Interspeech 2018
2-6 September 2018, Hyderabad

Should code-switching models be asymmetric?

Barbara E. Bullock[1], *Gualberto Guzmán*[1], *Jacqueline Serigos*[2], *Almeida Jacqueline Toribio*[1]

[1]University of Texas at Austin
[2]George Mason University

bbullock@austin.utexas.edu, gualbertoguzman@utexas.edu, jserigos@gmu.edu, toribio@austin.utexas.edu

Abstract

Since the work of Joshi [1], most models of code-switching (C-S) have assumed asymmetry of the participating languages. While there exist patterns of language mixing in which a dominant or matrix language (ML) may not be discernible, these more complex signatures are rarely modeled [2, 3]. We use a series of metrics to characterize the switching in corpora as asymmetrical (insertional C-S) or symmetrical (alternational C-S). We test the efficacy of a linguistic model that assumes no ML in predicting the syntax of C-S in three Spanish–English corpora that vary according to whether the ML is Spanish, English or indeterminate. Our results show that the same constraints on the grammatical junctures and on the directionality of switching hold irrespective of the symmetry of the data. The length of the alternating language spans varies according to POS with noun phrases comprising the shortest spans. This suggests that insertional C-S may be subsumed under alternational C-S, as spontaneous borrowing. These results invite researchers to reconsider the linguistic theories they adopt and to expand the typology of training data used in creating language models and processing tools for C-S.

Index Terms: code-switching, matrix language, language modeling, POS tagging

1. Introduction

Muysken (2000) distinguishes between the *insertion* of other-language tokens into the structure of another from the *alternation* between structures of two languages [2]. This approach to C-S patterns has been variously formalized, in the asymmetrical Matrix Language Frame (MLF) model [4] and by symmetrical models, including the linear Equivalence Constraint [5] and the hierarchical Functional Head Constraint [6]. Based on Joshi [1], 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) are inserted, as well as an asymmetry between system vs. content morphemes. Constraints on C-S follow from these asymmetries: the ML provides the grammatical elements and the EL merely content morphemes. An alternative approach is presented in the Functional Head Constraint (FHC), a generative account that draws on a principle of syntactic coherence: f-selection, proposed by Abney [6, 7]. The proposal rests on the dichotomy between functional and lexical elements: functional elements (e.g., DET, AUX/MOD, NEG, COMP) and their complements are assumed to share a strong syntactic relationship, hence switching between them is restricted. Importantly for our purposes, the FHC does not assume or prescribe a morphosyntactic base or matrix language.

While asymmetrical models have been widely assumed in NLP research [1, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21], Bhat et al. (2016) finds the MLF model to be incomplete and unsound, without a complementary symmetrical model [5], for building a working model of C-S [22]. They correctly point out that the MLF model is underspecified with respect to the categories that can or cannot be switched, a criticism that has also been leveled by linguists. Researchers who do not necessarily espouse the MLF model have operationalized the ML by various methods, including word frequency [23, 24], language of the verb [2], and language of verbal inflection [1, 25, 26], highlighting the difficulty in establishing clear criteria by which the ML is to be identified.

The notion of a ML is attractive because insertional C-S is commonplace, as bilinguals present *nonce borrowings*, short, other-language strings that are inserted into a ML [27]. This insertional-type mixing is said to be prevalent in colonial and immigrant settings in which there is an asymmetry between the dominant and minority languages. But in situations of sustained bilingualism, bilinguals may also alternate between grammatical systems, in which case, a single ML cannot be ascertained. For instance, in the United States, alternational-type C-S is widespread among Spanish-English bilinguals who move freely between languages in what is popularly coined *Spanglish*, as in the excerpt in (1) from Jewish-Latina writer Susana Chávez Silverman.

1. Pero siempre entro a los <shops, especially> los más <trendy> y <fancy, feeling> medio <cowed and pale, definitively out of place>
   'But I always enter those shops, especially the most trendy and fancy, feeling sort of cowed and pale, definitely out of place'

The distinction between the two types of language mixing is revealed in Adamou's (2016) analysis of data from the Pangloss collection of endangered European languages [3]. On the basis of token counts, she separates corpora into those in which <5% of tokens are inserted from a contact language into the endangered language and the other in which 20-35% are embedded, representing insertional and alternational patterns, respectively.

Because insertional and alternational C-S are both attested language practices in bilingual speech communities, the value of a model that is predicated on only one type of C-S is unclear. Models that assume asymmetry, like the MLF, make no predictions about more complex data, where the languages are more equally balanced and the 'language' of the utterance or conversation less certain. The question that arises is whether a model that does not assume asymmetry can make the correct predictions for both types of language mixing.

We begin by first presenting metrics that allow us to characterize the C-S in a corpus as insertional or as alternational. For our analysis, we assume that an insertional C-S signature implies a ML. Using three manually POS-tagged Spanish–English datasets remapped to a Universal tagset [28], we identify the

 10.21437/Interspeech.2018-1284

type of switching they entail and define the ML if there is one. Having confirmed that each corpus has a different profile, we model C-S across the data sets. Our results demonstrate that switching is overwhelmingly intersentential but that when it occurs intrasententially (within a clause), the same factors guide C-S regardless of whether or not there is a ML. But DET-NP switches show an asymmetry that cannot be predicted by a matrix language. To further examine this asymmetry, we perform additional tests on the length of the NP spans that follow a DET under the hypothesis that the NPs are likely to be nonce borrowings in the sense of Poplack et al. rather than switches [27].

The organization of the paper is as follows. Section 2 describes the metrics used to quantify the nature of C-S at the corpus level. Section 3 introduces the datasets and the application of the metrics. Section 4 presents the preprocessing necessary for our experiments and provides the results of the same. We discuss these results in Section 5 together with their implications for the processing of C-S data for language models.

2. Metrics for characterising CS type

To model the language mixing in our corpora, we used the automatic Language Identification (LID) procedure developed in Guzmán et al. [29, 30], which produces two tiers of annotation. The Language tier includes tags for Spanish, English, number, and punctuation, and the Named Entity tier includes labels recognized by the Spanish or English version of the Stanford Named entity recognizer [31]. The model works as follows: numbers and punctuations are identified by an algorithm and a character n-gram (5 gram) model identifies the language of the remaining tokens. The language models are trained on subtitle data that better match the word frequencies encountered in speech than does CALL-type data. For Spanish, we used the 3 million word ACTIVE-$_{\text{ES}}$ corpus and, for English, the SUBTLEX-$_{\text{US}}$ corpus [32, 33] . To disambiguate the homographs that occur across Spanish and English, the LID uses a HMM. The model achieves high accuracy of 97%.

The LID procedure returns a sequence of language tags that are the basis for metrics for characterizing the nature of switching across the corpora. The Multilingual-Index, or *M-Index*, calculates the proportion of languages represented in a corpus [34]; it is bounded between 0 (all tokens are from the same language) and 1 (every language is equally represented). The Integration-Index, or *I-Index*, calculates the probability of switching languages between any two tokens within a corpus [29]; it is bounded between 0 (no switching) and 1 (each token bears a language tag different from the previous one). Finally, *Burstiness* yields information on whether switching events occur regularly or aperiodically; it is bounded between -1 (periodic, regular events) and 1 (aperiodic, bursty events)[35].

3. Data and application of the metrics

Our data include three Spanish-English bilingual data sets of similar size (approx 8,000 words): *S7*, the transcript of a conversation among three Spanish-English bilinguals [36, 37]; *M40*, a file of recorded conversation from the Miami Corpus of the BilingBank repository [38]; and *KC*, an excerpt from the novel *Killer Crónicas* [39]. In order to be consistent, we processed each of the corpora using our LID. Table 1 presents the results of the application of the metrics. KC, with an M-index of .99, contains nearly equal proportions of Spanish (47%) and English (53%). In contrast, in *S7* and *M40* the number of tokens from one language far exceeds the other; English represents the majority in *S7* (75%), while Spanish represents the majority in *M40* (74%). The I-index reflects the higher probability of switching in *KC* relative to *S7* and *M40*, and the Burstiness calculation indicates that switching is more regular in *KC* than in *S7* and *M40*, where switching occurs sporadically.

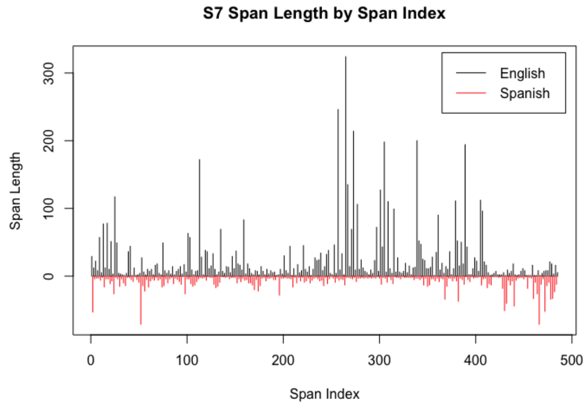

Figure 1: *Language Spans through S7*

Table 1: *Corpus Metrics*

Corpus	M-Index	I-Index	Burstiness
S7	.60	.06	.32
M40	.63	.10	.26
KC	.99	.17	-.06

Given these findings, we regard *KC* as symmetric, or alternational C-S since it has relatively short, regular switching between monolingual language spans, and *S7* and *M40* as asymmetrical, or insertional, C-S. The latter are characterized by sporadic insertions of an EL into long spans of a ML. English serves as the ML of *S7* and Spanish as the ML of *M40*.

Additionally, we captured the length of each monolingual span by demarcating the beginning and end of every sequence of words bearing the same language label. These spans, when arranged chronologically through each corpus, allow us to visualize the essential insertional and bursty natures of the Spanish intrusions in the English-dominant *S7* in Fig. 1 and of the English spans in the Spanish-dominant extract *M40* (Fig. 2) juxtaposed to the periodic, alternating spans of both languages throughout the extract of *KC* in Fig. 3. Note that relative to the asymmetric corpora, *KC* has shorter spans in each language (30 tokens or less) and as a consequence there are at least twice the number of spans in this 8,000 token text than there are in the others of the same length. The visualization also indicates that the ML of the asymmetric texts may change throughout the conversation. For instance, there seems to be a noticeable uptick in English spans in the Spanish-dominant *M40* conversation at or around the 350th span.

4. Syntactic experiments

In order to undertake a syntactic analysis of these data, we manually tagged *KC* with the 12 labels of the Universal POS Tagset [28] that has been used in corpus-based analyses of C-S in other language pairings, including Hindi-English [40] and

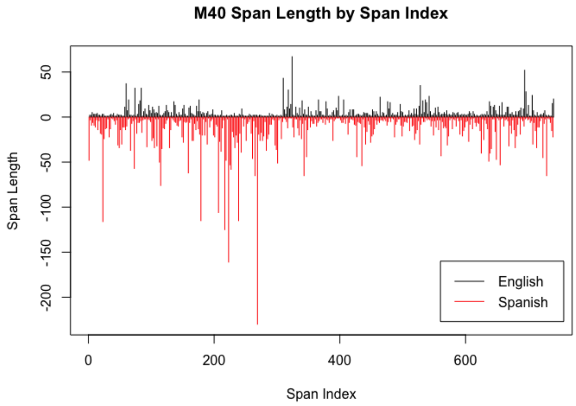

Figure 2: *Language Spans through M40*

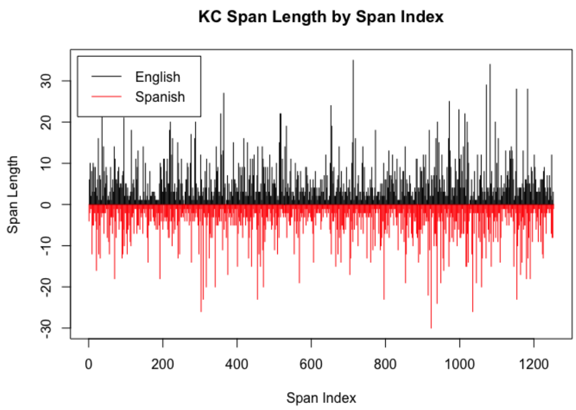

Figure 3: *Language Spans through KC*

Latin-Middle English [41], and so permits cross-corpus comparisons. We wrote an algorithm to remap the POS tags of *S7* from the Penn Treebank tags used by the corpus creators [36, 37] to the Universal set and to similarly convert the *M40* tags from the specialized set provided by its authors [38] to the same set. Following these steps, every token in each corpus bears a LID and POS tag that conforms to the same parameters.

We fit logistic regressions to the data using the rms package of R [42] with Switch/No-Switch as the dependent variable. The independent variables were the POS tag of the token preceding the switch (PreviousPOS), the language of the token preceding the switch (PreviousLang), and the corpus in which the token was observed. We set the reference levels of the predictor variables as follows: PreviousLang defaulted to English alphabetically. We releveled the corpus factor to the alternational dataset *KC* as it was the most distinct from the other two and we releveled the PreviousPOS factor to VERB, on the reasoning that, under this POS coding scheme, it was neutral with respect to C-S. Some verbs (auxiliaries and modals) should block switching while lexical verbs should permit it.

Most linguistic models predict that intrasentential switching should be blocked after functional elements, including conjunctions, determiners, and auxiliaries/modals. Switching after pronouns is also excluded on the syntactic grounds that they are syntactically weak heads rather than phrases. A symmetric

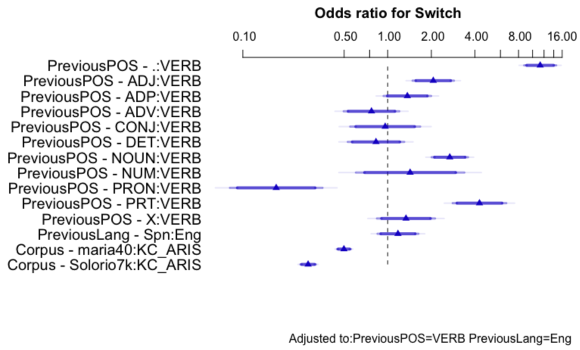

Figure 4: *Odds ratios results for language switching.*

Table 2: *C-S per total Verb-Verb transitions*

	KC		S7		M40	
Main Verb Lang.	Eng	Span	Eng	Span	Eng	Span
Eng	57	2	154	0	59	1
Span	2	29	0	27	3	102

model makes no prediction with regard to language.

5. Results

The results of the best fitting model are shown in Fig. 4, where the odds ratios of the model terms are given with their confidence levels (.90, .95, .99) given in progressively darker colors. Because of the dramatically different distributions of languages in these corpora, models with interactions between PreviousLang and PreviousPOS failed to converge. The results presented here are, then, main effects.

The vertical line marks the odds ratio of 1 so that odds ratio lines that fall to the right of the vertical line are significant positive predictors, while those that fall to the left are significant negative predictors. The abbreviations for the factors can be read as follows: '.' all punctuation, 'ADJ' adjectives, 'ADP' prepositions, 'VERB' auxiliary, modals and lexical verbs, 'ADV' adverb, 'CONJ' subordinating and coordinating conjunctions, 'DET' definite, indefinite, and possessive determiners, 'NUM' ordinal and cardinal numbers, 'PRON' subject and object pronouns, 'PRT' particles, 'X' other, 'KC_Aris' *KC*, 'maria40' *m40*, 'Solorio7k' *S7*.

The general trends suggest that C-S, as expected, is much more likely in the alternational corpus,*KC*, than in either of the insertional corpora, *S7* and *M40*. Switching is also 11 times more likely to occur after punctuation, hence at phrasal boundaries, than it is elsewhere. At the other extreme, C-S after a pronoun is vanishingly rare, as predicted.

Switching should be prohibited after support verbs (auxiliaries and models). However, the tagset we used does not break down verbs by type, so a post-hoc test is required to examine the effect of this grammatical juncture. The contingency table for these transitions is shown in Table 2. As predicted by models like the FHC, there are very few instances of C-S between contiguous verbs, the first of which would be expected to be a support verb in either language of this pairing. The values also suggest that there is no directional asymmetry at this juncture: C-S between an auxiliary and a VERB is rare, if not entirely absent as in *S7*, irrespective of the language of the support verb.

The FHC would also predict that switching should be dis-

Table 3: *C-S per total DET-NP transitions*

	KC		S7		M40	
DET Lang.	Eng NP	Span NP	Eng NP	Span NP	Eng NP	Span NP
Eng	498	60	399	3	133	11
Span	138	353	22	122	56	460

favored after DET, a functional element. While this trends negatively, it is not a statistically significant negative predictor in this model, either because much of the variance in the model is accounted for by punctuation as a predictor or because there is an interaction with language. The raw data provides insight into this interaction.

The crosstabulation of the DET-NP examines every instance in which a DET is followed by a NOUN or by an ADJ; it is broken down by language in Table 3 for each corpus. Though switching after the DET is not blocked, asymmetries in the distribution of other-language determiners are apparent. For instance, in the most linguistically permissive of the corpora, *KC*, 29% of the Spanish determiners appear with English NPs while only 10.7% of the English determiners occur with Spanish NPs. This same asymmetry is observed in the insertional corpora particularly in the English-dominant *S7* with less than 1% switches after an English DET compared to 7.6% after Spanish DET.

The effect of language on switching at the determiner cannot be accounted for in an asymmetric model because the ML varies across these corpora but the directionality of the effect of language remains unchanged; switching is disfavored when the determiner is English even if the ML is English. Switching after the DET was intentionally set aside by Belazi et al. in formulating the FHC because nouns are frequently borrowed in situations of bilingualism and language contact [6]. Our span data allow us to empirically test the possibility that switching after DET results from borrowing, which entails the insertion of relatively shorter spans after DET than those that occur after other parts of speech. We propose to test span length as a proxy for nonce borrowing.

5.1. Span length as a proxy for nonce borrowing

A follow-up experiment investigates the length of spans following a switch as a function of part of speech. We use span length, calculated as described above, as the dependent variable and the PreviousPOS and PreviousLang tags of the tokens immediately preceding the span as the predictor variables. Because each of our corpora follows a different span distribution, we fit separate models for each corpus.

The results of the best fitting linear regression to predict the length of span based on the immediately preceding POS tag and on previous Language for *KC* finds a significant regression in which length of span is predicted by the main effects of previous POS (reference level: verb) and of previous Language (reference level: English) (R^2 = 0.02133, F(13, 1255) = 3.126, p-value: 0.0001301). Specifically, the average length of span decreased significantly, albeit slightly, in token length (- 1.54) only after a DET. This decrease was significant at the .95 confidence level. There is also a significant main effect for Language ($p < .01$) whereby the average length of span *increases* slightly (+ .73) when the language preceding a switch is Spanish. Recall that, in general, the spans in *KC* are shorter than in the other two corpora.

The best fitting model for *S7* also shows significant main effects of previous POS and of previous language (R^2 = 0.07241, F-statistic (12, 471) = 4.142, p-value: 3.454e-06). The average length of span in S7 decreases significantly at the .95 confidence level only if the previous POS tag is a DET; the average decrease in length (20.69 tokens) is much greater than that of *KC*. In contrast, a highly significant main effect for Language ($p < .001$) shows that the average length of span again *increases* by 19.83 if the previous language of a span is Spanish.

A model fit for *M40* with both POS and Language as main effects returns significance only for Language, where a switch to English results in a decrease in the length of span (- 8.5) as would be expected in a corpus that is predominantly Spanish with English insertions. A model fit with only POS as a main effect shows a significant decrease ($p < .05$) in average span length length (- 7.48) again exclusively after the category of DET (R^2 = 0.0291; F(11, 728) = 3.014, p-value: 0.0006057).

6. Discussion

Despite the differences in the distribution of the languages within them and the sporadic versus regular nature of how the languages alternate, the Spanish–English corpora all show similar patterns; switching is mostly phrasal and is least likely to occur after functional and bound elements. Most intriguing, all three corpora demonstrate that the span lengths following DET, and only following DET, show a statistically significant decrease in length relative to the reference level. Additionally, as *S7* and *KC* on average have longer English span than Spanish, the observed insertions of shorter English NPs exclusively after Spanish DETs goes against the general trend in these two corpora of having longer English span lengths following a Spanish token. This suggests that English noun insertions are nonce borrowings. The asymmetry in borrowing English rather than Spanish nouns, irrespective of the ML, may be due more to the linguistic capital of English as the culturally and socially dominant language in the U.S. than to any linguistic variable [22].

7. Conclusions

In summary, our work presents procedures for quantifying and modeling CS that can be applied to any corpora that has been tagged for language and POS. An implication of this work is that mixed corpora are of two types – alternational or insertional – confirming the analysis of Adamou on endangered languages in Europe [3]. But across types, switching patterns are similar, and the same asymmetries in language appear to arise, potentially due to cultural reasons. Future research with different language pairings and a more granular universal tagset will be able to inform us if these observations hold across different language pairings in diverse social contexts.

8. Acknowledgements

The authors are grateful to Thamar Solorio for sharing her data, to Susana Chávez-Silverman who granted us permission to scan her remarkable text, and to Joseph Ricard.

9. References

[1] A. K. Joshi, "Processing of sentences with intra-sentential code-switching," in *Proceedings of the 9th conference on Computational linguistics-Volume 1*. Academia Praha, 1982, pp. 145–150.

[2] P. Muysken, *Bilingual speech: A typology of code-mixing*. Cambridge University Press, 2000, vol. 11.

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

[4] C. Myers-Scotton, *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press, 1997.

[5] S. Poplack, "Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1," *Linguistics*, vol. 18, no. 7-8, pp. 581–618, 1980.

[6] H. M. Belazi, E. J. Rubin, and A. J. Toribio, "Code switching and x-bar theory: The functional head constraint," *Linguistic inquiry*, pp. 221–237, 1994.

[7] S. P. Abney, "The english noun phrase in its sentential aspect," Ph.D. dissertation, Massachusetts Institute of Technology, 1987.

[8] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury, "Pos tagging of english-hindi code-mixed social media content," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 974–979.

[9] K. Bali, J. Sharma, M. Choudhury, and Y. Vyas, ""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*, 2014, pp. 116–126.

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

[11] 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.

[12] H. Adel, N. T. Vu, and T. Schultz, "Combination of recurrent neural networks and factored language models for code-switching language modeling," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2013, pp. 206–211.

[13] S. Maharjan, E. Blair, S. Bethard, and T. Solorio, "Developing language-tagged corpora for code-switching tweets," in *Proceedings of The 9th Linguistic Annotation Workshop*, 2015, pp. 72–84.

[14] J. a. Gebhardt, "Speech recognition on english-mandarin code-switching data using factored language models," 2011.

[15] H. Elfardy, M. Al-Badrashiny, and M. Diab, "Code switch point detection in arabic," in *International Conference on Application of Natural Language to Information Systems*. Springer, 2013, pp. 412–416.

[16] ——, "Aida: Identifying code switching in informal arabic text," in *Proceedings of The First Workshop on Computational Approaches to Code Switching*, 2014, pp. 94–101.

[17] Y. Li, Y. Yu, and P. Fung, "A mandarin-english code-switching corpus." in *LREC*, 2012, pp. 2515–2519.

[18] Y. Li and P. Fung, "Code-switch language model with inversion constraints for mixed language speech recognition," *Proceedings of COLING 2012*, pp. 1671–1680, 2012.

[19] A. Jamatia, B. Gambäck, and A. Das, "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*, 2015, pp. 239–248.

[20] S. Mandal, S. K. Mahata, and D. Das, "Preparing bengali-english code-mixed corpus for sentiment analysis of indian languages," *arXiv preprint arXiv:1803.04000*, 2018.

[21] S. Ghosh, S. Ghosh, and D. Das, "Complexity metric for code-mixed social media text," *arXiv preprint arXiv:1707.01183*, 2017.

[22] G. Bhat, M. Choudhury, and K. Bali, "Grammatical constraints on intra-sentential code-switching: From theories to working models," *arXiv preprint arXiv:1612.04538*, 2016.

[23] B. Gambäck and A. Das, "Comparing the level of code-switching in corpora." in *LREC*, 2016.

[24] A. Das and B. Gambäck, "Identifying languages at the word level in code-mixed indian social media text," 2014.

[25] J. Treffers-Daller, *Mixing two languages: French-Dutch contact in a comparative perspective*. Walter de Gruyter, 1994, vol. 9.

[26] F. Meakins, *Case-marking in contact: The development and function of case morphology in Gurindji Kriol*. John Benjamins Publishing, 2011, vol. 39.

[27] S. Poplack, D. Sankoff, and C. Miller, "The social correlates and linguistic processes of lexical borrowing and assimilation," 1988.

[28] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," *arXiv preprint arXiv:1104.2086*, 2011.

[29] G. A. Guzman, J. Serigos, B. E. Bullock, and A. J. Toribio, "Simple tools for exploring variation in code-switching for linguists," in *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, 2016, pp. 12–20.

[30] G. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, and A. J. Toribio, "Metrics for modeling code-switching across corpora," *Proc. Interspeech 2017*, pp. 67–71, 2017.

[31] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[32] J. C. Francom, "Activ-es: a novel spanish-language corpus for linguistic and cultural comparisons between communities of the hispanic world," 2013.

[33] J. Francom, M. Hulden, and A. Ussishkin, "Activ-es: a comparable, cross-dialect corpus of'everyday'spanish from argentina, mexico, and spain." in *LREC*, 2014, pp. 1733–1737.

[34] R. Barnett, E. Codó, E. Eppler, M. Forcadell, P. Gardner-Chloros, R. van Hout, M. Moyer, M. C. Torras, M. T. Turell, M. Sebba *et al.*, "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. 131–271, 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, "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.

[37] ——, "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.

[38] K. Donnelly and M. Deuchar, "The bangor autoglosser: a multilingual tagger for conversational text," *ITA11, Wrexham, Wales*, 2011.

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

[40] A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Srivastava, R. Mamidi, and D. M. Sharma, "Shallow parsing pipeline for hindi-english code-mixed social media text," *arXiv preprint arXiv:1604.03136*, 2016.

[41] S. Schulz and M. Keller, "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*, 2016, pp. 43–51.

[42] F. E. Harrell Jr, "rms: Regression modeling strategies. r package version 4.0-0," *City*, 2013.