# 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 22, 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 refined the unsupervised topic modeling to get ten topics that mostly seem to correspond to actual topic areas and show promise in having varying language and code-switching distributions.

I have also tried applying significance testing through directly using chi-squared testing on contingency tables of word counts, but the word counts were too high to provide meaningful p-values (they were all extremely close to zero). I researched some other possible tests but I would likely need to implement them in order to be able to use it. Based on advice from my advisor, I will likely focus on developing visualizations of the differences rather than on significance testing (which is apparently not often used in NLP in practice, at least for corpus analysis).

In addition, I have looked at distributions of words in the windows around some select verbs, across "academic" and "non-academic" topic labels (determined by the presence or lack thereof of manually defined keywords). In particular, I found that the word "take" is essentially only used in the sense of "taking courses" when discussing academics, but uses many other senses for other topics. In contrast, most other words had different topic-related words in their windows, but the senses in which they were used were the same across these topics.

Finally, I also started analyzing the amount of switching as a raw count, as well as the number of switches per line and the length of each segment by language, though I have not yet made any observations.

## 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

I plan to finish applying the switching analysis across the predefined categories and unsupervised topic labels. Other than that, I will concentrate on making my poster and creating visualizations to support the observations I previously made.

## 6 Milestone Revisions

I will not make any milestone revisions at this time.

## 7 Resources

So far, I have all the resources I need.