Aggregating small data sets to explore code-switching in a big way

**Barbara E. Bullock, Elizabeth Green, Jacqueline Serigos,
Vivek Sharath and Almeida Jacqueline Toribio**

The University of Texas

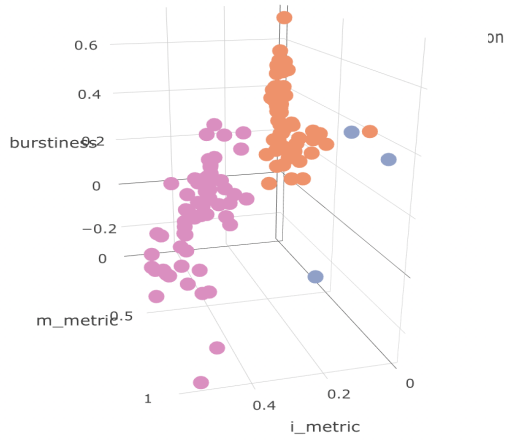
The Problem: Linguistic research on multilingual code-switching (CS) has been primarily concerned with identifying the grammatical sites of language alternation within a single clause (Poplack 1980; Sankoff & Poplack 1981; Woolford 1984; DiSciullo et al. 1986; Mahootian 1993; Belazi et al. 1994; Myers-Scotton 1993, 2002; Dussias 2003; MacSwan 1999, 2009; Cantone & Mueller 2008; Liceras et al. 2008; Herring et al. 2010; González-Vilbazo & López 2011; López 2018). While such a fine-grained view has its merits in providing insights about discrete aspects of morphosyntax, this narrow lens is incapable of testing broader claims about CS, such as the mixing typology theorized in Muysken (2000) that includes two main types: *insertion,* where a guest language sequence is embedded into the grammar of a matrix language; *alternation,* where the grammars of each language frame the structure of the utterance; and *congruent lexicalization*, a combination of the two, said to occur in typologically similar pairings (Muysken 2014). While claims about the factors that condition CS patterns are widely assumed, they have not been tested in a replicable way due to the paucity of multilingual data and the lack of metrics to numerically gauge CS. Here, we present methods to computationally characterize CS to compare between corpora and we demonstrate how small data sets can be aggregated to overcome data shortages (Guzmán et al. 2016, 2017; Bullock et al. 2017, 2018). Our goal is to quantitatively compare CS across bilingual Romance corpora and to model whether or not typological similarity or other factors affect the degree of language mixing.

Methods: For this analysis, we have compiled available mixed language Romance corpora including: 60 Macaronic verses (3723 lines) of various authors spanning from the 15th - 20th centuries compiled by Demo (2018), where Latin is mixed with Romance, Germanic, or Slavic languages; two contemporary 'Spanglish' novels, ~40K word *Killer Crónicas* (Chávez-Silverman 2004) and ~58K word *Yo, Yo, Boing!* (Braschi 1998); the transcript of the ~13K word bilingual French+English Quebecois film, *Bon Cop Bad Cop* (2006); and three transcripts of US-Spanish+English oral conversations: ~8K word *S7* (Solorio & Liu 2008), each of the subcorpora in the ~240K word *Miami* corpus (Deuchar 2010), and ~507K word Spanish in Texas Corpus (Bullock & Toribio 2013). Each word token in every corpus is labelled with the language from which it is drawn and consecutive words bearing the same language label are combined into spans of length *n* via a python script. From the count of word labels and the distribution of span lengths, we compute measures of the ratio of languages (i), the degree of CS between them (ii), and the intermittency of CS (iii) as:

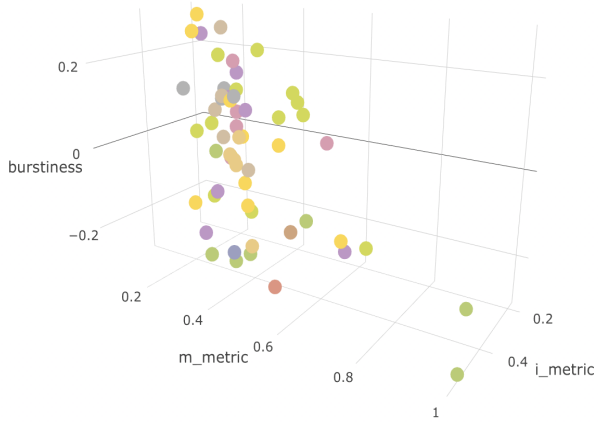
(i) the inequality of distribution of languages in a text, or *M-Index* (Barnett et al. 2000);
(ii) the probability of switching between languages, or *I-Index* (Guzmán et al. 2016, 2017);
(iii) the randomness of the switching, or *Burstiness,* computed from the mean and standard deviation of span lengths (Bullock et al. 2017, 2018, based on Goh & Barabási 2008).

Results: We fit separate linear models to predict the value of each of the metrics above with independent factors of Corpus, matrix language in contact and typology of embedded language for the Macaronic subcorpora (Romance, Germanic, Other), and Genre (Poetry, Fiction, Conversation). Of these predictors, only Genre served to group these corpora visualized in the 3D plot of the Full Corpus below. As illustrated, poetry is significantly more likely to show higher I-Index and less Burstiness than either the conversations or the fiction. The conversations are significantly less likely to show an even distribution of languages than the other genres, as reflected by the lower values along the M-Index dimension.

3-D Plot of Full Corpus by Metrics



3-D Plot Macaronic Sub Corpora by Metrics



As can be observed from the 3D plot of the Macaronic subcorpora, neither the language nor the language family embedded in Latin serves to separate out the corpora neatly. In other words, it is not the case that Romance languages mix more freely with the Latin matrix, as would be predicted by Muysken's congruent lexicalization hypothesis.

Discussion and Implications: Overall, the results demonstrate that there is a continuum along each dimension of CS rather than discrete categories and only genre, not language typology, contributes in a significant way to predicting the degree of mixing expected in a corpus. The methods and approaches to quantifying mixed Romance data that we have outlined here, and available on our github page, are still under evaluation and testing. Ongoing work seeks to determine if there are expected 'constants' of conversational versus performative (poetry, fiction) language mixing.