ANALYZING INDIA'S LINGUSTIC LEGACY: A SIMILARITY COMPARISON OF INDIAN SCRIPT FAMILIES

Vaishnavi Annabhemoju
School of Computing
University of Nebraska, Lincoln
vannabhemoju2@huskers.unl.edu

Bala Subramanyam Duggirala
School of Computing
University of Nebraska, Lincoln
bduggirala2@huskers.unl.edu

Abstract **---- This project aims to investigate the evolution and influence of the Brahmi script on various Indian languages, with the objective of understanding the degree of similarity and divergence between them. Brahmi is one of the oldest writing systems in India, and it has played a crucial role in the development of many Indian scripts, including Devanagari, Tamil, Telugu, and Kannada. Despite its declining usage in contemporary India, the Brahmi script still holds immense historical and cultural significance. By exploring the relationship between Brahmi and Indian languages, this project aims to shed light on the evolution of Indian scripts and their cultural and linguistic diversity.**

Keywords **---- Feature Analysis, Heat maps, Similarity Matrices, Indian languages, Brahmi script.**

1. INTRODUCTION

India is a country that boasts an incredible linguistic diversity, with the second highest number of languages in the world, i.e., 780, out of which 22 are officially recognized and most widely spoken across the country. The country's linguistic heritage is deeply intertwined with its rich history and culture, and each language and dialect tell a unique story of its own.

Indian languages can be broadly classified into four main categories: Indo-Aryan, Dravidian, Sino-Tibetan, and Austro-Asiatic. Each category has its own distinct linguistic features and is spoken in specific regions of the country. For instance, Indo-Aryan languages are primarily spoken in North and Central India, while Dravidian languages are spoken in South India. In addition to their linguistic diversity, Indian languages are also known for their diverse scripts. The scripts used for writing Indian languages are as varied as the languages themselves, with each language having its own unique writing system. These scripts have evolved over centuries and have been influenced by various factors, including historical events, cultural practices, and religious beliefs.

The Brahmi script is an ancient writing system that was used to write various languages in India, including Sanskrit and Prakrit, as early as the 3rd century BCE. It is believed to have originated in the Maurya period in eastern India and was used across the Indian subcontinent for centuries. The Brahmi script is known for its distinctive rounded characters, which are thought to have been inspired by early Aramaic and Phoenician scripts. Over time, the Brahmi script gave rise to several other writing systems that are still used in India today.

For instance, the Devanagari script, used primarily for writing Sanskrit and Hindi, evolved from the Brahmi script around the 11th century BCE. The Devanagari script is known for its distinctive horizontal line that runs across the top of the characters. Similarly, the Tamil script, used primarily for writing Tamil, evolved from the Brahmi script in the 3rd century CE. The Tamil script is known for its distinctive curved characters, which are written from left to right. Other scripts that evolved from the Brahmi script include the Telugu

script, used primarily for writing Telugu and other Dravidian languages, and the Kannada script, used primarily for writing Kannada. Each of these scripts has its own unique features and is widely used for writing different Indian languages today. For example, the Devanagari script is used for writing Hindi, Marathi, and Gujarati, among other languages. In addition to their use for writing Indian languages, these scripts have also been used for writing religious texts and literature. For example, the Devanagari script is used for writing the ancient Hindu texts, such as the Vedas and Upanishads, while the Tamil script has been used for writing Tamil literature for centuries.

Despite the declining usage of some Indian scripts in contemporary India, they continue to hold immense cultural and historical significance. The study of Indian languages and their scripts provides a fascinating insight into the country's cultural and linguistic diversity. This project aims to explore the relationship and similarities between Brahmi script and various Indian languages, with the objective of understanding the degree of commonality between them. By doing so, it aims to shed light on the evolution of Indian scripts and their unique cultural and linguistic heritage.

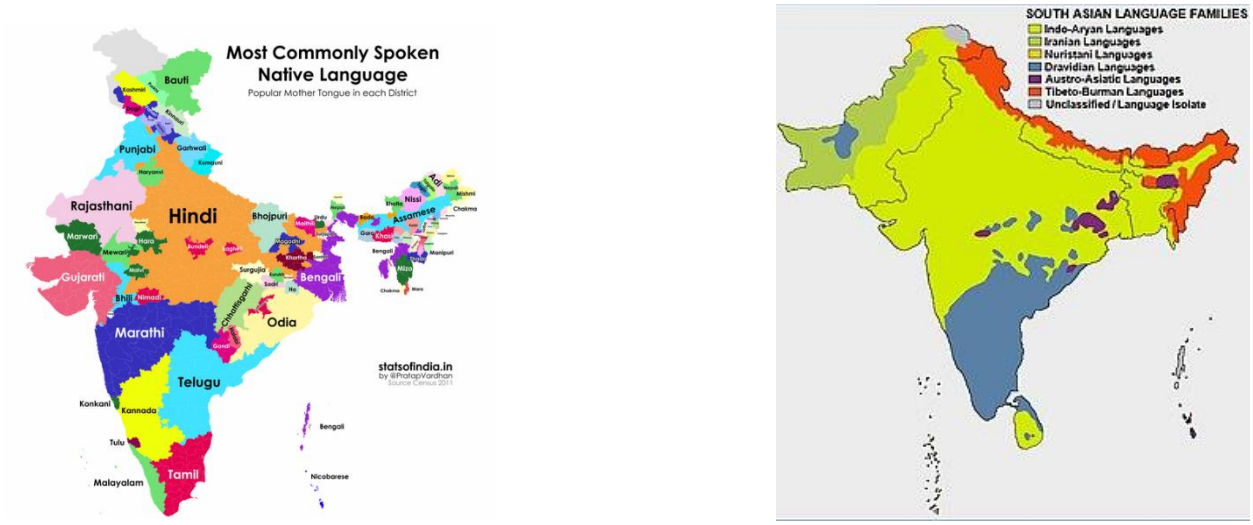

Fig 1: Indian Language Diversity[1]

2. BACKGROUND

The earliest known complete inscriptions in the Brahmi script are written in Prakrit and date back to the 3rd to 1st centuries BCE. These inscriptions, such as the famous Edicts of Ashoka from around 250 BCE, provide valuable historical evidence. Prakrit records dominate the epigraphic discoveries in the Indian subcontinent until approximately the 1st century CE. Epigraphical Hybrid Sanskrit, a combination of Sanskrit and Prakrit languages, has been found in ancient inscriptions across various sites in North and Central India, with occasional discoveries in South India as well. These inscriptions provide valuable evidence of the linguistic and cultural interactions that took place in different regions during ancient times.

Through the span of a thousand years, the Brahmi script underwent a process of evolution, resulting in the emergence of several regional scripts. Gradually, these regional scripts became closely associated with the local languages spoken in those areas. A variant of Brahmi known as Northern Brahmi eventually gave rise to the Gupta script, which gained prominence during the Gupta Empire and is sometimes referred to as "Late Brahmi." During the Middle Ages, the Gupta script further diversified, giving rise to various cursive scripts such as the Siddhaṃ script (6th century) and the Sarada script (9th century).

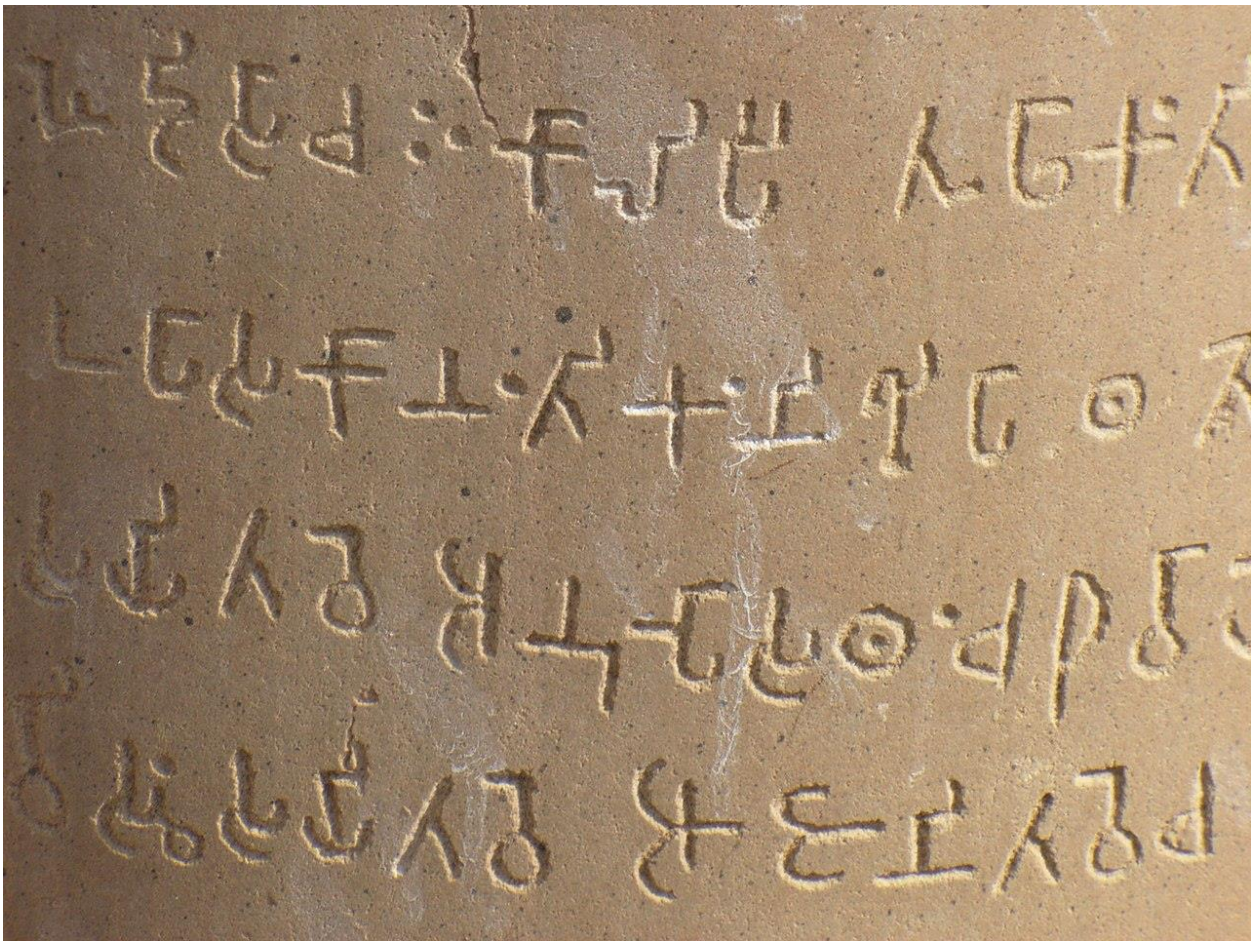

Fig 2: Brahmi Script on Ashoka Pillar (250 BCE)[2]

The Southern Brahmi script played a significant role in the development of various scripts in different regions. From Southern Brahmi, the Grantha alphabet emerged in the 6th century, followed by the Vatteluttu alphabet in the 8th century. Moreover, the influence of Hinduism in Southeast Asia during the early centuries CE led to the creation of other scripts like Baybayin in the Philippines, the Javanese script in Indonesia, the Khmer alphabet in Cambodia, and the Old Mon script in Burma. These diverse scripts were a direct result of the interaction between Southern Brahmi and the local languages and cultures in these regions.

The Brahmi script has not only influenced scripts within the Indian subcontinent but has also extended its impact to Central Asian scripts. Among these are the Tibetan script, the Tocharian script (also known as slanting Brahmi), and the script used to write the Saka language. Within the Brahmi script family, the Nagari script developed, which further transformed into the Devanagari script and the Nandinagari script. Initially, both Devanagari and Nandinagari were used to write Sanskrit, but later Nandinagari merged into Devanagari. This resulting script is widely adopted in India to write Sanskrit, Marathi, Hindi, its dialects, and Konkani.

3. DATA SOURCE

This project aimed to explore the similarities and relationships between the Brahmi script and its descendants, shedding light on their evolution and connections. The scripts examined in this project include Brahmi, Devanagari, Nandinagari, Gujarati, Telugu, Meitei, Modi, Gurmukhi, Bengali, Kannada, Malayalam, and Tamil. We have chosen only a selective number of letters from each script, roughly 20, by selecting only the letters which were substantially different from other letters in the script.

By examining the Wikipedia pages dedicated to these scripts and utilizing resources such as the Omniglot website, we were able to gather valuable information and insights. These resources provided a wealth of knowledge on the historical development, characters, phonetics, and writing systems of the scripts under investigation. Throughout the project, we delved into the unique characteristics of each script, studying their distinctive shapes, letter formations, and the linguistic traditions associated with them. We aimed to identify commonalities and patterns, such as shared script elements, to better understand the interconnections between these scripts.

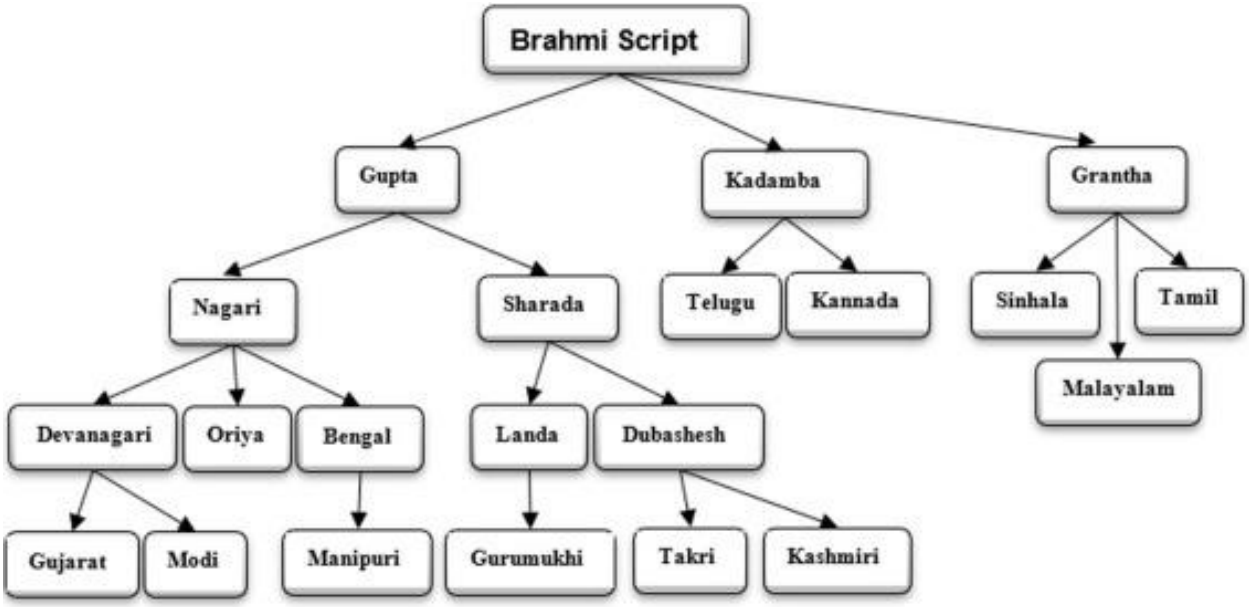

Fig 3: Family tree of Indian Scripts[3]

4. APPROACH

The objective of this project was to create a similarity matrix that would enable the comparison of any two languages based on a set of 17 features as shown in Fig. 4. To achieve this, our approach involved a series of steps. Firstly, we conducted a comprehensive manual analysis of the features present in each script under consideration. This step allowed us to identify and document the unique characteristics and traits of each script.

Next, we generated feature matrices for each script, capturing the presence or absence of the selected features. If the feature is present it is denoted by 1 or else 0. These matrices served as a representation of the script's distinctive attributes, enabling us to quantify and compare them.

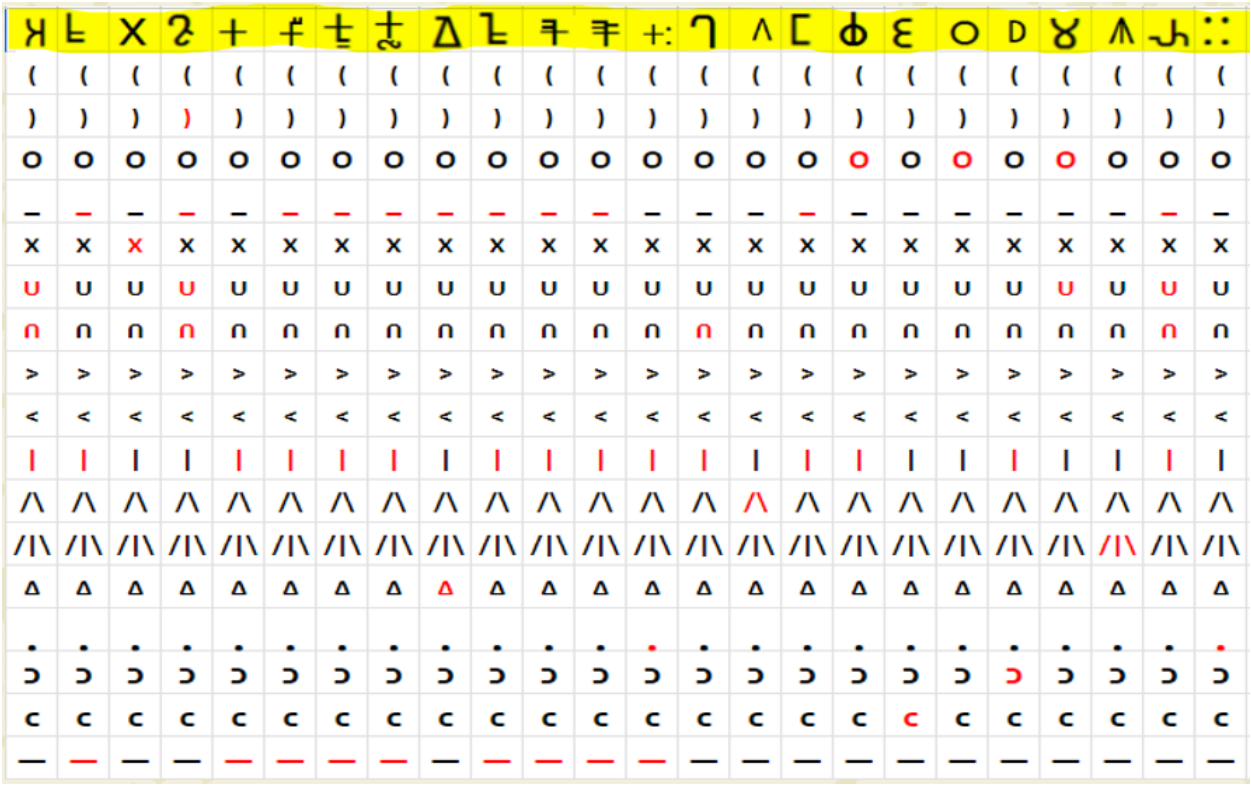

Fig. 4: Feature analysis for the Brahmi script.

After the feature matrices were prepared, they were stored as CSV files, which served as inputs for our logic which is shown in Fig. 5. CSV files are a widely used format for storing tabular data, making them suitable for our purposes. By storing the feature matrices in CSV files, we ensured that the data was organized and easily accessible for further analysis. The CSV format allowed us to store the feature values in a structured manner, with each row representing a script and each column representing a specific feature.

The subsequent step involved generating similarity matrices for all possible combinations of two scripts. This process entailed comparing the feature matrices of each script pair and calculating the degree of similarity between them. By examining the patterns and relationships within these matrices, we could identify the level of resemblance and affinity between different scripts.

To enhance the visual understanding of the results, we produced heatmaps based on the similarity matrices. These heatmaps provided a graphical representation of the script similarities, with warmer colors indicating stronger resemblances and cooler colors indicating greater differences.

By following this approach, we were able to create a comprehensive similarity matrix that allowed for a quantitative assessment of the relationships between various scripts. This matrix, along with the corresponding heatmaps, provided valuable insights into the interconnectedness and shared characteristics among the languages examined in this project.

```
def similarity_matrix_calc(init_matrix_1, init_matrix_2):
    similarity_matrix = np.zeros([init_matrix_1.shape[0], init_matrix_2.shape[0]])
    for i in range(init_matrix_1.shape[0]):
        for j in range(init_matrix_2.shape[0]):
            count = 0
            for k in range(init_matrix_1.shape[1]):
                if init_matrix_1[i,k]==init_matrix_2[j,k]:
                    count += 1
            similarity_matrix[i,j] = count

    return similarity_matrix
```

Fig. 5: Similarity matrix logic.

5. RESULTS

For our analysis, we had chosen twelve scripts out of the twenty-two scripts from the Brahmi script's family tree. During this selection, we made sure that at least one script was chosen from the three main branches of the *Brahmi* script's family tree, i.e., *Gupta*, *Kadamba*, and *Grandha* scripts. The following Fig. 6 shows our selection of the scripts.

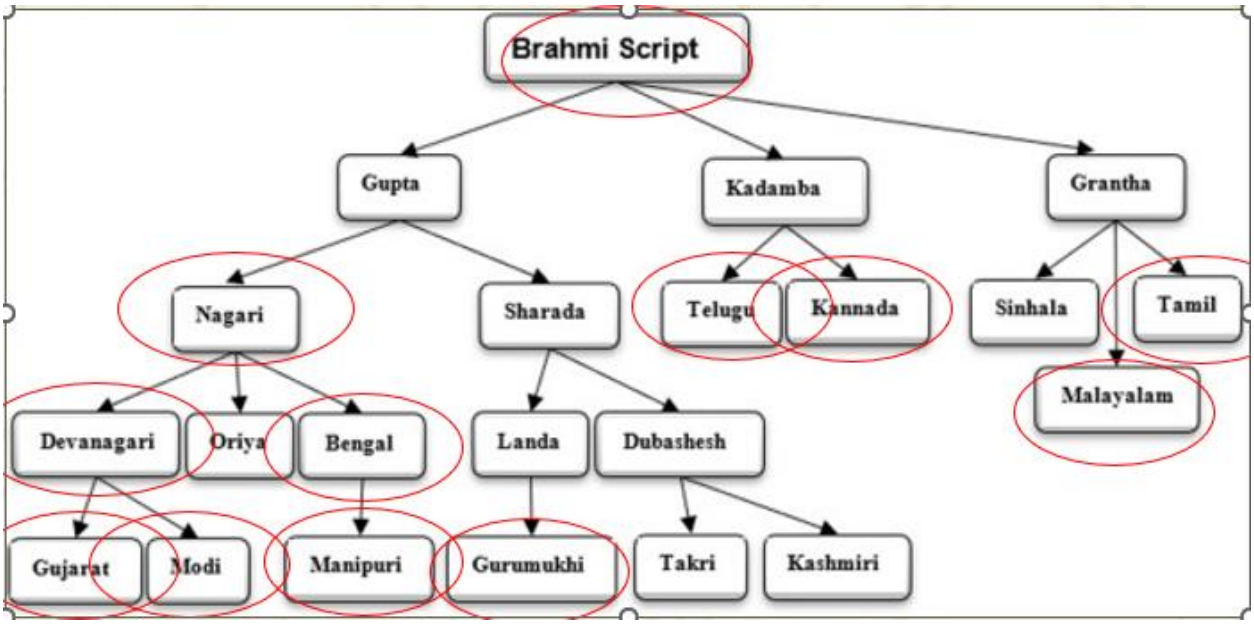

Fig. 6: Our selection of the scripts from *Brahmi* Tree

Comparting each of the above scripts with the rest of the scripts from the selection, we had generated $\mathbf{12C_2 = 66}$ total heat maps (All these heat maps can be found in the *Heat_Maps* folder of the project's artifacts. A heat map is a visual enhancement of the similarity matrix that highlights both the similarities and the

differences between each letter of the two scripts in comparison. A sample heat map from our results in shown below in Fig 7.

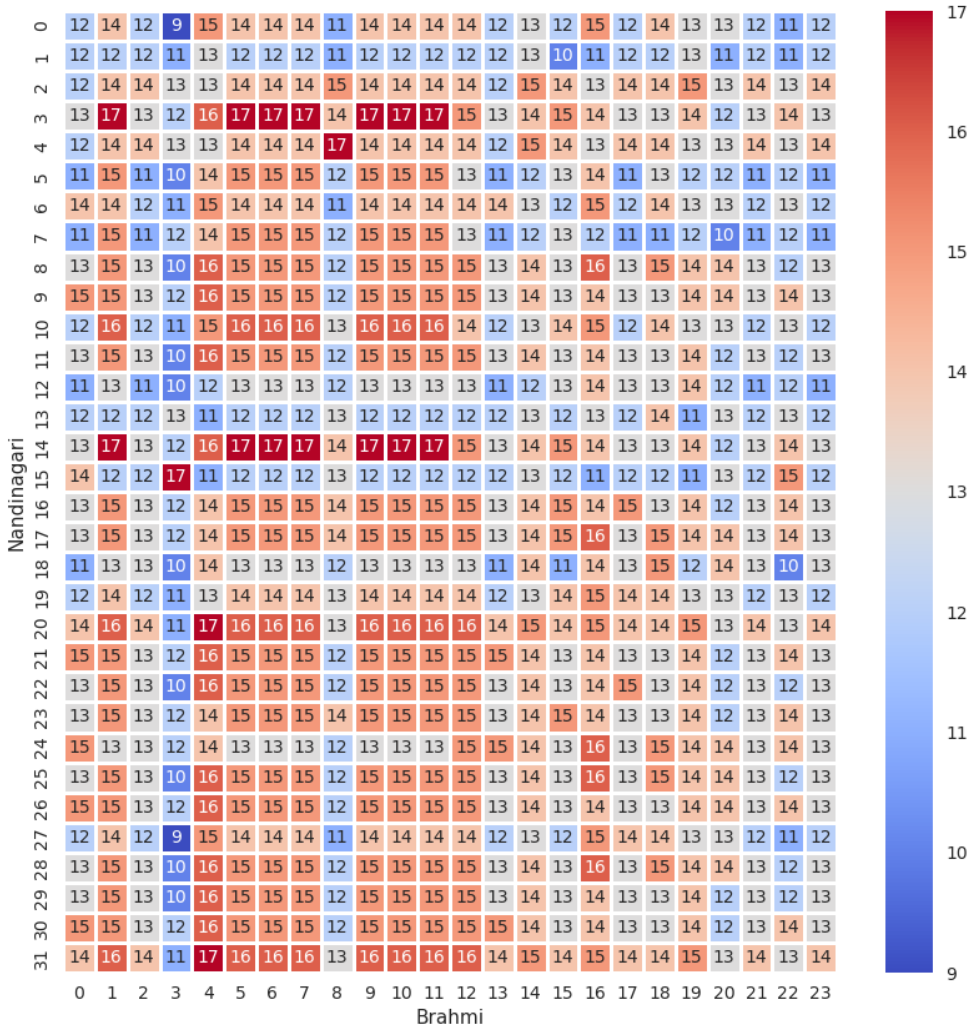

Fig. 7: Sample Heat Map for *Brahmi* vs. *Nandinagari*

From the above heat map that compares the languages *Brahmi* and *Nandinagari* scripts, we can see that the heat map is composed fundamentally with two colors, i.e., Red, and Blue. Darker shades of the Red color, in the similarity index that represents the degree of similarity between two letters of the two scripts, indicate a higher similarity, whereas the Bluer shades indicate lesser similarity. There are also numbers for each index that correspond to the number of the similar (present + absent) features between the two letters. Given this heat maps of two scripts, we can assess the degree of similarity between them, i.e., a predominantly Red heat map indicates a higher degree of similarity and vice-versa.

6. DISCUSSION

We use the family tree of the *Brahmi* script, shown in Figure 6, as the reference for our discussion. Our approach is to compare the different scripts from this family tree and check their similarities/differences based on their position in the tree. The closer two particular scripts are in the family tree (depth-wise and/or breadthwise), the more similar they must be, as indicated by a Red dominated heat map, and vice-versa with a Blue dominated heat map.

We have selected 6 of our observations to be included in this report. Out of these 6 scripts, 5 scripts support our naïve hypothesis that is based on the structure of the tree. We also include the 6th one as an example of observation that goes against our hypothesis. They are discussed as follows:

6.1 Depth

This observation comes from comparing two scripts from the family tree that are at different depth than each other. We look at a comparison between *Brahmi* script, that is at the top of the tree with two other scripts, i.e., *Nagari* (a.k.a. Nandinagari), and *Gujarati* scripts that are two, and four levels deeper, respectively, from th*e Brahmi* script. The respective heat maps are given below in Fig 8, and Fig 9.

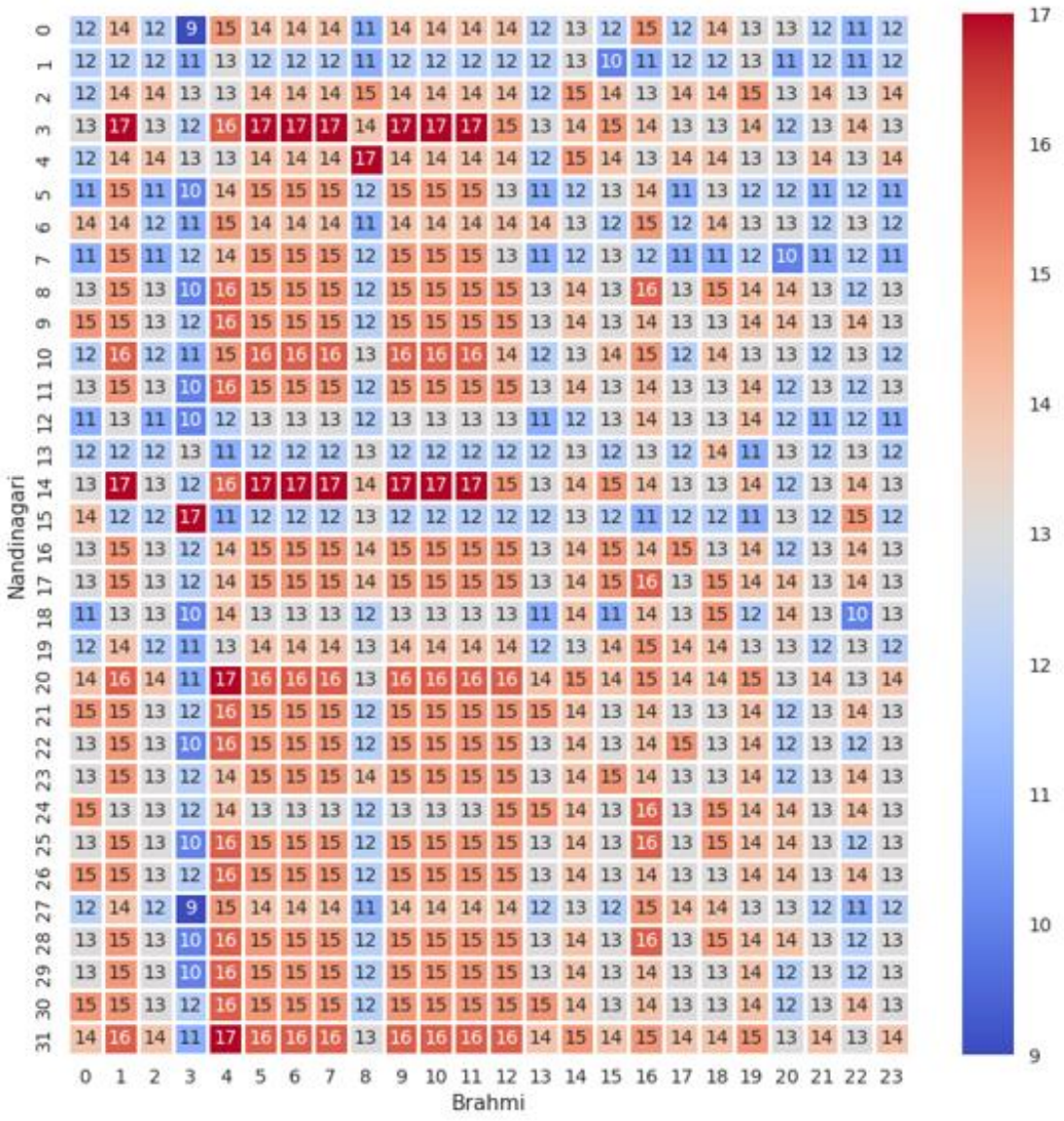

Fig. 8: Comparison 6.1.1 Heat Map for *Brahmi* vs. *Nandinagari*

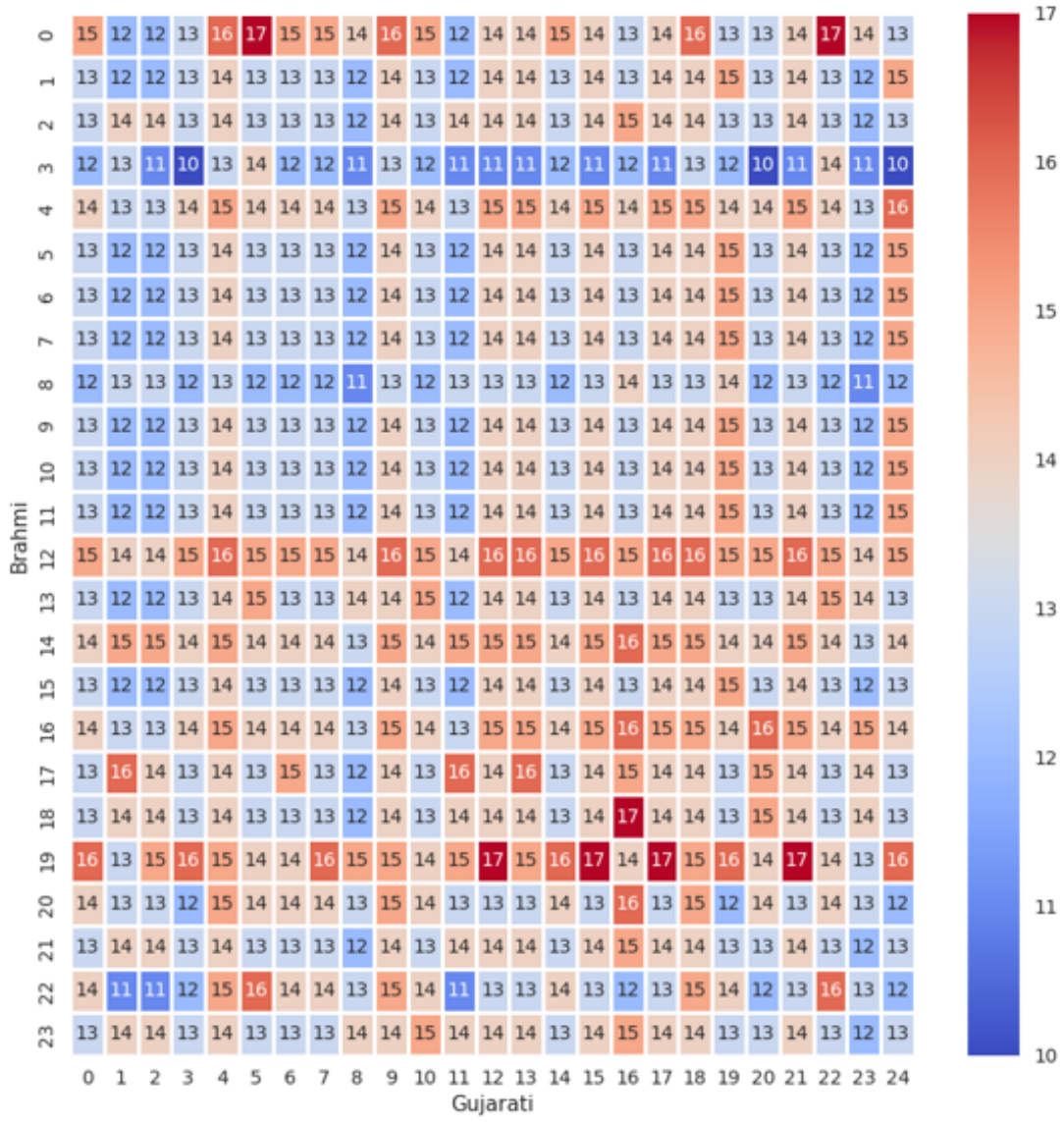

Fig. 9: Comparison 6.1.2 Heat Map for *Brahmi* vs. *Gujarati*

From the above heat maps, we can see that Comparison *6.1.1* that is between closer scripts of the family tree is predominantly Red in color compared to *Comparison 6.1.2*, corroborating our hypothesis about closer scripts being more similar.

6.2 Breadth + Depth

In this comparison, we consider a pivot script *Telugu*, and compare it with the *Nagari* that is from another sub-family (different breadth), and with *Gujarati* that is from another sub-family but is also at a lower level in the tree (different depth). The heat maps are shown below.

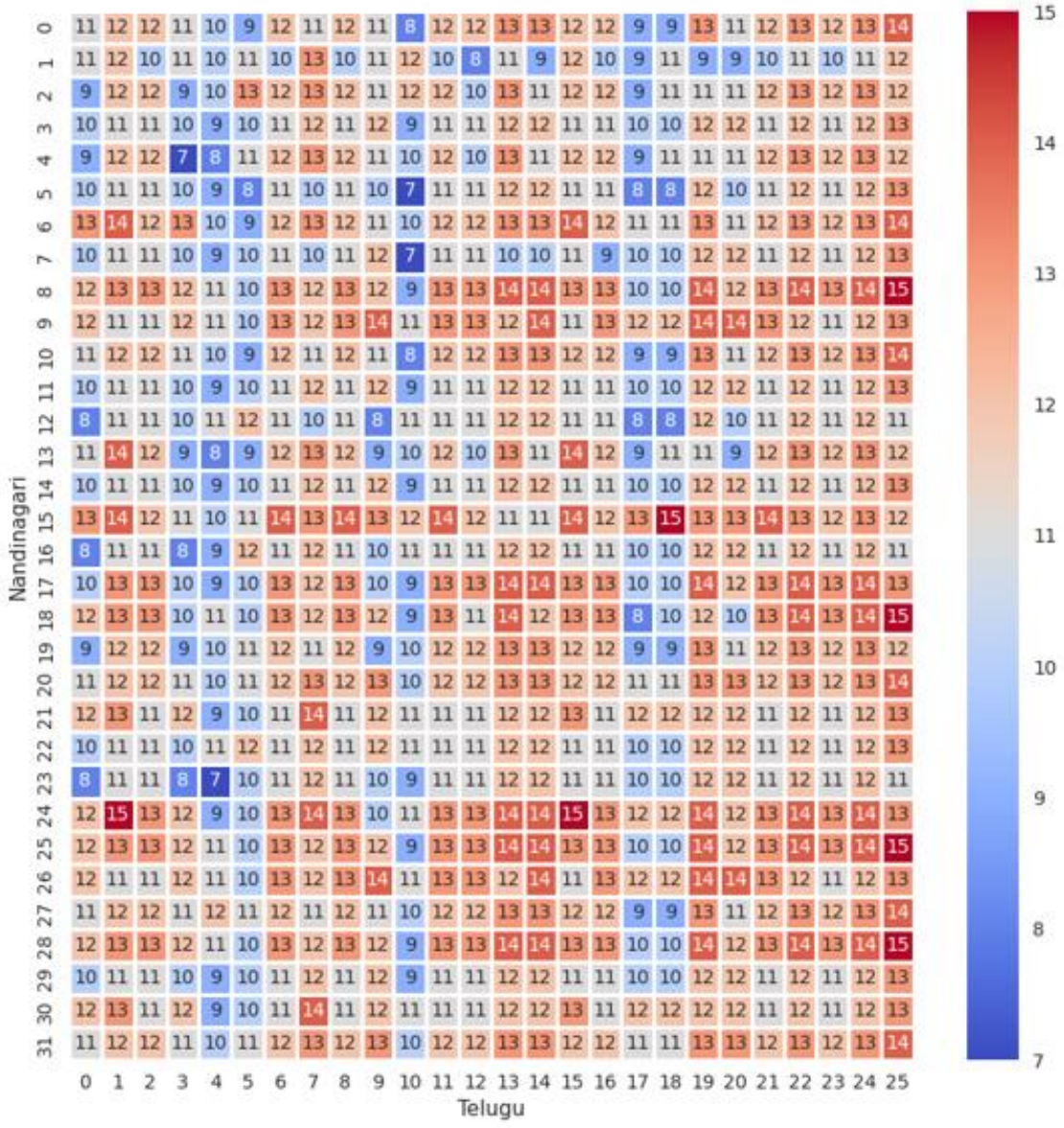

Fig. 10: Comparison 6.2.1 Heat Map for *Telugu* vs. *Nagari*

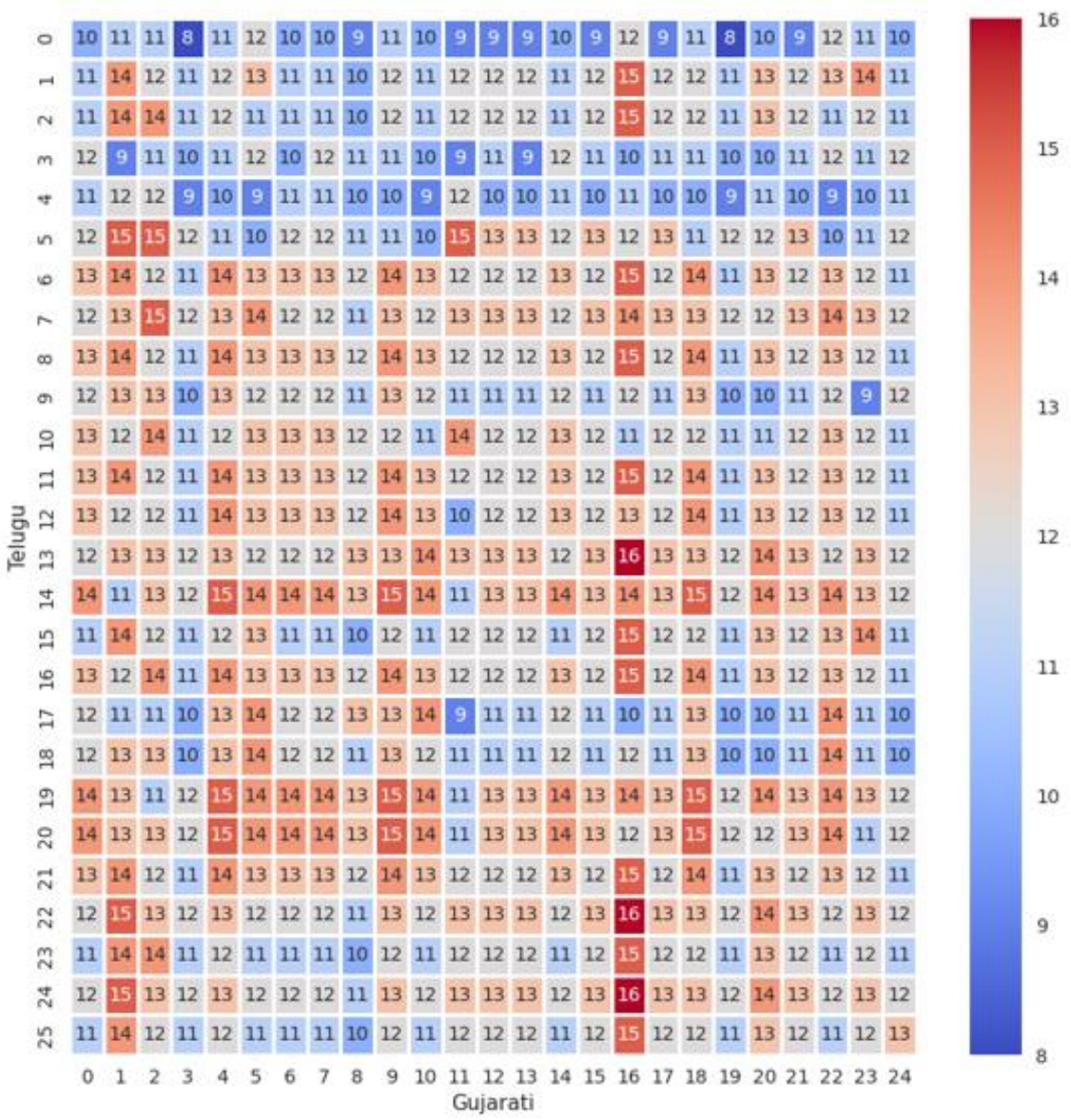

Fig. 11: Comparison 6.2.2 Heat Map for *Telugu* vs. *Gujarati*

From the above, we can see that Comparison 6.2.1 which deals with two scripts at only a different breadth, indicates more similarity than Comparison 6.2.2 which deals with scripts that are at different breadth as well as depth in the family tree.

6.3 Opposite Ends of the Tree

This is the case where we consider scripts from two opposite ends of the tree, i.e., *Gujarati*, and *Malayalam* scripts. These two modern and in-use scripts, although have their common root as the *Brahmi* script, belong to two completely different sub-families of the *Brahmi* script, i.e., *Gupta*, and *Grantha*. We hypothesize that these scripts, given their considerable separation in the family tree, should not be very similar. The heat map of similarity between them is shown as follows:

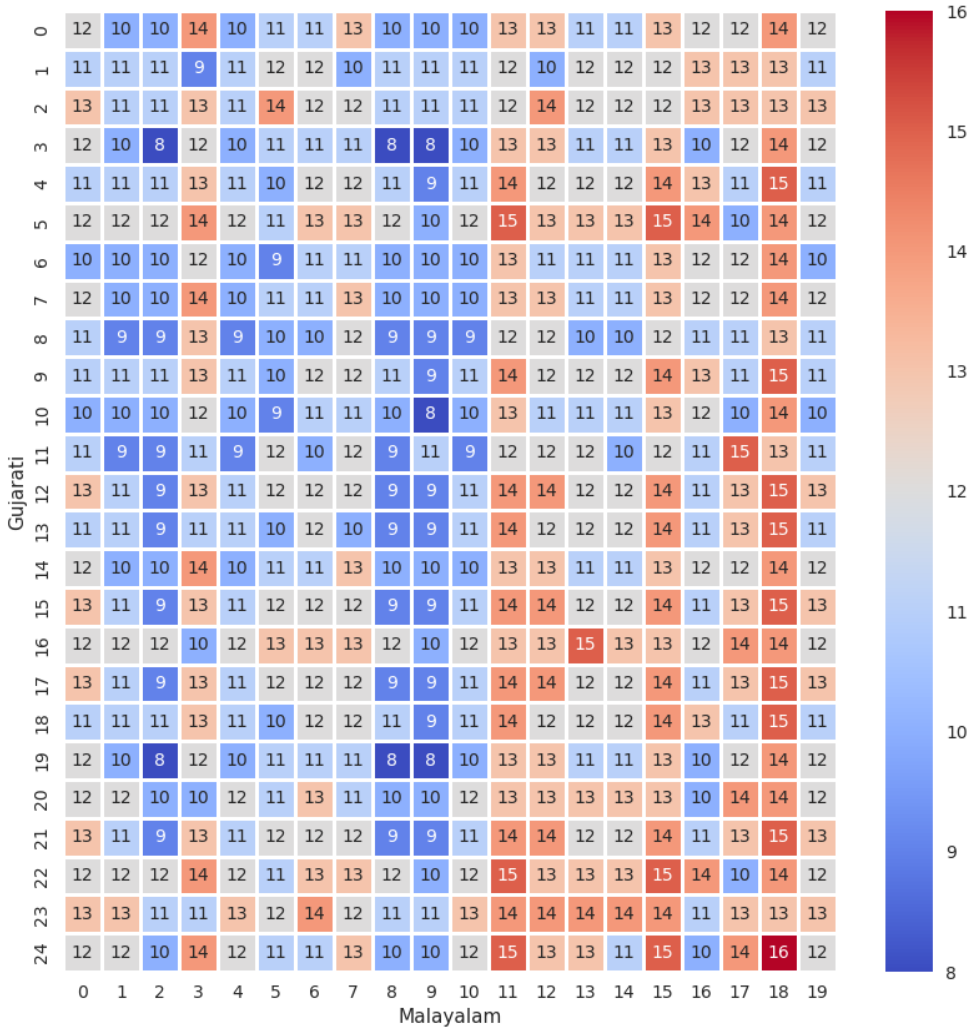

Fig. 12: Comparison 6.3.1 Heat Map for *Malayalam* vs. *Gujarati*

The heat map clearly indicates that the scripts are not very similar to each other, as expected.

6.4 Sibling vs. Distant Relative

For this case, we compare a pivot script *Kannada*, with a. it's sibling *Telugu*, and b. a distant relative *Gujarati*, and hypothesize that the pivot script should be more similar to its sibling compared to the distant relative. The heat maps are shown as follows:

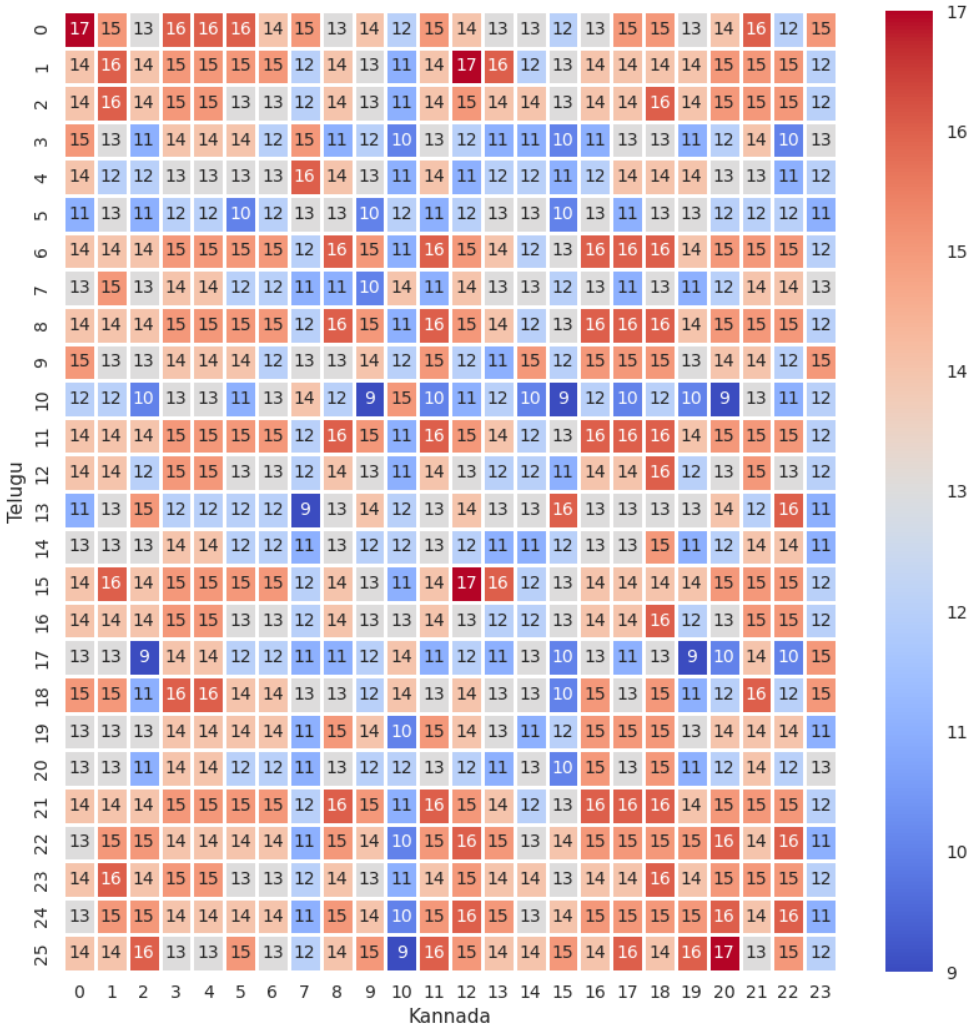

Fig. 13: Comparison 6.4.1 Heat Map for *Kannada* vs. *Telugu*

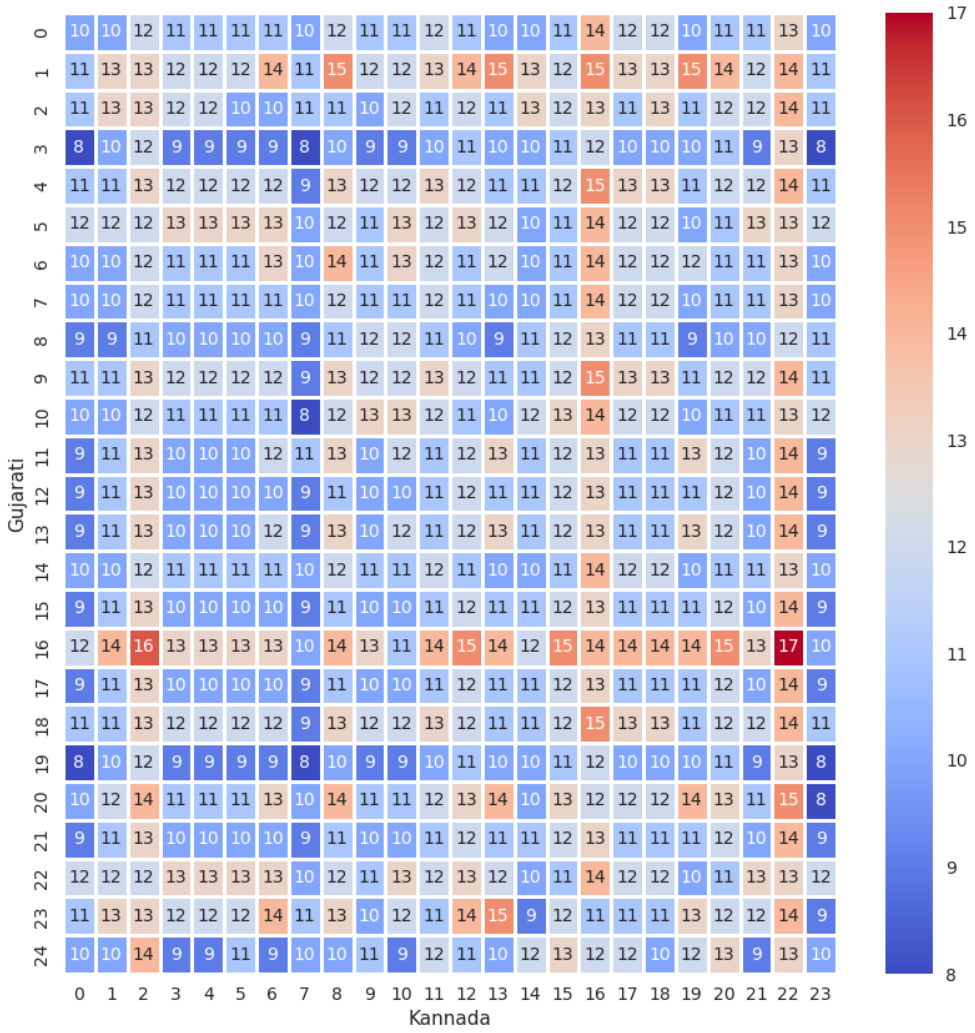

Fig. 14: Comparison 6.4.2 Heat Map for *Kannada* vs. *Gujarati*

Comparison 6.4.1, the sibling script similarity heat map indicates a stronger similarity than Comparison 6.4.2, the distant relative script, supporting our hypothesis.

6.5 Direct Descendants

This case talks about 3 scripts, *Nagari*, *Bengali*, and *Meitei*, where there is direct descendant chain from *Nagari* to *Bengali*, and *Bengali* to *Meitei*. We had checked to see if these scripts were similar to each other, and this turns out to be true. The heat maps are shown below:

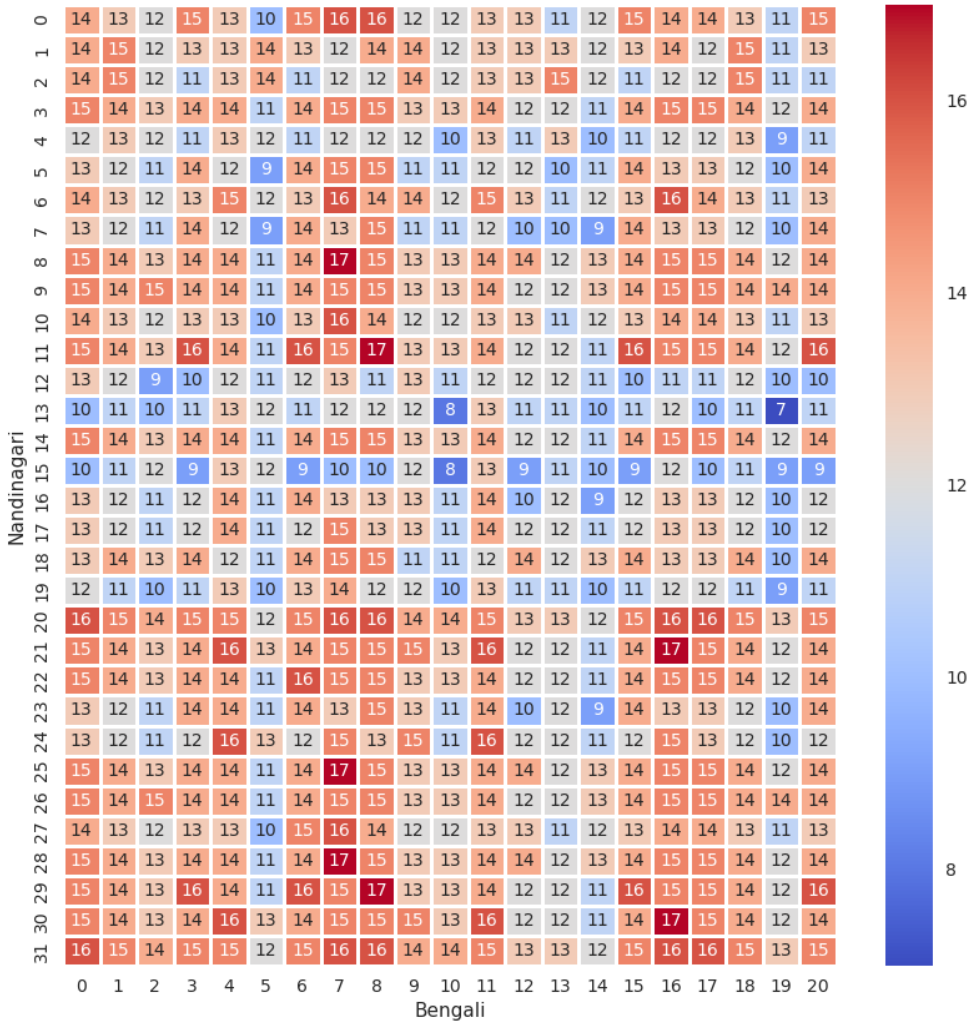

Fig. 15: Comparison 6.5.1 Heat Map for *Nandinagari* vs. *Bengali*

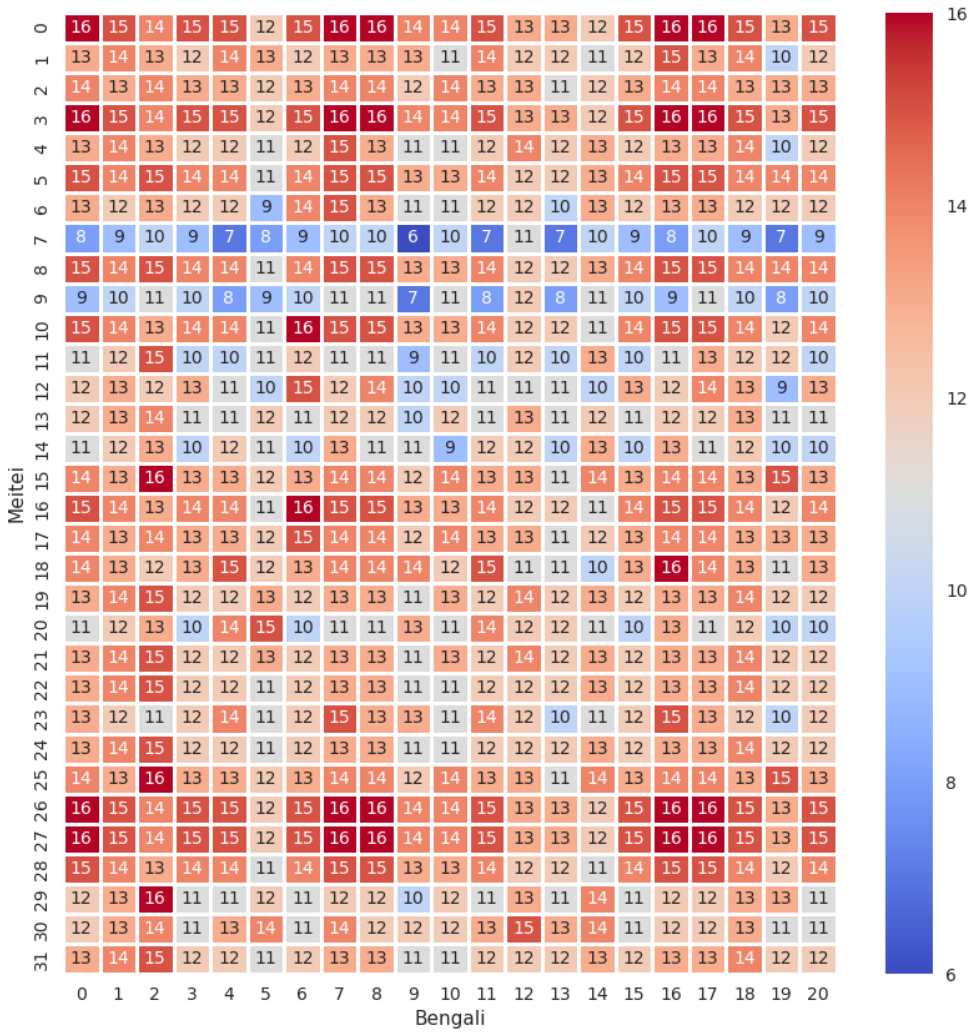

Fig. 16: Comparison 6.5.2 Heat Map for *Bengali* vs. *Meitei*

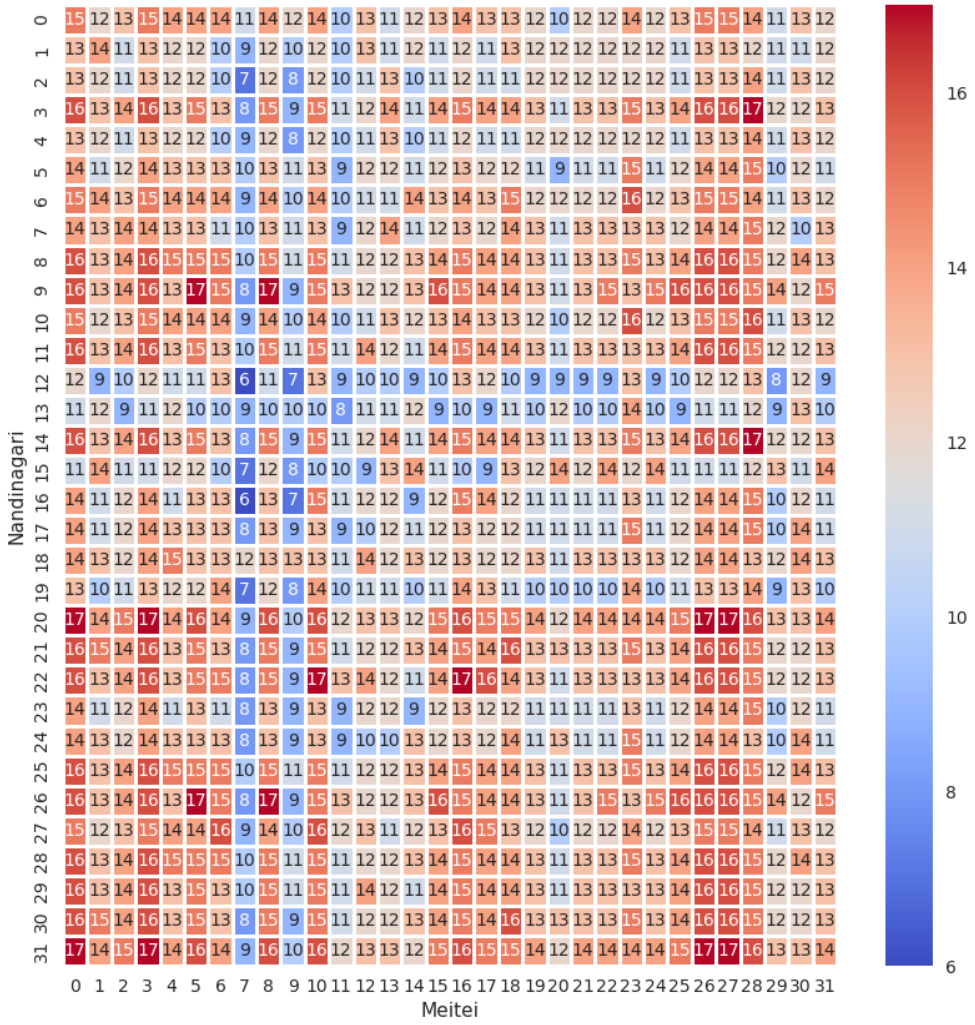

Fig. 17: Comparison 6.5.3 Heat Map for *Nandinagari* vs. *Meitei*

From the heat maps, we see that not only there is a strong similarity between the parent-child scripts, but also between the parent-grandchild scripts, i.e., *Nandinagari* and *Meitei* shown in Comparison 6.5.3.

6.6 Favorite Grandchild?

When comparing the similarities between the parent script Brahmi with three of its grandchildren, i.e., *Nagari*, *Telugu*, and *Malayalam*, that come from three different parents, our naïve hypothesis was that these three scripts should be equally similar to *Brahmi* script. But we observed a deviation. The heat maps of the comparisons are given below.

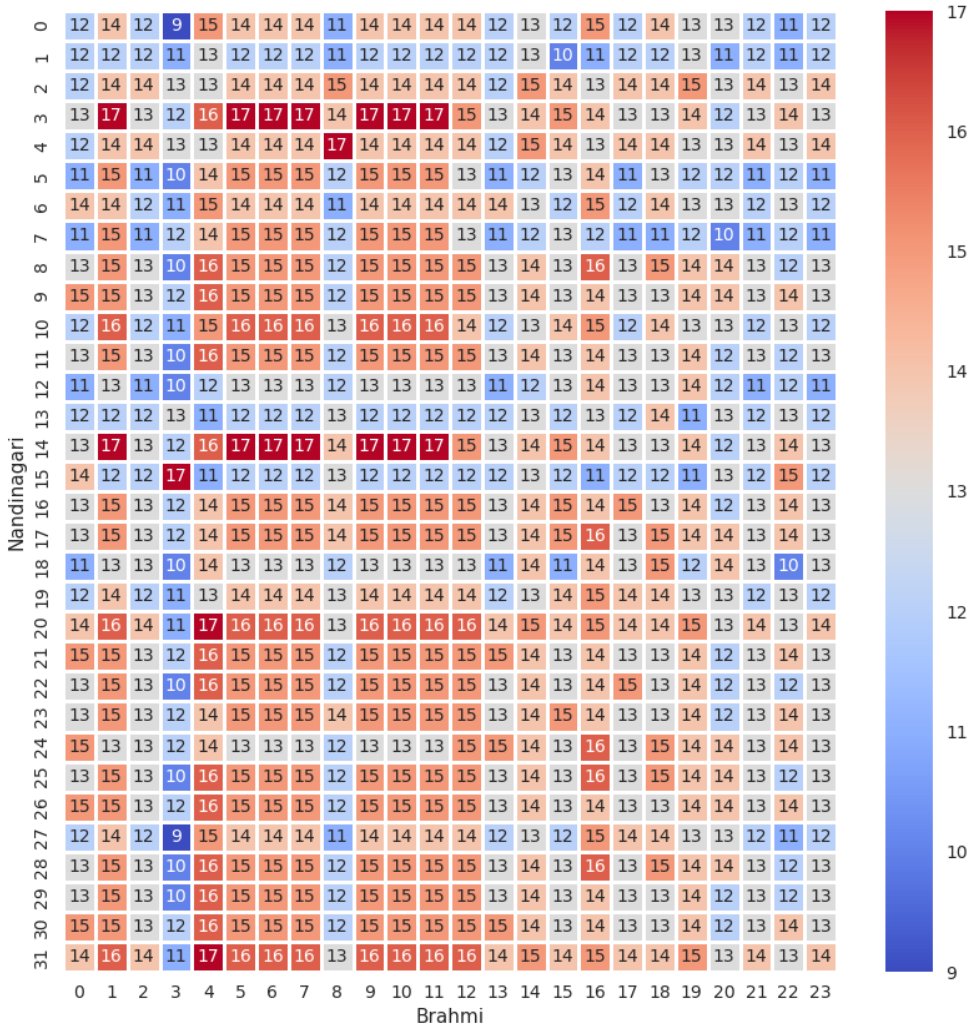

Fig. 18: Comparison 6.5.3 Heat Map for *Brahmi* vs. *Nandinagari*

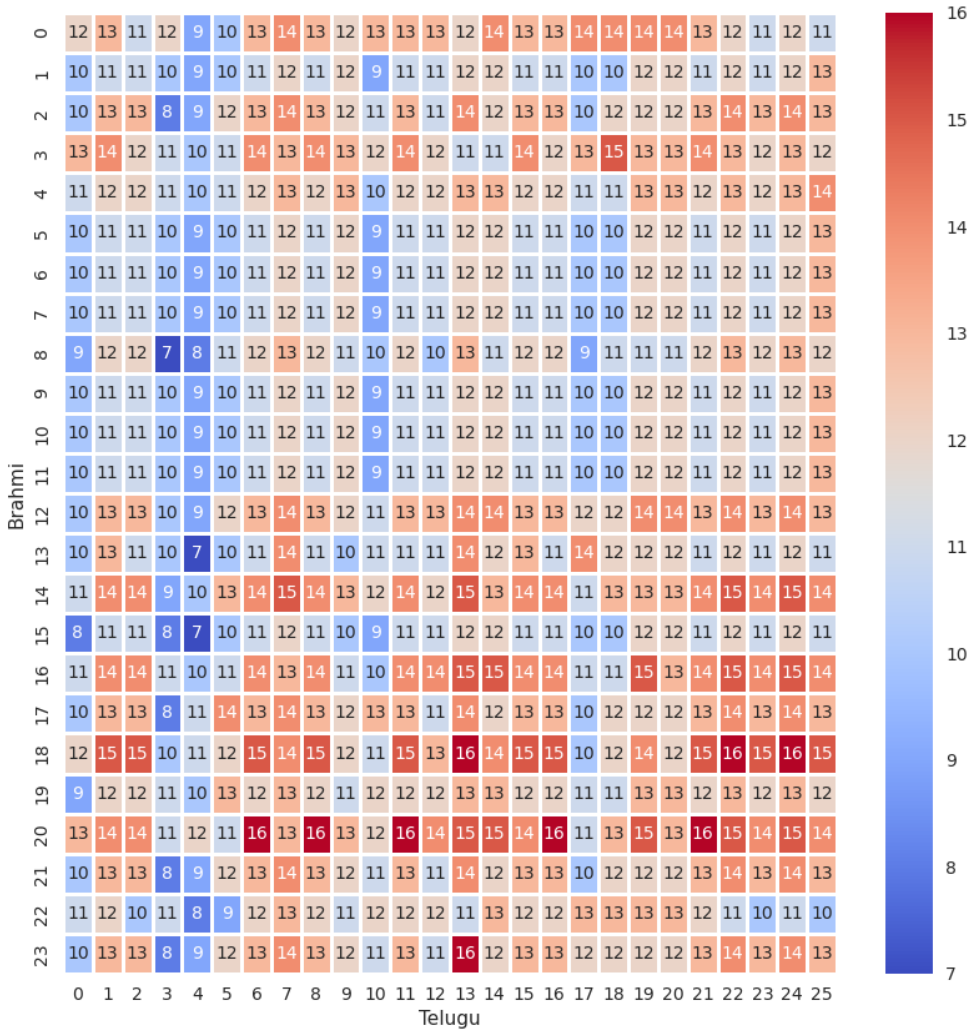

Fig. 19: Comparison 6.6.2 Heat Map for *Brahmi* vs. *Telugu*

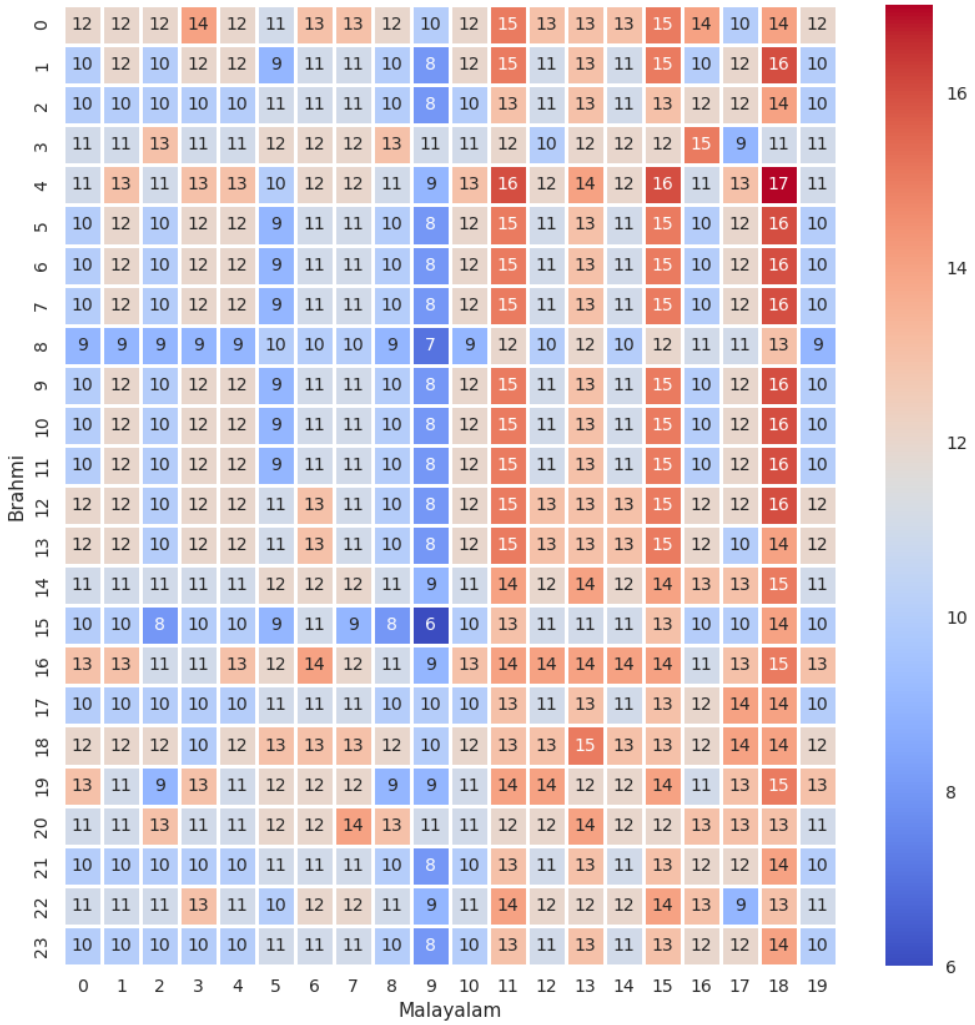

Fig. 20: Comparison 6.6.3 Heat Map for *Brahmi* vs. *Malayalam*

From the three comparisons above, we can see that Comparison 6.6.1, i.e., the heat map of *Brahmi-Nandinagari* indicated predominant similarity compared to *Brahmi-Telugu* and *Brahmi-Malayalam.* This case is an example observation that is against our hypothesis. To add to this, the similarity between *Brahmi-Telugu* is slightly stronger than *Brahmi-Malayalam.*

7. CONCLUSION AND FUTURE WORK

From Section 6 that discusses the results, although we could see from Observations 1, through 5, that there could be many instances where our hypotheses hold, Observation 6 gives an example of an outlier that vary significantly from our hypothesis. The rationale for the complying hypothesis is usually straightforward, i.e., given the general history of various scripts, and the fact that we are already looking at an organized

(formally?) tree. Outliers, however, could arise due to a number of reasons, including inaccurate sources, inaccurate history/research, insufficient feature sets that we used, etc., But most importantly, one can also argue that two scripts, just because they are closely related to each other, need not be similar, i.e., there could be cases where two scripts where one is adapted/derived directly from another, but can still somehow look very different because the borrowers had extensively modified their reference letters from the original scripts.

Given the above, we intend to do the following as our future work. We would first verify the accuracy of our source, i.e., the Brahmi script's family tree. Secondly, although we had modified our base feature list ad-hoc during the generation of heat maps, we would further expand it to extensively include all different shapes/patterns as possible. Thirdly, we would also dig deep into the history of each script and try to find clues that might support our empirical findings. Hopefully, this should give us a more in-depth analysis, and possibly patterns that were previously un-discovered.