NEW MEXICO SPANISH-ENGLISH CODESWITCHING

UCLA

Code-Switching Metrics Using Intonation Units

Rebecca Pattichis, University of California, Los Angeles, pattichi@g.ucla.edu
Dora LaCasse, University of Montana, dora.lacasse@mso.umt.edu
Sonya Trawick, Pennsylvania State University, sonyatrawickpsu@gmail.com
Rena Torres Cacoullos, Pennsylvania State University, rena@psu.edu

Introduction

Code-switching (CS): going back and forth between languages within a speaker turn.

NLP word-level metrics allow, for any four-word sequence:

- 1 switch: $w_{L1}w_{L1}w_{L2}w_{L2}$
- 3 switches: $w_{L1}w_{L2}w_{L1}w_{L2}$

Problems:

1. CS is **not** equally likely between any two words, and
2. single-word incorporations and multi-word strings are **not** created equal.

it was a general store,	'it was a general store,
vendían de todo.	they sold everything.'
	[03, 30:10-30:12]

Example of Across IU CS

y para nosotros *it was a snap*,	'and for us *it was a snap*,'
	[10, 01:23-01:24]

Example of Within IU CS

Syntactic-Prosodic CS Patterns

Equivalence Constraint (EC): CS is avoided at points of word order incompatibility (Poplack, 2013:586; Sankoff, 1998).
Intonation Units (IUs): speech segments "uttered under a single, coherent intonation contour" (Du Bois et al., 1993:47).
IU-Boundary Constraint: CS is favored across IU boundaries (cf. Torres Cacoullos and Travis, 2018: 51).

Lone Items vs. Multi-Word Strings

Lone items: are disproportionately nouns, are placed according to the word order of the surrounding matrix language, and participate in the constructions of that language.
Multi-word strings: are placed at cross-language equivalence points, while the internal constitution of each string is consistent with the grammar of its respective language.

New Mexico Span-Eng Bilingual (NMSEB) Corpus

16,957 prosodic sentences (43% English, 42% Spanish, 15% both); 5 transcribed recordings (4.8 hrs, ~48k words)

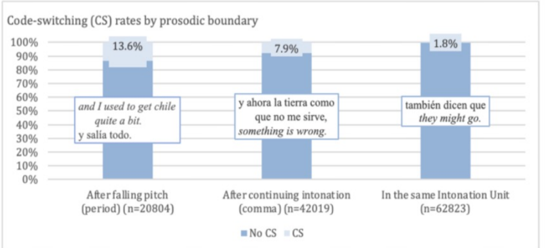


Fig. 1: IU-Boundary Constraint: CS is four times more likely at the boundary of IUs than within them for NMSEB *(adapted from Trawick, 2022: 3.4).*

CS Metrics

Multilingual Index (M-Index) (Barnett et al., 2000):
k is the number of languages,
p_j is the number of tokens in language j over the total number of tokens in the corpus:

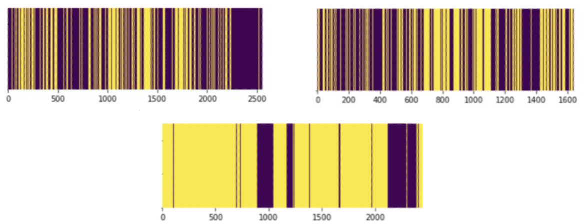
$$\text{M-Index} = \frac{1 - \sum p_j^2}{(k-1) \cdot \sum p_j^2}.$$

We only consider IUs eligible if they contain 'S' or 'E' language tags.
Integration Index (I-Index) (Guzman et al., 2017): the probability of CS in a corpus at any given token boundary. n is the number of tokens, and $S(l_i, l_j)$ is 1 if there's a switch, 0 otherwise:

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

Across-IU I-Index: We record the binary measure: is there a switch between the ith and jth IU?
Within-IU I-Index: We record the binary measure: is there a switch within the ith IU?
We also consider two perspectives: one with only eligible IUs containing 'S' and/or 'E', and one also including 'L' (lone items).

Results



Figures 2-4: IU-based language distribution graphs elucidate CS metrics. English IUs are in purple, Spanish are yellow. M-Index is depicted by extent of each color, I-Index (Across-IU) by width (or number) of bands. Compare speakers 03 and 10 (top) vs. 05 (bottom).

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	0.57	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 1: M- and I-Index for 5 bilinguals from NMSEB.

M- and I-Indexes are independent: compare 16 and 10.
Across-IU I-Index is greater than Within-IU I-Index for all speakers.
Lone items have little impact on Across-IU I-Index but substantial impact on Within-IU I-Index (as large as 300% for 03).

Conclusion

1. All speakers disfavor within-IU multi-word switching, regardless of speakers' M- or I-Indexes.
2. Bilinguals share the preference for CS to occur across IU boundaries (IU-Boundary constraint). This is blurred when lone items are not distinguished.
3. IUs are a vital unit of analysis in future development of CS datasets.

Acknowledgements

NSF support to Rena Torres Cacoullos and Catherine Travis (BCS-1019112/1019122) and to Rena Torres Cacoullos and Shana Poplack (1624966) is gratefully acknowledged.