Analyzing the Effects of Digital Language Contact on Hindi Language Degradation

Sodi Kroehler and **Sade Benjamin** and **Dr. Malihe Alikhani**
University of Pittsburgh

Abstract

Language degradation is a known phenomenon among linguists, however, there does not seem to be much research on understanding the role that digital language contact has on this. In this study, we analyze tweets across milestones of notable events in the field of digital language contact, and attempted to see if there is a noticeable effect on language degradation, measuring this through the amount of code-switched tokens per tweet. We did not find evidence to support such a claim sufficiently.

1 Introduction

Traditional linguistic analysis primarily utilizes manual investigation to uncover the relationships and evolution of natural human languages. Manual investigation is, however, rather limited in what kind of changes it can analyze. In addition, the amount of data being generated by users of a particular language (tweets/ social postings, internet news articles, etc) often exceeds what is possible to evaluate by hand and thus makes technologically-advanced methods much more desirable in this field.

To that end, we look at *digital language contact-* a phenomenon that occurs when speakers of different languages are not in close proximity geographically but still, through online communication and information sharing, exchange vocabularies or even grammatical sentence structures. While it is already established that the global prevalence of English has a great effect on the integrity of other languages, the study of this in linguistic and semantic studies has, at least in general, employed modern technologies rather infrequently. While we understand digital language degradation is a multi-faceted and complex issue, we posit that it may be possible to determine a measurement of direct change in the amount of degradation in a particular languages' primary speakers via analyzing tweets generated by them over a given length of time.

More concretely, we propose to measure this by analyzing code-switching in these gathered tweets. Code-switching is used here to "identify alternations of linguistic varieties within the same conversation" (Meyers-Scotton, 2006). Much research has already gone into training models to be able to understand or at least process instances of code-switching, especially in the intersection of Hindi and English. Hence, we were able to process a large volume of tweets quickly, as well as have efficient and comprehensive grading mechanics for the language decay expressed by them.

Our method to accomplish this involved us scraping Twitter for tweets originating in various cities around India, over regular intervals over the last decade, with particular focus around our chosen form of digital language contact - Amazon *Alexa*. Using NLP processes, we then determined the percentages of both English and Hindi tokens in each tweet, and analyzed them over time. We show here how, although there seems to be a trend around specific milestones of *Alexa's* introduction to the Indian society, it is not sufficient to count it as a basis for language degradation on it's own.

2 Related Work

As mentioned above, much research has already gone into developing NLP techniques for processing code-switched data. Among the most influential in our work were the work of (Patwa et. all, 2020) and the huggingface user sagorsarker.

3 Data Collection

To adequately measure the amount of code-switched data in a form that could reasonably be expected to illustrate language degradation, we chose a series of "milestones" which we would expect there to be a change in the rate of degradation. Specifically, we chose the introduction of Amazon's *Alexa* on November 6th, 2014, and their supporting of Hindi dialogue on October 4th, 2017.

Then, we generated data set A by using the *search all tweets* endpoint of the v2 Twitter API to poll all tweets from the last decade or so that had geo-tagging in a 3 mile radius of the largest city of each state [1]. We set Nov 6th, 2012 and 6th, 2020 as endpoints, polling at each city for 20 days around the end of every business quarter throughout the entire range.

Secondly, we obtained data set B - using the exact data set generated by Singh et.all - a corpus that had previously been tokenized and language-predicted using the same model and tuning as we did for the rest of our data sets. It had been generated from around a few historic events in India - ranging from September 2016 to August 2017.

Finally, for data set C, we pulled twitter histories for a list of users which we generated by searching twitter for users that regularly generated tweets classed as both Hindi and English. For this data set, we only polled from around 2018 - 2020, seeking to understand the language degradation around our second milestone more closely than dataset A was able to illustrate.

4 Model and Processing

We hypothesized that, given the announcement of the introduction, there would be a heightened percentage of English tokens per tweet around the first milestone, as potential users sought to learn more English to be able to use the technology, followed by a fall in the percentage when support for Hindi was extended, as users were once again comfortable to engage in their (more) native tongue. Data set B was already tokenized and run through a language identifier (LID), so no further processing was needed.

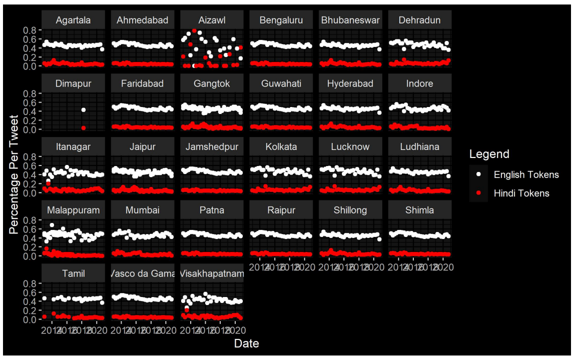
For data sets A and C, however, we used the tuning provided by sagorsarker on huggingface for the multilinugal-BERT model. We also measured the language classification of the tweet that Twitter provided, as well as any tokens with a confidence level < 75.

Using these results, we then tallied up the instances of English and Hindi tokens per tweet for all data sets, and graphed the average percentage of each as a function of time. We then fit a simple linear model to the data from all cities of the first data set to better understand the average trend.

5 Results

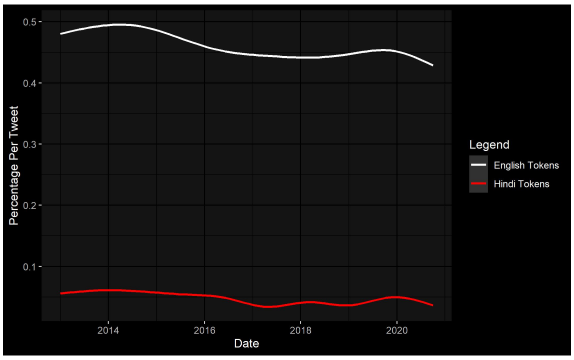
First, we just worked with data set A:

Figure 1: Percentage of Language Per Tweet in Data Set A, By City



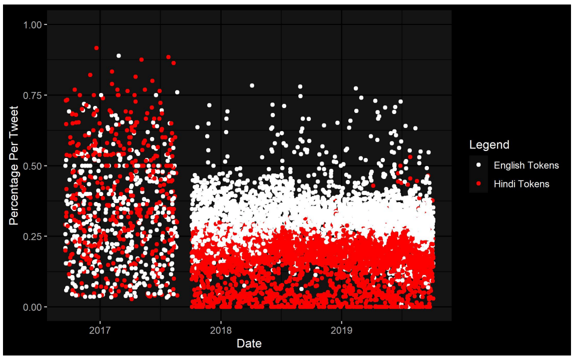
Averaged together, the trend is more apparent:

Figure 2: Average Percentage of Language in Data Sets A



Looking at data sets B and C, we chose to process them together, since they cover different age ranges:

Figure 3: Percentage of Language Per Tweet in Data Sets B and C

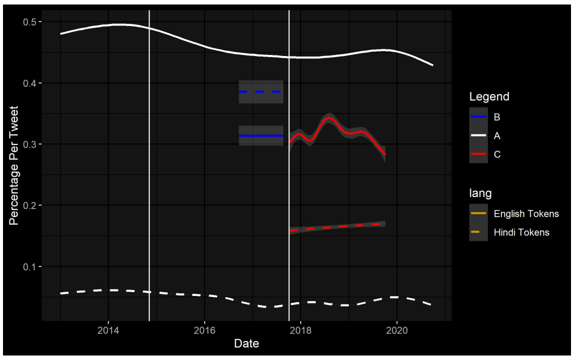


Finally, we visualized the average trends of all the data sets together, to understand how they varied. Given that we were interested in the effect

[1] Downloaded from here

of Amazon Alexa at specific milestones, we also graphed them as two vertical lines:

Figure 4: Average Behaviour of A, B, and C data sets



6 Conclusions

Even though there were some missing data, the average trend, at first, seemed to confirm our suspicions - there seemed to be an high average percentage of English tokens per tweet immediately prior and following the original release of Alexa, followed by a slump near when Alexa had support extended to the Hindi language. This seemed to signify that digital language contact may have an impact on language degradation.

However, upon closer inspection of the second and third data sets, there are further trends which do not seem to fit with this explanation. Furthermore, it is likely inconclusive to attribute a rise in language degradation in a specific medium to language degradation in an entire culture, as well as to attribute it to a single cause.

Thus it appears that our trend was likely due to either inexplicable randomness in the data or else tied to some other factor(s) of which we do not have knowledge. Hence, no conclusive statement can be made on the effect of virtual assistants on language degradation.

7 Future Work and Ethical Considerations

More work needs to be done to understand the causes of language degradation, so that work can progress to mitigate it. While this may not be an issue for Hindi, other languages and dialects are much more in danger of language degradation. In particular, languages that are spoken by increasingly few people are in danger of dying out if efforts are not made to preserve them. Although much research has already been done on it, the interplay between these facets of linguistics and modern computing techniques is somewhat sparse, and is it likely that great strides could be made in this field. Furthermore, with the creation and release of this data set, more work can be done to measure the predictive capability of both the tuning and Twitter's own language identification algorithms.

8 Acknowledgements

We would like to thank Dr. Malihe Alikhani for her instruction, encouragement, and technical guidance. We would also like to thank the Precog Research Group @ IIIT Hyderabad for generously allowing us to use their data set.

References

[1] Singh, K., Sen, I. and Kumaraguru, P. A Twitter Hindi English Code Mixed Dataset for POS Tagging. Workshop on Natural Language Processing for Social Media (SocialNLP 2018).

[2] Parth Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gamback, B., Chakraborty, T., Solorio, T., Das, A. SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets

[3] Myers-Scotton, C. (2006). In Social motivations for codeswitching: Evidence from Africa (pp. 1–2). introduction, Clarendon Press.