Predicting the presence of a Matrix Language in code-switching

Barbara E. Bullock, Gualberto Guzmán, Jacqueline Serigos
Vivek Sharath, Almeida Jacqueline Toribio
{gualbertoguzman,vivek.sharath}@utexas.edu
{bbullock,toribio}@austin.utexas.edu
{jserigos}@gmu.edu

Abstract

One language is often assumed to be dominant in code-switching (C-S), but this assumption has not been empirically tested. We operationalize the matrix language (ML) at the level of the sentence, using three common definitions. We test whether these converge and then model this convergence via a set of metrics that together quantify the nature of C-S. We conduct our experiment on four different Spanish-English corpora. Our results demonstrate that our model can separate some corpora according to whether they have a dominant ML or not but that the corpora span a range of mixing types that cannot be sorted neatly into an insertional vs. alternational dichotomy.

1 Introduction

From Joshi (1982) forward, many researchers assume that one of the participating languages in code-switching (C-S) is dominant. This notion is theorized in linguistics as the Matrix Language Frame model (MLF) (Myers-Scotton, 1997). The MLF assumes an asymmetry between the languages involved in C-S, with the matrix language (ML) providing the frame into which embedded language elements (EL) from the contact language are inserted, as well as an asymmetry between system vs. content morphemes. System morphemes in the MLF comprise a subset of closed class morphemes that neither assign nor receive a thematic role (e.g., determiners, quantifiers, auxiliaries, conjunctions). Constraints on language mixing follow from the asymmetry: The ML provides the grammatical elements and framing while the EL provides merely content morphemes. Nevertheless, there are two long-standing criticisms of the ML: (1) the criteria for the identification of the ML are not straightforward (Winford, 2003; Meakins, 2011; Bhat et al., 2016); and (2) the consistent identification of a single ML might not be possible (Auer and Muhamedova, 2005; Bhat et al., 2016; Liu, 2008; Adamou, 2016). To this we add a third concern: In ascertaining an ML, researchers often rely on selected, decontextualized example sentences. With some exceptions, most in NLP (Gambäck and Das, 2016; Bhat et al., 2016; Vyas et al., 2014), few scholars have calculated the ML for each sentence or utterance in a sizable dataset (Blokzijl et al., 2017). Thus, tests of the MLF using replicable methods are lacking, despite the fact that the determination of an ML has consequences for linguistic analyses and for accurate models of multilingual texts for language processing (Bhat et al., 2016; Solorio and Liu, 2008a,b) and for applications like TTS (Sitaram and Black, 2016; Sitaram et al., 2015) and ASR (Li et al., 2012).

In this paper, we attempt to quantify the nature of mixing using multiple measures and to operationalize the concept of the ML at the sentence level using code-switched Spanish-English corpora. We then test the concept of the ML and its applicability to different degrees of mixing as quantified by the ratio of languages represented in a sentence, by the probability of switching from one word to the next, and by the regularity vs. intermittency of switching as defined by the distribution of the interevent spans of each language. We operationalize the ML for the instances of intrasentential mixing identified in our corpora along three different parameters: numerically dominant language overall, numerically dominant language of all verbs, and numerically dominant language of a subset of system morphemes. We then predict the likelihood that these different calculations of the ML converge to the same language

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 68–75
Melbourne, Australia, July 19, 2018.

result (i.e., point to a unique ML) as a function of our corpus metrics. Our result is a model that classifies C-S data according to how likely it is that all three measures of the ML agree on the same language label. Our contribution is threefold: first, we show that one can ascribe a single ML with a high degree of likelihood given a particular pattern of C-S and that a simple word-count method is sufficient to do so; secondly, we empirically demonstrate that there is a cline of C-S such that corpora cannot always be neatly separated into insertional and alternational types as is generally claimed in the sociolinguistic literature; thirdly, we find that measures designed to assess the time-course of complex systems like C-S are lacking for small datasets.

2 Related Work

2.1 Debates about the MLF Model

Studies of C-S commonly distinguish between insertional and alternational patterns (Muysken, 2000). With insertional switching, speakers are said to know which one language an utterance is "coming from" (Joshi, 1982; Romaine, 1995; Sitaram and Black, 2016). In the MLF model (Myers-Scotton, 1997), this language is formalized as the ML. Insertional C-S, which may be indistinguishable from borrowing (Poplack et al., 1988), is encountered in many sociolinguistic settings irrespective of the typologies of the language pairing studied (Poplack et al., 1988; Muysken, 2000; Li and Fung, 2013; Vyas et al., 2014; Adamou, 2016). In Sentence 1, Marathi is the ML, identified by the relative ordering of words in the clause and by the language of the system or closed-class morphemes, such as the quantifier *kahi* and the light verb *kar*; English contributes only the EL lexeme *paint*. Hindi is argued to be the ML in Sentence 2 (Bhat et al., 2016) , which presents an EL Island (ELI), an English-language embedded constituent with its own internal structure.[1]

- Sentence 1
  mula kahi khurcya **paint** kartat
  [boys some chairs paint do+TNS]

- Sentence 2
  Shanivar neeras hai **from that perspective**
  [Saturday boring is from that perspective]

The means by which the ML of a clause or extended discourse is determined remains debated. The ML has been variously associated with the numerically dominant language, (Myers-Scotton, 1997; Gambäck and Das, 2016; Sharma and Motlani, 2015), with the language of the finite verb (Klavans, 1985; Treffers-Daller, 1994; Meakins, 2011), or with the first language in a left-to-right parsing (Doron, 1983). It should be noted that the ML, as defined by Myers-Scotton, operates over a unit she calls the CP, which is co-extensive with the clause. For the purpose of this paper, we define the ML over a sentence, which may contain more than one clause. Since the majority of the corpora to be examined are from natural conversations, it is likely that most sentences consist of a single clause, as sentences are known to be shorter and syntactically less complex in spoken language.

The ML is not argued to be applicable to alternational switching, because speakers move from one grammar to another within an utterance. But it is often not clear from cited examples whether a new language span constitutes the alternation of MLs, as appears to be the case in Sentence 3 from the bilingual memoire *Killer Crònicas*, or whether the span is an ELI inserted into an ML, as is argued to be the case for Sentence 2. For instance, examining natural Japanese-English data within the MLF, Namba (2012) could not determine the ML of C-S utterances such as Sentence 4, which accounted for 42% of the clauses in the corpus.

- Sentence 3
  Anyway, just leave him plantado, al taxista este, **or throw some money at him** y salir
  [stranded that taxi-driver ... and leave]

- Sentence 4
  I want to be goorukiipaa ni nari-tai
  [goalkeeper RESULTATIVE become]

2.2 Measuring the complexity of code-switching

Importantly, bilingual speech practices are complex and it is not clear that the traditional binary typology of insertional and alternational C-S, while useful as a heuristic, is adequate to characterize

[1] In the NLP literature, insertionanal mixing is often referred to as *code-mixing* (CM), following Gumperz (1982) (Vyas et al., 2014; Bali et al., 2014), though some researchers employ CM as an umbrella term for both insertional and alternational mixing (Sequiera et al., 2015). Others use CM for any switching that occurs within an utterance (and C-S for switching at or above the utterance level) (Gambäck and Das, 2016).

the nature of C-S (Auer and Muhamedova, 2005). There have been recent attempts to quantify mixing complexity with the aim of arriving at empirically reliable comparisons of C-S between corpora (Gambäck and Das, 2016; Das and Gambäck, 2014; Jamatia et al., 2015; Guzman et al., 2016; Guzmán et al., 2017a). Each aims to capture the fact that C-S may vary along multiple planes. We follow Guzmán et al. (2016, 2017a) who quantify mixing in terms of several parameters calculated from language labels at the word level: (1) the ratio of languages represented; (2) the probability of switching language between any two words; (3) the burstiness of switching as characterized by the distribution of the length of spans; and (4) the sequential ordering of alternating monolingual spans.

3 Data

Bhat et al. (2016) built models for generating C-S sentences based on input sentences and the constraints of the MLF and of the Equivalence Constraint, a symmetrical model for alternating C-S (Poplack, 1980). When the sentences were submitted to human evaluation, there was significant variance in acceptability, potentially attributable to discrepancies in the register of some of the words used, as C-S tends to be informal and conversational. For our study, we avoid confounds that can be introduced by generated C-S by drawing on C-S data generated by bilingual speakers themselves. We were restricted in our choice of data by the requirement that all data bear a language label and a POS tag. As is commonly observed, POS tagged bilingual data are rare because the accuracy of monolingual taggers is reduced when the context is broken by C-S (Vyas et al., 2014).

The corpora that we use reflect degrees of mixing so that we test the viability of the MLF hypothesis across varying types of C-S. Each corpus was previously tagged for language and POS by its creators. In order to be able to compare between, the original POS tags used for each datasets were mapped to the core POS tagset from the Universal Dependencies (UD) framework (Nivre et al., 2016). The corpora to be modeled are the following.

1. *S7* was created by Thamar Solorio (2008a; 2008b). It documents a conversation among three Spanish-English bilinguals, resulting in approximately 8,000 words. It was tagged for language and POS, using TreeTagger's English and Spanish parameters (Solorio and Liu, 2008a).

2. *Miami* consists of files from the Bangor Miami Corpus, transcripts of informal conversations between Spanish-English bilinguals in Miami. The data was automatically annotated for language and POS, using the Bangor Autoglosser (Donnelly and Deuchar, 2011).

3. *SpinTX* comprises selected transcripts of speakers from the Spanish in Texas Corpus, a set of recorded interviews between Spanish-English bilinguals residing in Texas (Bullock and Toribio, 2013; Toribio and Bullock, 2016). The corpus in its entirety was automatically tagged for POS using the English and Spanish versions of TreeTagger (Schmid, 1995) applied sequentially.[2]

4. *KC* is an excerpt of the epistolary work *Killer Crónicas: Bilingual Memoires* by Susana Chavez-Silverman. Nearly evenly balanced between English and Spanish, the POS annotated segment contains approximately 8,000 words. It was automatically tagged for language following Guzmán et al. (2016) and then manually annotated for POS using the UD tagset.

4 Procedures

In order to examine the viability and agreement of the MLF across the four corpora, we converted all POS labels to the core UD tagset. For *S7, Miami*, and *SpinTX*, we remapped the existing POS-tagset from either TreeTagger or the Bangor Autoglosser using a lookup table. In the case of *KC*, we manually tagged every token according to the UD framework since we had no previous tagging. The POS annotations were completed by a Spanish-English bilingual, professional linguist and then each annotation was checked by two others.

Each corpus was submitted to sentence tokenization, breaking on full or sentential stops. For *S7, Miami*, and *SpinTX*, we followed the existing sentence end markers, such as "SENT" and "FS", from the original POS tagging before conversion to UD. For *KC*, we performed a manual

[2]The Spanish in Texas Corpus is available through a creative commons license for non- commercial download in various file formats from http://corpus.spanishintexas.org/en.

Table 1: Anyway, **al taxista** right away **le noté un acentito**, not too specific

ML Definition	English	Spanish	ML
Word Count	6	6	Tie
Verb	0	1	Spanish
Functional words	2	3	Spanish

sentence-tagging since the UD tagset collapses all punctuation under the "." tag, which loses all sentence boundary information. There are currently no workable sentence tokenizers for C-S data.

As it is designed to permit the comparison of syntax in a language independent manner, the 17-tag core UD provides adequate POS annotations for capturing the system morphemes for Spanish-English. But the core level does not provide the level of granularity to distinguish finite verbs from non-finite ones (infinitives, participles and gerunds). Thus, we operationalized the ML for each sentence using three methods: the numerically dominant language of all tokens (TOTAL), the numerically dominant language of all verbs (VERB), and the numerically dominant language of functional elements (FUNC), i.e. DET, SCONJ, CCONJ, PRON, and AUX. Each of these three methods predicted "English", "Spanish", or "Tie" as the ML for each sentence in our datasets. We quantified agreement between these measures using the logical AND of all three. If at least one method predicted a different ML than the other two, then the agreement was 0 for DISAGREE.

As an example, consider Sentence 5 in Table 1. Since there is an equal number of tokens from English and Spanish, the word count or TOTAL method predicts a Tie. However, both VERB and FUNC predict a Spanish ML because of the higher number of Spanish verbs and functional elements. In this case, the sentence-level agreement is DISAGREE because the measures do not all concur.

We operationalize the nature of mixing via three metrics: M-Index, I-Index and Burstiness, each defined below.

1. The Mixing Index (M-Index), developed by the LIPPES Group (Barnett et al., 2000) from the Gini coefficient defines the ratio of languages in a text. It is bounded by 0 (a monolingual text) and 1 (a text with an even distribution of languages).

2. The Integration Index, (I-Index), created by Guzmán et al. (2016; 2017b; 2017a) describes the probability of switching. It ranges from 0 (monolingual text) to 1 (a text in which every other word is drawn from a different language).

3. Burstiness, proposed by Goh and Barabási for complex systems (2008), defines the regularity of switching. It is adapted here to apply to the interevent level of sequences of monolingual tokens, called spans, after every C-S. It is bounded within the interval of -1 (anti-bursty, periodic dispersion of switching) and 1 (predictable patterns of switching).

A fourth metric, Memory (Goh and Barabási, 2008), which models the temporal order of the spans, is desirable for examining C-S in larger corpora (Guzman et al., 2016; Guzmán et al., 2017b,a), and was calculated at the sentence level over the test corpora. We were forced to exclude it from further consideration because the sentences were short and often included spans of equal length, yielding a standard deviation of zero. Since the multiplicand and the multiplier in the divisor of the Memory function are standard deviations, our sentences yielded many divisors of 0.

We tested our ML methods only on the subsets of the datasets that contained C-S, i.e. we eliminated all monolingual sentences from our corpora. This has the consequence of removing conversational disfluencies such as restarts, which are unlikely to to demonstrate a C-S. In addition, we excluded all parts of the SpinTX and Miami corpora that did not contain a base-line amount of mixing. For Miami, we removed the herring11 and maria21 conversations. Similarly, in SpinTX we removed all conversations with an I-index of less than 0.1. The final test corpus contained 7,879 C-S sentences, each coded for the three sentence-level metrics described above, for the ML predictions from each of the three numerical methods TOTAL, VERB, and FUNC, and for whether the numerical ML predictions agreed or not.

Across all sentences, the three methods converge on an ML 58% of the time. There were notable differences in the range of convergence at the corpus-level. For *S7*, they agree 65%; for *Miami* 57%; for *SpinTX* 71%, and for *KC* only 45%.

Figure 1: Effects Plot for Corpus and I-index

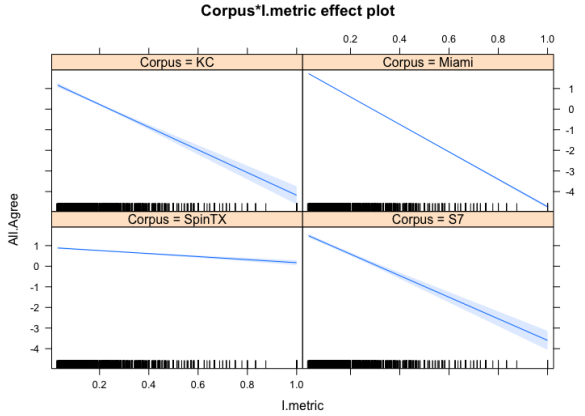
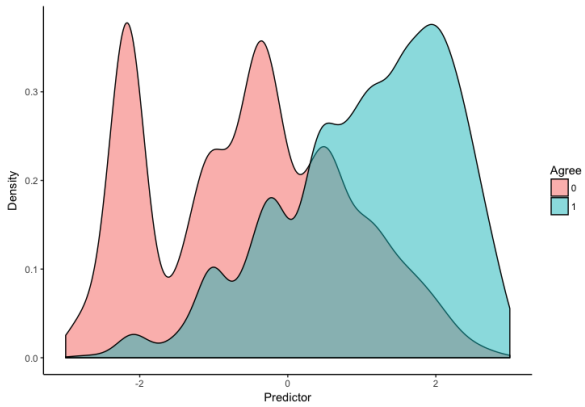


Figure 2: Agreement by Predictor Value



5 Methods

We fit a logistic regression to predict AGREE (i.e., there is an ML upon which all three measures agree) with three continuous predictor variables (M-Index, I-Index, Burstiness) and one categorical predictor (Corpus). An analysis of the model output revealed significant variability as to the effect of the I-index depending on the corpus, visualized in Figure 1. To capture this variability, an interaction between Corpus * I-Index was added to the model. The updated model is able to correctly predict AGREE or DISAGREE across all corpora with an F1-score of 69.3%, as shown in predictor density plot of Figure 2. All three metrics and the corpus as a categorical variable were significant in predicting agreement. The strongest predictors are the M-Index and Burstiness, with opposite effects, as seen in Figure 3. The M-Index inversely affects agreement; as the M-Index increases, the determinations of the ML are less likely to agree. Conversely, as the Burstiness increases, all three ML methods are more likely to agree. Plotting the predictors for each sentence yields Figure 4, which shows the model's prediction of agreement for the data from all four corpora.

Although the model does not cleanly split all sentences of *KC* and *SpinTX* by agreement, we do see a clear preference for predicting AGREE for *SpinTX* and DISAGREE for *KC*. However, we also find that the model predicts multi-modal agreement distributions for the S7 and Miami corpora. The small peaks around 0 indicate that the model does not have sufficient information to distinguish between predicting AGREE or DISAGREE for a small amount of data, which we discuss below.

Figure 3: Odds Ratio Plot

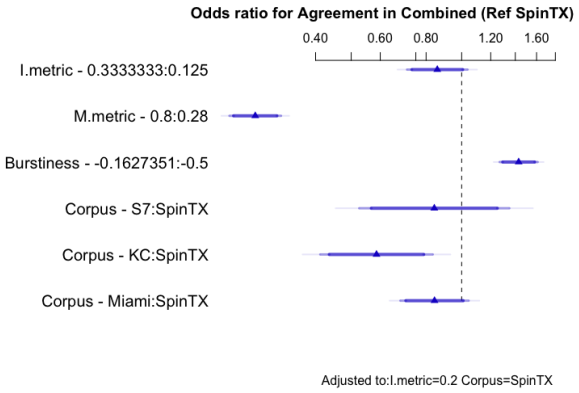
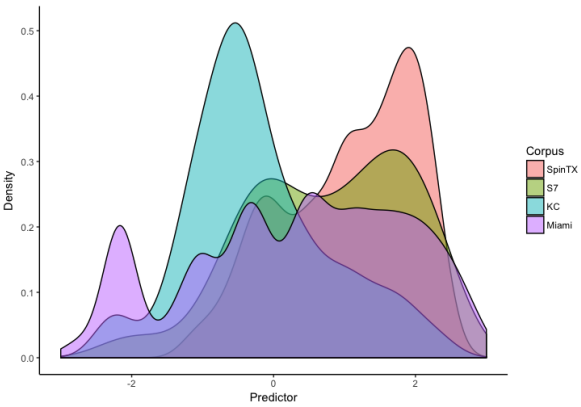


Figure 4: Agreement Density over all Corpora



6 Discussion

In this paper we found that the three different methods for determining the ML of a sentence agreed 58% of the time across different mixing types. Further, we found a clear distinction between the rate of agreement for corpora that appear to be more insertional versus others. We also

demonstrated that, collectively, these corpora span a range of types with some clearly intermediary between insertion and alternation. These intermediary patterns may correspond to instances of congruent lexicalization, a mix of insertion and alternation (Muysken, 2014).

Our model performance across all four corpora leads us to believe that language tagging is much more useful than previously thought and it may suffice in many cases for determining the ML. In fact, we can reliably predict the agreement of different ML methods with an accuracy of 69.3% using our metrics on language tags. The implication is that researchers in linguistics and in NLP could use word-count alone to determine the ML as a good first-approximation depending on the type of mixing in their data. Corpora with sporadic embeddings present an ideal case where the linguistic methods of determining ML often agree with word-count and these are likely to be prolific. In the Pangloss Collection of endangered Slavic languages in Europe, three of the six corpora contain less than 5% borrowed words (Adamou, 2016), a percentage that parallels the findings in other contact corpora of naturally produced speech (Treffers-Daller, 1994; Bullock et al., 2016; Cacoullos and Aaron, 2003; Varra, 2013). But, the performance of the model on the *S7* and *Miami* datasets indicate that our current metrics are not sufficient to predict agreement even when corpora have characteristics that indicate that they are largely insertional (low M-Index + high Burstiness). The uncertainty in the model predictions leads us to conclude that there is a continuum of mixing types within the existing typology of alternational and insertional mixing.

7 Future Research

In on-going work, we need to examine the methods of determining the ML in natural interactions in finer detail in order to determine which method is most likely to diverge from the other two. To further examine the viability of the MLF hypothesis, we are exploring other language pairings and analyzing the effectiveness of our current metrics at clustering and comparing across corpora, although we are hampered by the lack of POS-tagged bilingual data from natural speech. In addition, we are currently testing the performance of entropy-based measures (Guzmán et al., 2017a) as predictors for ML agreement. Finally, the performance of our model requires deeper syntactic analysis of the nature of mixing types and of the grammatical structures of the *S7* and *Miami* datasets in particular.

References

Evangelia Adamou. 2016. *A corpus-driven approach to language contact: Endangered languages in a comparative perspective*, volume 12. Walter de Gruyter GmbH & Co KG.

Peter Auer and Raihan Muhamedova. 2005. Embedded language'and 'matrix language'in insertional language mixing: Some problematic cases. *Rivista di linguistica*, 17(1):35–54.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "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*, pages 116–126.

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 preprint arXiv:1612.04538*.

Jeffrey Blokzijl, Margaret Deuchar, and M Couto. 2017. Determiner asymmetry in mixed nominal constructions: The role of grammatical factors in data from miami and nicaragua. *Languages*, 2(4):20.

Barbara E Bullock, Jacqueline Serigos, and Almeida Jacqueline Toribio. 2016. The stratification of english-language lone-word and multi-word material in puerto rican spanish-language press outlets. *Spanish-English Codeswitching in the Caribbean and the US*, 11:171.

BE Bullock and AJ Toribio. 2013. The spanish in texas corpus project. *Center for Open Education Resources and Language Learning (COERLL)*.

Rena Torres Cacoullos and Jessi Elana Aaron. 2003. Bare english-origin nouns in spanish: Rates, constraints, and discourse functions. *Language Variation and Change*, 15(3):289–328.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text.

Kevin Donnelly and Margaret Deuchar. 2011. Using constraint grammar in the bangor autoglosser to disambiguate multilingual spoken text.

Edit Doron. 1983. On a formal model of code-switching. In *Texas Linguistic Forum Austin, Tex.*, 22, pages 35–59.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *LREC*.

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

John J Gumperz. 1982. *Discourse strategies*, volume 1. Cambridge University Press.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017a. Metrics for modeling code-switching across corpora. *Proc. Interspeech 2017*, pages 67–71.

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara Bullock, and Almeida Jacqueline Toribio. 2017b. Moving code-switching research toward more empirically grounded methods. In *CDH@ TLT*, pages 1–9.

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.

Judith L Klavans. 1985. The syntax of code-switching: Spanish and english. In *Proceedings of the Linguistic Symposium on Romance Languages*, pages 213–231. Benjamins.

Ying Li and Pascale Fung. 2013. Language modeling for mixed language speech recognition using weighted phrase extraction. In *Interspeech*, pages 2599–2603.

Ying Li, Yue Yu, and Pascale Fung. 2012. A mandarin-english code-switching corpus. In *LREC*, pages 2515–2519.

Yu Liu. 2008. Evaluation of the matrix language hypothesis: Evidence from chinese-english code-switching phenomena in blogs. *Journal of Chinese Language and Computing*, 18(2):75–92.

Felicity Meakins. 2011. *Case-marking in contact: The development and function of case morphology in Gurindji Kriol*, volume 39. John Benjamins Publishing.

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

Pieter Muysken. 2014. Deja voodoo or new trails ahead. *Linguistic Variation: Confronting Fact and Theory*, page 242.

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

Kazuhiko Namba. 2012. Non-insertional code-switching in english–japanese bilingual children: alternation and congruent lexicalisation. *International Journal of Bilingual Education and Bilingualism*, 15(4):455–473.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation.

Suzanne Romaine. 1995. *Bilingualism*. Wiley-Blackwell.

Helmut Schmid. 1995. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Royal Sequiera, Monojit Choudhury, and Kalika Bali. 2015. Pos tagging of hindi-english code mixed text from social media: Some machine learning experiments. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 237–246.

Arnav Sharma and Raveesh Motlani. 2015. Pos tagging for code-mixed indian social media text: Systems from iiit-h for icon nlp tools contest.

Sunayana Sitaram and Alan W Black. 2016. Speech synthesis of code-mixed text.

Sunayana Sitaram, Sai Krishna Rallabandi, and SRAW Black. 2015. Experiments with cross-lingual systems for synthesis of code-mixed text.

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.

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

Jeanine Treffers-Daller. 1994. *Mixing two languages: French-Dutch contact in a comparative perspective*, volume 9. Walter de Gruyter.

Rachel Marie Varra. 2013. *The Social Correlates of Lexical Borrowing in Spanish in New York City.* ERIC.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979.

Donald Winford. 2003. *An introduction to contact linguistics*. Wiley-Blackwell.