Final Project Pre-Proposal Worksheet

Advait Sankhe

1. What is the research question you plan to answer in your Final Project?

   How does Marathi-English code-mixing differ across genres of discussion topics (sports, politics, entertainment, etc.) and also between informal, opinionated statements (tweets, YouTube comments) versus objective reports (newspaper articles).

   Another related topic is the comparison of code-mixing in Romanized Marathi texts versus texts in the Devanagari script.

2. List 1-2 resources you have read to inform your research question. For undergraduates, one of these could be a chapter or section from Brezina. For graduate students, these should be peer-reviewed journal articles.

   a) **A quantitative analysis of age-related differences in Hindi–English code-switching**

   b) **I may talk in English but gaali toh Hindi mein hi denge: A study of English-Hindi code-switching and swearing pattern on social networks** ***

   *** Contains profanity

3. Give a brief description of how the resources you have read inform your research question.

   Although both of these articles are about Hindi-English code-switching, the content is fairly applicable to some other Indian languages as well, including Marathi. The first article provides insight into the metrics that can be used for quantitative analysis of tokens for code-switching. It also brings to light possible reasons for favouring a certain language in the code-switching process that I could analyse based on the topic of discussion. The second article discusses ways in which twitter data could be collected, filtered and categorized and ways of labelling and matching transliterated words.

4. List 2-3 additional articles you plan on reading before submitting your Proposal.

   a) **Identifying and Analyzing Different Aspects of English-Hindi Code-Switching in Twitter**

b) My Boli: Code-mixed Marathi-English Corpora, Pretrained Language Models and Evaluation Benchmarks

c) **A diachronic investigation of Hindi–English code-switching, using Bollywood film scripts**

5. What corpus/corpora do you plan to use to explore your research question? If the corpus is on the server, which one is it? If it is not, how do you plan on accessing the corpus?

I will likely use the following corpora for my work. They contain categorized Marathi news articles and a collection of Romanized Marathi-English tweets. I might have to transliterate the news articles, which should be possible with reasonable effectiveness with the AI4Bharat python library.

https://github.com/AI4Bharat/indicnlp_corpus

https://github.com/l3cube-pune/MarathiNLP

If needed, I will also attempt to web-scrape the data based on the methods discussed by the creators of the above-mentioned corpora.

6. What Quantitative and Statistical methods will you use to analyze your data? Some examples from class include Chi-Square tests or Logistic regression.

Code-switching indices such as Mixing-metric, Integration-metric, Burstiness and Memory as discussed by Guzman et al. Moving code-switching research toward more empirically grounded methods.