INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Metrics for modeling code-switching across corpora

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos,
Barbara E. Bullock, Almeida Jacqueline Toribio

University of Texas at Austin, United States
{gualbertoguzman,joseph.ricard,jserigos}@utexas.edu
{bbullock,toribio}@austin.utexas.edu

Abstract

In developing technologies for code-switched speech, it would be desirable to be able to predict how much language mixing might be expected in the signal and the regularity with which it might occur. In this work, we offer various metrics that allow for the classification and visualization of multilingual corpora according to the ratio of languages represented, the probability of switching between them, and the time-course of switching. Applying these metrics to corpora of different languages and genres, we find that they display distinct probabilities and periodicities of switching, information useful for speech processing of mixed-language data.

Index Terms: multilingual, code-switching, burstiness

1. Introduction

Code-switching (C-S), exemplified by the alternation of Belizean Spanish-English in (1), has garnered much attention within the discipline of linguistics and allied speech sciences [1].

(1) Well <ése es uno de nuestros> major celebrations,
<y los >friends <que yo tengo ahorita, todos nosotros
somos> Carnival people.
'Well that is one of our major celebrations, and the friends
I have now, we're all Carnival people.'

A common practice in multilingual communities and one that is reflected across different media, linguists have sought to identify the lexical categories or clausal sites that are most predictive of C-S at the sentence level [2, 3]. Others have examined the conventionalization of C-S patterns as a function of individual, discursive, and social practices and explored the possibility that C-S may lead to the emergence of fused languages [4]. With minor exceptions, studies of C-S have been qualitative, focused on the features of the specific contact grammars or multilingual communities that promote or inhibit mixing. As a consequence, we lack a language-independent, typological perspective on how and how often languages are mixed. Information about the probability and degree of C-S within a particular language context (e.g., Hindi-English in the U.K. vs. India), medium (e.g., Twitter vs. talk), or time frame (e.g., leisure vs. work hours), could prove useful in improving the application of speech technologies for multilingual speech.

The global rise of social media such as Facebook, Twitter, SMS, and Usenet newsgroups has afforded large quantities of user-generated data that incorporates C-S [5, 6, 7, 8, 9]. However, the occurrence of multiple languages within a single text presents significant complexity for automated processing. Language identification and downstream tasks such as POS tagging and syntactic parsing, which are successfully performed for canonical monolingual data, are challenging when mixed text is encountered.Thus, C-S, though ubiquitous worldwide, remains a 'low resource' speech variety [10].

Because the internal characteristics of mixed-language corpora potentially affect tasks in different ways [11, 12, 13], it is useful to anticipate how the languages are distributed [14]: Does the document contain alternating monolingual texts from different languages, as in the Europarl parallel corpus [15], or is there more classic' C-S in the sense that a single speaker is employing multiple languages within an utterance [3]? It is additionally instructive to anticipate the typology of mixing in a corpus [16]: Is the mixing limited to lone lexical items and multi-word expressions, as in the Argentine newspress [17], or can it occur word-internally, as in Turkish-German Tweets [7]? Does mixing occur in bursts, as observed among working French women who engage in French-Arabic-Berber C-S significantly more in the morning than during working or evening hours [18]? Discriminating among different types of mixing, including its time course, can allow for the development and improvement of techniques for automatically processing mixed-language data. Here we present multiple methods for modeling corpora that respond to this need.

We introduce several metrics for quantifying and visualizing corpora according to the degree of mixing and the burstiness with which switching occurs. We use language identification at the word level to compute language frequency and probability for mixing, to calculate language span entropy over the time-course of the text, and to visualize the time-course of switching events. We exemplify the utility of these metrics by applying them to corpora of different genres that we know to be mixed differently: popular prose novels, spoken speech transcripts, and a film transcript.

2. Related Work

Predicting C-S is important for modeling multilingual speech in NLP [19, 20, 21], in TTS [22, 10], and in ASR [23, 24, 12, 25, 26, 27, 28, 29]. Of particular importance for our work are findings from ASR that indicate that individual speakers or speakers from different nationalities show different patterns of Mandarin-English switching. These differential patterns are referred to as "code-switching attitude" by Vu et al. [12], who demonstrate improved performance for adapted models. Similar findings with regard to number pronunciation have been reported by Molapo & Barnard [30], who show that Shona speakers used English for number pronunciation while other Bantu speakers used their lingua francas, information that can be used in normalizing numbers.

http://dx.doi.org/10.21437/Interspeech.2017-1429

While C-S attitude examines variation within corpora, Gambäck & Das [13, 31] examine what they refer to as complexity across corpora. They are motivated by the notion that more frequent switching is more complex and creates additional challenges for language processing technologies. Guzmán et al. [32] have similar aims, focusing on producing methods for visualizing language integration via a language signature or profile. Guzmán et al. [33] expand on that work by introducing measures that apply to the time-course of switching.

3. Metrics and Motivation

3.1. Ratio

3.1.1. M-Index

The Multilingual Index (M-index), developed by Barnett et al. [34] from the Gini coefficient, is a word-count-based measure that quantifies the inequality of the distribution of language tags in a corpus of at least two languages. The M-index is calculated as follows, where $k > 1$ 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:

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

The index is bounded between 0 (monolingual corpus) and 1 (each language in the corpus is represented by an equal number of tokens).

3.1.2. Language Entropy

In an effort to present metrics that are consistent with information theory, we also define the Shannon entropy of this language tag distribution as an alternative to the M-index. For our purposes, we describe this metric as the *language entropy* of a corpus. The language entropy returns how many bits of information are needed to describe the distribution of language tags.

Using the same conventions of notation as previously defined, language entropy is calculated as

$$LE = -\sum_{j=1}^{k} p_j \log_2(p_j) \tag{3}$$

and is bounded from below by 0 (representing a completely monolingual text) and bounded from above by

$$-\sum_{j=1}^{k} \frac{1}{k} \log_2(\frac{1}{k}) = \log_2(k), \tag{4}$$

which is the maximum entropy for a corpus with k languages (and, in such a case, each language is represented equally). In the case of two languages, the M-index and LE can be derived from one another.

3.1.3. Probability of Switching (I-index)

These metrics based on language ratio, while useful, do not describe the frequency of C-S behavior. As a supplement to the language ratio metrics, we created the Integration-Index, a metric that describes the probability of switching within a text [32] (see also [13, 31]). Let us define any token in the corpus that is preceded by a token with a different language tag as a switch point. Then the I-index is a proportion of how many switch points exist relative to the number of language-dependent tokens in the corpus. In other words, it is the approximate probability that any given token in the corpus is a switch point. Given a corpus composed of tokens tagged by language $\{l_i\}$ where j ranges from 1 to n, the size of the corpus, and $i = j - 1$, the I-index is calculated by the expression

$$\text{I-Index} \equiv \frac{1}{n-1} \sum_{1 \leq i = j-1 \leq n-1} S(l_i, l_j), \tag{5}$$

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.

This index has utility for differentiating between corpora that are similarly multilingual but contain different patterns of switching behavior. For example, a parallel corpus would return an I-index very close to zero, whereas a corpus containing classic C-S would return values relatively farther away from zero.

3.2. Time-course Measures

It is useful to develop metrics that go beyond simple word counts to include information about the temporal distribution of C-S across the corpus. Time series analyses like these are valid since corpora have a clear word sequence. In our time series analyses, adapted from the field of complex systems [35], we choose switch points at the word level as the object of interest. Let a *language span* be defined as the distance between switch points, in units of language-dependent tokens. These spans are thus lengths of monolingual discourse. The *language span distribution* is the aggregate of all such language spans into a discrete probability distribution. This distribution describes the approximate probabilities that a span of monolingual discourse will take on a given length.

3.2.1. Burstiness

Barabási [35] proposes *burstiness* as a metric to describe distributions of this type. In this context, burstiness measures the manner and extent to which observed C-S behavior differs from a Poisson process (i.e., a process in which switching occurs at random). Briefly stated, it quantifies whether switching occurs in bursts or has a more periodic character.

Let σ_τ denote the standard deviation of the language spans and m_τ the mean of the language spans. Burstiness is calculated

$$\text{Burstiness} \equiv \frac{(\sigma_\tau / m_\tau - 1)}{(\sigma_\tau / m_\tau + 1)} = \frac{(\sigma_\tau - m_\tau)}{(\sigma_\tau + m_\tau)} \tag{6}$$

and is bounded within the interval [-1, 1]. Corpora with anti-bursty, periodic dispersions of switch points take on burstiness values closer to -1. By contrast, corpora with less predictable patterns of switching take on values closer to 1.

3.2.2. Span Entropy

In keeping with our efforts to present metrics consistent with information theory, we also define the Shannon entropy of the language span distributions. The *span entropy* returns how many bits of information are needed to describe the distribution of the language spans. Because the language span distribution

Table 1: *Corpora Metrics*

Corpus	M-index	I-index	Burstiness	Memory	LE	SE
Yo-Yo Boing!	0.95	0.03	0.37	-0.12	0.98	5.27
Killer Crónicas	0.99	0.23	0.02	-0.03	0.99	3.36
Bon Cop, Bad Cop	0.87	0.10	0.44	-0.06	0.95	4.00
SpinTX	0.07	0.02	0.48	-0.11	0.22	5.58
Solorio7k	0.60	0.06	0.32	-0.11	0.81	4.85

contains many more possible states than the language tag distribution, the amount of information needed to describe this distribution is far higher.

Let M denote the total number of states within the language span distribution, and l denote a specific span within that distribution where p_l represents the sample probability of a span of length l. The *span entropy* is then defined as

$$LE = -\sum_{l=1}^{M} p_l \log_2(p_l) \tag{7}$$

and is bounded below by 0 (in which case all language spans are of equal length) and above by

$$-\sum_{l=1}^{M} \frac{1}{M} \log_2\left(\frac{1}{M}\right) = \log_2(M), \tag{8}$$

which is the maximum entropy for a corpus with M possible language span states (and in such a case, each possible span has probability $\frac{1}{M}$).

3.3. Memory

Although the burstiness and span entropy metrics take into account the time *spacing* between switch points, they cannot make claims about the time *ordering* of the language spans. It is possible for two corpora to have identical language span distributions – and thus the same Burstiness-index – that nonetheless appear very different due to how the switch points are ordered. Barabási & Goh [35] provide *memory* as a measure of average first-order autocorrelation between consecutive language spans. In simpler terms, *memory* quantifies the extent to which the length of language spans tend to be influenced by the length of spans preceding them.

Let n_r be the number of language spans in the distribution and τ_i denote a specific language span in that distribution ordered by i. Let σ_1 and m_1 be the standard deviation and mean of all language spans but the last, where σ_2 and m_2 are the standard deviation and mean of all language spans but the first. Memory is calculated as

$$\text{Memory} \equiv \frac{1}{n_r - 1} \sum_{i=1}^{n_r - 1} \frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1 \sigma_2} \tag{9}$$

and is bounded within the interval [-1,1]. Memory values close to -1 describe the tendency for consecutive language spans to be negatively autocorrelated, differing substantially in length; that is, long spans of discourse are followed by short spans of discourse, and short spans are followed by long spans. Conversely, memory values closer to 1 describe the tendency for consecutive language spans to be positively autocorrelated, meaning similar in length.

The distribution-based measures, burstiness and span entropy, and the time series measure, memory, complement one another to describe the intermittency of switching behavior. In concert with the other metrics, which give a sense of the extent of language mixing, these measures relay a comprehensive signature of C-S for any language-tagged corpora.

4. Experiments

4.1. Language Identification

Our language identification system, reported in Guzmán et al. [32, 33], is an adapted version of the language identification system of Solorio & Liu [19, 36]. It produces two tiers of annotation: (i) *Language* (e.g., English (ENG), Spanish (SP)/French (FR), Punctuation, or Number) and (ii) *Named Entity* (yes or no). We annotate Named Entities for language because they can be language dependent (e.g., Ciudad de México versus Mexico

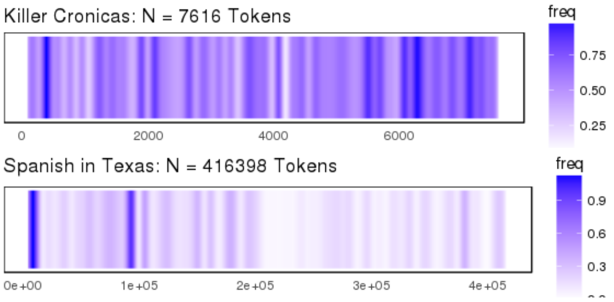

Figure 1: *Switch point density plots of SpinTX, KC*

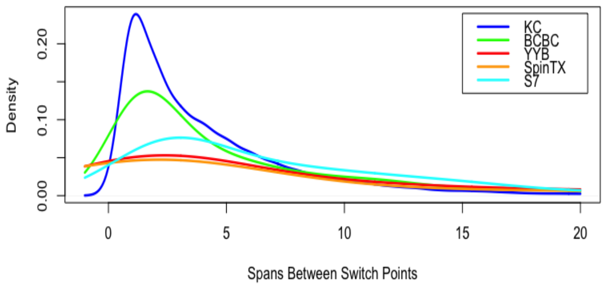

Figure 2: *Language span distributions of all corpora*

City), in which case they may act as triggers for switching [37]. 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 two film subtitle corpora of roughly equal sizes. The FR-ENG model was trained on a French Canadian newspaper corpus (*La Presse*) and a film subtitle corpus (*Bon Cop, Bad Cop*). When tested against our manually annotated gold-standards, our models achieved accuracy rates that do not deviate substantially from those of human annotators [38]: 95% for SP-ENG and 97% for FR-ENG.

4.2. Datasets

Given that our interest is in leveraging the power of accurate language identification to discover and detect patterns in C-S, we selected five different datasets known for their multilingualism, two written and three oral. The first two texts are literary works touted for their extensive Spanish-English C-S, which were written in two different styles. *Yo-Yo Boing!* (YYB) is a 58,494-word novel by Puerto-Rican author Giannina Braschi consisting of alternating chapters of poetry in English, Spanish and mixed Spanish-English [39]. *Killer Crónicas: Bilingual Memories* (KC) is a 40,469-word work written by Jewish Chicana author Susana Chávez-Silverman that is comprised of email messages entirely in 'Spanglish' [40]. These texts are available online and were used with the permission of the authors. Next, we selected the transcript of the French-Canadian film *Bon Cop, Bad Cop* (BCBC) directed by Eric Canuel. Both the French and English transcripts were downloaded from subtitles.com and manually combined into a final transcript of 13,502 words representing the actual language in the film. We also performed our analysis on the Spanish in Texas dataset (SpinTX) compiled by Bullock & Toribio [41] consisting of over 500,000 words of transcriptions from interviews with 97 heritage Spanish speakers across Texas. In addition, we chose the Spanish-English data (S7) collected by Solorio et al. [19] transcribed from a recorded conversation between three bilinguals, resulting in 8,011 words tagged for part-of-speech, punctuation, and language.

4.3. Results

As seen in Table 1, despite the close similarity in M-index for YYB and KC, there is a significant difference in their I-indices, reflecting a distinction in the switching probability of the two corpora. The film BCBC is less balanced than either KC or YYB, but it presents an intermediate I-index. Although S7 is much less balanced than the first three corpora, it still has a larger amount of C-S relative to YYB. Finally, SpinTX provides a lower bound for the M-index and I-index as an almost monolingual corpus. As a complement to the M-index, the language entropy (LE) of YYB, KC, BCBC, and S7 reflects the near balanced amount of C-S, since the maximum language entropy for two languages is $-log_2(1/2) = 1$ bit. In terms of burstiness, all but KC display relatively high values, indicating that the lengths of language spans throughout the texts are distributed in an irregular fashion. In contrast, KC's neutral value for burstiness implies that the C-S in the text is only weakly irregular. However, the negative values of memory suggest that longer language spans are slightly more likely to be followed by shorter spans and vice-versa in all corpora. This is exemplified by the heatmaps in Figure 1, which show the difference between the oral corpus SpinTX and the Spanglish novel KC, where darker bands indicate a higher incidence of language switching through the corpus. Lastly, the wide range in the span entropies of the corpora reveals a concrete difference in the distribution of the language spans throughout the texts. KC requires only 3.36 bits of entropy to describe its span distribution, revealing that KC relative to other corpora is comprised of small, highly-frequent language spans, as shown by the language span distributions in Figure 2. In contrast, the percentages of the other four corpora comprised of short language spans are lower, reflecting that they require more bits of entropy to encode their span distributions. Given these results, it appears that these datasets comprise two types of mixing, one essentially like insertional switching, where one language serves as the base language and elements from the other are inserted, and the other with no base language and regular, alternational switching. It remains to be seen whether other types of patterns might be discoverable.

5. Conclusion

Although the language tags of heavily-mixed texts are distributed such that they approach maximum entropy (i.e. the languages are nearly equally represented in the corpora), there is a wide variation in span entropy despite the similarity in language entropy. In particular, examples of highly-switched speech (such as the texts of KC or BCBC) are primarily composed of frequent spans of short length, reflecting the bursty nature of switching and the lower span entropy of frequent C-S. The three oral corpora (S7, BCBC, and SpinTX) present a low overall probability of switching and a high span entropy, perhaps reflecting an infrequent and irregular nature of spoken C-S. We expect that our entropy-based metrics will prove useful in future work on describing C-S behavior with maximum entropy models. We could also explore the complex network properties of C-S [42]. In this paper, we have applied the metrics at the word level for each corpus as a whole, but they can be applied at different levels of annotation (e.g., morpheme, POS tag) and to sub-corpora (e.g., individual speakers within a corpus). All of these measures are language-independent and can be used to compare corpora across language combinations. We view the resultant quantifications and visualizations as affording a comparative perspective that is important for linguists and NLP scholars, as the information regarding the probability and periodicity of C-S within a corpus can invite different types of analyses and motivate the development, improvement, or selection of technologies tailored to specific C-S typologies. For example, the predictability and probability of C-S in specific language combinations constrains the level of computation necessary for ASR systems in understanding spoken data.

6. Acknowledgements

We thank Thamar Solorio for sharing her annotated data, Rozen Neupane for compiling the bilingual version of the subtitled film, and Susana Chávez-Silverman and Giannina Braschi for allowing us to annotate their works.

7. References

[1] B. E. Bullock and A. J. Toribio, *The Cambridge handbook of linguistic code-switching*. Cambridge University Press Cambridge, 2009, vol. 1.

[2] H. M. Belazi, E. J. Rubin, and A. J. Toribio, "Code Switching and X-Bar Theory: The Functional Head Constraint," *Linguistic Inquiry*, vol. 25, no. 2, pp. 221–237, 1994.

[3] C. Myers-Scotton, *Duelling languages: grammatical structure in*

codeswitching. Oxford: Oxford University Press (Clarendon Press), 1993.

[4] P. Auer, "From Codeswitching via Language Mixing to Fused Lects: Toward a Dynamic Typology of Bilingual Speech," *International Journal of Bilingualism*, vol. 3, no. 4, pp. 309–332, Dec. 1999.

[5] K. Bali, Y. Vyas, J. Sharma, and M. Choudhury, "'i am borrowing ya mixing?" an analysis of English-Hindi code mixing in Facebook. In Proceedings of the First Workshop on Computational Approaches to Code Switching, EMNLP," pp. 116–126, 2014.

[6] D. Vilares, M. Alonso, and C. Gómez-Rodríguez, "EN-ES-EC: An English-Spanish code-switching twitter corpus for multilingual sentiment analysis," *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4149–4153, 2016.

[7] O. Çetinoglu, "A Turkish-German Code-Switching Corpus," *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4215–4220, 2016.

[8] D. Jurgens, S. Dimitrov, and D. Ruths, "Twitter users #codeswitch hashtags! #moltoimportante #wow," *EMNLP 2014*, pp. 51–61, 2014.

[9] Y. Samih, S. Maharjan, M. Attia, L. Kallmeyer, and T. Solorio, "Multilingual Code-switching Identification via LSTM Recurrent Neural Networks," in *ResearchGate*, Nov. 2016.

[10] S. Sitaram and A. W. Black, "Speech Synthesis of Code Mixed Text," in *Proc. LREC*, 2016, pp. 3422–3428.

[11] O. Çetinoglu, S. Schulz, and N. T. Vu, "Challenges of computational processing of code-switching," Austin, TX, 2016.

[12] N. T. Vu, H. Adel, and T. Schultz, "An investigation of code-switching attitude dependent language modeling," *Statistical Language and Speech Processing*, pp. 297–308, 2013.

[13] B. Gambäck and A. Das, "On Measuring the Complexity of Code-Mixing," in *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, 2014, pp. 1–7.

[14] B. King and S. Abney, "Labeling the languages of words in mixed-language documents using weakly supervised methods," in *Proceedings of NAACL-HLT*, 2013, pp. 1110–1119.

[15] P. Koehn and others, *Europarl: A multilingual corpus for evaluation of machine translation*, 2002.

[16] P. Muysken, "Language contact outcomes as the result of bilingual optimization strategies," *Bilingualism: Language and cognition*, vol. 16, no. 04, pp. 709–730, 2013.

[17] B. E. Bullock, J. Serigos, and Almeida Jacqueline Toribio, "The stratification of English-language lone-word and multi-word material in Puerto Rican Spanish-language press outlets: A computational approach," in *Spanish-English code-switching in the Caribbean and the U.S.*, R. Guzzardo Tamargo, C. M. Mazak, and M. C. Parafita Cuoto, Eds. Amsterdam ; Philadelphia: Benjamins, 2016, pp. 171–189.

[18] A. Troyansky, "Mother daughter tongue: The language use of North African women in France," Ph.D., University of Texas at Austin, 2016.

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

[20] E. Papalexakis, D.-P. Nguyen, and A. S. Doğruöz, "Predicting code-switching in multilingual communication for immigrant communities." Association for Computational Linguistics, 2014.

[21] M. Piergallini, R. Shirvani, G. S. Gautam, and M. Chouikha, "Word-level language identification and predicting codeswitching points in Swahili-English language data," *EMNLP 2016*, p. 21, 2016.

[22] H. Liang, Y. Qian, and F. K. Soong, "An HMM-based Bilingual (Mandarin-English) TTS," *Proceedings of SSW6*, 2007.

[23] J. Navratil, "Spoken language recognition-a step toward multilinguality in speech processing," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 678–685, 2001.

[24] T. I. Modipa, M. H. Davel, and F. De Wet, "Implications of Sepedi/English code switching for ASR systems," 2013.

[25] P. Fung and T. Schultz, "Multilingual spoken language processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 89–97, May 2008.

[26] H. Adel, K. Kirchhoff, D. Telaar, N. T. Vu, T. Schlippe, and T. Schultz, "Features for factored language models for code-Switching speech." in *SLTU*, 2014, pp. 32–38.

[27] H. Adel, N. T. Vu, K. Kirchhoff, D. Telaar, and T. Schultz, "Syntactic and semantic features for code-switching factored language models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 431–440, 2015.

[28] Y. Li and P. Fung, "Code-Switch Language Model with Inversion Constraints for Mixed Language Speech Recognition." in *COLING*, 2012, pp. 1671–1680.

[29] ——, "Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7368–7372.

[30] R. Molapo and E. Barnard, "Number pronunciation in a multilingual environment and implications for an ASR system," 2014.

[31] B. Gambäck and A. Das, "Comparing the level of code-switching in corpora," *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016*, pp. 1850–1855, 2016.

[32] G. Guzmán, J. Serigos, B. E. Bullock, and A. J. Toribio, "Simple Tools for Exploring Variation in Code-Switching for Linguists," *EMNLP 2016*, pp. 2–20, 2016.

[33] G. A. Guzmán, J. Ricard, J. Serigos, B. Bullock, and A. J. Toribio, "Moving code-switching research toward more empirically grounded methods," in *CDH2017: Corpora in the Digital Humanities*, p. 2017.

[34] R. Barnett, E. Codó, E. Eppler, M. Forcadell, P. Gardner-Chloros, R. v. Hout, M. Moyer, M. C. Torras, M. T. Turell, M. Sebba, M. Starren, and S. Wensing, "The LIDES Coding Manual A document for preparing and analyzing language interaction data Version 1.1—July, 1999," *International Journal of Bilingualism*, vol. 4, no. 2, pp. 131–132, Jun. 2000.

[35] K.-I. Goh and A.-L. Barabási, "Burstiness and memory in complex systems," *EPL (Europhysics Letters)*, vol. 81, no. 4, p. 48002, 2008.

[36] T. Solorio and Y. Liu, "Part-of-speech tagging for English-Spanish code-switched text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 1051–1060.

[37] M. Broersma and K. De Bot, "Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative," *Bilingualism: Language and cognition*, vol. 9, no. 01, pp. 1–13, 2006.

[38] C. Goutte, S. Léger, S. Malmasi, and M. Zampieri, "Discriminating similar languages: Evaluations and explorations," *CoRR*, vol. abs/1610.00031, 2016.

[39] G. Braschi, *Yo-yo boing!* Latin Amer Literary Review Press, 1998.

[40] S. Chávez-Silverman, *Killer crónicas: bilingual memories*. Univ of Wisconsin Press, 2004.

[41] B. E. Bullock and A. J. Toribio, "The Spanish in Texas Corpus Project," 2012.

[42] A. Behl and M. Choudhury, "A corpus linguistic study of bollywood song lyrics in the framework of complex network theory," in *International Conference on Natural Language Processing*. Macmillan Publishers, India, 2011.