new IU-based calculations of the CS probability metric to ten bilingual datasets featuring multi-word CS, which range from balanced to skewed proportions of the two languages, we probe the suitability of the word-token level, the standard in NLP. By comparing its output with the IU-token level on this CS metric, we reveal that the word-based CS values are compressed (i.e., a small range of values), while the IU makes it possible to discern similarities and differences when measuring CS rates across corpora. We additionally highlight the insufficiency of the word token at the probability metric's standard reporting precision of two significant figures through hierarchical clustering comparisons and an improved cluster similarity score using the IU.

2. Background

2.1 Syntactic and Prosodic CS Patterns

The first step towards suitable CS metrics is recognition that CS is structurally constrained, both syntactically and prosodically.

Syntactically, CS is avoided where there is a word placement mismatch between the two languages. The Equivalence constraint requires local equivalence of word order in the two languages around a switch point (Poplack 1980, page 586, 2015; Sankoff 1998; cf. Deuchar 2020, page 255; Lipski 1978, page 258, Muysken 2000, page 27, Pfaff 1979, page 291). For example, there is a mismatch between English and Spanish in the placement of attributive adjectives, which are postnominal often in Spanish but rarely in English; on the other hand, there is a match in the placement of articles, which are prenominal in both languages. Thus, the syntactic boundary between article and noun is an equivalence point. To illustrate, in the bilingual string "con los *big sizes*" (translated

to 'with the *big sizes*'), the switch point is after the article. Indeed, Spanish-English CS within a noun phrase is far more likely after the article than between the adjective and noun—at a 4:1 ratio (Torres Cacoullos and Vélez Avilés 2023, pages 624–625).

It is important that there is variability even among equivalence points. Consider bilingual clause combining. Bilinguals have choices as to the direction of CS (the language of the first vs. the second clause) and as to the positioning of CS with respect to the conjunction (switching at or after the conjunction). In Spanish-English main-and-complement clause combinations, for example, there is a strong preference for Spanish complementizer *que* over its English analogue *that* (89%, $N = 62$) (Torres Cacoullos and LaCasse 2025, n.4). This preference means, as to CS direction, switching from a Spanish main clause to an English complement clause rather than the reverse (58% to 42%), and as to CS positioning, switching after the complementizer rather than beginning with it (57% to 43%).[2] In sum, CS is not equally likely between any two words, depending on the direction of CS (order of languages) and the syntactic boundary of CS (position of the switch).

Prosodically, CS tends to occur at the boundary of Intonation Units (Torres Cacoullos and Travis 2018, page 51). Intonation Units (IUs) are melodic and rhythmic chunks that are the basic level of prosodic phrasing cross-linguistically (Chafe 1994, pages 53–70; Du Bois et al. 1993, page 47; Himmelmann 2022, pages 717–718; Inbar et al. 2023, pages 8189–8190). Speech is segmented into IUs perceptually, by acoustic cues such as a rising or high pitch at the beginning of the IU, lengthening of the last segment at the end of the IU, and (optionally) pauses between IUs. In corpora transcribed in IUs, each line represents one IU. Punctuation at the end of each line represents the transitional continuity between IUs. A comma indicates "continuing" intonation, projecting more to come, as in Example (1a–c), while a period marks "final" intonation (a fall to low pitch), as in Example (1d) (Du Bois et al. 1993, pages 52–55).

IU-based transcription demarcates major boundaries in speech, where the notion of the sentence is not readily applicable and the "utterance" remains poorly defined. The prosodic sentence is objectively defined as an IU or series of IUs containing at least one finite verb that ends in intonational completion (marked by a period) (Chafe 1994, page 139). The excerpt in Example (1) is an illustration of this. The prosodic sentence provides a speech unit within which to count CS direction. Example (1), for instance, hosts two multi-word CS points; the direction of CS is from Spanish to English between lines (a) and (b) and from English to Spanish between (c) and (d).

2.2 The IU-Boundary Constraint

According to the IU-Boundary constraint, CS is far more likely between two words at an IU boundary, rather than within the same IU (Torres Cacoullos and Travis 2018, pages 51–52; cf. Mettouchi 2008, page 195, Shenk 2006, page 189; Bullock et al. 2018, page 2536). Consider the instances of multi-word CS transcribed in IUs in Examples (2a–d). CS direction is indicated by the language tags, Spanish (S) and English (E). The

2 The distribution of the four distinct CS points in Spanish-English main-and-complement clause combinations is: 52% after the Spanish complementizer (me dijeron que *I was gonna run* 'they told me that *I was gonna run*'), 37% at the Spanish complementizer (*I told them* que iban a salir '*I told them* that they were going to be'), 6% at the English complementizer (ella me dijo *that she'd rather go* 'she told me *that she'd rather go*') and 5% after the English complementizer (*I don't even know that* la ∼Ivette no lo sabe '*I don't even know that* ∼Ivette doesn't know') (Torres Cacoullos 2020; Torres Cacoullos and LaCasse 2025).

(2)	IU-Boundary Constraint on CS		
	Distribution across prosodic positions [CS proportions], $N = 9,375$ & Rate [CS points over total IUs], $N = 70,882$		
(a)	Turn boundary (22% of CS & 26% of IUs)		
P:	*they had water fountains.*	*'they had water fountains.'*	E
R:	*... but that was it.*	*'... but that was it.'*	E
P:	pero no tenía excusado.	'but it didn't have a toilet.'	S
		[10, 04:20–04:23]	

(b)	Sentence boundary (30% of CS & 14% of IUs)		
	me trataron muy bien.	'they treated me very well.	S
	...(2.0) *the rent was cheap.*	...(2.0) *the rent was cheap.'*	E
		[10, 21:57–21:59]	

(c.)	Intonation Unit (IU) boundary (35% of CS & 8% of IUs)		
	you know the old school,	*'you know the old school,*	E
	...(0.8) era una escuela de adobe.	...(0.8) it was an adobe school.'	S
		[10, 01:22–01:24]	

(d.)	Within Intonation Unit (IU) (12% of CS & 2% of IUs)		
	y para nosotros *it was a snap,*	*'and for us it was a snap,'*	SE
		[10, 01:22–01:24]	

position of CS, or the CS point, may be at the speaker turn boundary (2a); at the prosodic sentence boundary (2b); at the IU boundary within the sentence (2c); and within the IU (2d).

The IU-Boundary constraint is seen in the distribution (proportions) of CS according to prosodic position. Eighty-eight percent of switches occur at some type of prosodic boundary—22% at speaker turn boundaries, 30% at prosodic sentence boundaries, 35% at IU boundaries within the sentence—and just 12% within the IU ($N = 9,375$). This means that the ratio of CS at IU boundaries vs. within the same IU is approximately 7:1. When we consider CS rate (CS points over total IUs) for each prosodic position it is 26% between turns, 14% between sentences, 8% between IUs, and just 2% within IUs ($N = 70,882$). Thus, the IU is an overarching delimiter of CS, which tends to occur precisely at prosodic boundaries.

2.3 Application of the IU to CS Metrics

Prosodic and syntactic structure are related. Neighboring words within the same IU tend to have a tighter syntactic relationship than words at the boundary of different IUs (Croft 1995, pages 849–864). For example, complements (object noun phrases) are more likely than adjuncts (postverbal prepositional phrases) to be within the IU of their verb (Croft 1995, page 863). The relevance for CS is that a looser syntactic relationship between two words at a boundary between IUs may correspond to more flexibility in word order, and thus a better chance that the boundary will be a switch point that satisfies the Equivalence constraint (Poplack 1980, page 586).

Circumscribing the universe of CS by syntactic boundaries at which there is equivalence of word order in the two languages has been described as "extremely onerous" (Poplack 1993, page 277; Sankoff and Poplack 1981). In addition, a large and growing number of studies of conversational data has revealed that assumed syntactic categories such as verb, complement or complement phrase are often unsuitable for the structure and processes of everyday language use (e.g., Ono and Thompson 2020, pages 1–2 and references therein).[3] As a major delimiter of CS, the IU provides a solution. Accurate transcription of spontaneous speech also requires a large time investment, especially if it is prosodically based (Torres Cacoullos and Travis 2018, pages 46–51). Nevertheless, IUs are reliably identified by phonetic cues (such as initial rise in fundamental frequency or pausing between IUs). Interrater agreement on IU boundaries is significant, including for languages unfamiliar to the annotator (Himmelmann et al. 2018, pages 241–242). Therefore, given the state of our knowledge of syntactic boundaries in everyday speech, the universe of CS in a bilingual corpus is more readily and reliably defined by IU boundaries. Here, we capitalize on a prosodically transcribed corpus.

3. Data

The data we use are from the everyday speech of bilinguals who spontaneously code-switch. The New Mexico Spanish-English Bilingual (NMSEB) corpus records members of a long-standing, non-immigrant community in northern New Mexico (Torres Cacoullos and Travis 2018, Chapters 2 and 3).[4] Spanish was spoken long before English in northern New Mexico—since the end of the 16th century. New Mexican Spanish has been in intense contact with English since the mid 19th century, especially with the expansion of the railroad and arrival of new settlers. Today, Spanish language loss is apparent in the proportion of U.S.-born New Mexicans identifying as Hispanic (or Latino) on the census who report speaking only English at home, nearly half in the state. Still, the complementary half resist the shift to English, speaking Spanish and also speaking English "very well" (Torres Cacoullos and Travis 2018, pages 21–25). The recorded speakers, drawn from this population, have birth years from 1922 to 1993 and occupations ranging from mineworkers and ranchers to teachers and service employees, most living in rural areas (Torres Cacoullos and Travis 2018, page 26). The recordings are of sociolinguistic interviews (Labov 1984) made by community insiders. Within this community, CS is globally a "discourse mode" in that, once the global situation is seen as appropriate, CS occurs with no change in interlocutor or topic, i.e., no external trigger (Poplack 2015, page 918; Poplack 1993, page 276; Torres Cacoullos and Travis 2018, pages 67–70).

Here, we work with ten transcripts, totaling 9 recorded hours, 78,600 words, and 26,600 IUs. Summary information of the ten datasets appears in Table 1. These bilingual recordings are largely monologic, in that fewer than one fourth of transcribed IUs were produced by interviewer(s) (IUs produced by speaker: mean 84%, range 76%–94%). This

3 This applies especially to traditional syntactic units smaller than the clause; there is evidence from neural responses that clause closure depends on IU closure (Inbar et al. 2023, page 8198).

4 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 and misinterpretation of local variants (Torres Cacoullos and Travis 2018, pages 47–49; cf. Poplack 2022, pages 212, 217).

Table 1
Summary information of bilingual speech recordings.

Corpus	Duration (min.)	# IUs	# words
5	61	3030	7858
16.1	45	2039	6317
8	35	2077	5643
26	46	2170	5692
27	64	3181	9038
11	64	2975	10340
15	47	2815	8287
3	75	3138	9700
10	42	2147	6216
23	62	3051	9551

Table 2
Distribution of languages and M-Index (0 = monolingual, 1 = balance of languages), IU-based and word-based (adapted from Barnett et al. 2000).

Corpus	IU-based			Word-based		
	# IUs Span/Eng	% IUs Span/Eng	M-Index	# words Span/Eng	% words Span/Eng	M-Index
5	1911/532	78/22	0.52	5859/1976	75/25	0.61
16.1	714/1058	40/60	0.93	2299/3974	37/63	0.87
8	631/918	41/59	0.93	2039/3574	36/64	0.86
26	1592/162	91/9	0.20	5039/602	89/11	0.24
27	616/2035	23/77	0.56	1970/7013	22/78	0.52
11	1412/1202	54/44	0.99	5212/4986	51/49	1.00
15	538/1723	24/76	0.57	1563/6678	19/81	0.44
3	1040/1501	41/59	0.94	3509/6079	37/63	0.87
10	737/894	45/55	0.98	2497/3686	40/60	0.93
23	1443/1007	59/41	0.94	5126/4267	55/45	0.98

allows us to discount the interlocutor as a potential trigger for CS (cf. Kootstra, Dijkstra, and van Hell 2020).

IUs are tagged for language, such that all-Spanish and all-English IUs are tagged "S" and "E", respectively (ex. 2a, 2b, 2c). IUs hosting both languages are tagged "SE" or "ES", or other combinations of "S" and "E" (as in 2d). Table 2 depicts the language distribution of the datasets, for IU tokens and word tokens.[5] On the right for the word-token level, referencing the 5th column, which reports the S and E tags (not counting lone items), our datasets average 7,795 word tokens each (min = 5,613 from transcript 8; max = 10,198 from transcript 11). Combined, our datasets total 77,948 word tokens eligible, in theory, to host CS. The sixth column gives the distribution of languages as

5 Table 2 token counts (2nd and 5th columns) are lower than in Table 1 (3rd and 4th columns), which include non-eligible tokens for CS (not tagged as Spanish or English), such as proper nouns and fillers (see Section 4.1).

%, ranging from a balanced 51/49 to a skewed 89/11. On the left, for the IU-based distributions, unsurprisingly, counts are smaller (2nd column). Despite the decrease, however, the language distributions from the word-based calculations are largely reproduced (with an average increase of just 3.5 percentage points in the % of S tokens, 3rd column).

The Multilingual Index, or M-Index, measures the multilinguality of a corpus, from 0 to 1, where 0 = monolingual and 1 = a perfect balance of languages (Barnett et al. 2000; adapted for the IU token in Pattichis et al. 2023, page 16845). Six datasets display near-balanced language distributions, three show asymmetry, and one is heavily skewed in favor of one of the languages. The two token levels yield fairly similar M-Indexes (shaded columns, Table 2). For the IU token, values above 0.9 correspond to IU distributions of 55%–60% of IUs in one language, values of 0.52–0.57 to 76%–78%, and the value of 0.2 to 91%. Correspondences between M-Index values and language proportions are similar for the word token: M-Indexes of 0.86-1 have one of the languages constituting 51%–64% of all words; 0.44–0.61, 75%–81%; and 0.24, 89%. These correspondences match ones previously reported; for example, three Spanish-English datasets have M-Index to language proportion values of 0.99 to 53%, 0.63 to 74%, and 0.60 to 75% (Bullock et al. 2018, page 2535), and for two Hindi-English speakers, 0.99 to 54% and 0.51 to 79% (speakers SRK and SN, Ellison and Si 2021, pages 1515, 1517). What is important is that the wide range in M-Index values here allows us to gauge and compare the output of CS probability metrics across different proportions of the two languages.

4. Methods

Below we outline the steps taken to conduct an empirical comparison of the IU and word-level for the Integration Index, or I-Index, a fairly widely used metric of CS probability (Guzmán et al. 2017).[6]

4.1 Automatic Tagging Rules

The original data are transcribed and tagged by IU; for example, a continuous multiword string in Spanish is given a single tag S. Some IUs (also) host potentially language-neutral items. These are proper nouns (tagged P), for example *California*, and discourse markers or backchannels (D), for example *so* (*so* appears both in English and in the otherwise monolingual Spanish of bilingual speakers in New Mexico [Aaron 2004]). Also potentially language-ambiguous is the question tag *no?* (N) when it occurs at a switch point. Finally, an IU may consist of a filler (filled pause) (F) such as *uh* or *uhm*. See Appendix B for language tagging examples.

Our first task, then, is to convert the IU-based language tags from the original transcriptions into a language tag for each word (delimited by spaces). All original text that is segmented by IU is first pre-processed to remove any speech-related transcription symbols that cannot be mapped to a word, i.e., '...' indicating pauses, @ laughter, % glottal stop, and descriptions within parentheses, such as ((COUGH)).

After pre-processing, we map each IU to its word tags. In the simplest case, there is only a single language tag for an IU. If the tag is S or E (for Spanish and English, respectively), then the word-level language tags are the IU tag replicated by the number

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

of words. Otherwise, for example, for an IU consisting of only a discourse marker, a single tag is left for all the words in that IU (e.g., D).

To deal with multiple language tags within a single IU (e.g., SDS), we must identify the boundaries of each language tag. We parse the string of tokens by identifying the tags in the following order: D, N, P. The order is important in that it deals with the most identifiable and possibly multi-word cases first (discourse markers D, such as *you know*). If there is an N tag, then we look for *no?* in the string.

In the case of a P tag, or proper noun, we first check if it is the last tag in the complete IU tag (e.g., SP), and whether or not it was anonymized with a preceding '∼' (Appendix B, a). If so, then the boundary is simple, because the beginning is immediately after the special transcription symbol '∼' up until the end of the IU. Otherwise, since the proper noun is not easily identifiable, we use a pre-trained Language Identification (LID) model trained on the LinCE dataset (Aguilar, Kar, and Solorio 2020) for Spanish and English code-switching.[7] The model tags each word with either spa (Spanish), en (English), or ne (Named Entity)—we use the latter for identifying the proper nouns, or P tag. Note that tagging might be incorrect in the case that the LID model is not identifying ne boundaries, such as considering a New Mexican town it has not seen before as simply Spanish (e.g., tagging "Los Lunas" as 'the' and 'moons').

(3) Lone nouns

(a.) English into Spanish

con la *flashlight* en una mano,	'with the *flashlight* in one hand,'	SLS
	[16.1, 25:04–25:06]	

(b.) Spanish into English

to my tia's *boss*.	'*to my* aunt's *boss*.'	ELE
	[05, 57:47–57:50]	

4.2 Distinguishing Multi-word CS and Lone Item Incorporations

The tagging further distinguishes single-word incorporations, or lone items, tagged L, such as *kid* in ex. 1a (cf. Poplack and Meechan 1998). Lone items are common nouns and other, mostly content, words incorporated into otherwise unilingual discourse. They are integrated into the grammar of the recipient language as nonce borrowings (Poplack 2017, page 9; cf. Bullock et al. 2018, page 2537). Examples are Spanish gender in 'the-FEM' *flashlight* (3a) (whereas grammatical gender is absent in English), and the English genitive possessive (with 's) in *tia's* (3b) (whereas Spanish uses the preposition *de* 'of'). Lone items have also been referred to in NLP as unassimilated lexical borrowings (Álvarez-Mellado and Lignos 2022, page 3870).

It is important that lone items differ from multi-word CS in their positioning. Whereas multi-word CS strings are positioned in compliance with the IU-Boundary and the Equivalence constraints, lone items are placed according to the word order of the language in which they are embedded. For example, in "unos *desks* muy grandes" ('some very big *desks*') English *desks* is placed according to Spanish word order, i.e., with a postnominal adjective (literally: 'desks big'). In contrast, the multi-word switch in "con

7 The LID model from Code Switch (an online NLP tool meant to process CS data) can be found at `https://github.com/sagorbrur/codeswitch`.

los *big sizes*" ('with the *big sizes*') is positioned after the article, a point of equivalence between these two languages.

Here, we distinguish the two mixing types in answer to the call from computational linguistics for the NLP field to offer more granular insight into CS (Doğruöz et al. 2021) and following previous methods (Pattichis et al. 2023; Wintner et al. 2023). To handle combinations of S, E, and L within a single IU language tag, we once more use the LID model to find the language boundaries when mapped onto words. For tags of S and E, we again replicate that tag for the amount of words. On the other hand, we handle L as a single token tag, even if orthographically it appears as two words (e.g., *high school* will count as one word tag).

4.3 I-Index and Modified IU I-Index

Using the language tags to identify CS points, we calculate the familiar I-Index (Guzmán et al. 2017) to conduct token-level comparisons. Mathematically, the numerator is a summation over $S(l_i, l_{i+1})$, which is equal to 1 if there is a switch at the boundary between the ith and $i+1$th boundary and 0 otherwise. The denominator is $n-1$, or the total locations for possible switching given n tokens:

$$\text{I-Index}_{word} = \frac{1}{n-1} \sum_{1 \leq i \leq n-1} S(l_i, l_{i+1}) \tag{1}$$

The I-Index was designed for the word-token level. Pattichis et al. (2023) adapt it to the IU level by calculating separate I-Indexes to measure CS at IU boundaries vs. within IUs, using binary measures to maintain token-level integrity. To make a direct comparison with word tokens, here we report an I-Index value that combines IU-boundary and within-IU CS values without double-counting an IU token. Namely, we calculate the overall switching rate by adding the IU-boundary and within-IU CS, and subtracting the number of IUs we estimate to host both:

$$\text{I-Index}_{IU} = \frac{1}{n}(\text{IU-Boundary}_{count} + \text{Within-IU}_{count} - (\text{IU-Boundary}_{count} * \text{I-Ind.}_{within})) \tag{2}$$

Here, $\text{IU-Boundary}_{\text{count}} = \sum_{1 \leq i \leq n-1} S(l_i, l_{i+1})$, or the number of IU-boundary switches. $\text{Within-IU}_{\text{count}} = \sum_{i=1}^{n} S(l_i)$, or the number of IUs hosting a within-IU switch. The last coefficient subtracts the number of IU tokens that are being double counted by asking: *Of the IUs hosting IU-Boundary CS, approximately how many are also hosting within-IU CS?* The estimated number of IUs hosting both types is calculated by multiplying the number of IU-boundary switches ($\text{IU-Boundary}_{count}$) by the rate of within-IU CS (I-Ind._{within}) for the given dataset (Appendix C).[8] For example, for dataset 3, we count 375 IU-Boundary switches and 30 within-IU switches, with a calculated within-IU I-Index of 0.01. Using the combined Spanish and English IUs counted ($n = 2,541$), we get $\text{I-Index}_{IU} = \frac{1}{2,541}(375 + 30 - (375 \cdot 0.01)) = 0.16$ (reported in Table 3). Thus, we use this combined I-Index_{IU} for our token-level comparisons.

In addition, to compare outcomes when lone items and multi-word CS are distinguished vs. merged, here we calculate the I-Index for each token two ways, where (1)

8 The Within-IU CS rate uses a modified Equation (1) to consider all IUs hosting at least one switch: instead of $n-1$ we consider all n tokens, and instead of a binary comparison between adjacent tokens, we look within the token for the switch. This leads to $\text{I-Ind.}_{within} = \frac{1}{n}\sum_i S(l_i)$ (Pattichis et al. 2023).

Table 3
CS frequency, or I-Index, IU-based and word-based (adapted from Guzmán et al. 2017, page 68).

Corpus	IU-based		Word-based	
	# CS_{IU}	I-$Index_{IU}$	# CS_{word}	I-$Index_{word}$
5	82	0.03	80	0.01
16.1	93	0.05	97	0.02
8	84	0.05	85	0.02
26	129	0.07	130	0.02
27	205	0.08	206	0.02
11	224	0.08	228	0.02
15	347	0.15	350	0.04
3	401	0.16	406	0.04
10	281	0.17	285	0.05
23	551	0.22	556	0.06

L's are treated as the language in which they are embedded vs. (2) where Ls 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 we 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). We also treat these mixing types differently in that we only count stepping into a lone item as CS since stepping out of the single word incorporation is merely a consequence of the first trigger (Wintner et al. 2023, page 1476).

4.4 Clustering Experiments

To probe the stability of token levels, we cluster the word and IU-based multi-word CS I-Indexes using hierarchical clustering with three clusters, experimenting with both the precision (i.e., number of significant figures kept in the I-Index value) and the linkage type (i.e., single, average, ward, and complete). In particular, we use the Davies-Bouldin index (DBI) (Davies and Bouldin 1979), meant to measure separability and infer the appropriateness of data partitioning. The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. The range of the metric is from 0 to infinity, with values closer to 0 signifying better data separation (through both tight clusters and distance between said clusters).

5. Results

Table 3 depicts results for multi-word CS (separate from lone items). Reported for each dataset are the number of CS points recorded for the I-Index (i.e., the numerator) and the I-Index values; the token level is the IU to the left, the word to the right.[9] The datasets are visualized in Table 4, which provides language distribution graphs for each dataset.[10] Table 5 provides the same information as Table 3 (i.e., number of CS points

9 The # CS points for the IU-based results are estimates, corresponding to the numerator from Equation (2).
10 The language distribution graphs, based here on graphing an ordered array of language tags per word, are not notably affected by the token level (see Pattichis et al. 2023, page 16846).

Table 4
Language distribution graphs for each dataset, where English is in dark blue (darker color) and Spanish is in light blue (lighter color). The *x*-axis represents the token number (here, the word).

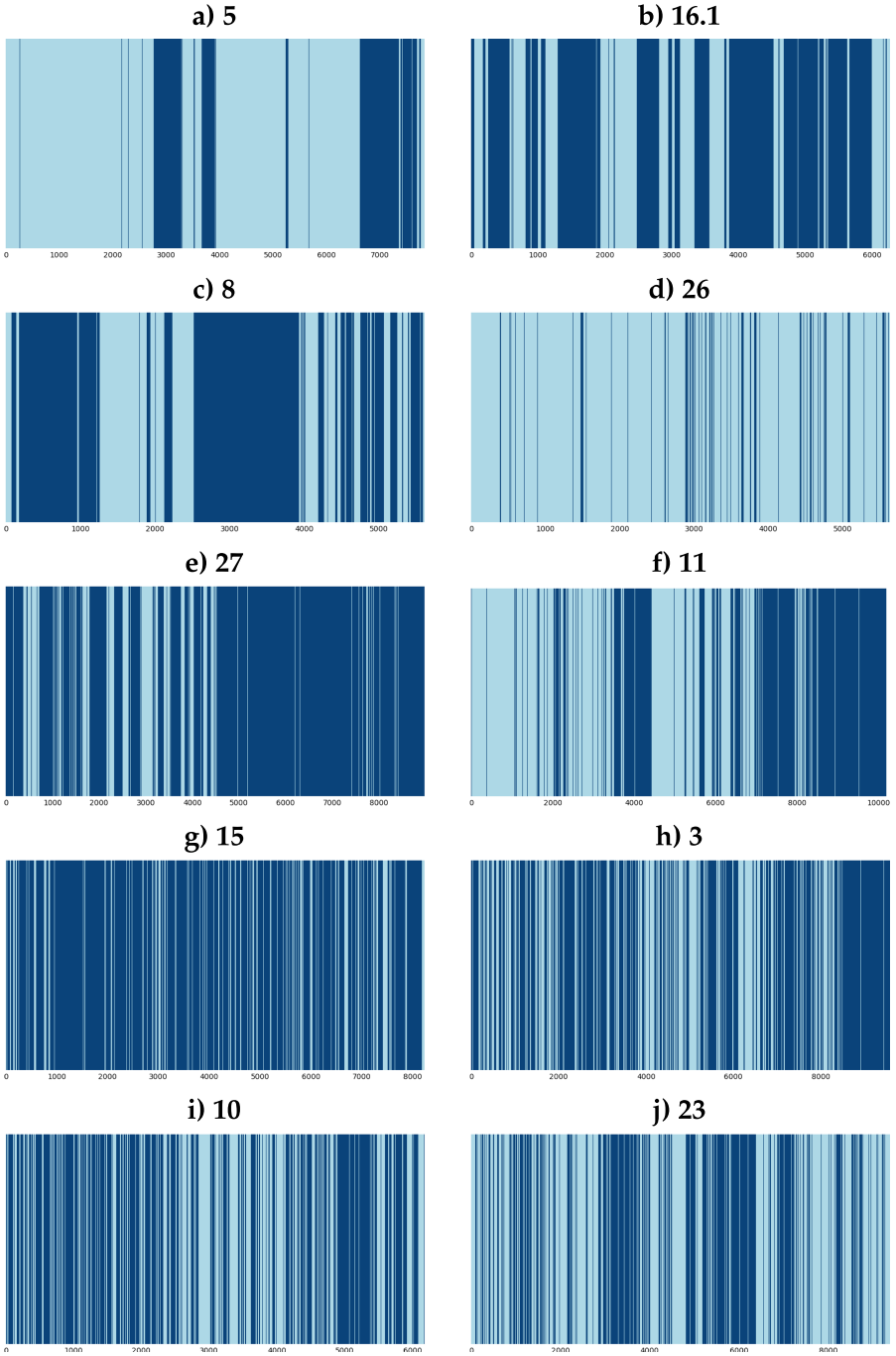


Table 5
CS frequency (I-Index) including lone items, IU-based and word-based.

Corpus	IU-based		Word-based	
	# CS	I-Index$_{IU}$	# CS	I-Index$_{word}$
5	101	0.04	101	0.01
16.1	132	0.07	136	0.02
8	107	0.07	109	0.02
26	174	0.10	177	0.03
27	249	0.09	251	0.03
11	343	0.13	357	0.03
15	378	0.17	385	0.06
3	482	0.19	500	0.05
10	297	0.18	304	0.05
23	656	0.26	683	0.07

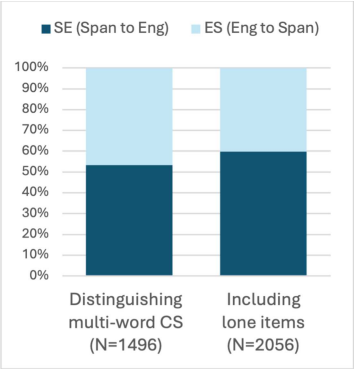
and I-Index) once lone items are not distinguished and instead are included as valid CS points. Table 6 shows, for both token levels, the direction of CS, from Spanish to English (as in Example (1), line a to line b) or the reverse (1cd); the percentages are not affected by the token level. For all tables and language graphs, datasets are ordered by increasing IU-based I-Index for multi-word CS.

5.1 Word-based Results: CS Rates are Generally Low

We observe, first, that the word-token I-Index ranges from 0.01 to 0.06, with half of our datasets assigned the value of 0.02 (Table 3, 5th column). Our range of values is starkly lower than those reported in the original experiments conducted on written data, though they approximate the I-Index reported for transcriptions of naturalistic speech (Bullock, Guzmán, and Toribio 2019, page 119; Guzmán et al. 2017, page 69). On the one hand, this emphasizes that literary CS data often exhibits unnaturally high CS rates,

Table 6
CS direction (order of languages), for multi-word CS and when lone items are included (word-based counts).

Corpus	Multi-word		including Ls	
	# SE/ES	% SE/ES	# SE/ES	% SE/ES
5	23/19	55/45	34/26	57/43
16.1	28/29	49/51	58/38	60/40
8	31/21	60/40	48/25	66/34
26	43/45	49/51	83/48	63/37
27	63/59	52/48	86/76	53/47
11	91/71	56/44	213/77	73/27
15	105/90	54/46	106/118	47/53
3	149/120	55/45	217/133	62/38
10	51/58	47/53	82/70	54/46
23	212/188	53/47	302/216	58/42



with switch points that are rarely present in everyday speech and ruled out by linguistic predictive models (Bullock, Guzmán, and Toribio 2019, page 119). On the other hand, and more fundamentally, the low values underscore the low frequency of CS at the word token level. This result is especially striking, since CS is an in-group discourse mode for the speakers in this bilingual community.

Second, recall that our datasets represent diverse language distributions (proportions of the two languages) (Table 2). The ordering of datasets by increasing I-Index (5, 16.1, 8, 26, 27, 11, 15, 3, 10, 23) does not match their ordering by M-Index (26, 5, 27, 15, 8, 16, 3, 23, 10, 11) (Pearson correlation of 0.36, n.s. for word-based I-Index and M-Index). This lack of a discernable pattern linking CS frequency to language distribution aligns with previous observations that the two metrics are independent (Pattichis et al. 2023, page 16845).

Third, despite differences in the distribution (proportion) of languages, the datasets display fairly balanced CS by direction, or order of languages. Recall that the IU token, which accommodates the prosodic structure of speech, inherently supplies the prosodic sentence (see Section 2.1) as the unit within which to assess switching direction.[11] Table 6 breaks up the number of CS points by language direction within a sentence. The direction of CS from Spanish to English (SE) and from English to Spanish (ES) is skewed by at most 10 percentage points (60/40, Table 6, 3rd column).

5.2 IU-based Results: IU Token Enables Data Granularity

The # CS columns cannot be directly compared across token levels, and therefore may not be identical. Nevertheless, the number of multi-word CS points is similar (2nd and 4th columns in Table 3), corroborating that, rather than switching between single words, bilinguals switch between multi-word strings.

The IU-based I-Index (Table 3, 3rd column) yields values that are 2.5 to 4 times greater than the word-based values, while arranging the datasets to reveal the relative ordering suggested by the word-based metric. The IU-based I-Index ranges from 0.03 to 0.22. For the word-based I-Index, there are five datasets with the same value of 0.02 (i.e., 16, 8, 26, 27, 11), which all have a higher I-Index than the lowest dataset (5, at 0.01) and a lower I-Index than the four highest datasets (15, 3, 10, and 23, at 0.04–0.06). These relationships are put in working order by the IU-based ordering of the datasets. Thus, the IU-based I-Index provides insightful granularity for distinguishing the datasets, in contrast with the generally contracted values of the word-based metric. The discrepancy in the range of I-Index values makes sense: While both token level numerators remain similar (Table 3, 2nd and 4th columns), the word-level calculations count far more tokens in the denominator as possible switch points (Table 2, 2nd and 5th columns).

At the same time, just as for the word-based results, the IU-based I-Index (Table 3, 3rd column) is only weakly to moderately associated with the M-Index (Table 2, 4th column) (Pearson's correlation of 0.35, n.s. for IU-based I-Index and M-Index). Here, for I-Index values 0.08 or lower, M-Indexes range from 0.2 (for 26) to 0.99 (for 11). For I-Indexes of 0.15 or higher, the range in M-Indexes is smaller, but it is still wide,

11 In calculating CS direction within the prosodic sentence (see Section 2.1), we 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); CS direction is calculated within the prosodic sentence for both the IU and the word token.

from 0.57 (for 15) to 0.98 (for 10). The language distribution graphs in Table 4 visualize the transcripts in bands representing language spans, depicted in the darker color for English and the lighter color for Spanish. The number of bands indicates the frequency of CS and the aggregate proportion of each color, the distribution of languages. These visualizations show that CS rates do not correspond to language proportions. Datasets 8 and 16.1 have near balanced language proportions (M- Index: 0.93; Table 2) but lower CS rates (I-Index: 0.05; Table 3), while 15 has a disproportion of one language (M-Index: 0.57) but a relatively high CS rate (I-Index: 0.15). In addition, also as for the word token, regardless of the proportions of the two languages (M-Index) and the rate of CS (I-Index), the datasets are fairly balanced by CS direction (Table 6).

5.3 Lone Items Exaggerate Asymmetry

Table 5 now shows the same information as Table 3, i.e. the number of CS points and the I-Index values when lone items, such as *kid* in Example (1a), *flashlight* (3a), and *tía* (3b), are conflated with multi-word strings. The IU-based results are on the left, word-based on the right.[12]

Naturally, by including lone items as CS, Table 5 increases the rate of CS for all datasets compared with Table 3. However, increases are uneven. For the word-based I-Indexes, four of the ten datasets experience no I-Index change (5, 16.1, 8, 10), one goes up by 17% (23, from 0.06 to 0.07), another by 25% (3, from 0.04 to 0.05), and four by 50% (26, 27, 11, from 0.02 to 0.03, and 15, from 0.04 to 0.06).[13] The impact is similar if somewhat greater for the IU token, for which increases range from 6% to 63% (in order of increase, 6% [dataset 10], 13% [27, 15], 18% [23], 19% [3], 33% [5], 40% [16.1, 8], 43% [26], and 63% [11]). Thus, conflating lone items with multi-word CS affects the ordering of the datasets by I-Index (switching rate). On the other hand, distinguishing between mixing types enables a more fine-grained view of bilingual practices, such as identifying which datasets (individuals, settings, or communities) contain more lone items than others.[14]

A striking distortion caused by conflating lone items with multi-word CS affects the apparent direction of switching (Table 6, using word-based percentages; IU-based % are mostly identical, see footnote n. 12). The question is whether there is a tendency to switch to one of the languages from the other one as the base language. For multi-word CS, the direction of "intra-sentential" CS (within the prosodic sentence) is quite balanced, ranging from 49/51 to 60/40 SE/ES (Table 6, 3rd column). Including switches into lone items (L's) exaggerates the asymmetry between the languages (Table 6, 5th column). The asymmetry shifts in favor of SE in all datasets (except for one (15)). The figure in Table 6 compares switching direction in the aggregate when distinguishing multi-word CS and including lone items. While multi-word CS is near-balanced at 53/47, including lone items shifts the proportions to 60/40. This skewing is consistent

12 The difference in the counts of CS points (# CS, Table 5), which are somewhat greater for words than for IUs, is because there may be multiple lone items within a single IU. CS direction (Table 6) is identical for the token levels (except for datasets 26, 27, and 10, for which % SE/ES are, for the IU and word token:, respectively: 64/36 and 66/37; 54/46 and 53/47; 51/49 and 54/46).

13 The uneven impact of lone items across datasets remains for I-Index values at a fuller precision of four decimal places (Appendix D).

14 Frequency of lone items may vary more by situational factors (topic, interlocutor) than by individual. For example, the same speaker recorded on two occasions had a normalized frequency of 7 lone English nouns per 10,000 words in one recording but 51 in the other (Aaron 2015, page 480).

with the robust community pattern of preferring to incorporate English lone words into Spanish rather than the reverse.[15]

The distinction between lone (single-word) items and multi-word CS operationalizes classifications that have been made under labels such as insertional versus alternational switches (e.g., Guzmán et al. 2017, page 70; Muysken 2015, pages 251–254). As we've seen, lone items unevenly affect switching rates (I-Index), while uniformly distorting switching direction, in favor of the language preferred in particular for lone items (which follows from extra-linguistic community norms [Torres Cacoullos et al. 2022, page 630]). Moving forward, distinguishing these two major mixing types in CS metrics responds to 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, page 2537, Wintner et al. 2023, page 1480).

5.4 Comparing Word and IU Token Levels

The chief difference between the two token levels is that the word token compresses the CS frequency metric (I-Index) compared with the IU-based metric: The values decrease by a factor of 2.5 to 4 and are thus less differentiated, as the range of the values contracts. This is because the denominator is higher in the word-based measure. Furthermore, implicit in a count of all words as the denominator is the assumption that CS is equally likely between any two words. This assumption is incommensurate with the phenomenon, the frequency of which is regulated by syntactic and prosodic constraints. Empirically, the false assumption yields a less discerning metric than that computed based on the IU.[16]

However, the word-based I-index, which is conventionally reported to two decimal places, could be computed with higher precision to compensate for the inflated denominator. Indeed, at four decimal places, the values are 0.0102, 0.0154, 0.0151, 0.0228, 0.0228, 0.0221, 0.0422, 0.0419, 0.0459, 0.0582 (in the same order of the datasets as in the preceding tables), leaving just two datasets with identical values (Appendix D). Still, as shown in Figure 1, the IU-based I-Index values effectively stretch the word-based values. This granularity achieves the disambiguation of seven distinct datasets rather than two at the standard precision used when reporting the I-Index (five at 0.02 and two at 0.04) (Table 3), and two datasets rather than one at four figures (at 0.0228) (Appendix D). These visual findings are quantitatively confirmed in Figure 2, which plots the difference in DBI scores for the clustering output between the IU-based and word-based I-Indexes (see Section 4.4). While after four significant figures the word- and IU-based measures are similar, at the standard two significant figures, the IU token level provides an improvement in both in-cluster tightness and between-cluster distance (0.19 and 0.29 for the IU and word token results, respectively). The linkage type used in hierarchical clustering does not impact these results.

15 In the New Mexico bilingual community corpus, lone items tend to be English incorporations into Spanish (72%, $N = 2991$) (Torres Cacoullos and Vélez Avilés 2023, page 614). In contrast, multi-word CS is fairly balanced by CS direction (42% Spanish to English, 29% English to Spanish, 29% both directions, $N = 2489$ prosodic sentences) (Pattichis et al. 2023, page 16842).

16 We thank a *Computational Linguistics* reviewer for noting that the word-based metric is additionally inadequate in ignoring conventionalized multi-word chunks, or prefabricated sequences (e.g., *pull strings*, *pick and choose*, *I don't know* [Bybee 2010, pages 34–37]). Furthermore, the word is unsuitable as a token level for comparisons across morphological types of languages (e.g., an isolating language, where each grammatical category is a separate word, vs. an agglutinating language, where morphemes are added to a root in a single word).

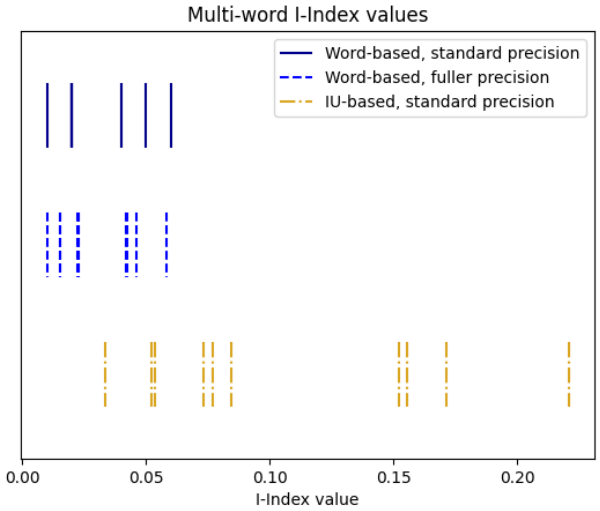


Figure 1
Visual comparison between the word-based and IU-based multi-word I-Index, showing that the IU-based token level provides more granularity in cases where the word-based token level flattens datasets' values (i.e., maps them to almost identical values).

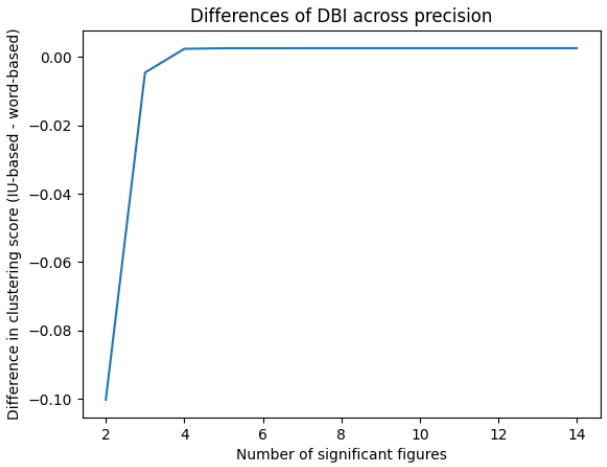


Figure 2
Graph of the difference in Davies-Bouldin index (DBI) score between the IU-based and word-based multi-word I-Index clusters generated from hierarchical clustering. Our results indicate that for the standard precision of two significant figures, the IU provides better separability. These results are identical across linkage types (single, complete, average, and ward).

The fact is that, despite the wide implementation of word-based CS metrics for several years now, NLP advances with bilingual data continue to lag behind monolingual data. The expediency of just segmenting a text into words is tempting, but this diminishes with more careful annotation for distinguishing non-eligible words, such as

fillers, other potentially language-neutral material, or words constituting prefabricated sequences. Moreover, despite any such expediency, given the present cautionary findings on the word token for currently standard probability metrics, the onus should fall on future work that fails to adopt the IU to justify that choice.

However, because corpora transcribed in IUs are not yet generally available, there is a practical minimal step to approximate IU-based measures.[17] By dividing the word-based universe by a constant representing the average length of IUs in words, calculated from a modest sample transcribed prosodically, one could estimate IU-based measures. For example, here, the constant for the Spanish datasets would be 3.3, for English, it would be 3.9 (35,113 words divided by 10,634 IUs and 42,835/11,032, respectively, from Table 2) (see Himmelmann et al. [2018, page 225] for mean IU length in four languages, ranging from 3.4 to 5.2, due to differences in function words). Dividing total word count by a language-specific constant could also be applicable to bilingual social media data used in NLP. Note that while scaling probability values by a constant is generally to be avoided because it risks pushing the values out of the valid 0 to 1 range, here this would ameliorate outcomes, precisely because the word token tends to compress CS probability values.

6. Conclusion

In sum, using the IU will enhance comparisons between bilingual individuals, settings, and communities. In bilingual datasets, the word-token level inflates the universe of CS by counting all words, regardless of position, as eligible CS hosts. The IU is a more apt token level than the word because it corresponds to the prosodic structure of CS in real bilingual speech, where CS is neither as frequent as its salience to researchers may convey; nor, crucially, is it equally likely between any two words. For this reason the IU provides a more suitable scale for CS probability metrics.

Our comparison has shown that the IU-based adaptation of a familiar CS probability metric at the standardly reported precision better groups bilingual datasets than the word-based measure. The present results thus suggest that normalizing word-based measures by using mean IU length can give more meaningful measures of CS probability. Importantly, the IU's ability to accommodate multi-word recognition also holds promise for predicting CS points with machine learning, toward developing automatic processing of mixed-language text (Solorio and Liu 2008). The present results, then, highlight the IU as a reliable token level that can better align CS datasets and metrics for NLP with bilingual speech patterns.

7. Limitations

CS probability metrics will have to consider additional factors. Clearly, transitional continuity between IUs affects CS rate (CS is more frequent following final than continuing intonation). CS models, especially concerning predictability, will also have to consider time course measures of the distance between CS points, or burstiness (cf. Goh and Barabási 2008). In addition, CS metrics have yet to be applied to datasets representing diverse bilingual community norms (cf. Poplack 1988) using the IU token level.

17 We thank the *Computational Linguistics* reviewers for encouraging such a practical interim solution.

Appendix A. Prosodically Based Transcription Conventions (Du Bois et al. 1993)

Table A.1

Each line represents an Intonation Unit (IU)*	
.	final intonation contour
,	continuing intonation contour
?	appeal intonation contour
..	short pause (0.2 secs)
...	medium pause (0.3–0.6 secs)
...(N)	timed pause (0.7 secs or longer)
∼	pseudonymized proper noun
Prosodic sentence	One or a series of IUs ending in final or appeal intonation contour and containing at least one finite verb.

* Where the IU does not fit on one line, the second line is indented.
** For readability, removed are vocal noises, laughter and vowel lengthening, as well as excerpt-initial pauses.

Appendix B. Language Tagging: Language-neutral Material

Table B.1

Language tagging: Language-neutral material			
(a)	Proper nouns (P)		
	It was a bigger school here in ∼***Chamisal***.	*It was a bigger school here in* ∼***Chamisal***.	EP
			[10, 01:45–01:46]
	regresábanos pa' acá pa' ∼**Chamisal**.	'we'd return here to ˜**Chamisal**.'	SP
			[10, 23:43–23:46]
(b)	Discourse markers or backchannels (D)*		
	so allí nos estuvimos todo lo demás del día,	'**so** we were there the whole rest of the day,'	DS
			[16.1, 19:17–19:20]
	pero estaba bien machucado todo adentro,	'but he was badly injured all inside,	S
	you know?	**you know**?'	D
			[03, 28:29–28:31]
(c)	Fillers (F)**		
	uh hubiera ríos,	'**uh** there would be rivers,	S
	.. montañas,	.. mountains,	S
	...(1.4) **eh**,	...(1.4) **eh**,	F
	... de todo.	... everything.'	S
			[27, 03:10–03:18]
(d)	Question tag *no?* (N)		
	∼*Kathleen used to work like that,*	'∼*Kathleen used to work like that,*	E
	no?	**no?**	N
	más antes.	before.'	S
			[09, 03:43–03:46]

* Discourse markers or backchannels are *so*, *you know*, *(oh) yeah*, *yep/yup*, *okay*, *anyway*, *hey*, *wow*, *oh*.
** Fillers are tagged F when they occur in their own IU.
*** Also tagged separately are IUs consisting of laughter or other nonlinguistic material and IUs consisting of unclear speech or including more than three syllables of unclear speech.

Appendix C. IU-based I-Index Numbers and Calculations

Table C.1
CS frequency, or I-Index, for Intonation Unit (IU), according to prosodic position and as a unitary measure (see Section 4.3 I-Index and Modified IU I-Index).

Corpus	# CS IU-Boundary	# CS Within-IU	I-Index IU-Boundary	I-Index Within-IU	I-Index$_{IU}$ (combined)
5	76	6	0.03	0	0.03
16.1	80	14	0.05	0.01	0.05
8	79	5	0.05	0	0.05
26	122	8	0.07	0	0.07
27	192	14	0.07	0	0.08
11	194	32	0.07	0.01	0.08
15	330	20	0.15	0.01	0.15
3	375	30	0.15	0.01	0.16
10	260	25	0.16	0.01	0.17
23	490	76	0.20	0.03	0.22

Appendix D. Higher-precision I-Index Values

Table D.1
CS frequency, or I-Index, at a precision of four significant digits for Intonation Unit (IU) and word. For each token level, we also report the I-Index when lone items (Ls) are included as CS.

Corpus	IU-based		Word-based	
	I-Index CS	I-Index incl. Ls	I-Index CS	I-Index incl. Ls
5	0.0334	0.0413	0.0102	0.0129
16.1	0.0521	0.0737	0.0154	0.0215
8	0.0536	0.0684	0.0151	0.0193
26	0.0732	0.0982	0.0228	0.0311
27	0.0767	0.0930	0.0228	0.0278
11	0.0846	0.1295	0.0221	0.0345
15	0.1522	0.1658	0.0422	0.0465
3	0.1554	0.1871	0.0419	0.0516
10	0.1713	0.1811	0.0459	0.0489
23	0.2206	0.2626	0.0582	0.0715

References

Aaron, Jessi Elana. 2004. "So respetamos un tradición del uno al otro": *So* and *entonces* in New Mexican bilingual discourse. *Spanish in Context*, 1(2):161–179. https://doi.org/10.1075/sic.1.2.02aar

Aaron, Jessi Elana. 2015. Lone English-origin nouns in Spanish: The precedence of community norms. *International Journal of Bilingualism*, 19(4):459–480. https://doi.org/10.1177/1367006913516021

Aguilar, Gustavo, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813.

Álvarez-Mellado, Elena and Constantine Lignos. 2022. Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3868–3888. https://doi.org/10.18653/v1/2022.acl-long.268

Barnett, Ruthanna, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland Van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, 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. https://doi.org/10.1177/13670069000040020101

Bullock, Barbara, Wally Guzmán, and Almeida Jacqueline Toribio. 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. https://doi.org/10.18653/v1/W19-2515

Bullock, Barbara E., Gualberto A. Guzmán, Jacqueline Serigos, and Almeida Jacqueline Toribio. 2018. Should code-switching models be asymmetric? In *Interspeech*, pages 2534–2538. https://doi.org/10.21437/Interspeech.2018-1284

Bullock, Barbara E. and A. Jacqueline Toribio. 2013. The Spanish in Texas Corpus project. *Center for Open Education Resources and Language Learning (COERLL)*.

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Croft, William. 1995. Intonation units and grammatical structure. *Linguistics*, 33(5):839–882. https://doi.org/10.1515/ling.1995.33.5.839

Davies, David L. and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Deuchar, Margaret. 2020. Code-switching in linguistics: A position paper. *Languages*, 5(2):22. https://doi.org/10.3390/languages5020022

Doğruöz, A. Seza, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 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. https://doi.org/10.18653/v1/2021.acl-long.131

Doğruöz, A Seza, Sunayana Sitaram, and Zheng-Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. *arXiv preprint arXiv:2310.20470*. https://doi.org/10.18653/v1/2023.findings-emnlp.382

Du Bois, John, Susanna Cumming, Stephan Schuetze-Coburn, and Danae Paolino. 1993. Outline of discourse transcription. In Jane A. Edwards and Martin D. Lampert, editors, *Talking Data: Transcription and Coding in Discourse*. Lawrence Erlbaum Associates, pages 45–89.

Ellison, T. Mark and Aung Si. 2021. A quantitative analysis of age-related differences in Hindi–English code-switching. *International Journal of Bilingualism*, 25(6):1510–1528. https://doi.org/10.1177/13670069211028311

Fricke, Melinda, Judith F. Kroll, and Paola E. Dussias. 2016. Phonetic variation in bilingual speech: A lens for studying the production–comprehension link. *Journal of Memory and Language*, 89:110–137. https://doi.org/10.1016/j.jml.2015.10.001, PubMed: 27429511

Gambäck, Björn and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

Goh, K.-I. and A.-L. Barabási. 2008. Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002. https://doi.org/10.1209/0295-5075/81/48002

Green, David W. and Li Wei. 2014. A control process model of code-switching. *Language, Cognition and Neuroscience*, 29(4):499–511. https://doi.org/10.1080/23273798.2014.882515

Guzmán, Gualberto A., Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Interspeech*, pages 67–71. https://doi.org/10.21437/Interspeech.2017-1429

Himmelmann, Nikolaus P. 2022. Prosodic phrasing and the emergence of phrase structure. *Linguistics*, 60(3):715–743. https://doi.org/10.1515/ling-2020-0135

Himmelmann, Nikolaus P., Meytal Sandler, Jan Strunk, and Volker Unterladstetter. 2018. On the universality of intonational phrases: A cross-linguistic interrater study. *Phonology*, 35(2):207–245. https://doi.org/10.1017/S0952675718000039

Inbar, Maya, Shir Genzer, Anat Perry, Eitan Grossman, and Ayelet N. Landau. 2023. Intonation units in spontaneous speech evoke a neural response. *Journal of Neuroscience*, 43(48):8189–8200. https://doi.org/10.1523/JNEUROSCI.0235-23.2023, PubMed: 37793909

Kootstra, Gerrit Jan, Ton Dijkstra, and Janet G. van Hell. 2020. Interactive alignment and lexical triggering of code-switching in bilingual dialogue. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.01747, PubMed: 32793070

Labov, William. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh and Joel Sherzer, editors, *Language in Use: Readings in Sociolinguistics*. Prentice Hall, pages 28–53.

Lipski, John M. 1978. Code-switching and the problem of bilingual competence. In M. Paradis, editor, *Aspects of Bilingualism*. Hornbeam Press, pages 250–264.

Mettouchi, Amina. 2008. Kabyle/French codeswitching: A case study. In M. Lafkioui and V. Brugnatelli, editors, *Berber in Contact: Linguistic and Sociolinguistic Perspectives*. Rüdiger Köppe Verlag, pages 187–198.

Muysken, Pieter. 2000. *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Muysken, Pieter. 2015. Dé jà voodoo or new trails ahead? Re-evaluating the mixing typology model. In R. Torres Cacoullos, N. Dion, and A. Lapierre, editors, *Linguistic Variation: Confronting Fact and Theory*. Routledge, pages 242–261.

Myers-Scotton, Carol. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press.

Ono, Tsuyoshi and Sandra A Thompson. 2020. *The 'Noun Phrase' Across Languages: An Emergent Unit in Interaction*, volume 128. John Benjamins Publishing Company. https://doi.org/10.1075/tsl.128

Pattichis, Rebecca, Dora LaCasse, Sonya Trawick, and Rena Torres Cacoullos. 2023. Code-switching metrics using intonation units. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16840–16849. https://doi.org/10.18653/v1/2023.emnlp-main.1047

Pfaff, Carol W. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, 55:291–318. https://doi.org/10.2307/412586

Poplack, Shana. 1980. Sometimes I'll start a sentence in Spanish y termino en espanol: Toward a typology of code-switching. *Linguistics*, 51(s1):11–14. https://doi.org/10.1515/ling-2013-0039

Poplack, Shana. 1988. Contrasting patterns of code-switching in two communities. Heller, M., editor, *Codeswitching: Anthropological and Sociolinguistic Perspectives*. Mouton de Gruyter, pages 215–244. https://doi.org/10.1515/9783110849615.215

Poplack, Shana. 1993. Variation theory and language contact. In Dennis R. Preston, editor, *American Dialect Research*. John Benjamins, pages 251–286. https://doi.org/10.1075/z.68.13pop

Poplack, Shana. 2015. Code switching: Linguistic. *International Encyclopedia of the Social & Behavioral Sciences*, 3:918–925. https://doi.org/10.1016/B978-0-08-097086-8.53004-9

Poplack, Shana. 2017. *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford University Press. https://doi.org/10.1093/oso/9780190256388.003.0004

Poplack, Shana. 2022. 16 Data management at the uOttawa Sociolinguistics Laboratory. *The Open Handbook of Linguistic Data Management*, page 209. https://doi.org/10.7551/mitpress/12200.003.0021

Poplack, Shana and Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International Journal of Bilingualism*, 2(2):127–138. https://doi.org/10.1177/136700699800200201

Sankoff, David. 1998. The production of code-mixed discourse. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. https://doi.org/10.3115/980451.980848

Sankoff, David and Shana Poplack. 1981. A formal grammar for code-switching. *Research on Language & Social Interaction*, 14(1):3–45. https://doi.org/10.1080/08351818109370523

Shenk, Petra. 2006. The interactional and syntactic importance of prosody. *International Journal of Bilingualism*, 10:179–205. https://doi.org/10.1177/13670069060100020401

Solorio, Thamar and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981. https://doi.org/10.3115/1613715.1613841

Torres Cacoullos, Rena. 2020. Code-switching strategies: Prosody and syntax. *Frontiers in Psychology*, 11:2130. https://doi.org/10.3389/fpsyg.2020.02130, PubMed: 33071841

Torres Cacoullos, Rena, Nathalie Dion, Dora LaCasse, and Shana Poplack. 2022. How to mix: Confronting "mixed" NP models and bilinguals' choices. *Linguistic Approaches to Bilingualism*, 12(5):628–656. https://doi.org/10.1075/lab.20017.tor

Torres Cacoullos, Rena and Dora LaCasse. 2025. Bilingual clause combining: A variable equivalence hypothesis for conjunction choice. *International Journal of Bilingualism*, 29(5):1202–1218. https://doi.org/10.1177/13670069241265587

Torres Cacoullos, Rena and Catherine E. Travis. 2018. *Bilingualism in the Community: Code-switching and Grammars in Contact*. Cambridge University Press. https://doi.org/10.1017/9781108235259

Torres Cacoullos, Rena and Jessica Vélez Avilés. 2023. Mixing adjectives: A variable equivalence hypothesis for bilingual word order conflicts. *Linguistic Approaches to Bilingualism*, 14:5:609–639. https://doi.org/10.1075/lab.22038.tor

Winata, Genta, 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978. https://doi.org/10.18653/v1/2023.findings-acl.185

Wintner, Shuly, Safaa Shehadi, Yuli Zeira, Doreen Osmelak, and Yuval Nov. 2023. Shared lexical items as triggers of code switching. *Transactions of the Association for Computational Linguistics*, 11:1471–1484. https://doi.org/10.1162/tacl_a_00613