Deriving Language Signatures for Bilingual Code-Switching

Keywords: *Code-Switching, Probabilistic Language Models, Sociolinguistics*

Research Question: How do bilingual speakers of the same language pairings code-switch between them differently? More specifically, what components can be extracted from bilingual data to differentiate speakers of the same languages?

Background: Linguistic scholars have observed that there is wide variation in code-switching (CS) due to social differences (Gardner-Chloros, 2009). For example, Post (2015) observed variations in frequency and type of CS as a function of gender among Arabic-French speakers in Morocco. Unfortunately, findings like these have been restricted to specific languages and small datasets and until recently, there have been no tools to classify CS at the level of large corpora (Gambäck & Das, 2016). Furthermore, there have been no attempts to distinguish the unique CS patterns presented by speakers, i.e., individual language signatures. The key problem is that CS can include small word-level insertions of single lexical items or long stretches of dialogue across several speakers, which make its study difficult as speakers can vary their patterns of speech considerably from one utterance to another.

I propose to address these gaps in the study of CS by applying statistical models to extract and extrapolate patterns from bilingual corpora. The crux of my approach is to look at CS as a sequence of language spans, in which a speaker remains in one language before switching to another. Solorio and Liu (2008) have previously exploited this idea to predict switch points and to tag for Part-Of-Speech, yet their approach made no attempt to distinguish or identify different patterns. Outside of current analyses at the level of corpora I do not know of any statistical approach to studying bilingual CS across speakers that exploits this idea of language spans. By adapting this concept to the alternation of language at the individual level and not corpora, I believe that it is possible to distinguish the CS of one bilingual speaker from others to produce a distinctive language signature, regardless of small changes in speech.

Hypothesis/expected findings: Based on my previous work with the *Killer Crónicas*, *Yo-Yo Boing!*, and *Bon Cop, Bad Cop* datasets, I hypothesize that CS can uniquely characterize individual speakers and that the relative frequencies of language spans across speakers are distributed differently, with some speakers preferring spans of one length to another. I expect that speakers of the same language pairings vary enough in the length of languages spans that there are statistically significant differences in the speech of two bilingual individuals. I anticipate that there are also several variables contributing to these differences such as region, attention paid to speech, and social factors. My methods to extract different features of CS and provide unique signatures of CS across bilingual speakers will be language-neutral and applicable across different language pairings.

Approach and Methods: The first component of this project will be the gathering of a large number of bilingual corpora involving CS in order to introduce as much variation as possible. Given my previous work in language annotation, I am free to work with larger, untagged datasets so long as enough training data exists. By far the largest collection of such datasets is the Linguistic Data Consortium (LDC), which charges a fee for unlimited use and access. The remaining resources are access to powerful computing resources (or XSEDE access) and the expertise of faculty members working on CS and statistical language models.

After preliminary analysis of the datasets, I expect distinct patterns of CS to emerge across speakers such as switching differently around breaks in speech, which will lead to differing patterns in the corresponding language signatures. At this point, I will work to examine

bilingual CS with statistical models by looking at the distribution of language spans per speaker and by examining the probability of switching in a discourse as a function of time, both of which necessitate large datasets.

Assuming the simplest case, the distribution of language spans of a speaker can be modeled as a stochastic exponential decay after normalization for length of text. I expect that different speakers will present different rates of decay in their language, resulting from varied use of language spans. In addition, by looking at bilingual text as a series of switches between monolingual spans, I will model CS as a Poisson process and find the most apt parameters for individual speakers by working backwards from their speech. I will tune different stochastic models to speakers in order to find statistically significant differences in speech patterns and I plan to subject the data to a rigorous analysis of different stochastic models to test my hypotheses. I will also perform a regression analysis to find correlations between the social and environmental factors mentioned above and any differences from the models.

Intellectual Merit/Broader impacts: A Graduate Research Fellowship will allow me to promote further research between the fields of Linguistics and Mathematics. My work will inspire the development of a general framework with which to examine cases of switching phenomena within Linguistics. It will have broad impacts in computational linguistics, sociolinguistics, and bilingualism both as a tool and as a theoretical construct. My model will contribute a new, language-neutral approach to examining bilingual CS, freeing researchers from being constrained by the availability of data in dominant languages like English. In addition, my proposal has potential contributions to the fields of linguistic methodology and linguistic anthropology. Its development may lead to the possibility of deconstructing seemingly homogenous language or CS use into discrete subgroups by geography, ancestry, or culture.

Finally, it must be noted that the development of my model need not be restricted to the study of switching between two languages. My ambition is to generalize my model to work with as many languages as needed. In addition, a refined version of my proposed model would be able to uniquely identify changes in style, dialect, or register given enough training data. As an example, learners of a second language could apply the principles of my model to pinpoint exactly where their usage differs from that of a native speaker, which provides a new possibility for accelerated language learning and for the study of second-language acquisition.

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.

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 and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1051-1060. Association for Computational Linguistics.

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.