Simple Tools for Exploring Variation in Code-Switching for Linguists

Gualberto A. Guzman Jacqueline Serigos Barbara E. Bullock Almeida Jacqueline Toribio
University of Texas at Austin
{gualbertoguzman,jserigos}@utexas.edu
{bbullock,toribio}@austin.utexas.edu

Abstract

One of the benefits of language identification that is particularly relevant for code-switching (CS) research is that it permits insight into how the languages are mixed (i.e., the level of integration of the languages). The aim of this paper is to quantify and visualize the nature of the integration of languages in CS documents using simple language-independent metrics that can be adopted by linguists. In our contribution, we (a) make a linguistic case for classifying CS types according to how the languages are integrated; (b) describe our language identification system; (c) introduce an Integration-index (I-index) derived from HMM transition probabilities; (d) employ methods for visualizing integration via a language signature (or switching profile); and (e) illustrate the utility of our simple metrics for linguists as applied to Spanish-English texts of different switching profiles.

1 Introduction

Sociolinguists who focus on CS have been reluctant to adopt automatic annotation tools in large part because of the Principle of Accountability (Labov, 1972), which demands an exhaustive and accurate report for every case in which a phenomenon (e.g., a switch) occurs or could have occurred. Thus, in order to encourage linguists to move beyond slow but accurate manual coding and to take benefit of computational methods, the tools need to be precise and intuitive and consistent with linguistic concepts pertaining to CS. Herein, we provide a means of quantifying language integration and of visualizing the language profile of documents, allowing researchers to isolate events of single-word other-language insertions (borrowing, nonce borrowing) versus spans of alternating languages (code-switching) versus lengthy sequences of monolingual text (translation, author/speaker change). Our methods differ from existing NLP approaches in attending to some issues that are relevant for linguists but neglected in other approaches, e.g., in classifying the language of Named Entities as they can trigger CS (Broersma & De Bot, 2006) , in using ecologically valid training data, and in not assuming that each text or utterance has a main language.

2 Related Work

2.1 Mixed Texts

Multilingual documents may comprise more than one language for various reasons, including translation, change of author/speaker, use of loanwords, and code-switching (CS). For this reason, the term bilingual (or multilingual) as applied to corpora can be ambiguous, referencing a parallel corpus such as Europarl (Koehn, 2002) as well as a speech corpus in which more intimate language mixing is present (e.g., the *BilingBank Spanish-English Miami Corpus* (Deuchar, 2010)). King & Abney (2013) have noted that it is desirable that a language identification annotation system operate accurately irrespective of whether it is processing a document that contains monolingual texts from different languages or texts in which single authors are mixing different languages (Das & Gambäck, 2013; Gambäck & Das, 2014, Gambäck & Das, 2016; Nguyen & Doğruöz, 2013; Chittaranjan et al., 2014) . For lin-

Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 12–20,
Austin, TX, November 1, 2016.

guists with interests in patterns of CS there is also a need to be able to classify types of mixed multilingual documents. CS is not monolithic—it can range from switching for lone lexical items and multiword expressions to alternation of clauses and larger stretches of discourse within an individual's speech or across speech turns—and different types of CS invite different types of analyses and reflect different social conditions and types of grammatical integration.

2.2 Mixing Typology

There is consensus that 'classic' or intrasentential code-switching, of all mixing phenomena, is most revealing of the interaction of grammatical systems (Joshi 1982, Muysken 2000, Myers-Scotton 1993, Poplack 1980). Muysken (2000) presents a typology of mixing, identifying three processes—insertion, alternation, and congruent lexicalization—each reflecting different levels of contributions of lexical items and structures from two (or more) languages and each associated with different historical and cultural embedding. Insertional switching, (Example 1, Rampton et al. 2006:1) involves the grammatical and lexical properties of one language as the Matrix Language (Myers-Scotton, 1993) which supplies the morphosyntactic frame into which chunks of the other language are introduced (e.g., borrowing and small constituent insertion). Insertion is argued to be prevalent in postcolonial and immigrant settings where there is asymmetry in speakers' competence of both languages. In alternational switching (Example 2, Nortier 1990: 126) the participating languages are juxtaposed and speakers are said to draw on 'universal combinatory' principles in building equivalence between discrete language systems while maintaining the integrity of each (MacSwan, 2000; Sebba, 2009). Alternation is purported to be most common among proficient bilinguals in situations of stable bilingualism. In a third type, congruent lexicalization (Example 3, Van Dulm, 2007:7; cited in Muysken 2014), the syntax of the languages are aligned and speakers produce a common structure using words from both languages; it is claimed to be attested among bilinguals who are fluent in typologically similar languages of equal prestige as well as in dialect/standard and post-creole/lexifier mixing. Muysken (2013) augments this tripartite taxonomy by incorporating a fourth strategy, backflagging (Example 4, DuBois & Horvath, 2002: 276), in which the grammatical and lexical properties of the majority language serve as the base language into which emblematic minority elements are inserted (e.g., greetings, kinship terms); speakers may select this strategy to signal ethnic identities once they have shifted to the majority language.

- Example 1, English/Punjabi
  I don't mix with <kɑɭe:> ('black boys')

- Example 2, Moroccan Arabic/Dutch
  <Maar 't hoeft niet> li-ʔanna ida seft ana ('But it need not be, for when I see, I . . . ')

- Example 3, English/Afrikaans
  You've got no idea how <vinnig> I've been <slaan-ing> this <by mekaar>
  ('You have no idea how quickly I've been throwing this together')

- Example 4, English/French in Louisiana
  <Ça va>. Why don't you rewire this place and get some regular light switches? ('It's okay.')

2.3 Mixing types as correlates of social differences

Social factors are the source of variation in CS patterns (Gardner-Chloros, 2009). The same language pairings can be combined in various ways and with varying frequency depending on a range of social variables. Post (2015) found gender to be a significant predictor of both frequency and type of switching among Arabic-French bilingual university students in Morocco. Vu, Adel & Schultz (2013) showed that syntactic patterns of Mandarin-English CS differ according the regional origin of the speaker (Singapore vs. Malaysia). Poplack (1987) observed that CS patterns reflected the differential status of French and English in the adjacent Canadian communities of Ottawa and Hull. Larsen (2014) demonstrated that there are significant differences in the frequency of English unigram and bigram insertions in Argentine newspapers destined for distinct social classes of readerships. In contrast, Bullock, Serigos & Toribio (2016) report that in Puerto Rico, where degree of language contact is

stronger, it is the presence of longer spans of English (3+gram but not uni- and bigram) that correlates with higher social prestige.

2.4 Matrix language

In linguistic CS research, the Matrix Language (ML) refers to the morphosyntactic frame provided by the grammar of one of the contributing languages as distinct from lexical items or spans (islands) from embedded languages (Myers-Scotton, 1993). The ML cannot be assumed to be the most frequent language, instead it must be discovered via grammatical analysis.

2.5 Multilingual Indexes

For sociolinguists, Barnett et al. (2000) created a mixing index M to calculate the relative distribution of languages within a given document. Values range from 0 (a monolingual text) to 1 (a text with even distribution of languages). The M-index is valuable in that it indicates the degree to which various languages are represented in a text; its limitation is that it does not show how the languages are integrated and, as a consequence, cannot provide an index of CS versus the wholesale shift from one monolingual text to another in a document. Methods of estimating the proportion of languages in large corpora like Wikipedia have been proposed by Lui, Lau & Baldwin (2014) and by Prager (1999).

2.6 Integration Index

Gambäck & Das (2014) created an initial Code-Mixing Index (CMI) based on the ratio of language tokens that are from the majority language of the text, which they call the matrix language. Like the M-index, CMI does not take account of the integration of CS, thus Gambäck & Das (2016) present a more complex formulation that enhances the CMI with a measure of integration that is applied first to the utterance level and then at the corpus level.

2.7 Language Signature of a document

In their description of the Bangor Autoglosser, a multilingual tagger for transcriptions of Welsh, Spanish, and English conversations in which languages are manually annotated, Donnelly & Deuchar (2011) underline the utility of their system for visualizing the shifting of languages during the course of a conversation, but they make no attempt to quantify language integration, a central point of interest for linguists and one we address here.

2.8 Language Identification

Language identification in multilingual documents continues to present challenges (Solorio et al. 2014 for the first shared task on language identification in CS data). Researchers have tested a combination of methods (dictionaries, n-grams, and machine learning models) for identifying language or for predicting switching, mostly at the word level, with varying degrees of accuracy (Elfardi & Diab, 2014; King & Abney, 2013; Solorio & Liu, 2008a, 2008b; Nguyen & Doğruöz, 2013; Rodrigues, 2012). Because transcriptions of spoken CS are rare, researchers have drawn on social media, particularly Twitter (Bali et al., 2014; Vilares, Alonso, & Gómez-Rodríguez, 2016; Çetinoglu, 2016; Jurgens, Dimitrov & Ruths, 2014; Samih & Maier, 2016), as well as artificially generated texts to develop NLP tools that support the processing of mixed-language data (Lui, Lau & Baldwin, 2014; Yamaguchi & Tanaka-Ishii, 2012) .

3 Language Model

Our language model produces two tiers of annotation: language (Spanish, English, Punctuation, or Number) and Named Entity (yes or no). For the language tier, two heuristics are applied first to identify punctuation and number. For tokens that are not identified as either, a character n-gram (5-gram) and first order Hidden Markov Model (HMM) model, trained on language tags from the gold standard, are employed to determine whether the token is Spanish or English. Two versions of the character n-gram model were tested. One is trained on the CALLHOME American English and CALLHOME Spanish transcripts. The second n-gram model is trained on two subtitle corpora: the $\text{SUBTLEX}_{\text{US}}$ corpus representing English and the ACTIV-ES representing Spanish. Though in its entirety, the $\text{SUBTLEX}_{\text{US}}$ corpus contains 50 million words, only a 3 million-word section was used to remain consistent with the ~3 million words in the ACTIV-ES corpus. Both these corpora provide balanced content as they include subtitles from film and television covering a broad range of genres. The

validity of film and television subtitle corpora to best represent word frequency has been successfully tested by Brysbaert & New (2009). For the Named Entity tier, we use the Stanford Named Entity recognizer with both the English and Spanish parameters. If either Entity recognizer identified the token as a named entity, it was tagged as a named entity. Unlike other taggers where named entities are viewed as language neutrals, our named entities retained their language identification from their first tier of annotation (Çetinoglu, 2016).

4 Integration Index

In order to quantify the amount of switching between languages in a text, we offer the I-Index, which serves as a complement to the M-index (Barnett et al., 2000). It is a computationally simpler version of the revised CMI index of Gambäck & Das (2016) and one which does not require the segmentation of the corpus into utterances or require computing weights. Consistent with principles of CS, our approach does not assume a matrix language. Consider the two examples below.

- Example 5 (Spanish-English, KC)
  Anyway, al taxista right away le noté un acentito, not too specific.
- Example 6 (Spanish-English, YYB)
  Sí, ¿y lo otro no lo es? Scratch the knob and I'll kill you.

Ex.	1	2	3	4	5	6
Lg. 1	4	4	8	2	6	7
Lg. 2	1	5	4	12	6	7
CS	1	1	5	1	4	1
M	0.47	0.98	0.8	0.32	1	1
I	0.25	0.125	0.45	0.08	0.36	0.08

Table 1: Spans for Examples

Examples 5 and 6 contain perfectly balanced Spanish/English usage, reflected in their M-index of 1. However, the two languages are much more integrated in the first sentence, with four switch points, when compared to the second sentence, with just one switch point. Their respective integration, or I-index, captures this difference. Additionally, Example 2 and 3 each have high M-index values but differ in the I-index values in ways that might be predicted by social context: English-Afrikaans contact lends itself to congruent lexicalization, while Moroccan-Arabic-Dutch shows low integration insertion common of immigrant settings. The I-index is calculated as follows. Given a corpus composed of tokens tagged by language $\{l_i\}$ where i ranges from 1 to n, the size of the corpus, the I-index is calculated by the following expression:

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

where $S(l_i, l_j) = 1$ if $l_i \neq l_j$ and 0 otherwise.The factor of $1/(n-1)$ reflects the fact that there are $n-1$ possible switch sites in a corpus of size n. The I-index can also be calculated using the transition probabilities generated from a first-order Hidden Markov Model on an annotated corpus (ignoring the language independent tags) by summing only the probabilities where there has been a language switch. The I-index is an intuitive 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. For a 10-word sentence in which each word is contributed by a different language, Gämback & Das's (2016) maximum integration is .90 rather than 1.

5 Language Signature

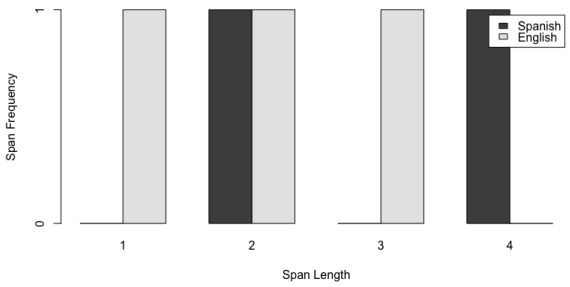

Figure 1: Example 5 Span Distribution

In order to visualize the level of language integration along with language spans, we offer the concept of a language signature that takes into account span length and frequency to derive a unique pattern per document. For Example 5,

	ACTIV-ES & SUBTLEX$_{US}$			CALLHOME		
Language	Accuracy	Precision	Recall	Accuracy	Precision	Recall
English	0.9507	0.9332	0.9729	0.9343	0.8931	0.9893
Spanish	0.9479	0.9021	0.9853	0.9442	0.9286	0.9422

Table 2: Accuracy of language detection on KC using different training corpora

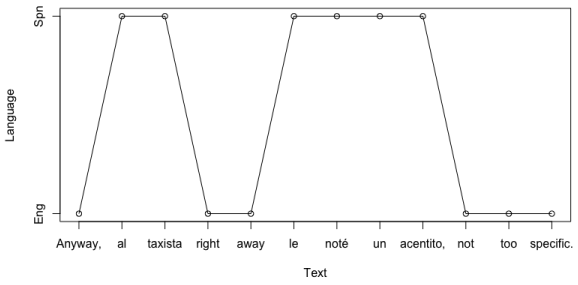

Figure 2: Example 5 Span Plot

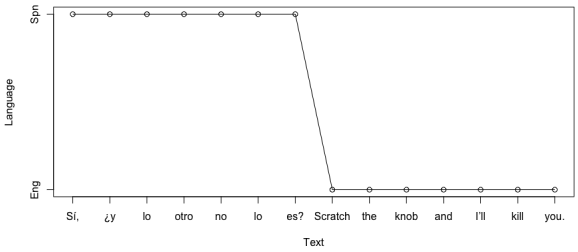

Figure 3: Example 6 Span Plot

there are spans in English of length one, two and three words. In Spanish, there are language spans of length two and four words. Although not particularly revealing with such a short segment, these distributions result in a histogram plot as shown (Figure 1).

In combination with the I-indices, these plots (Figures 1, 2, and 3) display a unique insight into the nature of language mixing and the extent of integration. In contrast to the singular data point of the I-index, the span plots provide a multi-level view of how and to what extent CS occurs in the text.

6 Experiments

6.1 Dataset

Because our interest here is in exploiting language identification for the purpose of detecting variable CS patterns, we draw on two literary texts that we know to employ extensive Spanish-English CS but in two distinctly different styles. Killer Crónicas: Bilingual Memoires (KC), by the Jewish Chicana writer Susana Chávez-Silverman (2004), is a 40,469-word work of creative nonfiction that chronicles the author's daily life through a series of letters that began as email messages written entirely in 'Spanglish'. Yo-Yo Boing! (YYB), by the Puerto Rican writer Giannina Braschi (1998), is a 58,494-word hybrid of languages and genres, incorporating Spanish, English, and 'Spanglish' monologues, dialogues, poetry, and essays. These popular press texts are available online and were used with the permission of the authors.

6.2 Evaluation

The effectiveness of our model was evaluated on a gold standard of 10k words from KC. The segment was selected from the middle of the text, beginning at token 10,000. It was tagged for language by a Spanish-English bilingual professional linguist and 10% was inspected by a second bilingual professional linguist for accuracy. The annotators agreed on all but 2 of the 1000 tokens. The gold standard includes the following tags: Spanish, English, Punctuation, Number, Named Entity, Nonstandard Spanish, Nonstandard English, Mixed along with three other language tags (French, Italian, Yiddish). The Nonstandard tags included forms such as *cashe ~ calle* 'street' to represent dialectal differences, and the Mixed tags included any tokens with morphology from two or more languages such as *hangueando ~ hanging out*. Since Spanish, English, Punctuation, Number and Named Entity account for over 98% of the gold standard, only those tags were used in our model. For the evaluation, we relabel the nonstandard tags to their respective languages and the other languages were ignored.

6.3 Results

As seen in Table 3, despite the close similarity in M-index for the two corpora, the I-index demonstrates

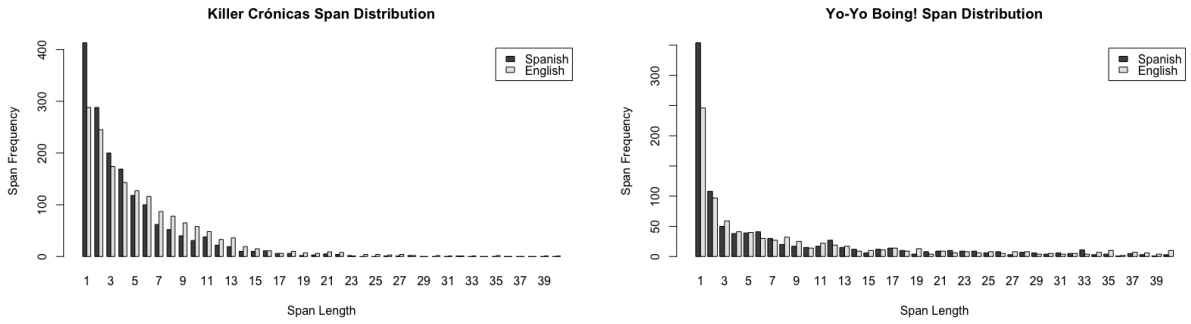

Figure 4: Span Distributions

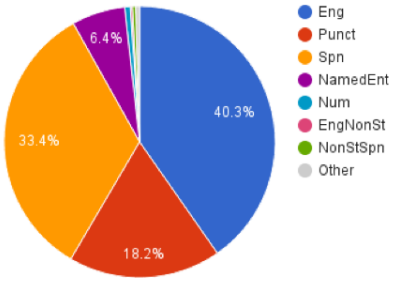

Figure 5: Distribution of Tags

the difference in CS between them; KC has a higher integration of languages than YYB. In Figure 4, we see that even though both corpora contain switches, KC has a much higher incidence of short, switched spans in both languages, increasing its I-index relative to YYB.

As shown in Figure 4, KC displays a rapid exponential decay in span length vs. frequency, whereas YYB does not. In addition, YYB displays a heavy tail, indicating a higher frequency of large spans in both languages compared to KC, which has very few spans longer than twenty words.

Table 2 shows our model's performance on language tagging the 10k gold standard of KC using two separate sets of training corpora.

However, as shown in Table 2, our original results

Corpus	M-index	I-index
Killer Crónicas	0.96	0.197
Yo-Yo Boing!	0.95	0.034
EN-ES	0.72	0.067

Table 3: Language Integration and Mixing

	Accuracy	Precision	Recall
Same	96.72%	79.19%	65.30%
Opposite	88.92%	33.24%	74.85%
English	96.65%	83.94%	58.08%
Spanish	89.00%	34.42%	82.06%

Table 4: NER Classification Performance

using the CALLHOME corpus reflect a lower performance due to the disparity in size of the corpora (3.1M Eng, 976K Mex). Using these corpora resulted in recurring mistakes such as identifying the word "ti" as English due to the overabundance of the acronym "TI" in the English CALLHOME corpus relative to the Mexican Spanish corpus. Additionally, common words in both languages such as "a" or "me" were initially tagged as English for similar reasons. In contrast, changing to equal-size corpora of 3.5M words (ACTIV-ES and $\text{SUBTLEX}_{\text{US}}$) resulted in a quantitative increase of 1% in language accuracy for both languages as seen in Table 2 and better tagging of "ti", "me" and "a" in mixed contexts.

Furthermore, we chose four different methods of classifying named entities as shown in Table 4: using the classifier in the same language as the token, the opposite language, only the English classifier, and only the Spanish classifier. The Stanford Named Entity Recognizer clearly over identifies named entities, reflected in its low precision scores. We found that using only the English classifier in all cases recognized named entities with the highest precision, but the Spanish classifier resulted in the highest recall rate. Finally, using the classifier in the same language as the token is only marginally better in

accuracy than relying purely on the English classifier.

7 Conclusion

In this paper we provided an intuitively simple and easily calculated measure—the I-index—for quantifying language integration in multilingual texts that does not require weighting, identification of a matrix language, or dividing corpora into utterances. We also presented methods of visualizing the language profile of mixed-language documents. To illustrate the I-index and how it differs from a measure that shows the ratio of languages mixed in a text (the M-index), we created an automatic language-identification system for classifying Spanish-English bilingual documents. Our annotation system is similar to that of Solorio & Liu (2008a, 2008b), which includes an n-gram method and a bi-gram HMM model for probabilistically assigning language tags at the word level. We improved accuracy by 1% in our model by using training corpora that reflected natural dialogue and we used different methods of classifying Named Entities in an attempt to reduce the greediness of the Named Entity Recognizer. Our automatic procedure achieves high accuracy in language identification, and although the texts examined proved to be equally bilingual, our analysis demonstrated that the languages are integrated very differently in the two data sets, a distribution that can be intuitively depicted visually.

The implication is that though texts might be mixed, only some texts are suitable for the study of intrasentential CS. For instance, the I-index metric indicates that any random selection from KC, but not YYB, would likely contain intersentential CS. As linguists move to exploit larger multilingual datasets to examine language interaction (Jurgens et al., 2014; Bali et al., 2014), it is crucial to have an uncomplicated metric of how the languages are integrated because different types of integration correlate with different social contexts and are of interest for different domains of linguistic inquiry.

References

Kalika Bali, Yogarshi Vyas, Jatin Sharma, and Monojit Choudhury. 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, EMNLP. pages 116–126.

R. Barnett, E. Codo, E. Eppler, M. Forcadell, P. Gardner-Chloros, R. van Hout, M. Moyer, M. C. Torras, M. T. Turell, M. Sebba, M. Starren, and S. Wensing. 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):131–132, June.

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

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

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990, November.

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: A computational approach. In Rosa Guzzardo Tamargo, Catherine M. Mazak, and M. Carmen Parafita Cuoto, editors, *Spanish-English code-switching in the Caribbean and the U.S.*, pages 171–189. Benjamins, Amsterdam ; Philadelphia.

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

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

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using crf: Code-switching shared task report of msr india system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 73–79.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text: the last language identification frontier. *Traitement Automatique des Langues (TAL): Special Issue on Social Networks and NLP*, 54(3).

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. *11th International Conference on Natural Language Processing*.

Margaret Deuchar. 2010. BilingBank Spanish-English Miami Corpus.

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

Sylvie Dubois and Barbara M. Horvath. 2002. Sounding Cajun: The Rhetorical Use of Dialect in Speech and Writing. *American Speech*, 77(3):264–287.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. A hybrid system for code switch point detection in informal Arabic text. *XRDS: Crossroads, The ACM Magazine for Students*, 21(1):52–57, October.

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.

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

Penelope Gardner-Chloros. 2009. Sociolinguistic factors in code-switching. In Barbara E. Bullock and Almeida Jacqueline Toribio, editors, *The Cambridge handbook of linguistic code-switching*, pages 97–113. Cambridge University Press, Cambridge, UK.

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.

David Jurgens, Stefan Dimitrov, and Derek Ruths. 2014. Twitter users# codeswitch hashtags!# moltoimportante# wow. *EMNLP 2014*, pages 51–61.

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

Philipp Koehn. 2002. *Europarl: A multilingual corpus for evaluation of machine translation*.

William Labov. 1972. Some Principles of Linguistic Methodology. *Language in Society*, 1(1):97–120, April.

Jacqueline Rae Larsen. 2014. *Social stratification of loanwords: a corpus-based approach to Anglicisms in Argentina*. Masters Thesis, University of Texas at Austin.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

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

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

Pieter Muysken. 2013. Language contact outcomes as the result of bilingual optimization strategies. *Bilingualism: Language and cognition*, 16(04):709–730.

Pieter Muysken. 2014. DÉJÀ VOODOO OR NEW TRAILS AHEAD? *Linguistic Variation: Confronting Fact and Theory*, page 242.

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

Dong-Phuong Nguyen and A. Seza Dogruoz. 2013. Word level language identification in online multilingual communication. Association for Computational Linguistics.

Jacomine Nortier. 1990. *Dutch-Moroccan Arabic code switching among Moroccans in the Netherlands*. Foris, Dordrecht.

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.

Shana Poplack. 1987. Contrasting patterns of code-switching in two communities. In Erling Wande, Jan Anward, Bengt Nordberg, Lars Steensland, and Mats Thelander, editors, *Aspects of multilingualism: Proceedings from the Fourth Nordic Symposium on Bilingualism, 1984*, pages 51–77. Borgströms, Uppsala.

Rebekah Elizabeth Post. 2015. *The impact of social factors on the use of Arabic-French code-switching in speech and IM in Morocco*. Ph.D. thesis, University of Texas at Austin.

John M. Prager. 1999. Linguini: Language identification for multilingual documents. *Journal of Management Information Systems*, 16(3):71–101.

Ben Rampton, Roxy Harris, and Lauren Small. 2006. The meanings of ethnicity in discursive interaction: Some data and interpretations from ethnographic sociolinguistics. *ESRC Identities and Social Action Programme*, pages 1–14.

Paul Rodrigues. 2012. *Processing highly variant language using incremental model selection*. Ph.D. thesis, Indiana University.

Younes Samih and Wolfgang Maier. 2016. An Arabic-Moroccan Darija Code-Switched Corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4170–4175.

Mark Sebba. 2009. On the notions of congruence and convergence in code-switching. In Barbara E. Bullock and Almeida Jacqueline Toribio, editors, *The Cambridge handbook of linguistic code-switching*, pages

40–57. Cambridge University Press, Cambridge, UK ; New York.

Thamar Solorio and Yang Liu. 2008. 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, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and others. 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. Citeseer.

Ondene van Dulm. 2007. *The grammar of English-Afrikaans code switching: A feature checking account*. Utrecht: LOT.

David Vilares, Miguel Alonso, and Carlos Gómez-Rodríguez. 2016. 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)*, pages 4149–4153.

Ngoc Thang Vu, Heike Adel, and Tanja Schultz. 2013. An investigation of code-switching attitude dependent language modeling. *Statistical Language and Speech Processing*, pages 297–308.

Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. 2012. Text segmentation by language using minimum description length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 969–978. Association for Computational Linguistics.