The Development of Hindi and Marathi

Vansh Singh - CE
Sam Parker - LIGN
Gates Zeng - CE
Rammy Issa - CS

Abstract

We conducted experiments to investigate the similarities that the modern indian languages Hindi and Marathi share with the ancestral languages, Prakrit and Sanskrit. To do this, we established binary similarity scores based on certain features between the modern and ancient languages to better understand how the modern languages have diverged from the ancient ones.

1 Intro:

We are researching the history of language development in India. In particular, we aim to investigate how the modern languages Hindi and Marathi diverged from the ancient Prakrit languages and Sanskrit by quantitatively determining orthographic similarities. Sanskrit, traces it's roots back to Indo-European languages. Around 250 BCE, more simple, derivative forms of Sanskrit evolved. This time period was known as the Middle Indo-Aryan stage, and these languages, more friendly for commoners, were known as Prakrit. As time went on, these languages that were considered Prakrit eventually gave rise to, among others, the languages Hindi and Marathi, which are spoken today. In our research, we used the Pali form of Prakrit written by Siddhartha when he wrote his Buddhist scriptures. All current North Indian languages currently ascribe their roots to one of the many Prakrit family of languages. Notwithstanding, Sanskrit never disappeared. Many religious texts were written and passed on in Sanskrit. This allowed Sanskrit to have a continuing influence on the development of languages of India. Thus, we aim to discover how Sanskrit and Prakrit differently influenced the creation of the Hindi and Marathi languages, and better understand how each of the modern languages relates to Sanskrit, or the intermediate Prakrit language, Pali.

2 Approach:

We first had to acquire source data for all of our languages. Our source data for Sanskrit is comprised of texts from the epics Mahabharata and Ramayana from the Classical Language Toolkit (CLTK). Our source data for Prakrit, is it's earliest form, known as Pali, and is comprised of the teachings of Buddha from Vipassana Research Institute. Our source data for Hindi is comprised of history articles from the Resource Centre for Indian Language Technology Solutions. Our source data from Marathi is comprised of a novel and speculations on astrology also from Resource Centre for Indian Language Technology Solutions (CFILT). Overall, we had assembled 62 MB of data.

After acquiring our corpora, we had to make sure that it was appropriately prepared for analysis. So, after downloading our data, we first removed all punctuation that was apparent in the corpora using a custom bash script ./clean.sh. Afterwards, we wanted to normalize our corpora such that all of the data samples had the same representations for the same character and such that our Indic NLP applications could handle the data in a consistent manner. Thus, we created a bash script, ./normalize.sh to normalize all of our corpora. After normalization, our corpora was still spread across hundreds of files. So, in order to simplify our analysis, we merged all of our corpora files for each language into its own single file using a bash script called ./combine.sh. This combined our entire corpora down into four files.

At this point we were prepared to start analyzing our corpora. Our idea was the following. We would first select a feature of the languages. Then, on that feature, we would train four NLTK Naive Bayes Classifiers: Hindi/Sanskrit, Marathi/Sanskrit, Hindi/Pali, Marathi/Pali. From each classifier, we would find an epsilon value calculated by 1 - accuracy. We assumed that this epsilon value was proportional to the similarity between pairs of languages. The reasoning was that if a classifier was able to distinguish between pairs of languages X percent of the time then that means it was confusing the languages (100-X) percent of the time. Hence we let the variable epsilon, designated by 1 - accuracy, represent the similarity score between pairs of languages. From the differences between $\varepsilon_{h/s}$ and $\varepsilon_{m/s}$ we can quantitatively estimate which of the languages shares a feature more with Sanskrit. Likewise, from the differences between $\varepsilon_{h/p}$ and $\varepsilon_{m/p}$ we can estimate which language shares a feature more from Pali.

We then ran our naive bayes classifiers on the following features: words, 1-char grams, 2-char grams. 3-char grams, 4-char grams, 5-char grams, and 6-char grams. Due to the shallow orthography of these languages, the n-char grams features give insight in to phonological similarities, while the word-only features keep track of any identical words and any homographs. For each of these classifiers, we used a 20% test, 80% training split and performed cross-validation such that each sample had been tested once. For each of these classifiers, we then calculated the mean and standard deviation. The results of our data are in Appendix A. The python code that we ran for our project is located on `https://github.com/VSingh98/indic-lang-development` .

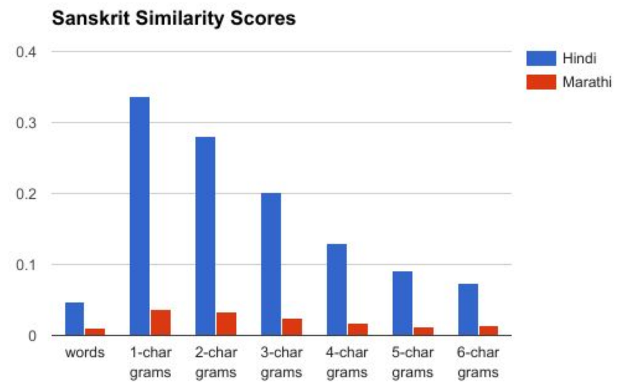
3 Results:

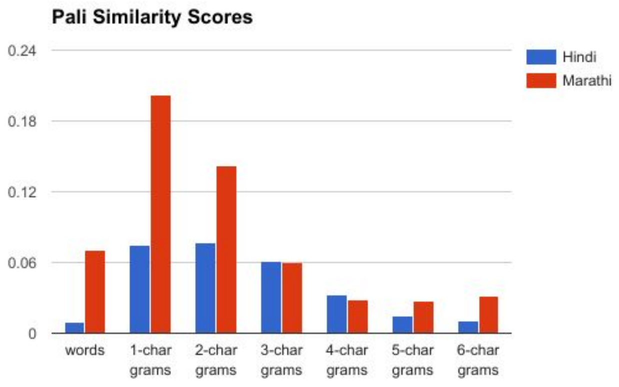
Upon looking at our results, in Appendix A, our data showed that overall Marathi is more similar to Pali than Hindi is, while Hindi is more similar to Sanskrit than Marathi is.

Let us define similarity distributions as $(\varepsilon,\ \sigma)$, where epsilon is the mean similarity score and sigma is the standard deviation. When our features chosen were words themselves, Hindi/Sanskrit had a similarity score of (0.048, 2.2e-4) compared to Marathi/Sanskrit's score of (0.011, 8e-5). This demonstrated that Hindi, on a word basis, shares more with Sanskrit than Marathi. On the other hand, Hindi/Pali had a similarity score of (.009, 1e-4) as compared to Marathi/Pali's similarity score of (.071, 6.1e-4). As expected, the error rate was very low for whole words, and of course this metric merely scratches the surface of lexical similarity. That being said, it still offers us an interesting result because we see that Hindi lines up with Sanskrit and Marathi with Pali, in agreement with the rest of our findings.

Our results for our N-char gram feature classifiers corroborate what we see from the results from our word-feature classifiers. The table on Appendix A shows that for 1-char, 2-char, and 3-char grams, Hindi and Sanskrit have a similarity score greater than 0.2. Meanwhile, for these three grams, Marathi and Sanskrit have a similarity score lower than .04. On the other hand, Marathi goes to share a lot more with the Prakrit language, Pali than Hindi does.

Overall, we noticed that as we increased the N-char grams from 1 to 6, Marathi's similarity to Pali and Hindi's similarity to Sanskrit dramatically lowers. As you can see, the 2-char gram shows Hindi is more similar to Sanskrit than Marathi is with a similarity score of 28% while Marathi and Pali have a similarity score of 14%. After seeing this pattern in the 3-char gram, we continue to see this pattern in the 4-char gram. In 4-char grams we see Hindi is more similar to Sanskrit than Marathi is with a similarity score of 13% while Marathi's and Pali's similarity score continues to drop to 2%. The following two bar charts help summarize our results:





4 Discussion:

Our system took massive amounts of Hindi, Marathi, Pali, and Sanskrit and put them each into 1 separate file. Using the Indic NLP library, we were able to normalize and tokenize our data. Then, using NLTK, we were able to compare the languages across a variety of features. These features spanned 1-char grams to 7-char grams and included individual words as well.

4.1 One factor we found challenging:

Preparing the data was a big challenge for us. We had to correctly format the data. This included taking out the special characters that do not determine a language like parentheses, newline character and carriage return. Leaving these in the file comprise our results. One detailed example of why we needed to correctly do this was that the Hindi and Sanskrit data sets were made on a windows machine, which has the carriage return character which mac does not. We caught this confounding variable when we were analyzing the most informative features of our classifier. We saw that if a classifier saw a carriage return, it was able to predict a language with extremely high frequency. Future investigations found the same with question marks, commas, and other special characters. To resolve this, we created a bash utility to properly clean our corpora.

4.2 Another factor we found challenging:

The size of our corpora for each language varied widely. We had 1MB of data for Marathi, 3MB for Pali, 21MB for Sanskrit and 37MB for Hindi.

The problem arises when we are comparing languages whose corpora are of different sizes. For example, for a Hindi/Pali classifier, on average, 37/40 of the data in our test sets would be comprised of Hindi. Thus, if our Naive-Bayes Classifier actually over-classifies samples as Hindi, despite not being good at classifying Pali, then it would skew the overall accuracy of the classifier towards a high result, and correspondingly low similarity score. Thus, we are aware that having such various sizes of corpora can affect our results. However, we can still form valid comparisons of similarity scores because that skewed-ness is constant between classifiers.

5 Conclusion

By training our classifiers on words, we got an idea for the lexical overlap between the languages. We picked up any identical words, and of course any potential homographs. As expected, the differences were small.

When we looked at the n-grams, we got an idea for similarities shared between the sounds of the language. Because of the shallowness of the orthography of these languages, we are able to glean some information about phonological similarities between the languages. Similarities in things like consonant combinations would have helped contribute to the overlap we see in the stats.

Generally speaking, our results show that Sanskrit had more lexical and phonological overlap with Hindi than Marathi, whereas Pali had more lexical and phonological overlap with Marathi than Hindi.

We could expand upon our data by seeing just what were the most useful features were when we trained the classifier on n-char-grams. This could have given us a better idea of just which consonant combinations were unique / indicative of a certain language. Also, upon a slight modification of the NLTK NaiveBayesClassifier class, we could also expand upon this bey seeing which n-char-grams were the least useful in identifying a language - in other words the n-char-grams most similar across the languages. Overall, after gaining the insights we did from our project, so many doors opened with what we could do to expand upon this study, we regret that we were not able to go through all of them.

Appendix A

epsilon = 1 - accuracy

80% training, 20% test. Used Cross-Validation to rotate our training and test sets five times for each classifier.

```
features: words
        hindi/sanskrit   mean epsilon: 0.048     stdev: 0.00022
 marathi/sanskrit         mean epsilon: 0.011     stdev: 0.00008
        hindi/pali        mean epsilon: 0.009     stdev: 0.00010
  marathi/pali            mean epsilon: 0.071     stdev: 0.00061

==================================================================
features: 1-char grams
        hindi/sanskrit   mean epsilon: 0.336  stdev: 0.00022
 marathi/sanskrit         mean epsilon: 0.037  stdev: 0.00012
        hindi/pali        mean epsilon: 0.075  stdev: 0.00024
  marathi/pali            mean epsilon: 0.202  stdev: 0.00046

==================================================================
features: 2-char grams
        hindi/sanskrit   mean epsilon: 0.281  stdev: 0.00008
 marathi/sanskrit         mean epsilon: 0.034  stdev: 0.00010
        hindi/pali        mean epsilon: 0.077  stdev: 0.00019
  marathi/pali            mean epsilon: 0.142  stdev: 0.00079

==================================================================
features: 3-char grams
        hindi/sanskrit   mean epsilon: 0.202     stdev: 0.00024
 marathi/sanskrit         mean epsilon: 0.025     stdev: 0.00013
        hindi/pali        mean epsilon: 0.061     stdev: 0.00027
  marathi/pali            mean epsilon: 0.060     stdev: 0.00077

==================================================================
features: 4-char grams
        hindi/sanskrit   mean epsilon: 0.130     stdev: 0.00021
 marathi/sanskrit         mean epsilon: 0.018     stdev: 0.00020
        hindi/pali        mean epsilon: 0.033     stdev: 0.00028
  marathi/pali            mean epsilon: 0.028     stdev: 0.00024

==================================================================

features: 5-char grams
        hindi/sanskrit   mean epsilon: 0.092     stdev: 0.00031
  marathi/sanskrit        mean epsilon: 0.013     stdev: 0.00025
        hindi/pali        mean epsilon: 0.015     stdev: 0.00014
  marathi/pali            mean epsilon: 0.027     stdev: 0.00082

=================================================================
features: 6-char grams
        hindi/sanskrit   mean epsilon: 0.073     stdev: 0.00032
  marathi/sanskrit       mean epsilon: 0.014     stdev: 0.00015
        hindi/pali       mean epsilon: 0.011     stdev: 0.00021
  marathi/pali           mean epsilon: 0.032     stdev: 0.00097

======================================================================
```