The Limits of Spanglish?

Barbara E. Bullock Gualberto Guzmán Almeida Jacqueline Toribio
The University of Texas at Austin
{bbullock,toribio}@austin.utexas.edu
{gualbertoguzman}@utexas.edu

Abstract

Linguistic code-switching (C-S) is common in oral bilingual vernacular speech. When used in literature, C-S becomes an artistic choice that can mirror the patterns of bilingual interactions. But it can also potentially exceed them. What are the limits of C-S? We model features of C-S in corpora of contemporary U.S. Spanish-English literary and conversational data to analyze why some critics view the 'Spanglish' texts of Ilan Stavans as deviating from a C-S norm.

1 Introduction

Code-switching (C-S), the alternating use of languages in a single conversation, is a vernacular practice of U.S. Spanish-English bilinguals. Latinx authors use C-S in their writing for various functions and at varying rates in addressing different readers. The occasional insertion of a Spanish word or expression into English language texts can appeal to monolingual and bilingual readers alike. Alternatively, the languages can co-occur in more complex patterns that engage only the most bilingual reader (Torres, 2007). The question then arises: What are the limits to the stylistic choices available to bilingual writers? To attempt to answer this question, we submit extracts of 'Spanglish' literature to experiments that allow us to model the features that identify the contour of an author's mixing. These results are, in turn, compared with naturally produced Spanish-English C-S conversation corpora.

C-S language data complicates NLP tasks like language identification, POS tagging, or language modeling (Solorio and Liu, 2008b,a; Solorio et al., 2014; Çetinoğlu et al., 2016; Barman et al., 2014; Vilares et al., 2016; Jamatia et al., 2015; Lynn et al., 2015; Elfardy et al., 2014; Molina et al., 2016; Rijhwani et al., 2017). Therefore, our experiments rest on language identification at the word level, coupled with analyses of syntactic and lexical features that do not require POS tagging. Our contributions are the following: (1) We compare the complexity of C-S in the prose of Ilan Stavans to that in other 'Spanglish' texts; (2) We introduce a new method of normalizing the probability of C-S in a corpus scaled according to the distribution of languages in a corpus; (3) We extract linguistic features of Stavans's writing – out-of-vocabulary items and syntactic transitions – and manually review them for grammatical analysis; (4) We assess the degree to which C-S in literature conforms to features that are attested in speech and that are predicted by linguistic principles and constraints.

2 Related Work

Research into C-S in spontaneously-produced and elicited spoken speech has offered insights into the social, cognitive, and structural dimensions of this multilingual phenomenon (Bullock and Toribio, 2009). The analysis of C-S in written discourse has garnered substantially less attention and, with some exceptions reviewed below (Montes-Alcalá, 2001; Callahan, 2004, 2002), it has centered largely on C-S in historical texts as a genre (Latin macaronic poetry, medieval Castilian Spanish-Hebrew *taqqanots* 'ordinances', personal letters) (Demo, 2018; Schulz and Keller, 2016; Miller, 2001; Gardner-Chloros and Weston, 2015; Swain et al., 2002; Nurmi and Pahta, 2004).

Spanish-English C-S is integral to the U.S. Latino experience, and Latino authors such as Gloria Anzaldúa and Junot Díaz, to name but two, have given authentic expression to this bilingual, bicultural reality and, in so doing, have brought legitimacy to literary C-S. The C-S crafted by Ilan Stavans stands as a point of contrast, a Spanish-English composite employed in rendering Spanglish renditions of *Don Quixote*, *Hamlet*, *Le Petit*

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 115–121
Minneapolis, MN, USA, June 7, 2019.

Prince, and The United States Constitution (excerpted in Example 1 below). The creative texts incorporate word-internal switches such as *joldeamos, unalienables, suddenmente*, which violate the Free Morpheme Constraint (Poplack, 1980), and *tinkleada*, whose phonotactic sequence of English syllabic [ɬ] followed by a Spanish bound morpheme is ruled out by the PF Disjunction Theorem (MacSwan, 2000). Stavans also employs the switching of lone function words, flaunting the Matrix Language Hypothesis (Myers-Scotton, 1997), which proscribes switching to a functional element and then immediately returning to the base language (Joshi, 1982). These properties led Lipski to characterize Stavans's Spanglish as "grotesque" (Lipski, 2004) and Torres to describe it as "unlikely" and "implausible" (Torres, 2005).

1. Nosotros joldeamos que estas truths son self-evidentes, que todos los hombres son creados equally, que están endawdeados por su Creador con certain derechos unalienables, que entre these están la vida, la libertad, y la persura de la felicidad. (Stavans, 2004)

2. Este asteroid ha sido glimpseado solamente una vez through un telescopio, y eso fue por un turco astrónomo en 1909. Él gave una impressive presentación de su discovery en una international astronomía conferencia. Pero nadie believed him por la manera en que él dessed up. Así es como son los grown-ups. Luckymente pa' el Asteroid B612, un Turco dictador made que sus people dressed con European estilo, on amenaza de death. Usando un very elegante traje, el astrónomo dio su presentatión again, en 1920. This time todos estaban convinced. (Stavans, 2017)

The parallel between literary and conversational C-S with respect to syntactic structure has been investigated. Callahan (2002; 2004) analyzed a corpus of 30 bilingual texts — novels and short stories published in the U.S. between 1970-2000 — totaling 2954 pages (word count unknown), with the goal of testing whether the Matrix Language Frame model (MLF), developed for oral speech, could be predictive of literary C-S. In broad terms, the asymmetric MLF model holds that one language provides the grammatical frame into which other-language material is inserted. Callahan manually annotated for Matrix language (ML) and Embedded language (EL) concluding that, in general, the C-S in the literary corpus can be accounted by the principles of the MLF model.

Human judges of automatically generated C-S have been shown to converge in their agreements that certain syntactic switches, such as the switching between subject pronoun and verb or between auxiliary and main verb, are dispreferred (Bhat et al., 2016; Solorio and Liu, 2008a). These findings are confirmed in linguistic research eliciting intuitions on constructed stimuli (Toribio, 2001). There are also observed directional effects in natural C-S, most notably with respect to the DET-N boundary; a switch generally follows a determiner in only one of the component languages (Joshi, 1982; Mahootian and Santorini, 1996; Blokzijl et al., 2017; Parafita Couto and Gullberg, 2017). In Spanish-English switches at this syntactic juncture, Spanish DET is consistently followed by an English bare noun regardless of which language is the ML (Bullock et al., 2018).

While we know much about the grammatical co-occurrence restrictions on intrasentential C-S, patterns of mixing in a broader sense remain to be explored. It is frequent to encounter claims that a vernacular is 'highly mixed' or to classify mixing according to a typology of complexity, e.g., from *insertion* to *alternation* or *congruent lexicalization*, where there is a single grammar into which words from more than one lexicon are inserted (Muysken, 2000). Metrics that aim to quantify C-S complexity in order to compare between corpora have been proposed to characterize the nature of language mixing (Das and Gambäck, 2014; Barnett et al., 2000; Gambäck and Das, 2016, 2014). In this paper we use and expand upon the metrics proposed by Guzmán et al. (2017), which are designed to quantify patterns of switching within and between corpora, to compare the C-S in the writings of Stavans against other literary works as well as against conversational C-S.

3 Methods

Four short extracts of stories rendered in Spanglish by Stavans, totaling 10,051 words, were downloaded from the web and converted from pdf format to text files. Additional data include the text of two other novels recognized for their sustained C-S: *Yo-Yo Boing!* by Nuyorican author Giannina Braschi (1998) and *Killer Crónicas: Bilingual Memories* by Chicana writer Susana

Chávez-Silverman (2004), both used by permission from the authors. Data representing natural, oral C-S include a Spanish-English transcription of a bilingual conversation in Texas (S7), collected and shared by Thamar Solorio (Solorio and Liu, 2008a) and a conversation, *maria40* (M40), extracted from the Miami Corpus, deposited in the Bilingual Bank (Donnelly and Deuchar, 2011). Each data set was processed using the word-level language identification system for Spanish-English available on github https://github.com/Bilingual-Annotation-Task-Force/python-tagger and described in Guzmán et al. (2016). In post-processing, punctuation and numbers were given the language tag of the previous token so that they were not counted as switches. Named Entities are tagged for Spanish or English within the language identification system used.

The sequence of language tags output from the system is used as input to the python script that calculates metrics for C-S (https://github.com/Bilingual-Annotation-Task-Force/Scripts/blob/master/lang_metrics.py): the M-Index (Barnett et al., 2000), or the ratio of languages represented in a corpus, bound between 0 (monolingual) and 1 (perfectly bilingual); the I-Index (Guzman et al., 2016), the probability of switching between any two n-grams, also bound by 0 (no switching) and 1 (switching at every token); and Burstiness (Goh and Barabási, 2008), which provides a probability distribution of how many tokens will appear in a sequence in a given language before a switch to another, bound between -1 (periodic) and 1 (aperiodic). These results of application of these metrics to our corpora are shown in Table 1.

3.1 Normalized I-Index

One of the drawbacks of the I-Index developed by Guzmán et al. (2016) is that it does not account for the underlying language distribution of a text. For example, a text with an M-Index of 0.01, i.e. a text dominated by one language, could never achieve an I-Index of 1 because there are insufficient tokens to incorporate more switching. In fact, the only way to reach an I-Index of 1, linguistic constraints on switching aside, is if the M-Index were near 1, or if the languages were almost equally distributed. As a result, values of the I-index are not directly comparable across corpora from different language distributions. To correct for this, we have developed an improved version of the I-Index normalized to account for these bounds. In a text of N tokens, with k languages, each with n_i tokens, then the following equation can be used to compute a normalized I-Index, which we will refer to as I_2:

$$I_2 = \frac{I - L}{H - L} \tag{1}$$

where I represents the I-index described in (Guzman et al., 2016), and the lower and upper bounds, L and H, respectively, are defined by the following formulas:

$$L = (k - 1)/(N - 1) \tag{2}$$

$$H = \min\left(\frac{2 \cdot (N - \max_i n_i)}{N - 1}, 1\right) \tag{3}$$

The lowest amount of switching possible, L, outlined in Eq. 2 occurs when all n_i tokens of each language are concatenated together, leading to $k-1$ switches between all monolingual chunks. However, the highest amount of switching possible, H, which we compute in Eq. 3, occurs if we alternate tokens from each of the languages and intersperse them between the tokens of the most common language. An issue that our I_2 presents is that, for a highly-skewed corpus, the difference between the H and L values is minuscule, which can cause numerical problems. In other words, this metric performs poorly for corpora where the vast majority (>95%) is in one language.

Note that our I_2 scales I according to the language distribution and allows for direct comparison across different corpora. An I_2 of 0 or 1 now corresponds to a text with the absolute minimum and maximum, respectively, of switching possible given a fixed underlying language distribution. This new metric, in a manner of speaking, controls for a varying M-Index. In fact, as a rough estimate, one can think of I_2 as being approximately equal to I/M, where M is the M-Index.

3.2 Results of Metrics

The three literary works (Stavans, *Killer Crónicas* and *Yo-Yo Boing!*) are distinguished from the conversations (M40, S7) by the M-Index, as seen in Table 1, indicating that the balance of languages

in these texts is more even than in the conversations, where one language predominates (English in S7 and Spanish in M40). Within the literary corpora, the Stavans subcorpora stand out as having a higher probability of switching (I-Index) than the others, even more than *Killer Crónicas*, which is the most bilingual of all the datasets, with an M-Index of .99. This is reflected best by the Normalized I-Index, which is a valid measure of comparison here since none of the corpora are highly-skewed.

The quantitative models of these corpora indicate that the Stavans excerpts exhibit extreme switching relative to the other datasets. Contrary to prior work by Guzmán et al. (2016), the values of I_2 demonstrate that KC is not that much different from M40 and S7. The largest differences observed in I and I_2 are with the M40 and S7 corpora due to the skewed language distributions of the texts, which exaggerate the measurement of the amounts of switching.

A plot representing the densities of monolingual spans in the corpora, a visualization of Burstiness, is shown in Figure 1, where it can be seen that language mixing in Stavans and Killer Crónicas occurs more regularly throughout the text, whereas *Yo-Yo Boing!*, M40, and S7 show a long-tailed signal, indicating that C-S is a sporadic occurrence.

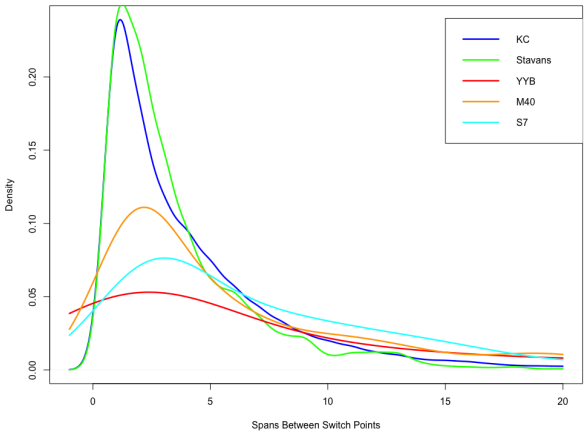

Figure 1: Span Densities

3.3 Lexical and Grammatical Analyses

As a second step toward modeling C-S, we compared the structural profiles of the Stavans extracts to *Killer Crónicas*, the texts in which C-S is the least bursty, to calculate the rate of word-internal switching. We filtered out words using the `aspell` command on Linux for English and Spanish. The mixed words were manually selected based on intra-word switching and NOT typographical errors, variable spellings (e.g., *cashe* to represent the Argentine pronunciation of *calle*), or non-words. We retained in our mixed-word list cross-linguistic phoneticizations such as *livin* in which English words are given a Spanish-like phonological representation. The results are given in Table 2 relative to the number of unique words in the corpora. The frequency of mixed words in *Killer Crónicas* is negligible relative to the proportion of the unique words in Stavans that are mixed. This difference is highly significant (χ^2 = 109.26, df = 1, p-value $<$ 2.2e-16) with a Cramer's-V test of .129 indicating a small effect size.

We investigated patterns of grammatical constraints by searching and tagging all subject pronouns and determiners in Spanish and in English according to their lexical entries (*la, the, yo, I,* etc.) and listing them alongside the word that followed in the text. We manually reviewed the lists of PRON + word and DET + word to eliminate any errors or any cross-linguistic homographs (e.g., Spanish *he* is an auxiliary verb). These were tabulated according to the language of the token and the language of the next word for each corpus. The proportion tables for DET-NOUN transitions is found in Table 3.

The asymmetry in directionality discussed above is evident in both literary corpora; Spanish determiners are more frequently found with English nouns than vice versa. However, Stavans shows a much higher mixing rate at this juncture, in general: .36 relative to .17 for *Killer Crónicas*: (χ^2 = 32.249, df = 1, p-value $<$ 1.356e-08) with a Cramer's-V test of .199 indicating a small effect size. The results for switching at the PRON-V juncture are shown in Table 4. While switching after a PRON is rare in *Killer Crónicas*, Stavans switches after a subject pronoun at a rate of about 13%, particularly if the pronoun is Spanish (χ^2 = 17.547, df = 1, p-value $<$ 2.803e-05) with a Cramer's-V test of .174). These analyses inform us that the Stavans corpora is qualitatively different from Killer Crónicas and distinguished by unusual C-S within words and across tightly knit syntactic boundaries.

Table 1: Metric results

Corpus	Length	Switches	M-Index	I-Index	I_2	Burstiness
Stavans	12405	4880	0.96	0.27	0.32	-0.03
Killer Crónicas	7002	2127	0.99	0.17	0.19	-0.06
Yo-Yo Boing!	75679	5339	0.97	0.04	0.05	0.36
M40	7638	1250	0.63	0.10	0.18	0.26
S7	8011	894	0.60	0.06	0.12	0.32

Table 2: Frequency of word-internal C-S

Corpus	Unique	Mixed	Freq
Stavans	4000	254	0.635
Killer Crónicas	2524	24	0.009

Table 3: Determiner-NP switching

	Stavans		*Killer Crónicas*	
Det	EnNP	SpNP	EnNP	SpNP
Eng	0.109	0.075	0.339	0.052
Span	0.278	0.538	0.050	0.560

Table 4: Pronoun-VP switching

	Stavans		*Killer Crónicas*	
Pro	EnVP	SpVP	EnVP	SpVP
Eng	0.474	0.099	0.653	0.005
Span	0.067	0.360	0.005	0.338

4 Discussion

We have observed that literary texts present more C-S than what is manifested in natural speech. However, different authors manifest different patterns of C-S, even when they employ more or less the same ratio of languages in their writings. While the M-index for the Stavans and *Killer Crónicas* corpora are nearly identical, demonstrating a near perfect balance of Spanish and English, with *Yo-Yo Boing!* close behind in terms of balance, the texts present distinct switching profiles. Specifically, Stavans, whose switching is criticized as unnatural, shows a higher probability of alternating between the languages, quantified by the I_2 and visualized as short spans of one language followed for short spans of the other. The C-S in Stavans also differs qualitatively from that in *Killer Crónicas*, the other literary text in our sample to show a similar anti-bursty distribution of C-S, in the preponderance of switching within the word (e.g., *adrifteando, astonisheado, askeó, wistfulmente*), switching at the DET-N boundary (e.g., *the casa*), and switching after PRON (*él slept*), all sites that are very rarely attested junctures of mixing in oral speech, and that are ruled out by predictive linguistic models.

Note that the effect of switching on functional words, such as pronouns and determiners, while in itself odd, will also lead to increased rates of C-S and to short language spans. Thus, we cannot know if it is the frequency of switching, the decision to switch after functional elements and within words, or a combination of these features that lead critics to characterize Stavans's 'Spanglish' texts in negative terms. In future work, we seek to determine whether there are expected constants of C-S for Spangish literature versus for natural speech. This will help determine the degree to which an observed C-S contour is an outlier.

We have presented methods for comparing between corpora that rest on multiple features easily gleaned from small corpora, but our conclusions can only be tentative. Language models that would permit direct comparisons of the statistical distribution of C-S between corpora would be desirable for establishing the limits of mixed vernaculars like so-called 'Spanglish'.

References

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23.

Ruthanna Barnett, 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.

Gayatri Bhat, Monojit Choudhury, and Kalika Bali. 2016. Grammatical constraints on intra-sentential code-switching: From theories to working models. *arXiv:1612.04538*.

Jeffrey Blokzijl, Margaret Deuchar, and M Carmen Parafita Couto. 2017. Determiner asymmetry in mixed nominal constructions: The role of grammatical factors in data from Miami and Nicaragua. *Languages*, 2(4):1–12.

Giannina Braschi. 1998. *Yo-yo boing!* Latin Amer. Literary Review Press.

Barbara E Bullock, Gualberto Guzmán, Jacqueline Serigos, and Almeida Jacqueline Toribio. 2018. Should Code-switching Models Be Asymmetric? *Proc. Interspeech 2018*, pages 2534–2538.

Barbara E. Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge handbook of linguistic code-switching.* Cambridge University Press.

Laura Callahan. 2002. The matrix language frame model and Spanish/English codeswitching in fiction. *Language & Communication*, 22(1):1–16.

Laura Callahan. 2004. *Spanish/English codeswitching in a written corpus*. John Benjamins Publishing.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of Computational Processing of Code-Switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Susana Chávez-Silverman. 2004. *Killer Crónicas: bilingual memories*. Univ. of Wisconsin Press.

Amitava Das and Björn Gambäck. 2014. Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.

Šime Demo. 2018. Mining macaronics. In *Multilingual Practices in Language History: English and Beyond*, pages 199–221. De Gruyter Mouton.

Kevin Donnelly and Margaret Deuchar. 2011. The Bangor Autoglosser: a multilingual tagger for conversational text. *ITA11, Wrexham, Wales*, pages 17–25.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. AIDA: Identifying code switching in informal Arabic text. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 94–101.

Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, pages 1–7. Citeseer.

Björn Gambäck and Amitava Das. 2016. Comparing the Level of Code-Switching in Corpora. In *LREC*, pages 1850–1855.

Penelope Gardner-Chloros and Daniel Weston. 2015. Code-switching and multilingualism in literature. *Language and Literature*, 24(3):182–193.

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

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for Modeling Code-Switching Across Corpora. In *Interspeech*, pages 67–71.

Gualberto A Guzman, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 239–248.

Aravind K Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 145–150. Academia Praha.

John M Lipski. 2004. Is "Spanglish" the third language of the South?: truth and fantasy about US Spanish. In *3rd Language Variation in the South (LAVIS III) conference, Tuscaloosa, AL*.

Teresa Lynn, Kevin Scannell, and Eimear Maguire. 2015. Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 1–8.

Jeff MacSwan. 2000. The architecture of the bilingual language faculty: Evidence from intrasentential code switching. *Bilingualism: language and cognition*, 3(1):37–54.

Shahrzad Mahootian and Beatrice Santorini. 1996. Code switching and the complement/adjunct distinction. *Linguistic Inquiry*, pages 464–479.

Elaine R Miller. 2001. Written code switching in a medieval document: A comparison with some modern constraints. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 46(3-4):159–186.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49.

Cecilia Montes-Alcalá. 2001. Written codeswitching: Powerful bilingual images. *Trends In Linguistics Studies AND Monographs*, 126:193–222.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Arja Nurmi and Päivi Pahta. 2004. Social stratification and patterns of code-switching in early English letters.

M Carmen Parafita Couto and Marianne Gullberg. 2017. Code-switching within the noun phrase: Evidence from three corpora. *International Journal of Bilingualism*, page 1367006917729543.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish y termino en Español: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1971–1982.

Sarah Schulz and Mareike Keller. 2016. Code-switching ubique est-language identification and part-of-speech tagging for historical mixed text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 43–51.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Thamar Solorio and Yang Liu. 2008a. 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.

Thamar Solorio and Yang Liu. 2008b. 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.

Ilan Stavans. 2004. *Spanglish: The making of a new American language*. Harper Collins.

Ilan Stavans. 2017. *El Little Principe*. Edition Tintenfass.

Simon Swain, James Noel Adams, Mark Janse, et al. 2002. *Bilingualism in ancient society: Language contact and the written text*. Oxford University Press on Demand.

Almeida Toribio. 2001. Accessing bilingual code-switching competence. *International Journal of Bilingualism*, 5(4):403–436.

Lourdes Torres. 2005. Don Quixote in Spanglish: traducttore, traditore? *Romance Quarterly*, 52(4):328–334.

Lourdes Torres. 2007. In the contact zone: Code-switching strategies by Latino/a writers. *Melus*, 32(1):75–96.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2016. EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis. In *LREC*, pages 4149–4153.