Moving code-switching research toward more empirically grounded methods

Gualberto A. Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara Bullock &
Almeida Jacqueline Toribio
University of Texas at Austin

{gualbertoguzman,joseph.ricard,jserigos}@utexas.edu
{bbullock,toribio}@austin.utexas.edu

Abstract

As our world becomes more globalized and interconnected, the boundaries between languages become increasingly blurred (Bullock, Hinrichs & Toribio, 2014). But to what degree? To date, researchers have no objective way to measure the frequency and extent to which languages might be mixed. While Natural Language Processing (NLP) tools process monolingual texts with very high accuracy, they perform poorly when multiple languages are involved. In this paper, we offer an automated language identification system and intuitive metrics–the Integration, Burstiness, and Memory indices–that allow us to characterize how corpora are mixed.

1 Introduction

When multilinguals are in interaction with one another, some degree of language mixing is likely to take place (talk & Choudhury, 2016). Indeed, the phenomenon has been attested since the ancient world (Adams, 2002) and is prevalent in contemporary societies worldwide. Code-switching (C-S), defined as the alternation of languages within the same speech event (Bullock & Toribio, 2009) is generally an oral practice that occurs in informal speech (Example 1), but it is increasingly found in written form on social media platforms (Example 2) and has gained acceptance in prose (Example 3) and on television and film (Example 4).

1. I guess, mi closest companion siempre ha sido Raúl [Spanish in Texas Corpus, Bullock & Toribio, 2013; Toribio & Bullock, 2016]

2. diana@Dianier1019 Oct26 for some reason I'm starting to talk Spanglish like I'll start off talking American despues mi mexicana quiere salir [Twitter]

3. . . . but she had the posture and speech (and arrogance) of una muchacha respetable [Junot Díaz, Brief Wondrous Life of Oscar Wao]

4. Après, c'était bien easy de l'embarquer pour tuer les autres fuckers... qui ont détruit notre great game. [*Bon Cop Bad Cop*] (Ball et al. 2015)
   'After, it was real easy to set out to kill the other fuckers... who destroyed our great game.'

For those interested in the forms, meanings, and dispersion of multilingual language use, observing variation in C-S in reliable, reproducible, and language independent ways is essential. In seeking to understand C-S, it would be advantageous to have the ability to compare the frequency and the degree to which the languages represented in different corpora are intermingled. Herein we present our methods for quantifying and visualizing language mixing in corpora and apply our methods to the analysis of mixed language texts of various genres and of different language pairings. Our contributions in this paper are as follows: (i) to provide a brief explanation of the models that we built to identify the language of each word token in a corpus; (ii) to describe the metrics that we use to calculate and to visualize the frequency and degree of language mixing found in a corpus and sub-corpora; (iii) to describe the corpora that we model; and (iv) to demonstrate the application of the metrics to these corpora to quantify and visualize the results. We conclude with implications for future work in the digital humanities, in linguistics, and in NLP research.

2 Methods

Language Identification. Corpora may contain more than one language for a variety of reasons, including a change in author from one sub-corpus to the next (King & Abney, 2013) or the presence of classic or composite C-S (Myers-Scotton, 1993) as illustrated by examples 1 through 4, a challenge for NLP approaches. Language identification systems were originally built to automatically recognize the language of a text and work best when the text is assumed to contain one and only one language. For this reason, more complex language identification systems must be employed to process texts in which languages are mixed by a given author or speaker.

Our system is an adapted version of the language identification system of Solorio & Liu (2008a, 2008b). It produces two tiers of annotation—*Language* (English(ENG), Spanish(SP)/French(FR), Punctuation, or Number) and *Named Entity* (yes or no). In accord with Çentinoğlu (2016), we annotate Named Entities for language because they can be language dependent (e.g., Ciudad de México versus Mexico City) in which case they may act as triggers for code-switching (Broersma & DeBot, 2006). For tokens not identified as punctuation or number, we use a 5-gram character n-gram trained at the character level and a first order Hidden Markov Model (HMM) trained on language token bigrams to determine the most probable language of the token. Our SP-ENG model was trained on film subtitle corpora of roughly equal sizes.The FR-ENG data were trained on a French Canadian newspaper corpus (La Presse). When tested against our manually annotated

gold-standards, our models achieved accuracy rates of 95% for SP-ENG and 97% for FR. These accuracy ratings do not deviate substantially from those of human annotators (Goutte et al., 2016).

The Integration Index. Barnett et al. (1999) developed the Multilingual Index (M-Index) to quantify the ratio of languages in oral speech corpora based on the Gini coefficient to measure inequality of a distribution [1]. The values range from 0 (monolingual) to 1 (perfectly balanced between two or more languages), permitting a measure of how 'monolingual' or, for present purposes, 'bilingual' a given text is. The M-index is calculated as follows where k is the total number of languages represented in the corpus, p_j is the total number of words in the language j over the total number of words in the corpus, and j ranges over the languages present in the corpus:

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

To supplement the M-Index, we created the Integration-index, a metric that describes the probability of switching within a text (Guzmán et al. 2016) (see also Gambäck & Das (2014, 2016)). We calculate the I-index from summing up the probabilities that there has been a language switch (from Lang1 $\rightarrow$ Lang 2 or vice-versa). The values of the I-Index range from 0 (a monolingual text in which no switching occurs) to 1 for a text in which every other word comes from a different language, i.e. every word represents a switch in language. Given a corpus composed of tokens tagged by language $\{l_i\}$ where i ranges from 1 to n, the size of the corpus, and j ranges from $i+1$ to n, the I-index is calculated by the following expression:

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

where $S(l_i, l_j) = 1$ if $l_i \neq l_j$ and 0 otherwise, and the factor of $1/(n-1)$ reflects the fact that there are $n-1$ possible switch sites in a corpus of size n.

Muysken (2000) presents a typology of mixing, identifying three types of patterns: insertion, in which an other-language item is inserted within the a string of a base language (A A A B A A), alternation, in which the base language changes (A A A B B B), and congruent lexicalization, in which the structures of the two contributing languages overlap (A\B A\B A\B) so that either language can occupy a position in a string. The M-index and the I-index are calculated at the lexical level, which does not capture the contribution of syntax. Nonetheless, we use the I-index as a proxy measure of how much CS is in a document, where the value 0 represents a monolingual text with no switching and 1 a text in which every word switches language, a highly unlikely real-world situation. It is an empirical question whether or not there is a threshold of integration beyond which a C-S is perceived as inauthentic.

[1] A reviewer points out that Shannon entropy may also be used for measuring diversity in text.

Intermittency. To refine our profile of C-S within a corpus, we utilize measures of intermittency from research on complex systems (Goh & Barabási, 2008). Measures of burstiness and memory together provide a picture of the frequency and the time order of C-S. We define a *switch point* as an instance when there is a switch between languages and a *language span* as a stretch of discourse between switch points. The language span distribution, an aggregate of all the spans in the corpus, approximates a probability distribution that returns the probability of how long a speaker/text will stay in one language before switching to the next. This distribution can be compared to the Poisson distribution in which the likelihood of a switch is assumed to be random. Burstiness defines how much the language span distribution differs from the Poisson distribution; in other words, how non-random the switching activity is. In simple terms, burstiness describes whether switching occurs in spurts or more regularly. The Burstiness-index is bounded within [-1,1]: An anti-bursty signal that repeats regularly, like a heartbeat, receives a value closer to -1, whereas a bursty signal is irregular and appears closer to 1. Burstiness is calculated as:

$$\text{Burstiness} \equiv \frac{(\sigma_\tau/m_\tau - 1)}{(\sigma_\tau/m_\tau + 1)} = \frac{(\sigma_\tau - m_\tau)}{(\sigma_\tau + m_\tau)},$$

where σ_τ is the standard deviation of the language spans and m_τ is the mean of the language spans.

Burstiness, by considering the length of these language spans, provides one measure of the intermittency of C-S. However, the ordering of these language spans in time is important, as it is possible for two corpora to have identical language span distributions—and thus the same Burstiness-index—that nonetheless appear very different to a reader due to how the switch points are ordered in each corpus. In Goh & Barabási's system, this is measured as Memory, a measure of first order autocorrelation between the language spans. The computation of memory involves going through the language spans in order, measuring the extent to which the length of one language span is influenced by the length of the previous language span. Memory is calculated as:

$$\text{Memory} \equiv \frac{1}{n_r - 1} \sum_{i=1}^{n_r - 1} \frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1 \sigma_2},$$

where n_r is the number of language spans in the distribution, τ_i is the current language span, τ_{i+1} is the language span after τ_i, σ_1 is the standard deviation of all language spans but the last one, σ_2 is the standard deviation of all language spans but the first, m_1 is the mean of all language spans but the last, and m_2 is the mean of all language spans but the first. Memory is bounded within [-1, 1]: a signal closer to -1 indicates that the language spans are negatively autocorrelated, meaning that spans of discourse in one language tend not to be similar in length to the span of discourse in the language preceding it. That is, long spans are followed by short spans and short spans are followed by long spans. Conversely, a signal closer to 1 indicates that the language spans are positively autocorrelated, meaning that

the span of discourse in one language tends to be similar in length to the span of discourse in the language preceding it. In summary, Memory and Burstiness are mechanisms that give a complete signature of the intermittency—the time order and frequency—of C-S for a corpus, allowing for meaningful comparison of C-S behavior between corpora. It is important to note that this method is not exclusive to C-S behavior, and is a time series analysis that may be applied more generally to any type of *events* that may occur in corpora. The crux of the strategy is to iterate over the corpus, marking when the events occur, thereby generating the distribution of time spans between the events. The memory and burstiness metrics then describe the intermittency of that event.

Data and Analysis. The data that we analyzed comprises three texts of distinct genres and languages, each of which is touted for its bilingualism. The first is the film transcript of the FR-ENG bilingual buddy movie *Bon Cop Bad Cop* (BCBC) (2006). The French and English versions of the transcript were downloaded from subtitles.com and the final transcript (n = 13,502 words) was pieced together by watching the film frame by frame and choosing the appropriate language from the subtitles. The other two are Spanish—English written texts that are available online: *Killer Crónicas* (KC) is an 40,469-word novel by Susana Chávez-Silverman in multilingual (and multi-dialectal) emails that present extensive SP-ENG C-S, and *Yo-Yo-Boing* (YYB) is 58,494-word novel comprised of alternating and mixed SP-ENG poetry, and essays by Giannina Braschi. We annotated each text for language and quantified the switching as outlined above.

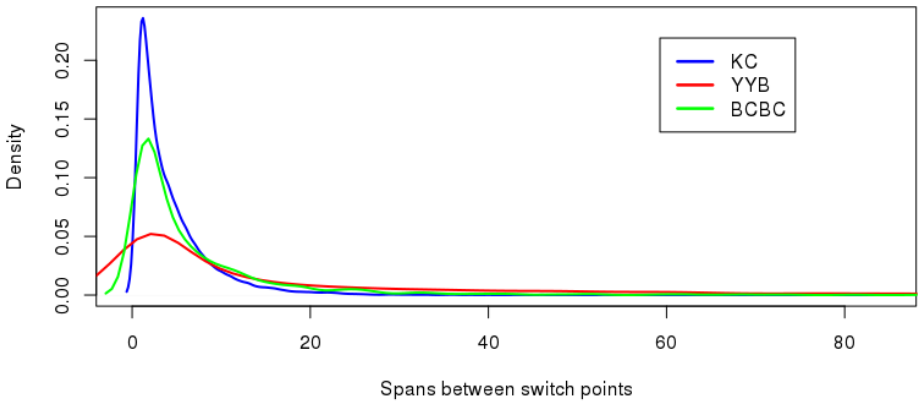
3 Results

Table 1: Language Span Density Metrics by Corpus

Corpus	M-Metric (Mixing)	I-Metric (Integration)	Burstiness	Memory
KC	0.9868	0.2298	0.0156	-0.0280
YYB	0.9528	0.0345	0.3695	-0.1194
BCBC	0.8651	0.1039	0.4362	-0.0581

The results of these metrics as applied to the three texts are found in Table 1. A comparison of the M-index for these texts reveals that the novels YYB and KC are nearly equally balanced between SP and ENG, with M-index values that are close to 1; the film, BCBC, with an M-Index of 0.86, is less balanced between languages than the novels. The I-index serves to differentiate the two balanced texts and indicates that the languages are more closely integrated in KC than in YYB despite their similar M-indices. BCBC shows an integration value that is intermediate between YYB and KC. In terms of burstiness, BCBC has the highest value of the three texts, indicating that there is not a regular pattern to the C-S but rather there

Figure 1: Language Span Density by Corpus



are moments in the film in which characters switch languages frequently followed by moments where little switching occurs. Overall, the Spanish—English novels are very different from one another; while YYB shows bursts of C-S throughout the text, the low Burstiness value for KC shows that C-S occurs with regularity throughout the text. Finally, both KC and BCBC, texts in which the probability of C-S is relatively high compared to YYB show a neutral value for memory, which appears to be the normal complexity measure for texts (Altmann et al. 2009), whereas YYB shows a more negative memory-index entailing that the spans between switching repeat at more regular intervals. The nature of mixing in the three texts can be visualized by the density plot in Figure 1. KC's Integration-index reflects the highest incidence of short, switched spans in each language, relative especially to YYB, and KC's low Burstiness-index suggests that this type of C-S remains constant throughout. YYB's low Integration-index and high Burstiness-index follows from the alternation of monolingual-English, monolingual-Spanish, and mixed-language chapters, and its higher negative Memory-index depicts a sequencing of long and short periods between switch points compared to the more neutral, regular pattern of bursts in KC and in BCBC.

4 Discussion & Conclusion

The metrics that we have proposed and tested here are useful for distinguishing the types of mixing patterns found in corpora. They tell us, for instance, that any random selection from KC, but not YBB, is likely to contain frequent switching events since the text is characterized by short spans between switching events that recur regularly. The Canadian movie, BCBC, would also be a good candidate for the study of C-S but because switching is burstier relative to KC, one would need a larger sample of that text than of KC in order to capture language alternation. Finally it is much less probable that choosing a random section from YYB would

yield any switching phenomenon because there are long spans within the book in which no C-S occurs. These methods and models can be applied to any language-tagged corpora in which more than one language appears. This would allow us to compare patterns of language mixing across various corpora in a standard and reliable way, a task that cannot currently be achieved in a straightforward fashion. Additionally, these metrics enable scholars from any discipline in the humanities to visualize their data before they begin to analyze or model it. Since these measures quantify the actual frequency and degree to which languages are intermixed in a sample, they may aid in dispelling popular (and sometimes scholarly) misconceptions about the nature and extent of C-S among multilingual societies, communities, and individuals. In our future work, we intend to compare across corpora produced with the same language pairings, for example, to quantify and visualize the differences between the 'Spanglishes' of Miami, El Paso, Los Angeles, and New York, and to compare these, in turn, to 'Hinglish' (Hindi—English) corpora in India and England and to French—Arabic in Europe and the Maghreb with the intent to model the variation inherent in code-switching worldwide.

References

[1] James N. Adams, Mark Janse, and Simon Swain. *Bilingualism in ancient society: Language contact and the written word*. Oxford University Press on Demand, 2002.

[2] Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *CoRR*, abs/0901.2349, 2009.

[3] Kalika Bali and Monojit Choudhury. NLP for code-switching: why more data is not necessarily the solution. In *Empirical Methods in Natural Language Processing (EMNLP)*. The Association of Computational Linguistics, 2016.

[4] Kelsey Ball, Barbara E. Bullock, Gualberto Guzmán, Rozen Neupane, Kristopher S. Novak, and Jacqueline L. Serigos. Bon cop, bad cop: A tale of two cities. In *Transcultural Urban Spaces*, 2015.

[5] Ruthanna Barnett, Eva Codo, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. The LIDES Coding Manual: A document for preparing and analyzing language interaction data Version 1.1–July, 1999. *International Journal of Bilingualism*, 4(2):131–132, June 2000.

[6] Mirjam Broersma and Kees De Bot. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and cognition*, 9(01):1–13, 2006.

[7] Barbara E. Bullock, Lars Hinrichs, and Almeida J. Toribio. World Englishes, code-switching, and convergence. *The Oxford Handbook of World Englishes, Oxford University Press, Oxford, England*, 2014.

[8] Barbara E. Bullock and Almeida J. Toribio. *The Cambridge handbook of linguistic code-switching*. Cambridge University Press, 2009.

[9] Ozlem Cetinoglu. A Turkish-German Code-Switching Corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4215–4220, 2016.

[10] Junot Díaz. *The brief wondrous life of Oscar Wao*. Penguin, 2007.

[11] Björn Gambäck and Amitava Das. On Measuring the Complexity of Code-Mixing. In *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, pages 1–7, 2014.

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

[13] K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008.

[14] Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. Discriminating similar languages: Evaluations and explorations. *arXiv preprint arXiv:1610.00031*, 2016.

[15] Gualberto Guzman, Barbara E. Bullock, Jacqueline Serigos, and Almeida J. Toribio. Simple tools for exploring variation in code-switching for linguists. In *Empirical Methods in Natural Language Processing (EMNLP)*. The Association for Computational Linguistics, 2016.

[16] Ben King and Steven Abney. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL-HLT*, pages 1110–1119, 2013.

[17] Pieter Muysken. *Bilingual speech: a typology of code-mixing*. Cambridge University Press, Cambridge, 2000.

[18] Carol Myers-Scotton. *Duelling languages: grammatical structure in codeswitching*. Oxford University Press (Clarendon Press), Oxford, 1993.

[19] Thamar Solorio and Yang Liu. Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics, 2008.

[20] Thamar Solorio and Yang Liu. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics, 2008.

[21] Almeida J. Toribio and Barbara E. Bullock. A new look at heritage Spanish and its speakers. *Advances in Spanish as a Heritage Language*, 49:27–50, 2016.