# Biweekly Milestone Report: Identifying Context-Based Clusters of Mandarin-English Code-Switching 07-400 Fall 2022

Anna Cai
Mentor: Professor Alan Black, LTI
https://www.andrew.cmu.edu/user/annacai/07400.html

April 15, 2022

## 1 Major Changes

There have not been any major changes to the goals or implementation of the project.

## 2 Progress Report

Since the last status update, I have grouped the data by language use and looked at the vocabulary and code-switching usage, but did not find anything particularly useful. I also looked at function words, verbs, and nouns as percentages of total words in each language and found that the amount of English function words seems to be heavily correlated with the amount of English use, but also slightly correlated with less code-switching.

I have also found that manually selecting additional stopwords seems to greatly improve the unsupervised topic modeling methods I tried previously. While the "topics" do not all seem to correspond to actual discussion topics, I believe unsupervised topic modeling (particularly NMF with KL divergence) shows promise, so I will likely explore this further rather than try clustering the BERT sentence vectors.

Additionally, I read through several conversations (approximately 6 hours) to get a more extensive understanding of the variance in the data. I also looked into grouping the data by the ages of the speakers, but unfortunately most speakers are too close in age (19-24) for a meaningful analysis.

## 3 Milestone Evaluation

I believe that I have been able to meet my milestone goal for this week.

## 4 Surprises

There have not been any major surprises in the project so far.

## 5 Future Plans

Based on the conversations I read, I found some generic words that could be used differently when discussing particular topics (academics), so I plan to compare the windows around words across different topics and manually determine whether the words are used in different senses.

I also plan to try confirming some of the observations I previously made about differences in distributions through significance testing.

## 6 Milestone Revisions

I originally wrote that I would potentially try to apply some of my methods to code-switching datasets for other language pairs or to monolingual datasets, but this is likely too time-consuming to be feasible and other datasets are more likely than not to have different grammatical and vocabulary distributions (due to different data collection methods and a host of confounding factors). I will instead concentrate my time and energy on further analysis of SEAME.

## 7 Resources

So far, I have all the resources I need.