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.

Personal, Relevant Background, and Future Goals

Personal Statement: My family often recounts that as a child I was not to be left alone with anything that had screws. I was about six when I first discovered my father's tool set and one day he returned to find me amid the disassembled remains of the family VCR. It would not be the last time I would take something apart, and it was clear even then that I was obsessed with breaking things down into their component parts. Fortunately, I had parents who tolerated my mischievous hobbies and supported my curiosity. However, as I grew older I recognized the limitations of the understaffed and overworked Rio Grande Valley schools. I enrolled in the University of Texas at Austin, declared a double major in Computer Science and Electrical Engineering and I continue to deconstruct the world around me, albeit in a much less destructive manner. I plan to become the second in my family to obtain an undergraduate degree, the first to complete a doctoral degree, and the first to work as an academic research professor.

In Austin, I found that my linguistic practice as a Spanish–English bilingual was different from others on campus because I engage in a behavior known as code-switching (CS). As a first-generation Mexican-American, I often switched between English and Spanish with friends and family as a part of my regular speech and self-expression just as the overwhelming Hispanic population in the Valley did. Although I did not know it at the time, CS in border cities such as my hometown of Pharr, Texas constitutes a research focus in linguistics. Overcome with curiosity as to why few of my peers at UT code-switched despite knowing other languages, I began using my knowledge of natural languages and my abilities as a computational scientist to break down speech into the smaller units common to all languages. At first, I applied my programming skills to implement small command-line programs, like tokenizers for English or nonce-word generators for Spanish to sate my curiosity. Then, I started taking courses in Mathematics during my first year in order to better understand the theoretical underpinnings of natural language processing (NLP). I reasoned that addressing the issue of CS might best be studied by combining Mathematics and Computer Science, with the former providing a concrete description of language switching and the latter a means to explore it. To that end, I resolved to complete the requirements for a degree in Mathematics in addition to my studies. I also committed to self-study in French and Italian to obtain a firm grounding in natural languages beyond my own.

Searching high and low for research opportunities involving CS, I learned that the linguists Dr. Barbara E. Bullock and Dr. Almeida J. Toribio were researching the Spanish-English CS in Texas that I displayed. Eager to work with them, I accelerated my pace of study in French to test out of the introductory courses and secured a spot in Dr. Bullock's upper-division French Phonetics course, which required a thorough understanding of written and spoken French. After excelling in her course, I discussed my background, skills, and motivations with her, which led to an offer for a position in her research group, the Bilingual Annotation TaskS force (BATS: http://sites.utexas.edu/bats/). I realized that my previous forays into studying the CS of my speech were an initiation into the limited but growing body of NLP research in bilingual CS. The knowledge that researchers around the world were working on CS at the cutting-edge of Computational Linguistics and NLP provided me with the motivation to pursue a Ph.D. after graduation to continue my studies.

As a first-generation immigrant, I am appreciative of the mentorship of Drs. Bullock and Toribio, but it strikes me that others do not have the same guidance, opportunities, or familial

support that I have, and fewer still are contemplating a future in academia. As I look around now for classmates who share my background, I realize that many have fallen behind; I came to UT with over thirty friends and classmates from the Valley, but by my senior year I can count those remaining on the fingers of one hand. This disparity seems to extend to every discipline and level of academia but the most prominent effects are felt in the STEM fields. Low numbers of professional Hispanic scientists and mentors lead to a reinforcement of the cultural barriers that prevent Hispanic students from achieving success in those fields. I feel that my duty to the Latino community does not stop at completing a doctoral degree, but extends to becoming a community leader who can address the difficulties faced by students like me.

The efforts of Drs. Bullock and Toribio in mentoring underrepresented minorities drove me to successfully compete for the inaugural class of the Mellon-Mays Undergraduate Fellowship Program (MMUF) and to apply for a Teaching Assistant (TA) position at UT, with the aim of funding my research and advancing towards my new career objective. As a Mellon-Mays Fellow and as a TA in Engineering, I have seen first-hand the positive effects of a supportive environment on minority students and I hope to continue that same practice into graduate school as a mentor and eventually as a professor. I aim to convince the next generation of students like me that there are no barriers preventing them from studying what they wish and I hope to attain a position from which I can advocate for Computer Science education to the general public, specifically to underrepresented minorities.

Relevant Background/Intellectual Merit: My drive to pursue a career in research comes from my voracious desire for learning and from my experiences as a student, teacher, and researcher. In collaboration with the other members of BATS, my responsibilities involved designing and writing the code necessary for our investigations of CS. Over the course of the spring semester of my senior year, I taught myself the Python programming language in order to further develop a hybrid of a character N-Gram Model and a word-level Hidden Markov Chain that was based on code written in Scala. By looking at the orthography of individual words and their context with this algorithm, we have been able to successfully determine the language of words in a mixed-language text with a 96% degree of accuracy, far higher than with other methods. We used this tool to extract and extrapolate patterns in CS from mixed-language data such as the locations of switch points and length of language spans, two highly important aspects of study in CS. Thus far, we have applied our analysis to *Killer Crónicas* and *Yo-Yo Boing!*, two nonfiction texts written in a combination of English, Spanish and Spanglish, and to *Bon Cop, Bad Cop*, a bilingual French-Canadian film that incorporates French, English, and Franglais.

A year and a half of work has resulted in co-authorships for the 2016 Empirical Methods of Natural Language Processing Conference and for the Linguistic Society of America, at which I will be co-presenting our algorithm and findings on novel methods to visualize and quantify CS in multilingual documents. Specifically, I have co-developed a new statistical metric to quantify the integration of CS in a document based on the number of concrete language switches and created original methods for visualizing the variation in CS at the level of sentences and corpora. I am currently contributing to two research articles on related work, targeted for refereed linguistics journals. In addition, last October I co-presented our research findings on the CS in *Bon Cop, Bad Cop* at the Transcultural Urban Spaces conference at the University of Bern in Switzerland. We demonstrated that the film dialogue supports the claim that there is an individual and social component to variation in CS, which we will continue to explore.

Broader Impacts: While the research aim of BATS was to identify distinct patterns in the CS, our results have proved to be much more fruitful. As mentioned, we are confident that different speakers switch between the same pairs of languages in different ways, which suggests a sociolinguistic component to code-switching, regardless of languages spoken. I plan to exploit these findings as the basis for my research proposal.

As a Mellon-Mays Fellow, I independently submitted a journal article titled "A Survey of Automatic Annotation Methods for Multilingual Corpora," which was accepted for the *Mellon-Mays 2016 Undergraduate Journal*. In it, I demonstrate that automatic annotation through machine learning is the best method for annotation, citing diminished cost, time, and effort involved. In the past, researchers have resorted to automatic annotation through crowdsourcing or combining the annotations of experts in various ways, but current work using machine learning far surpasses the accuracy of other methods. As an example, my current work with language annotation in BATS outperforms other methods by at least 10%. An added benefit is that our code and algorithms are completely open-source and shared on GitHub.

I am working on another project to visualize different types of CS across several corpora. One problem linguists face is that the study of CS often requires an analysis at multiple levels. I hope to address this difficulty by providing simple and open-source programming tools to aid researchers in examining CS in a language-agnostic manner at the level of sentences, speakers, dialogues, or corpora. I will be presenting this work at the MMUF Southeastern Regional Conference in Atlanta; it represents the initial idea that I develop further in my research proposal.

Despite my heavy interdisciplinary course load and research work, I elected to become a teaching assistant (TA) in the Electrical Engineering department to share my passion for learning and to better prepare myself for life as a graduate student and as an academic research professor. I find that students especially enjoy my approach to taking apart difficult concepts like virtual memory or process scheduling and I believe that such experience is necessary for gaining new perspectives on learning as well as teaching. I am confident that my varied experiences as a student, teacher, and researcher have prepared me well for graduate school by feeding my obsession for deconstructing complex material and by teaching me to draw from and communicate with varied disciplines.

Future Goals: Although my focus over the years has evolved from the disassembly of electronics to the deconstruction of CS, my passion for dissection and analysis remains the same. I am committed to continuing my studies using the tools and techniques from related fields as necessary and I wish to become an interdisciplinary Computer Science faculty member and further the field of Linguistics through a deep study of CS as it relates. Finally, I desire to increase interest among Hispanic students for pursuing advanced education and careers in academia. The NSF fellowship will permit me to reach my aspirations of attaining a doctoral degree, continue my work as a postdoctoral researcher, and achieve my final ambition of becoming a faculty member, where I can share my experiences and drive for learning with students like me. My continued involvement with the BATS research group and the Mellon-Mays Undergraduate Fellowship Program will provide a platform from which I can openly address minority students both as a positive example and as a mentor. Additionally, I believe that my current and future work will contribute to the study of CS by creating tools and methods that will advance Linguistics as a whole.