Identification and Analysis of Conversational Codeswitching Triggers

Total Funds Requested: $6,000

August 30, 2017

Student Co-PIs:

Jocelyn Alvarado
joalvar1@student.uiwtx.edu
Department of Mathematics
University of the Incarnate Word
Advisor: Dr. Theresa Martines

Dr. Chen Chen
carachenchen@gmail.com
Department of Earth, Atmospheric, and Planetary Sciences
Purdue University

Rouzbeh Shirvani
rouzbeh.asghari@gmail.com
Department of Electrical Engineering and Computer Science
Howard University
Advisor: Dr. Mohamed Chouikha

Dr. Yanina Shkel
yshkel@princeton.edu
Department of Electrical Engineering
Princeton University

Taylor Williams
taylorhallwilliams@gmail.com
Department of Bioengineering
University of California, San Diego
Advisor: Dr. Todd Coleman
Todd Cole

1 Problem Statement

In language, codeswitching occurs when a speaker uses two or more languages in the context of one conversation. Speculation on motivation for switching is multifaceted; it is possible that a person may switch languages to hide certain information from listening native speakers, to better express themselves because certain words cease to exist in a given language, or to accommodate the person with whom they are speaking. In the proposed project, Team SEALS (Swahili And English Language Switch) plans to use theoretical modeling and data analysis techniques to better understand the mechanisms of codeswitching. The team is particularly interested in using data mining and machine learning methods to develop algorithms which can reliably predict when codeswitching will take place, and to accurately classify the type of codeswitching trigger. From the team's preliminary research, it is apparent that very few scholars have investigated prediction methods related to codeswitching [1],[2]. Though it has been proposed that conversation dynamics affect factors of codeswitching, even fewer articles exist on the mathematical quantification of the hypothesized correlations between two individuals' switching frequency [3]. The team aims to use its unique interdisciplinary composition to tackle this problem with the assistance of the Center for Science of Information and NSF project grant.

2 Proposed Activity

To accomplish the proposed research, the team will use its interdisciplinary expertise to combine theoretical approaches with numerical simulations and data analysis. To develop theoretical models for codeswitching the team will draw from computational linguistics, information theory, machine learning, neuroscience, and statistics. For numerical data analysis the team will utilize data analysis toolboxes in R and Python: most natural language processing (NLP) scholars utilize Python and open source libraries, such NLTK and Stanford CoreNLP, are available.

The team activity will be governed along the following three directions: (1) **Data** acquisition and analysis, (2) **Model** selection and identification, and (3) **Algorithm** design. In particular, the proposed activity is organized by the tasks outlined below:

2.1 Task 1: Literature Review

All team members will research the basic linguistic theory behind and the current perspectives on codeswitching. Part of this task is to further refine the correct questions to ask about codeswitching triggers. A preliminary literature review has been done, but the team will do a more in-depth review here to best guide our next actions.

2.2 Task 2: Data Acquisition

The team has the Swahili-English interview scripts dataset and have performed preliminary speculation and analysis on this dataset. It comprises of around 10,000 utterances which is equivalent to 188,000 tokens. The team also has access to the English-Spanish Twitter dataset which could be of significant value in terms of the transferability of findings across multiple languages and conversational contexts. This dataset consists of 15,000 tweets which is equivalent to 170,000 tokens. Similarly, to further explore codeswitching behavior, the team plans to look into additional datasets or construct a new dataset, such as mining public information from other social media websites and online platforms which facilitate multilingual communication [4].

2.3 Task 3: Data Mining

In this task, the team will perform fundamental statistical and linguistic analyses to identify potential trends, such as ordering the switched words by frequency and identifying the dialogue contexts for codeswitching by translating the switched words.

2.4 Task 4: Model Selection and Identification

Here, the team will implement and critically analyze current models for codeswitching and codeswitching triggers while implementing our own nuances to best answer the proposed questions. This approach will likely incorporate models frequently used in the NLP literature, such as Naive Bayes classifiers or Hidden Markov models [1] as well as recurrent neural networks [5]. It may also include the development of new information theoretic models for interactive communication.

2.5 Task 5: Model Validation

The five-member team will use acquired datasets to validate selected models. Outside of our immediate network, we will likely use information theoretic, mathematical, and statistical expertise from our mentors for the theoretical analysis of our models as supplementary validation.

2.6 Task 6: Algorithm Design

The team plans to use insight from previous tasks [1],[2] to propose and design rigorous algorithms that predict switching and/or classify switching styles. One of the team's main interests is developing an algorithm to automatically detect and identify the influential members of the conversation. In other words, the team would like to explore the dynamics of the conversation and examine which members have more influence on the conversation in terms of switching and vocabulary selection.

The proposed activity is inline with the core mission of the Center's Communication and Life Sciences thrusts: the team seeks to utilize communication models and knowledge extraction from data to predict and classify the behaviour of bilingual individuals.

3 Goals and Outcomes

The team's primary goals are to gain experience completing data analysis research in an interdisciplinary, collaborative environment, as well as to develop innovative approaches to NLP. In terms of tangible outcomes over the year-long duration of this collaboration, the team aims to first develop novel models for codeswitching and secondarily produce prediction and/or classification algorithms for codeswitching. As an overarching outcome, the team plans to present their results at a peer-reviewed conference as a measurable method of contributing to the NLP field. This conference is expected to be directly in the NLP field, such as the Conference on Empirical Methods on Natural Language Processing (EMNLP). In the case when the model validation task yields a significant theoretical component, the team will also consider a conference in the information theory realm, such as the International Symposium on Information Theory (ISIT).

The team's work will contribute to a better understanding of language in the codeswitching context while extending syntax and semantics comprehension further in the multilingual realm. Largely through the exploration of machine learned grammatical structures, the team aims to advance research related to codeswitching prediction. In a broader scope, perspectives on sociolinguistics will also be affected as the team's research may help explain certain human language interactions such as power dynamics or group membership.

4 Proposed Work Statement

To accomplish the specific tasks and goals outlined above, the team will utilize each team member's unique skillset.

Dr. Shkel brings a strong background in information theory, coding and machine learning. Her expertise will be especially helpful in Tasks 4-6. As a senior member of the team she will be able to help guide team's progress.

Dr. Chen has extensive experience analyzing seismic data and using this data to validate geophysical models. This expertise will be transferred to the current problem and will be especially helpful in Tasks 2, 3, and 5. As a senior member of the team she will be able to help guide team's progress.

Rouzbeh Shirvani brings a strong expertise in machine learning, artificial intelligence, and computational linguistics. His advisor, Dr. Mohamed Chouikha, is an expert in the fields of signal and information processing, machine learning, and hardware and software security. Rouzbeh's expertise will be especially helpful in Tasks 1, 4, and 6.

Taylor Williams and her faculty mentor, Dr. Todd Coleman, specialize in causality problems related to the field of information theory. In addition, Taylor has working knowledge of computational neuroscience algorithms. This expertise will be especially helpful in Tasks 3-5.

Jocelyn Alvarado and her faculty advisor, Dr. Theresa Martines, specialize in applied mathematics, with particularly strong backgrounds in nonlinear partial differential equations, applied analysis, scattering and spectral problems, and inverse problems. This complements the mathematical background of other team members and will be especially helpful in Tasks 2, 3, and 5.

The team will meet monthly for virtual meetings over Skype and communicate regularly over both email and Google Hangout. All data will be shared on Google Drive, and working algorithms will be sent over email. The team will follow the prospective schedule put forth in the Proposed Activity section above to ensure all data acquisition and subsequent analysis is efficiently and effectively completed. One annual team meeting will occur on the East Coast where the majority of team members reside, with the exact location to be decided.

5 Diversity Statement

The team actively promotes diversity, equity, and inclusion through the proposed project. In terms of the technical aspect of codeswitching, the project will serve as a catalyst for understanding diverse, mixed-language regions and the intent behind the use of multiple languages within a single conversation. More specifically, the team is preliminarily working with Swahili-English codeswitching data, with Swahili considered a low-resource language yet one of the most prominent languages in central and eastern Africa. Though the current data is simply Swahili-English conversations, the team aims expand its database to include codeswitching across multiple regions and languages in congruence with the Center's diversity mission. The project also stands to support diversity through its five-member team consisting of mixed genders, ethnicities, disciplines, and regions of residence, as well as by providing opportunities for undergraduate research experience to two of the team members. Through its promotion of diversity, the team will not only render a more effective solution to the proposed problem, but it will also strengthen the Center's mission to support women, United States citizens, and numerous members of other underrepresented groups in the field of science of information.

6 References

[1] Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 973–981. Association for Computational Linguistics.

[2] Mario Piergallini et al. 2016, Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data. In Proceedings of the Second Workshop on Computational Approaches to CodeSwitching, pages 21-29. Association for Computational Linguistics.

[3] Carol Myers-Scotton. 1993b. Social Motivations for Codeswitching: Evidence from Africa. Oxford University Press, Oxford, UK.

[4] "JamiiForums." JamiiForums | The Home of Great Thinkers. Web.

[5] Sepp Hochreiter, Jürgen Schmidhuber. 1997, Long Short-Term Memory. Journal of Neural Computation, pages 1735-1780.

7 Budget and Justification

The team requests that the Center fund one in-person team meeting on the East Coast which includes domestic air travel and lodging costs. With a primary goal of this project being the presentation of results at a peer-reviewed conference geared towards NLP or information theory, the team also seeks funding for three members to travel to and stay in the city where the conference will be held. Individual expenses and respective amounts requested are outlined in Table 1.

Table 1: Budget Breakdown

Item	Amount Per Item	Total Amount Requested
Air Travel to Meeting	2 x $500 3 x $150	$1,450
Lodging at Meeting	4 x $125	$500
Conference Registration Fee	3 x $600	$1,800
Air Travel to Conference	3 x $600	$1,800
Lodging at Conference	3 x $150	$450
	Total	**$6,000**