The Role of Cognate Words, POS Tags, and Entrainment in Code-Switching

Victor Soto, Nishmar Cestero, Julia Hirschberg

Columbia University, New York, USA

vsoto@cs.columbia.edu, nishi@cs.columbia.edu, julia@cs.columbia.edu

Abstract

The linguistic or contextual stimuli that elicit code-switching are largely unknown, despite the fact that these are of key importance to understanding mixed language and building tools that can handle it. In this paper, we test the following hypotheses proposed in linguistics literature: first, that cognate stimuli are directly correlated to code-switching; second, that syntactic information facilitates or inhibits code-switching; and third that speakers *entrain* to one another in code-switching in conversation between bilinguals. In order to test these hypotheses, we built a lexical database of cognate pairs for English-Spanish. Using statistical significance tests on a corpus of conversational code-switched English-Spanish, we found that a) there is strong statistical evidence that cognates and switches occur simultaneously in the same utterance and that cognates facilitate switching when they precede a code-switch, b) there is strong statistical evidence of the relationship between part-of-speech tags and code-switching and c) speakers tend to show converging entrainment behavior with respect to their rate of code-switching in conversation.

Index Terms: code-switching, speech analysis

1. Introduction

Code-switching (CS) is the phenomenon by which multilingual speakers switch between languages in written or spoken communication. For example, an English-Spanish speaker might say “El teacher me dijo que Juanito is very good at math.” CS can be observed in various linguistic levels: phonological, morphological, lexical, and syntactic and can be classified as *intra-sentential* if the switch occurs within the boundaries of a sentence or utterance, or *inter-sentential* if the switch occurs between two sentences or utterances.

Very little research has been done to develop NLP approaches to CS, due largely to the lack of sufficient corpora of high-quality annotated data to train on. Yet CS presents serious challenges to all language technologies, including part-of-speech (POS) tagging, parsing, language modeling, machine translation, and automatic speech recognition, since techniques developed on one language quickly break down when that language is mixed with another. One of the often asked but unresolved questions regarding CS is whether there are particular conditions that facilitate or “trigger” its occurrence. In this paper, we study the influence that *cognate words* (words that exist in two different languages with the same etymological origin and that share similar spelling and meaning), part-of-speech tags and *entrainment* (the tendency of conversational partners to begin behaving like each other) have on CS behavior. Answers to these questions will aid with the design of CS grammars, the developing of features for CS detection, the creation of better language models, and the generation of CS content.

The remainder of the paper is organized as follows. Section 2 describes previous work on the relationship between cognate words, part-of-speech tags, entrainment, and CS. In Section 3, we describe the Miami Bangor (MB) corpus, which is used throughout this paper, and the list of English-Spanish cognate words that we collected from the Internet. Section 4 presents the experimental studies. Section 4.1 describes the analysis of cognate influence on CS, Section 4.2 discusses the role of POS tags in CS, and Section 4.3 discusses entrainment in the MB corpus. Finally, Section 5 presents our conclusions.

2. Previous Work

Work on computational approaches to modeling CS has been increasing in the last few years. Most efforts have focused on language identification and CS detection [1, 2], but there has also been some research on language modeling [3, 4, 5], part-of-speech tagging [6, 7, 8, 9] and even speech recognition [10, 11]. While some of this research has tried to incorporate existing linguistic theories of CS [3, 5], the vast majority have focused on standard machine learning approaches. Ultimately, even if some of these models successfully solve the task they are trained for, they shed little insight into the intrinsic mechanics of CS and why and how it takes place.

On the topic of eliciting code-switching, Michael Clyne proposed his triggering hypothesis which has been reformulated during the years [12, 13, 14]. This hypothesis claims that CS can be facilitated by words that exist in both languages with similar form and meaning if those words occur immediately preceding or immediately following a CS. Those words are said to include lexical transfers, bilingual homophones and proper nouns. Clyne’s triggering hypothesis states that trigger words facilitate code-switching but does not imply direct causality, since it has also been observed that syntactic, prosodic and sociolinguistic factors also play a role. Broersma and Bot [15] evaluated this triggering hypothesis on a corpus of Dutch-Moroccan Arabic transcribed conversations and proposed alternative hypotheses based on modern speech production models. Although they were able to confirm and reject aspects of Clyne’s hypothesis, the corpus used in their analysis is severely limited by its size: 3 speakers, 318 clauses, 1,723 words, of which 60 include instances of CS.

In this paper, we test the triggering hypothesis for CS on a much larger corpus of English-Spanish speech following the methodology proposed in [15]. Our findings confirm some aspects of the hypothesis with much higher statistical power than Broersma and Bot’s findings [15].

There has been much research on the topic of syntax and CS, mainly focusing on the study of how multiple monolingual grammars interact to produce mixed speech [16] and whether they work together in a symmetric relationship [17] or whether one is subsumed by the other, matrix language [18, 19]. POS tags have played a role in many of these theories, typically being used to identify constraints that researchers have observed in their data. In this paper, we test the significance of the statistical relationship between CS and POS tags and inspect the role of different part-of-speech tags in the triggering process. The final contribution of this paper is an analysis of speaker entrainment on the CS rate we observe in the MB corpus. While [20] have investigated lexical priming in entrainment, no research has been done on longitudinal entrainment and CS.

3. Corpus

The Miami Bangor (MB) corpus is a conversational speech corpus recorded from bilingual Spanish-English speakers living in Miami, Florida. It includes 56 recordings of conversational speech from 84 speakers, including 242,475 words (transcribed) and 35 hours of recorded conversation. The manual transcripts include the beginning and end times of utterances and per word language identification done manually. The dominant language in this corpus is English (53.48% of the tokens), followed by Spanish (27.78%). However, the composition of the subset of CS utterances is different: In this subset, Spanish becomes the dominant language in CS utterances, comprising 46.12% of tokens compared to 38.98% of the English tokens. Other annotation labels include the ‘ambiguous’ label for words that were difficult to tag as either English or Spanish due to lack of context, the ‘mixed’ label for words that are formed by morphemes and roots from both languages (e.g. “ripear”), and ’other’ category for untranscribed tokens. We created an additional punctuation tag for the corpus.

The original MB corpus was automatically glossed and tagged with POS tags using the Bangor Autoglosser [21, 22], but the version of the corpus used here makes use of the POS tags crowdsourced by [23], using the universal POS tagset; these tags were collected with high inter-annotator agreement. The Universal POS tagset includes 17 categories: adjective (ADJ), adposition (ADP), adverb (ADV), auxiliary verb (AUX), coordinating and subordinating conjunction (CONJ and SCONJ), determiner (DEAT), interjection (INTJ), noun (NOUN), numeral (NUM), proper noun (PROPN), pronoun (PRON), particles (PART), verbs (VERB), punctuation, symbol, and other. Notice that we will not include results on the last three annotations due to space reasons. The version of the corpus used here contains 4,193 CS instances. We use the following naming convention throughout the rest of the paper: CS word is the first word where a change of language occurs, the word preceding a CS is the word that occurs immediately before a CS word. Similarly, the word following a CS is the word that occurs immediately afterwards. For example, in the sentence ’Mis papás were so happy to see you’, ‘were’ is the code-switched word and ‘papás’ and ‘so’ are the words immediately preceding and following the code-switch respectively.

A list of English-Spanish cognate pairs were collected from a variety of online sources[1]. We preprocessed the list of cognates first automatically and then manually to remove determiners, break cognate compound words into single words and remove duplicates. Not counting masculine/feminine duplicates, a total of 3,432 cognate word pairs were collected, of which 1,305 appear on the MB corpus.

4. Experiments

4.1. Code-Switching and Cognate Words

In this section we analyze the statistical relationship between CS and cognate words on the MB corpus testing the triggering hypothesis. First, we observe that there is a strong statistical relationship between CS utterances and the presence of cognates in those utterances: Table 1 shows the contingency table for all the utterances in the corpus split in utterances with and without cognates and monolingual and code-switched utterances. The results of a χ^2 test returns a highly significant p-value that rejects the hypothesis that both distributions are independent. The percentage of CS utterances in each group (last row of the table) confirms that cognated utterances are more likely to be in CS utterances than non-CS utterances.

$\chi^2 = 309.63$		Cognate	
$p < 10^{-68}$		no	yes
CS	no	20,029	18,767
	yes	1,037	1,937
	% yes	4.92	9.36

Table 1: *Number of code-switched and monolingual utterances split by utterances that contain a cognate or not.*

Next, we replicate the experiments from [15] in Tables 2, 3, 4, 5 and 6. For all tables, we present contingency tables for the two groups being compared (one always CS words and the other some aspect of immediately adjacent cognates), plus the percentage of CS words for the second group, and the results of a χ^2 test on the contingency table, including the test’s statistic value (χ^2) and its p-value p. Table 2 shows that there is no significant statistical relationship between words that precede a cognate and CS when compared to words that do not border on cognates.

$\chi^2 = 0.14$		Cognate	
$p = 0.71$		No bordering	Precedes
CS	no	206,005	28,901
	yes	3,256	466
	% yes	1.56	1.59

Table 2: *Number of code-switched words and percentage of code-switched words split by words preceding a cognate and words not bordering cognates.*

Table 3 shows that there is a strong statistical relationship between CS words and words that follow cognates, when compared to words that do not border on cognates. Furthermore, it can be seen that the percentage of CS words increases for the group of words that immediately follow cognates.

A variation of the same test, Table 4, shows that there is a strong statistical relationship between CS and words that follow cognates when compared to words that do not follow cognates.

$\chi^2 = 26.55$		Cognate	
$p < 10^{-6}$		No bordering	Follows
CS	no	206,005	26,812
	yes	3,256	540
	% yes	1.56	1.97

Table 3: *Number of code-switched words and percentage of code-switched words split by words following a cognate and words not bordering cognates.*

[1] http://nlp.cs.berkeley.edu/projects/historical.shtml; http://spanishcognates.org/; http://www.colorincolorado.org/sites/default/files/Cognatelist.pdf; https://www.duolingo.com/comment/5508808/The-Most-Useful-Spanish-Cognates

Ignoring the restriction that words are not followed by cognates, the result of the test is the same, which suggests that cognates that follow CS have no effect on them. This is further confirmed in Table 5, which shows that there is no statistical relationship between CS and the disjoint sets of words that border on cognates and words that only follow cognates.

$\chi^2 = 26.63$		Follows a Cognate	
$p < 10^{-6}$		no	yes
CS	no	230,768	26,812
	yes	3,653	540
	% yes	1.56	1.97

Table 4: *Number of CS words and percentage of CS words split by words following and not following a cognate.*

$\chi^2 = 2.67$		Cognate	
$p = 0.1$		Follows	Bordering
CS	no	22,674	4,138
	yes	471	69
	% yes	2.03	1.64

Table 5: *Number of CS words and percentage of CS words split by words that border on two cognates and words that only follow a trigger word.*

From this experiments we can confidently conclude that cognates immediately preceding CS help facilitate the switch and cognates immediately following CS do not have a meaningful impact on it. Furthermore from Table 5 we conclude that CS does not occur significantly more often when words are immediately preceded and followed by cognates. Overall, it can be observed that the same results obtained for Dutch-Moroccan Arabic in [15] translate to the English-Spanish MB corpus with much higher statistical power. We also checked the statistical relationship between CS words being cognate words (Table 6) and found that there is a strong statistical relationship between both variables, but surprisingly we found that CS words are overall less likely to be cognates than other words.

$\chi^2 = 26.23$		Cognate	
$p < 10^{-6}$		no	yes
CS	no	222703	34877
	yes	3740	453
	% yes	1.65	1.28

Table 6: *Number of CS words and percentage of CS words split by cognate and non cognate words.*

4.2. Code-Switching and Part-of-Speech Tags

The second set of experiments examines the relationship between CS and part-of-speech tags. Here we examine the role that POS categories play when immediately preceding and following a CS, and when they are themselves a CS. We start by measuring the statistical relationship between the tagset and the CS words. In order to do so, we create three contingency tables for the counts of all POS tags and whether they occur in one of the mentioned positions, and run a χ^2 test on them. Results for the three tests are shown in Table 7. It can be observed that, in the three cases, the null hypothesis that the POS tag distribution and the CS distribution are independent can be rejected. Specifically, POS tags seem to be have a statistically strong relationship to the words preceding a CS and the CS words themselves.

	POS		
	Preceding	Current	Following
χ^2	1,817.8	795.0	35.39
p-value	0.0	$< 10^{-158}$	< 0.01

Table 7: *Statistical significance results of running the χ^2 test of all the part-of-speech tags in three pairs of groups: words preceding a CS, CS words, and words following a CS.*

In order to study the role that specific tags play in eliciting CS, we start by comparing the tagging distribution over the whole corpus (top subtable on Table 8) with the tagging distribution of the words neighboring a CS (second subtable on Table 8). Some things are immediately clear: auxiliary verbs are very unlikely to precede CS, determiners and interjections are very likely to precede a CS, nouns appear more frequently as CS or neighboring a CS than on the rest of the corpus, particles are unsurprisingly not involved in CS; pronouns very rarely precede a CS word, and verbs are less likely to be CS.

We also study which tags are more likely to precede, be in or follow a CS by examining the rows from the third subtable in Table 8. It can be observed that Proper Nouns, Nouns and Interjections are the tags most likely to trigger a CS; Nouns and Subordinating Conjunctions are the two categories that are more often switched. Moreover, we observe that the tags following a CS are all comparably likely to be switched (third row).

To end this section, we study the statistical relationship between specific tags and CS by running the χ^2 test on the contingency tables populated by the counts of specific tags when preceding, on or following a CS. These results are shown in the bottom subtable of Table 8, where ✓indicates that the p-value of the statistical test is significant. The remaining (empty) cells have p-values larger than 0.01. The first observation we make from the first row of the subtable is that most of the POS tags have a strong statistical relationship when preceding a CS (first row), whether because they precede a code-switch more often (DET, INTJ, NOUN, PROPN, SCONJ) or less often (ADJ, ADP, AUX, NUM, PART, PRON). With respect to the CS words themselves, the second row shows that ADJ, NOUN and SCONJ significantly increase their presence on CS compared to AUX, DET, PART and VERB.

Some of these results might be expected: a CS between an auxiliary verb and another verb would be highly disruptive. Similarly, a switch is not likely to occur right after a pronoun, since most often pronouns are followed by verbs that need to agree on person and number. Indeed, the statistical relationship between verbs and CS words is very strong; the percentage of verbs that are switched is much smaller than the overall percentage of verbs in the corpus.

Both pronouns and nouns have a strong relationship with CS when immediately preceding the switch, albeit in different ways. Whereas nouns are very likely to precede a switch (18.89% of the tokens preceding a switch in the corpus are nouns), pronouns are much less likely to occur before a switch than in general (4.46% of the words before a switch are pronouns, compared to their percentage of 15.98% throughout the corpus). This fact is counterintuitive since pronouns substitute for nouns and noun phrases and both must agree with following verbs in person and number. So, it is not immediately clear why they behave so differently with respect to CS. However this finding agrees with previous research on pronoun-verb CS [16]

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	SCONJ	VERB
POS (%)	4.1	6.97	8.11	3.25	4.4	8.81	5.94	11.04	1.51	2.58	15.98	2.49	3.88	20.00
POS(t-1)|CS(t) (%)	3.84	5.08	7.66	**0.33**	5.29	**13.90**	9.23	**18.89**	0.79	**0.60**	**4.46**	**4.41**	**6.01**	17.72
POS(t)|CS(t) (%)	5.03	7.23	7.51	2.12	4.89	7.27	5.13	21.23	1.48	0.38	17.10	2.89	6.32	**11.42**
POS(t+1)|CS(t) (%)	4.17	6.78	6.98	3.28	**2.59**	10.09	2.27	14.71	1.67	2.21	15.66	2.93	3.59	22.21
CS(t)|POS(t-1) (%)	2.37	1.21	1.86	0.17	1.99	2.57	3.94	4.44	1.01	0.38	0.50	4.36	2.57	1.63
CS(t)|POS(t) (%)	1.97	1.66	1.48	1.05	1.78	1.32	1.38	3.08	1.57	0.24	1.71	1.85	2.61	0.91
CS(t)|POS(t+1) (%)	1.43	1.39	1.34	1.52	1.82	1.65	1.41	1.81	1.61	1.16	1.62	1.78	1.58	1.61
CS(t), POS(t-1)	✓	✓✓		✓✓✓		✓✓	✓✓✓	✓✓✓	✓	✓✓	✓✓✓	✓✓✓	✓✓	
CS(t), POS(t)	✓			✓✓		✓✓		✓✓✓		✓✓			✓✓	✓✓✓
CS(t), POS(t+1)			✓					✓		✓✓				

Table 8: *First subtable shows the percentage of POS tags in the MB corpus. Second shows the % POS preceding, on, and following a CS word. Third shows the percentage of words that are CS for each POS tag category preceding, on or following CS words. Bottom subtable shows the significance of running* χ^2 *statistical tests on each group of POS tag and CS words. One ✓ indicates* $p < 0.01$, *two indicate* $p < 10^{-4}$ *and three indicate* $p < 10^{-18}$.

that states that even though most often such switches are banned [24] they can still occur [17], depending, among other things on the length of the noun phrase they represent.

Another unexpected observation comes from the disparity between coordinating and subordinating conjunctions. We observe from the second subtable that the fraction of subordinating conjunctions that appear preceding or on a CS is higher than the number in the corpus as a whole, and, while the same can be said about coordinating conjunctions, the increase is not significant. Indeed conjunctions seem to be the ideal place to facilitate a switch, since they can often start a new sentence. We hypothesize that the reason for this difference is that the "and/y" coordinating conjunctions, which make up the majority of that tag category, are most often used for pairing objects in which case the switch could be disruptive.

4.3. Code-Switching and Entrainment

In this section, we analyze the MB corpus for evidence of entrainment in CS between conversational partners throughout the conversation. Entrainment is the phenomenon of conversational partners becoming similar to each other in their behaviors in dialogue. It has been found to occur in multiple dimensions of spoken language, including acoustic-prosodic [25], linguistic style [26], and syntactic structure [27]. Importantly, entrainment has been associated with positive conversation outcomes, such as likability [28], naturalness, and task success [29]. To measure entrainment in CS, we measure *convergence* (becoming more similar over time) between the conversational partners in the frequency of their CS behavior.

Earlier work on priming in CS [20] investigated structural priming effects as they relate to CS, also in the MB corpus. They found that the probability of an utterance featuring CS was higher when the previous utterance contained a CS. We further analyze the MB corpus for evidence of entrainment in CS behavior beyond utterance-to-utterance priming. We measure convergence between the conversational partners frequency of CS, the degree to which the **amount** of code switched segments produced by each speaker becomes more like that of their partner through the conversation as a whole. In total, we analyzed 37 conversations from the MB corpus, excluding those with more than 2 speakers, conversations for which we only have the dialogue of 1 speaker, and conversations lacking CS entirely.

Convergence was calculated by using a Pearson-R correlation analysis on each speakers CS ratio (total number of CS normalized by total number of tokens) for each speaker turn. A significant positive correlation is indicative of convergence, the pairs code switching frequency becomes more similar to each other, while a significant negative correlation is indicative of divergence, the pairs CS frequency becomes more different over the course of the conversation. Out of 37 pairs, 32 show significant correlations in convergence or divergence of CS ratio. A total of 28 conversations are converging, of which 10 were weakly converging $(0 < r < 0.5)$, 7 were moderately converging $(0.5 \leq r < 0.7)$, and 11 were strongly converging $(r \geq 0.7)$. The other 4 conversations showed diverging patterns: 2 weakly diverging $(0 > r > -0.5)$ pairs and 2 moderately diverging $(-0.7 < r \leq -0.5)$ pairs.

We know from previous studies that the introduction of CS may immediately prime a CS in the following utterance. Here we find interlocutors adapt to each others CS rates over the course of a conversation. Fricke and Kootra [20] speculate that other factors must be prompting CS beyond mechanistic language priming due to the infrequency of CS in MB. Based on our findings, we propose entrainment as one such high-level mechanism driving CS behavior.

5. Conclusions and Future Work

In this paper we have presented a thorough analysis of the relationship between code-switching and cognate words, and code-switching and part-of-speech tags for English and Spanish. We confirmed that cognate words facilitate code-switching when immediately preceding the code-switch, but have no effect on it when they immediately follow the switch. We presented statistical evidence that there is a strong relationship between code-switching and part-of-speech tags, and we examined the specific tags that occur more and less frequently in the vicinity of a switch. We also demonstrated that speakers entrain to one another in the rate at which they code-switch. This latter finding may provide further socio-linguistic insight into the social aspects of code-switching.

For future work we plan to study how cognate-based features (including pronunciation, semantic, and spelling features of cognate word pairs) impact part-of-speech tagging, language modeling and language ID on code-switched language.

6. Acknowledgements

The authors would like to thank Lesley Cordero for collecting an initial set of cognate words and Ismael Villegas for productive research discussions and insight. We would also like to thank Melinda Fricke for bringing the importance of cognate words to code-switching to our attention.

7. References

[1] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang *et al.*, "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*, 2014, pp. 62–72.

[2] G. Molina, F. AlGhamdi, M. Ghoneim, A. Hawwari, N. Rey-Villamizar, M. Diab, and T. Solorio, "Overview for the second shared task on language identification in code-switched data," in *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, 2016, pp. 40–49.

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

[4] H. Adel, N. T. Vu, and T. Schultz, "Combination of recurrent neural networks and factored language models for code-switching language modeling." in *Proc. of ACL*, 2013, pp. 206–211.

[5] Y. Li and P. Fung, "Language modeling with functional head constraint for code switching speech recognition." in *Proc. of EMNLP*, 2014, pp. 907–916.

[6] T. Solorio and Y. Liu, "Part-of-speech tagging for English-Spanish code-switched text," in *Proc. of EMNLP*, 2008, pp. 1051–1060.

[7] P. Rodrigues, "Part of speech tagging bilingual speech transcripts with intrasentential model switching." in *AAAI Spring Symposium*, 2013, pp. 56–63.

[8] A. Jamatia, B. Gambäck, and A. Das, "Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages." in *Proc. of Recent Advances in Natural Language Processing*, 2015, pp. 239–248.

[9] F. AlGhamdi, G. Molina, M. Diab, T. Solorio, A. Hawwari, V. Soto, and J. Hirschberg, "Part of speech tagging for code switched data," in *Proc. of the Second Workshop on Computational Approaches to Code Switching*, 2016, pp. 98–107.

[10] B. H. Ahmed and T.-P. Tan, "Automatic speech recognition of code switching speech using 1-best rescoring," in *Proc. of IALP*, 2012, pp. 137–140.

[11] T. Lyudovyk and V. Pylypenko, "Code-switching speech recognition for closely related languages." in *Proc. of SLTU*, 2014, pp. 188–193.

[12] M. G. Clyne, *Transference and triggering: Observations on the language assimilation of postwar German-speaking migrants in Australia*. Martinus Nijhoff, 1967.

[13] ——, "Triggering and language processing." *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 34, no. 4, p. 400, 1980.

[14] ——, *Dynamics of language contact: English and immigrant languages*. Cambridge University Press, 2003.

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

[16] E. Woolford, "Bilingual code-switching and syntactic theory," *Linguistic inquiry*, vol. 14, no. 3, pp. 520–536, 1983.

[17] D. Sankoff and S. Poplack, "A formal grammar for code-switching," *Research on Language & Social Interaction*, vol. 14, no. 1, pp. 3–45, 1981.

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

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

[20] M. Fricke and G. J. Kootstra, "Primed codeswitching in spontaneous bilingual dialogue," *Journal of Memory and Language*, vol. 91, pp. 181–201, 2016.

[21] K. Donnelly and M. Deuchar, "The Bangor Autoglosser: a multilingual tagger for conversational text," *ITA11, Wrexham, Wales*, 2011.

[22] ——, "Using constraint grammar in the Bangor Autoglosser to disambiguate multilingual spoken text," in *Constraint Grammar Applications: Proceedings of the NODALIDA 2011 Workshop, Riga, Latvia*, 2011, pp. 17–25.

[23] V. Soto and J. Hirschberg, "Crowdsourcing universal part-of-speech tags for code-switching," in *Interspeech*, 2017, pp. 77–81.

[24] F. Barkin and A. Rivas, "On the underlying structure of bilingual sentences," in *Linguistic Society of America, 54th Annual Meeting, Los Angeles, Calif*, 1979.

[25] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*. Association for Computational Linguistics, 2012, pp. 11–19.

[26] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words!: linguistic style accommodation in social media," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 745–754.

[27] D. Reitter and J. D. Moore, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation," in *Proceedings of the Cognitive Science Society*, vol. 28, no. 28, 2006.

[28] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[29] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*. Association for Computational Linguistics, 2008, pp. 169–172.