Deciphering Indus Scripts through Clustering Techniques and Frequency Analysis

Geetha Ramani
Professor
Department of
Information Technology,
Madras Institute of Technology
Anna University,
Chennai 600044, India

Joseph Samuel
Department of
Information Technology,
Madras Institute of Technology
Anna University,
Chennai 600044, India

ABSTRACT

In this research work, the deciphering of Indus scripts is undertaken through a comprehensive methodology that integrates Clustering analysis, comparative analysis with Tamil Brahmi, identification of primary, secondary, and composite symbols, N-gram analysis and Grammar Analysis. Commencing with Clustering analysis, four algorithms: K-means, Agglomerative, Birch, and Spectral Clustering are employed. By combining the outputs of these algorithms through voting, coherent patterns within the Indus script are identified, paving the way for a deeper understanding of its structural and semantic properties. Following this, a comparative analysis of the Indus script symbols with those of Tamil Brahmi is conducted, exploring potential linguistic connections and cultural influences. Subsequently, primary, secondary, and composite symbols within the Indus script corpus are identified, shedding light on their hierarchical usage and contextual coherence. This hierarchical classification enhances the understanding of the script's semantic organization and usage patterns, providing valuable insights into its communicative capabilities and linguistic conventions. Finally, through N-gram analysis, the predictive modeling of symbol sequences is undertaken, aiming to uncover underlying structures and linguistic patterns encoded within the script's corpus. This analysis yields a list of influential signs, offering fresh perspectives on the script's symbolic and cultural significance. This research also employs a comprehensive approach to analyze the grammatical aspects of the Indus scripts. Utilizing Frequency Analysis, meticulous examination of the co-occurrence of symbols within the script corpus uncovers recurring patterns and potential grammatical markers. Subsequently, through Pattern Recognition and Contextual Analysis, a deeper understanding of the structural and semantic properties of the script is achieved by identifying linguistic patterns. By contextualizing these patterns within the inscriptions and comparing them with known linguistic structures, the aim is to decipher the underlying grammar encoded within the script. Overall, this interdisciplinary approach represents a significant milestone in the ongoing quest to decipher the Indus script, providing innovative methodologies and insights for future research in ancient linguistics and archaeology. Notably, this research marks the first application of Clustering algorithms to the Indus script, thereby pioneering a novel approach to decipherment. It's pertinent to mention that the analysis is conducted on the Interactive Corpus of Indus Texts (ICIT) comprising 694 symbols, providing a robust foundation for the investigations.

Keywords

Indus script, Deciphering, predictive modeling, N-gram analysis, Comparative analysis, Tamil Brahmi, Clustering analysis, Grammar Analysis.

1. INTRODUCTION

The Indus script remains one of the most intriguing and enigmatic writing systems of the ancient world. Dating back over 4,000 years, it was utilized by the Indus Valley Civilization, one of the world's earliest urban societies flourishing in the northwestern regions of the Indian subcontinent, primarily in what is now modern-day Pakistan and northwest India. The script was predominantly inscribed on seals, tablets, and other artifacts discovered at various archaeological sites associated with the civilization, such as Harappa and Mohenjo-Daro.

Despite numerous attempts over the years, the Indus script has yet to be fully deciphered, presenting a tantalizing puzzle for linguists, archaeologists, and historians alike. Its complex symbols and elusive grammar have defied easy interpretation, sparking debates and conjectures about its linguistic and cultural significance.

The Indus script holds immense historical and cultural significance, offering valuable insights into the social organization, trade networks, and religious practices of the Indus Valley Civilization. Its decipherment has the potential to unlock a wealth of knowledge about this ancient society, including its language, literature, and interactions with neighboring civilizations.

Efforts to decode the script have been ongoing for over a century, employing diverse methodologies ranging from comparative linguistics to computational analyses. While progress has been made in identifying patterns and potential linguistic features, the script's full meaning and purpose remain elusive, contributing to its enduring mystique and scholarly fascination. In this research work, various analytical techniques such as Clustering, comparative analysis, evolutionary studies, and N-gram analysis are employed, each offering unique insights into the script's intricate symbolism and linguistic structure. These methodologies aid in categorizing symbols based on intrinsic similarities, shedding light on their potential meanings and contextual usage, and provide valuable clues to potential linguistic connections and cultural exchanges with contemporaneous scripts. However, the script's decipherment continues to pose challenges, emphasizing the need for interdisciplinary collaboration and innovative approaches to unlock its secrets and deepen the understanding of ancient civilizations and human history.

This research recognizes the pressing need for innovative approaches to tackle this ancient script's complexities.

Traditional methods have fallen short in fully unlocking its meaning. Hence, a comprehensive and interdisciplinary approach is adopted, integrating clustering analysis, comparative analysis with Tamil Brahmi, evolutionary studies, N-Gram Analysis and Grammar Analysis.

Clustering analysis is pivotal in categorizing symbols, providing insights into their potential meanings and contextual usage. To achieve this, four robust clustering algorithms are employed—K-means [7], Agglomerative [8], Birch [9], and Spectral Clustering [10]. Notably, this is the pioneering attempt to utilize clustering algorithms in deciphering the Indus script, marking a significant departure from previous methodologies.

2. RELATED WORK

This section presents a comprehensive approach, combining Frequency analysis and Clustering analysis to decode the Indus scripts. While Frequency analysis reveal patterns and cultural insights, Comparative Analysis offers a broader perspective by juxtaposing the script with other ancient writing systems.

2.1 Frequency Analysis

The exploration of n-grams for deciphering the Indus script has a rich history, with earlier studies laying the groundwork for recent advancements in this approach. Gong, H., & Zhang, Y. [1] introduced an innovative approach that blends convolutional neural networks (CNNs) with n-grams, departing from conventional methods to enable a more thorough examination of the script's complexities. Their integration of CNNs allowed for intricate image recognition, providing insights into the script's symbolic and cultural significance. Building upon this foundation, Shan, X., et al. [2] undertook a profound reevaluation of the complexity inherent in the Indus script by employing principles of information theory. Through their pioneering effort, they explored the intricacies of the script, aiming to elucidate its structural characteristics and information density. By analyzing the information content of n-grams and comparing it with other writing systems, they shed new light on the potential underlying structure and complexity of the Indus script, deepening the understanding of its linguistic properties and ancient communication systems.

In a complementary vein, Bhadoria, A., et al. [3] provided a thorough examination of machine learning techniques, including n-grams, applied to decipher the Indus script. They underscored the challenges posed by the limited corpus and the absence of a bilingual reference point while advocating for interdisciplinary approaches. Despite obstacles, their study delved into potential future directions, emphasizing the evolving landscape of Indus script research and the need for integration with other decipherment strategies. This literature survey highlights the progressive nature of research in deciphering the Indus script, showcasing a continuous effort to unlock its secrets through innovative methodologies and interdisciplinary collaboration.

Bahata Ansumali Mukhopadhyay et al. [11] presents a pioneering study delving into the intricacies of the undeciphered inscriptions of the ancient Indus Valley civilization. Their research unveils a sophisticated system of meaning conveyance within these inscriptions, challenging conventional interpretations of the script. By analyzing the structural and syntactic features of the Indus inscriptions and categorizing the signs into content-morphemes and functional-morphemes, the authors demonstrate that the script primarily functioned as a logographic system, representing complete ideas rather than phonetic sounds. Their categorization of Indus logograms into nine functional classes sheds light on the unique roles played by different sign-classes in conveying messages. Moreover, their identification of numerical and metrological signs and analysis of their collocation with specific lexemes provide insights into the quantification mechanisms employed in the inscriptions. This groundbreaking work not only deepens the understanding of the compositional semantics of Indus inscriptions but also sets the stage for further exploration into the socio-cultural and economic aspects of the ancient civilization.

The research addresses missing aspects by considering the position of occurrence of each n-gram, which is utilized to predict missing symbols within the script. Additionally, the study conducts a thorough analysis of grammar patterns within the script, shedding light on previously unrecognized structures. This innovative approach sets the study apart, contributing to the advancement of decipherment techniques for the Indus script.

2.2 Clustering Analysis

The exploration of comparative analysis in decipherment not only unlocks linguistic mysteries but also unveils the interconnectedness of ancient civilizations. By investigating similarities and differences between scripts, researchers can unravel the shared heritage and cultural exchanges that have shaped human societies throughout millennia. Shu-Ming Hsieh and Chiun-Chieh Hsu [4] introduced a graph-based model named SRG (spatial relation graph) designed for image retrieval, revolutionizing the field by offering a novel means to represent the semantic information of image objects and their spatial relationships without the need for file annotation. Through meticulously crafted experiments utilizing synthetic symbolic image databases and existing datasets, the authors demonstrate the efficacy of their approach, contributing significantly to the advancement of image retrieval techniques, paving the way for enhanced semantic understanding and spatial analysis in image processing and related fields. Building upon this foundation, Guanglin Huang, Wan Zhang, and Liu Wenyin [5] introduced a discriminative representation for symbolic image similarity evaluation, addressing the fundamental challenge of visual similarity evaluation in intelligent graphics systems. Focusing specifically on symbolic image recognition, the authors proposed the Directional Division Tree representation as the cornerstone of their algorithm, contributing significantly to the advancement of intelligent graphics systems, offering promising avenues for enhanced image recognition and retrieval capabilities in various applications.

Extending the exploration of comparative analysis into linguistics, Sarat Sasank Barla, Sai Surya Sanjay Alamuru, and Peter Zsolt Revesz [6] delve into the intriguing question of the similarity between the Indus Valley script and scripts utilized for writing Dravidian languages such as Kannada, Malayalam, Tamil, and Telugu. Their research, published in the Proceedings of the 26th International Database Engineered Applications Symposium (IDEAS '22), embarks on a comprehensive feature analysis of each sign present in these scripts. By leveraging this analysis, they generate similarity matrices that quantitatively measure the likeness between any pair of signs across different scripts, shedding light on potential linguistic and cultural connections between the ancient Indus Valley script and contemporary Dravidian languages, offering valuable insights into the evolution and interplay of writing systems in South Asia.

However, a notable gap in previous research lies in the absence of systematic Clustering methodologies to identify patterns and associations within the Indus script. This research introduces a pioneering approach by employing four cluster algorithms: k-means, Agglomerative Clustering, BIRCH, and Spectral Clustering. It represents the first application of clustering techniques to the Indus script, aiming to uncover underlying structures and relationships among its symbols. This innovative methodology offers a fresh perspective on decipherment techniques for ancient scripts. By leveraging Clustering algorithms, this study seeks to identify coherent groups of symbols and explore their potential linguistic and cultural significance, contributing to the ongoing efforts to unravel the mysteries of the Indus script.

3. 3. METHODOLOGY

The decipherment of the Indus script requires a meticulous and multi-faceted approach that combines traditional archaeological methods with cutting-edge computational techniques. This study outlines the methodologies utilized, focusing on frequency and clustering analyses. Frequency analysis examines the occurrence and distribution of symbols within the script, identifying patterns and commonalities in symbol usage to make inferences about the structure and potential meanings of the script, such as common words, grammatical markers, or phonetic elements. Clustering analysis, a statistical method, groups symbols or sequences of symbols exhibiting similar characteristics, helping to identify possible relationships between symbols and their contextual usage, such as identifying symbols that may represent similar sounds or meanings. Together, these methodologies offer complementary insights into the structural and semantic properties of the Indus script, providing a broad overview of symbol usage and distribution while delving deeper into the relationships and patterns within the script. The integration of frequency and clustering analyses, alongside traditional archaeological methods, provides a robust framework for deciphering the Indus script, enhancing the understanding of its structure and semantics and opening new avenues for exploring the linguistic and cultural heritage of the Indus Valley Civilization.

3.1 System Architecture

Integrating Frequency analysis and Clustering techniques offers a comprehensive understanding of the Indus script's complexity, leveraging computational tools for decipherment.

3.1.1 Frequency Analysis

Symbol segmentation is a crucial initial step in the analysis of symbol sequences from the ICIT Indus Corpus. This process involves separating individual symbols from the corpus text and assigning them unique identifiers based on their position within the sequence. To achieve this, a segmentation algorithm designed specifically for the characteristics of the Indus script was employed. Despite the script's complexity, the methodology successfully segmented symbols with high accuracy, enabling precise analysis of symbol sequences. Symbol segmentation is foundational for subsequent analyses, as it ensures that each symbol is distinctly recognized and accurately placed within the overall sequence. By meticulously identifying and cataloging each symbol, researchers can more effectively conduct frequency and clustering analyses, leading to deeper insights into the structure and meaning of the Indus script. This precise segmentation is essential for building a comprehensive understanding of the script's linguistic and cultural context, ultimately contributing to the broader goal of decipherment and interpretation.

The methodology for symbol segmentation primarily relied on pattern recognition techniques tailored to the distinctive features of the Indus script. This involved parsing the corpus text and identifying recurring patterns indicative of individual symbols. Challenges encountered during segmentation included ambiguous or overlapping symbols, which required careful consideration and manual verification to ensure accuracy. By employing a combination of automated processes and manual oversight, and through iterative refinement and validation, reliable segmentation results were achieved. This precise symbol segmentation lays the groundwork for subsequent analysis, such as frequency and clustering analyses, enabling a deeper understanding of the structural and semantic properties of the Indus script.

The results present the segmented symbols from the ICIT Indus Corpus along with their corresponding index numbers. Each symbol is uniquely identified, allowing for easy reference and analysis in subsequent stages of the project. An example of the segmented symbols demonstrates the effectiveness of the methodology in accurately capturing the structure of symbol sequences from the corpus. N-gram analysis offers valuable insights into the frequency and patterns of symbol sequences within the ICIT Indus Corpus. By examining sequences of symbols of varying lengths (n-grams), recurring patterns can be identified, and their significance assessed in the context of the script. The methodology involved analyzing n-grams ranging from single symbols (1-gram) to longer sequences of up to 15 symbols. This comprehensive approach enables a detailed understanding of the script's structure, revealing important linguistic and cultural patterns that contribute to the ongoing efforts to decipher the Indus script.

The n-gram analysis methodology employed in this study encompasses several key steps. First, symbol sequences of varying lengths were extracted from the corpus, and their frequency of occurrence was calculated. Next, the n-grams were sorted based on their frequency counts, prioritizing those with higher frequencies for further analysis. This approach allowed for the identification of prevalent patterns within the corpus and the exploration of their potential meanings or functions. By focusing on the most frequent n-grams, the analysis could pinpoint significant recurring sequences that might indicate common words, grammatical structures, or other meaningful elements in the script. This systematic method of n-gram analysis provides a robust framework for uncovering the underlying structure and semantics of the Indus script, contributing to the broader goal of its decipherment and understanding.

Through detailed examination and interpretation of the n-gram analysis findings, a deeper understanding of the linguistic and semantic aspects of the Indus script is achieved. The analysis results illustrate the distribution of symbol frequencies and patterns within the corpus, facilitating further examination and interpretation. By closely studying these figures, researchers can identify significant trends and anomalies, providing clearer insights into the structure and usage of the symbols. This approach highlights the most common symbol sequences and sheds light on their possible functions and meanings within the script, advancing the overall effort to decipher and understand the Indus script’s complex linguistic and cultural dimensions.

Grammar analysis in the study of the Indus script employs several key methodological approaches, including frequency analysis, pattern recognition, and contextual analysis. Frequency analysis involves examining the occurrence rates of various linguistic elements, such as individual signs or combinations of signs, to identify usage patterns that may reveal grammatical rules or structural norms. Pattern recognition focuses on detecting recurring structures or motifs within the inscriptions, aiming to uncover consistent sequences that might indicate grammatical functions or syntactical conventions. Contextual analysis examines the surrounding context of linguistic elements to understand their specific function or meaning within the text. By systematically applying these methods, researchers can uncover underlying grammatical patterns and structures within the Indus script, advancing the decipherment and interpretation of its linguistic and cultural dimensions.

Filling missing symbols in symbol sequences is essential for reconstructing complete and coherent texts from fragmented or damaged inscriptions. This process involves predicting the most likely values for missing symbols based on the surrounding context, which is crucial for making sense of incomplete or partially preserved texts. The approach to filling missing symbols leverages the results of n-gram analysis, which identifies patterns and sequences of symbols within the corpus. By analyzing these patterns, the algorithm developed for this purpose can systematically predict missing symbols. This prediction is based on the frequency and context of symbol sequences, ensuring that the interpolated symbols fit logically and cohesively within the existing text. The systematic process involves detecting recurring patterns and applying them to infer the most probable missing symbols, thereby reconstructing the text with a high degree of accuracy. This method not only helps in piecing together fragmented inscriptions but also enhances the overall interpretation and understanding of the script, contributing to the broader effort of deciphering and analyzing the Indus script.

The implementation of the missing symbol prediction algorithm involved developing a user-friendly web page interface that facilitates easy interaction and use. Users can input symbol sequences with missing values, and the algorithm automatically predicts and fills in the missing symbols based on the analysis of the surrounding context. The interface is designed to be intuitive, allowing users to quickly input data and receive predictions. The algorithm leverages the results of n-gram analysis to identify patterns and predict the most likely values for the missing symbols. By iteratively applying the algorithm, users can gradually reconstruct the entire sequence, effectively minimizing the impact of missing or damaged portions. This iterative process ensures that each predicted symbol is contextually appropriate, enhancing the coherence and completeness of the reconstructed text. The web interface not only streamlines the process of filling in missing symbols but also makes the advanced computational techniques accessible to researchers and enthusiasts alike, contributing significantly to the decipherment and understanding of the Indus script.

The results have been verified using a next-symbol predicting LSTM (Long Short-Term Memory) model, which was trained on the segmented symbol sequences to predict the next symbol based on the preceding symbols. The LSTM model's ability to handle sequences and capture long-term dependencies makes it well-suited for this task. By validating the results from the segmentation algorithm and n-gram analysis with the LSTM model, the robustness and accuracy of the findings are ensured. This validation process involves comparing the LSTM model's predictions with the actual symbols, allowing for the identification and correction of discrepancies. The LSTM model serves as an additional verification step, enhancing the reliability of the analytical approach and providing further confidence in the interpretation of the linguistic and semantic aspects of the Indus script. The integration of traditional analytical methods with advanced machine learning techniques like the LSTM model offers a robust framework for exploring the script's complex linguistic and cultural dimensions.

3.1.2 Clustering Analysis

The Clustering analysis framework integrates evolutionary analysis, clustering, and similarity comparison to elucidate the complex dynamics of the Indus script symbols. At its core, the architecture is designed to uncover patterns of symbol evolution, identify distinct symbol clusters, and assess their similarity with other ancient scripts, such as Tamil Brahmi. The evolutionary analysis component traces the development and transformations of symbols over time, revealing insights into how the script evolved. Clustering identifies groups of symbols that share common features, helping to classify and organize the symbols based on their structural or functional characteristics. Similarity comparison assesses the resemblance between Indus script symbols and those from other ancient scripts, providing context for possible linguistic and cultural connections. Each stage of the architecture is meticulously crafted to ensure a comprehensive understanding of the script's historical context and linguistic connections. This integrated approach not only clarifies the structure and usage of the Indus script but also contributes to broader efforts in deciphering and understanding ancient scripts.

The evolutionary analysis phase begins with a detailed comparison of keypoints and descriptors for each image in the dataset, focusing on identifying distinctive features to track changes in symbol morphology. The cv2.BFMatcher() algorithm matches descriptors between image pairs to identify similarities, with a ratio test applied to filter out incorrect matches and ensure accuracy. Once filtered, similar images are identified based on the number of good matches exceeding a minimum threshold. These similar images serve as key indicators of evolutionary relationships between symbols, revealing how symbols have changed over time and contributing to a deeper understanding of the script's historical and linguistic development.

In the Clustering analysis stage, a variety of clustering algorithms, including K-means, Agglomerative, Birch, and Spectral Clustering, are employed to group Indus script symbols based on the similarities identified during the evolutionary analysis phase. These algorithms help organize symbols into clusters by evaluating their structural and functional characteristics. To facilitate accurate clustering, feature extraction techniques such as VGG16 are utilized to extract relevant features from the symbol images. This deep learning model processes the images to identify and extract key features that are crucial for distinguishing between different symbols. Symbols are then clustered into groups according to these extracted features, enabling the identification of distinct clusters that represent various symbol categories or evolutionary stages.

The similarity comparison stage further extends the analysis by assessing how Indus script symbols compare to those from other ancient scripts, such as Tamil Brahmi. Structural similarity comparison is conducted using metrics like the Structural Similarity Index (SSIM), which quantifies similarities in brightness, contrast, and overall structure between symbols. By analyzing these structural similarities, researchers can identify shared symbols or potential linguistic and cultural connections between the Indus script and other ancient scripts. This comparative analysis provides valuable insights into historical interactions, cultural exchanges, and potential influences between civilizations, enhancing the understanding of the Indus script’s context within the broader spectrum of ancient writing systems.

This methodology enables a comprehensive exploration of the Indus script's evolution, clustering patterns, and similarities with other ancient scripts, leading to a deeper understanding of the script's significance and cultural context. By integrating evolutionary analysis, various clustering algorithms, and similarity comparisons, the approach offers a detailed view of how symbols have developed over time, how they can be grouped into distinct categories or stages, and how they relate to other ancient scripts like Tamil Brahmi. This thorough analysis enhances the understanding of the script's structure and usage, providing valuable insights into its historical and cultural role and supporting its decipherment within a broader context.

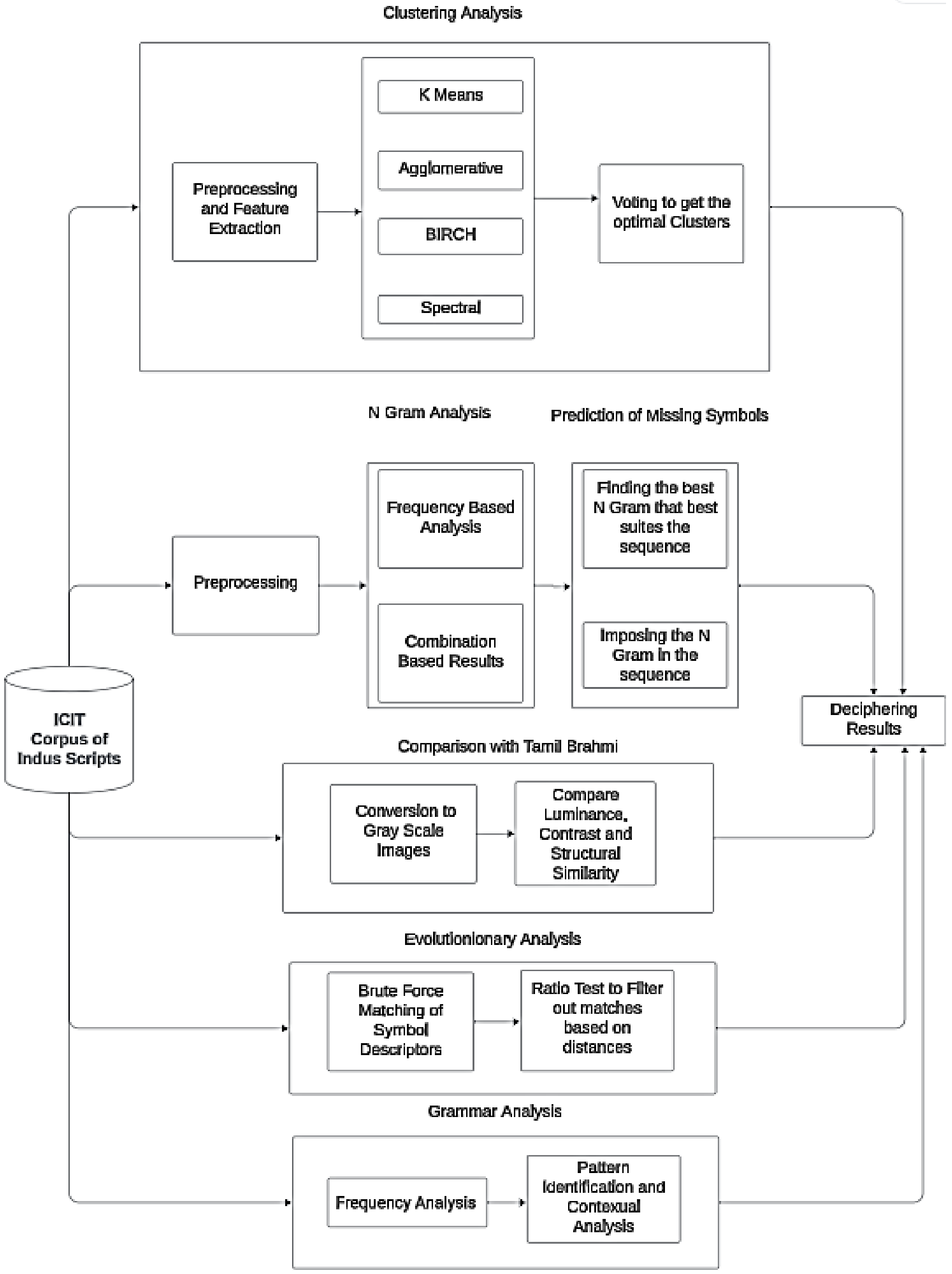

Clustering Analysis
Preprocessing and Feature Extraction
K Means
Agglomerative
BIRCH
Spectral
Voting to get the optimal Clusters
N Gram Analysis
Prediction of Missing Symbols
Preprocessing
Frequency Based Analysis
Combination Based Results
Finding the best N Gram that best suites the sequence
Imposing the N Gram in the sequence
ICIT Corpus of Indus Scripts
Deciphering Results
Comparison with Tamil Brahmi
Conversion to Gray Scale Images
Compare Luminance, Contrast and Structural Similarity
Evolutionionary Analysis
Brute Force Matching of Symbol Descriptors
Ratio Test to Filter out matches based on distances
Grammar Analysis
Frequency Analysis
Pattern Identification and Contexual Analysis

Figure 1: Overall System Architecture

3.2 Algorithms

3.2.1 Algorithm of Frequency Analysis

1. Initialization:

1.1 Initialize the dataset containing symbol sequences extracted from the ICIT Corpus.

1.2 Set parameters such as the minimum support threshold and minimum confidence threshold for linguistic analysis.

1.3 Preprocess the dataset to prepare it for symbol segmentation, n-gram analysis, and missing value prediction.

1.3.1 Clean the dataset to remove any irrelevant or noisy data.

1.3.2 Tokenize the symbol sequences into individual symbols for further analysis

1.3.3 Handle any missing or incomplete data to ensure completeness.

2. Symbol Segmentation:

2.1 Develop an algorithm to segment symbols from the ICIT Corpus, accounting for the script's unique characteristics

2.1.1 Research existing segmentation techniques and adapt them to the Indus script.

2.1.2 Design specific rules for segmenting symbols based on visual and contextual features.

2.2 Implement symbol segmentation by iteratively parsing the dataset and identifying recurring patterns indicative of individual symbols.

3. N-Gram Analysis:

3.1 Design a methodology for n-gram analysis to identify patterns and frequencies of symbol sequences within the segmented dataset.

3.1.1 Determine the range of n-gram sizes to be considered (eg, 1-gram to 15-gram).

3.2 Implement n-gram analysis by extracting sequences of varying lengths and calculating their frequency of occurrence.

3.2.1 Extract n-grams from the segmented symbol sequences using sliding windows.

3.2.2 Count the frequency of each n-gram and store the results for further analysis.

3.3 Sort n-grams based on frequency counts and generate top-N lists for 1-gram and bi-gram occurrences.

3.3.1 Identify the top occurring 1-grams and bi-grams to understand the most frequent symbol sequences.

3.3.2 Visualize the frequency distribution of n-grams to identify patterns and anomalies.

4. Grammar Analysis:

4.1 Frequency Analysis:

4.1.1 Quantify the occurrences of linguistic elements to assess their significance.

4.1.2 Determine the frequency distribution to identify usage patterns.

4.2 Pattern Recognition:

4.2.1 Utilize algorithms to detect recurring motifs or structures.

4.2.2 Employ clustering and sequence analysis for pattern detection.

4.3 Contextual Analysis:

4.3.1 Scrutinize surrounding signs and layout to infer roles.

4.3.2 Account for spatial and temporal variations influencing interpretation.

4.4 Interpretation:

4.4.1 Analyze patterns within linguistic frameworks to deduce meanings.

4.5 Validation

4.5.1 Cross-validate with findings from previous research.

5. Filling Missing Values:

5.1 Develop an algorithm to predict missing symbols within symbol sequences based on surrounding context and n-gram analysis results.

5.1.1 Define rules or heuristics for inferring missing symbols based on neighboring symbols and frequent patterns.

5.1.2 Implement the algorithm to iteratively fill in missing symbols until no further predictions can be made.

5.2 Evaluate the accuracy of missing value prediction through comparison with known symbol sequences and manual validation.

5.2.1 Assess the accuracy of filled sequences by comparing them with ground truth data.

6. Output and Visualization:

6.1 Output the segmented symbols, n-gram analysis results, and filled symbol sequences for further linguistic analysis.

6.1.1 Store the processed data in a suitable format for easy access and analysis.

7. Interpretation and Analysis:

7.1 Interpret the segmented symbols and n-gram analysis results to discern linguistic patterns and structural characteristics of the Indus script.

7.1.1 Analyze the distribution of symbols and their co-occurrence patterns to infer linguistic features.

7.2 Analyze the filled symbol sequences to infer missing symbols and reconstruct complete texts for linguistic study. 7.2.1 Examine the filed sequences to identify common sequences and transitions between symbols.

7.3 Evaluate the effectiveness of the linguistic analysis methodology in deciphering and understanding the Indus script.

3.2.2 Algorithm of Clustering Analysis

1. Data Acquisition and Preprocessing:

1.1. Obtain Symbols from the ICIT Corpus.

1.2. Preprocess the data by cleaning and normalizing images, ensuring consistency and uniformity.

1.2.1. Remove noise, artifacts, and background clutter from the images.

1.3. Extract features from the images using VGG16 to represent symbols in a umerical format

1.3.1. Tokenize the symbols and convert them into numerical vectors using embedding techniques.

2. Evolutionary Analysis:

2.1. Implement keypoint and descriptor comparison for each image to identify distinctive features.

2.1.1. Use feature detection algorithms like SIFT or SURF to identify keypoints.

2.2. Utilize a brute-force matcher (eg, cv2 BFMatcher()) to match descriptors between pairs of images.

2.2.1. Compute descriptors (eg, SIFT, ORB) for each keypoint in the images.

2.3. Apply a ratio test to filter out incorrect matches, retaining only high-confidence matches.

2.3.1. Set a threshold ratio (eg.,0.7) to accept matches based on the ratio of distances.

2.4. Determine similarity between images based on the number of good matches exceeding a predefined threshold.

2.4.1. Establish a minimum number of matches (eg. 10) required for images to be considered similar

3. Clustering Analysis:

3.1. Select 4 Clustering Algorithms – K-Means, Agglomerative, BIRCH and Spectral Clustering.

3.1.1. Assess dataset characteristics such as size, dimensionality, and density distribution.

3.2. Perform feature extraction using techniques such as VGG16 to enhance clustering accuracy.

3.2.1. Utilize pre-trained convolutional neural networks (CNNs) to extract high-level features from symbol images.

3.3. Cluster Indus script symbols into groups based on similarities identified during evolutionary analysis.

3.3.1. Explore different clustering

parameters and initialization strategies to optimize cluster formation

3.4. Evaluate clustering performance using metrics like Jaccard index.

3.4.1. Quantitatively assess cluster quality and separation using established clustering metrics.

3.2.3 Algorithm for Similarity Comparison with Tamil Brahmi

1. SimilarityComparison:

1.1. Conduct structural similarity comparison with other ancient scripts (eg. Tamil Brahmi) using SSIM.

1.1.1. Compute SSIM scores between pairs of images to quantify structural similarity.

1.2. Implement algorithms to quantify similarities in brightness, contrast, and structure between symbols.

1.2.1. Define algorithms to compute similarity metrics based on pixel intensity and spatial amangement

1.3. Identify shared symbols or linguistic cultural connections based on structural similarities.

1.3.1. Analyze clusters of similar symbols to identify patterns and relationships with other scripts.

4. RESULTS AND INTERPRETATION

The Interactive Corpus of Indus Texts (ICIT) stands as a pivotal asset for researchers endeavoring to decode the mysterious Indus script. Crafted by Bryan Wells and Andreas Fuls, this digital repository offers an expansive and readily accessible compilation of Indus inscriptions. This centralized platform facilitates global collaboration among scholars, enabling the exchange of data, cross-referencing of interpretations, and collective efforts in unraveling the enigmas concealed within the Indus script. The corpus encompasses data on the symbols found in 1125 seals and 694 distinct seals identified within their comprehensive database.

4.1 Dataset

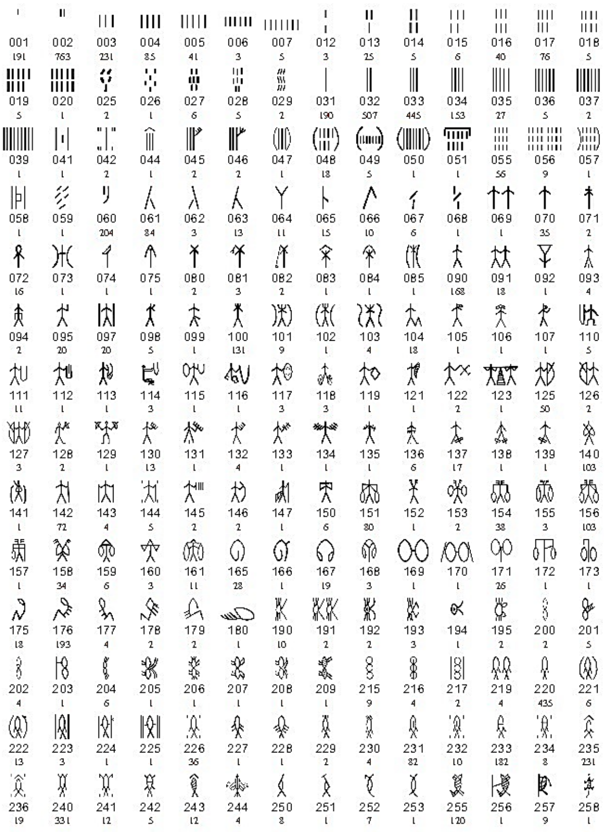
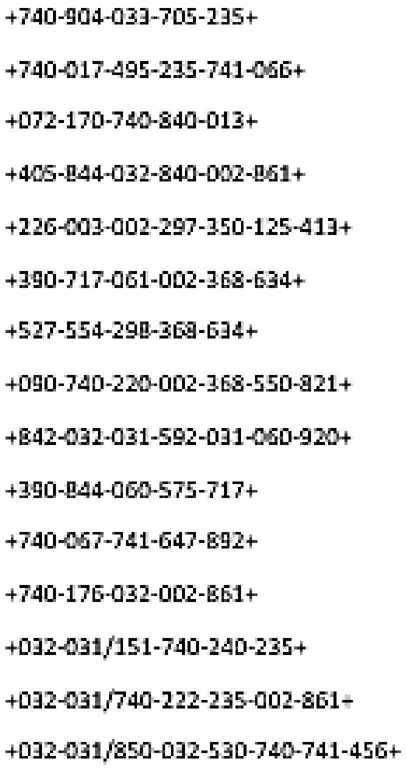


Figure 2: Sample Symbols present in the corpus


+740-904-033-705-235+

+740-017-495-235-741-066+

+072-170-740-840-013+

+405-844-032-840-002-861+

+226-003-002-297-350-125-413+

+390-717-061-002-368-634+

+527-554-298-368-634+

+090-740-220-002-368-550-821+

+842-032-031-592-031-060-920+

+390-844-060-575-717+

+740-067-741-647-892+

+740-176-032-002-861+

+032-031/151-740-240-235+

+032-031/740-222-235-002-861+

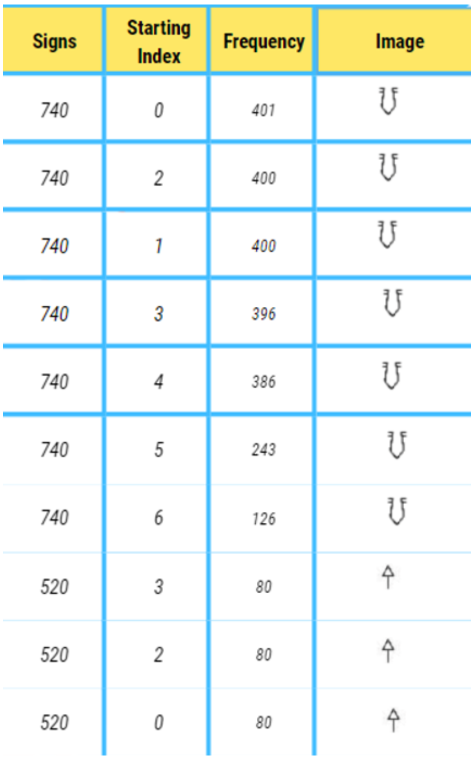
+032-031/850-032-530-740-741-456+

Figure 3: Sample single line sequences present in the seals

4.2 Frequency Analysis

Through an exhaustive exploration employing N-gram analysis, recurring symbol patterns within the Indus script have been unveiled. The investigation spans a wide range, from 1-grams to 15-grams, meticulously sorting these patterns based on their frequency within the corpus. This approach presents an unparalleled depth of insight into the script's linguistic structures and potential semantic meanings, revealing important patterns and trends that contribute to a better understanding of the script's complexity and significance.

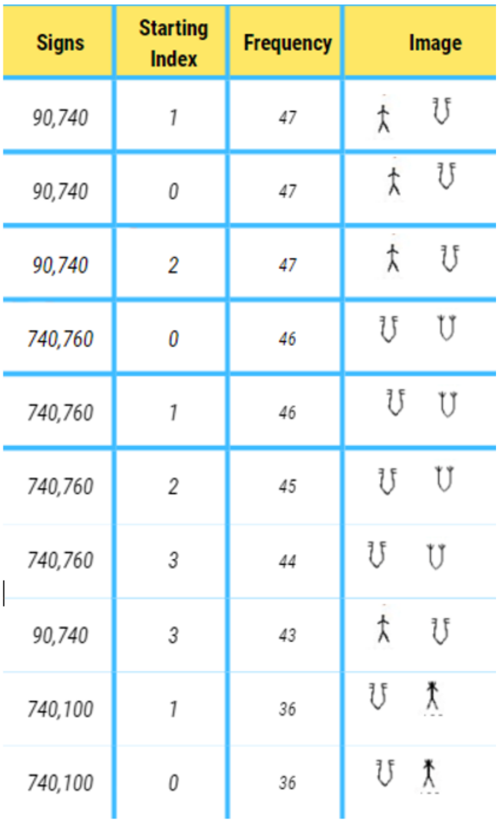
In Figure 4.3, the top 10 occurring 1-grams are showcased, providing an intricate portrayal of the individual symbols' frequency within the script. This nuanced analysis illuminates the prevalence of specific symbols and offers invaluable guidance for unraveling and interpreting the complexities of the Indus script. By highlighting the most common symbols, the analysis contributes to a deeper understanding of their usage and significance, aiding in the broader effort to decipher this ancient writing system.



Signs	Starting Index	Frequency	Image
740	0	401	
740	2	400	
740	1	400	
740	3	396	
740	4	386	
740	5	243	
740	6	126	
520	3	80	
520	2	80	
520	0	80	

Figure 4: Top 10 unigrams

Similarly, in Figure 4.4, the top 10 occurring bi-grams are explored, uncovering pairs of symbols that frequently co-occur within the corpus. These revelations shed light on recurring symbol combinations and hint at underlying patterns of association, unlocking profound insights into the script's linguistic and semantic intricacies. By identifying these frequent pairs, the analysis enhances the understanding of how symbols interact within the script, providing key information for interpreting its structure and meaning.



Signs	Starting Index	Frequency	Image
90,740	1	47	
90,740	0	47	
90,740	2	47	
740,760	0	46	
740,760	1	46	
740,760	2	45	
740,760	3	44	
90,740	3	43	
740,100	1	36	
740,100	0	36	

Figure 5: Top 10 bigrams

Table 1: Symbols frequently used in last position

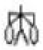
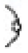
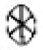
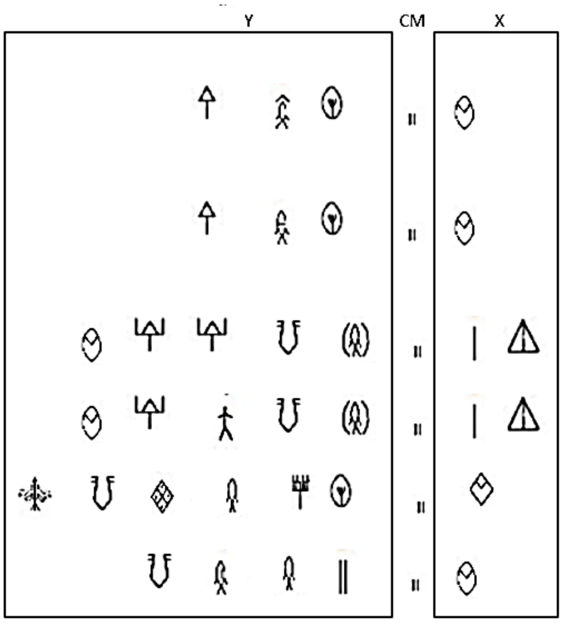
Symbol	Symbol Image	Frequency
740		401
520		80
90		49
400		48
390		30
527		24
700		21
151		20
407		14

Table 2: Symbols used in 10 or more unique positions

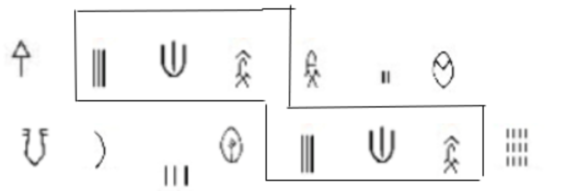
Symbol	Symbol Image
('032',)	
('002',)	
('861',)	
('060',)	
('920',)	
('820',)	

4.2.1 Grammar Analysis

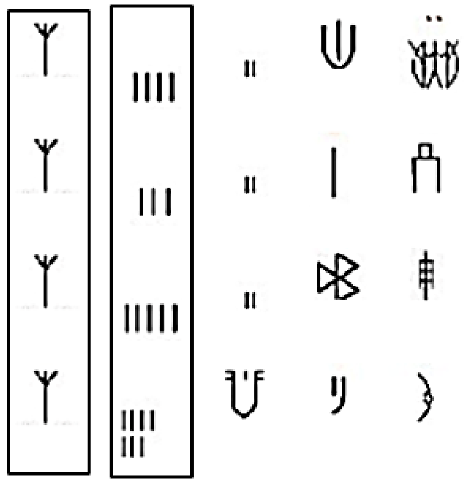
The analysis revealed recurring patterns of Y connective morpheme X, indicating a systematic syntactic structure within the corpus. This finding aligns with previous research, further validating the identified grammatical patterns.



The examination uncovered various sub-sequences serving as pre-connective and post-connective elements, suggesting a structured syntactic framework within the Indus script. This observation highlights the intricate grammatical organization of the script, where specific sequences fulfill distinct syntactic roles, contributing to a deeper understanding of its linguistic characteristics.

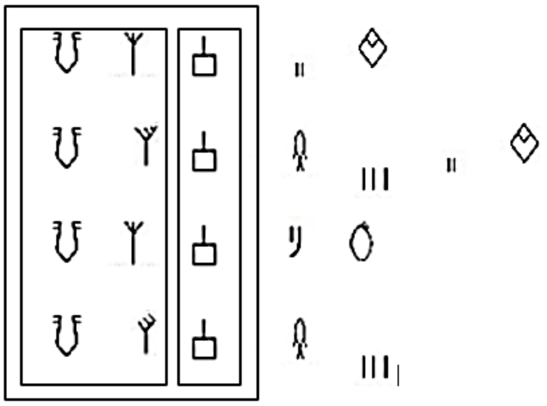


When the crop symbol occupies the terminal position within inscriptions, it consistently precedes stroke symbols. This consistent pattern indicates a systematic syntactic relationship between the crop symbol and stroke symbols, suggesting a deliberate linguistic convention within the Indus script.

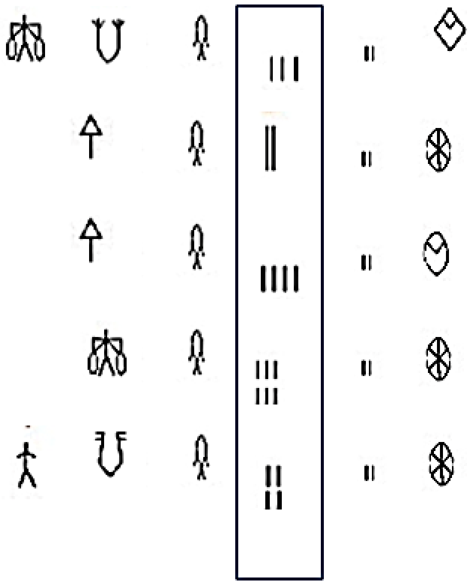


CROP +NUMERALS

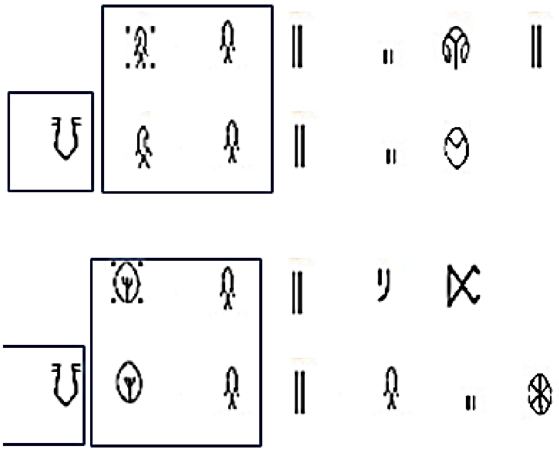
When the jar symbol follows a crop symbol within Indus script inscriptions, it's frequently observed that the preceding symbol is consistently



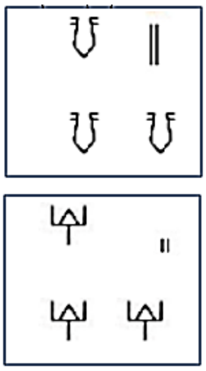
The stroke sign shares a common inscriptional context with other stroke numerals within the Indus script. This observation indicates a contextual similarity or shared usage between the stroke sign and other numerical symbols characterized by strokes.



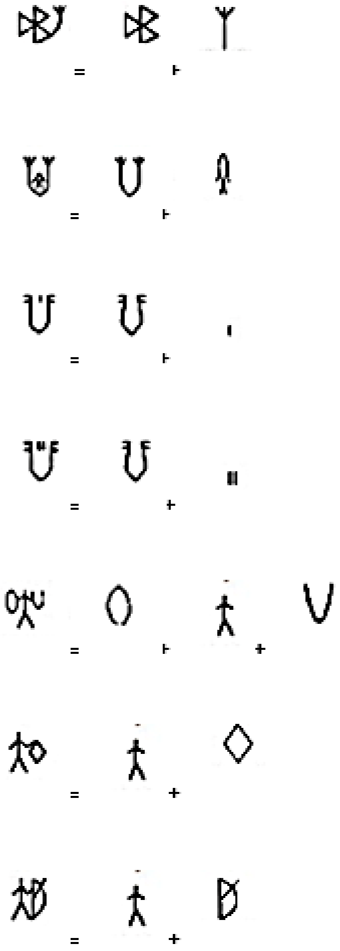
Encapsulated signs serve as replacements for basic signs and symbols, primarily appearing at terminal positions within the Indus script. This phenomenon suggests a functional role for encapsulated signs in concluding or summarizing inscriptions.



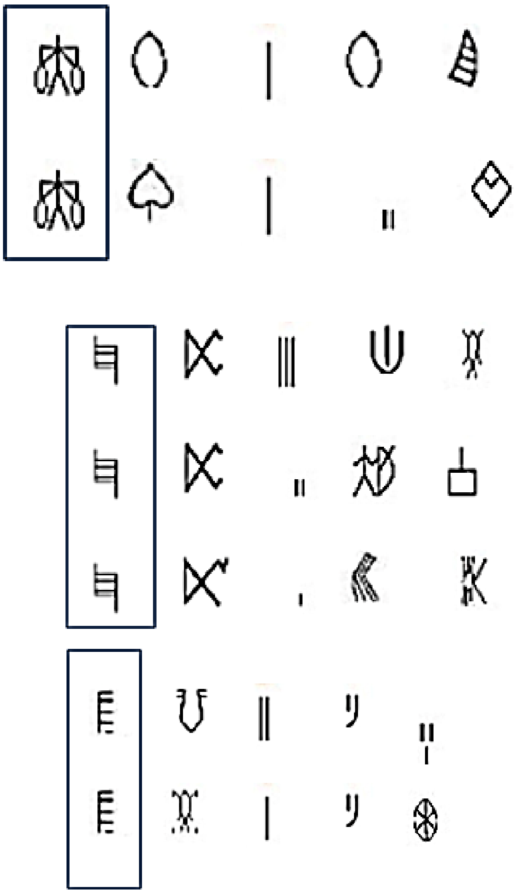
When a lexeme is preceded by a numerical symbol in one sequence, indicating quantity or repetition, it often occurs precisely that number of times in another sequence. This consistent correlation between numerical symbols and the subsequent repetition of lexemes suggests a systematic pattern in the Indus script, where numerical symbols serve as quantifiers or multipliers for subsequent linguistic elements.



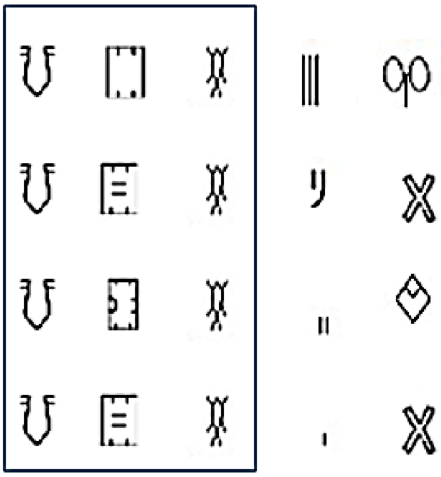
Individual signs found within composite signs often retain their original meaning or contribute to the semantic composition of the composite sign. This observation suggests that composite signs in the Indus script may possess a hierarchical structure, where constituent signs maintain their individual significance while also contributing to the overall meaning of the composite sign.



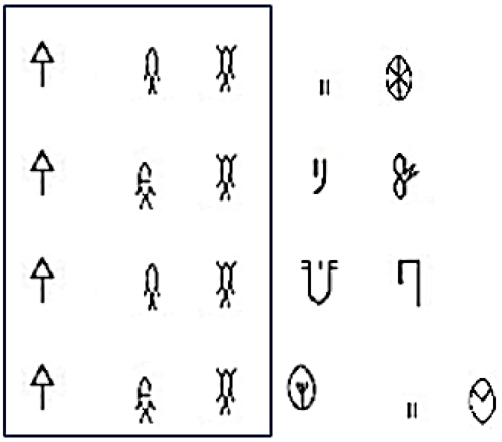
Symbols appearing in the final position without any connective signs may have standalone meanings. This observation suggests that certain symbols in the Indus script can function independently to convey specific concepts or ideas. By analyzing the contextual usage of these symbols and examining their occurrences in various inscriptions, researchers can infer their individual meanings and semantic significance.



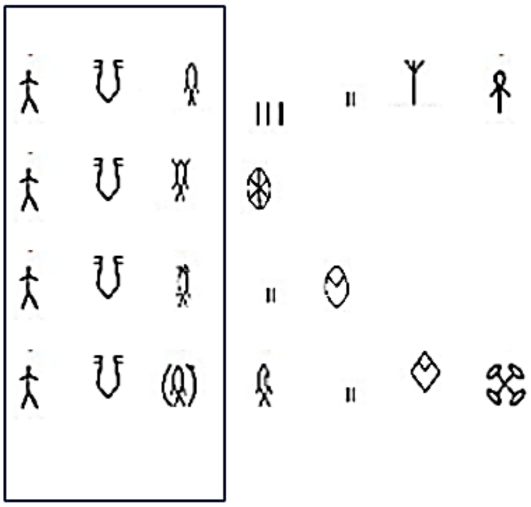
When a rectangle symbol is followed by a jar symbol in the Indus script, it frequently appears that the preceding symbol is a fish symbol. This consistent pattern suggests a potential syntactic or semantic relationship between the rectangle, fish, and jar symbols.



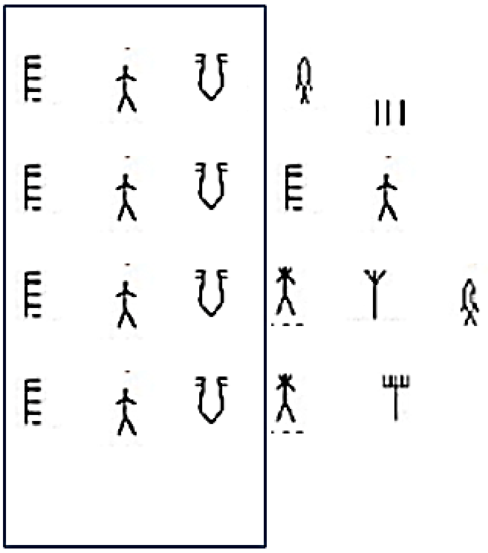
In the Indus script, a noteworthy pattern emerges where an arrow symbol, when preceded by a fish symbol, is often further preceded by another fish symbol. This recurring sequence implies a consistent syntactic or semantic association between these symbols within the script.

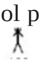


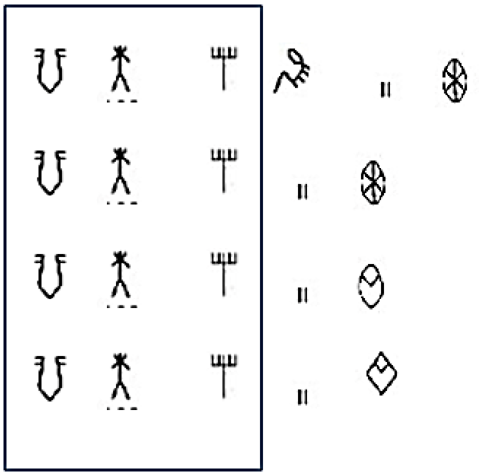
In the Indus script, a notable pattern is observed where the depiction of a man symbol, when preceded by a jar symbol, frequently occurs alongside a preceding fish symbol. This consistent sequence suggests a systematic relationship or semantic connection between these symbols within the script's context.



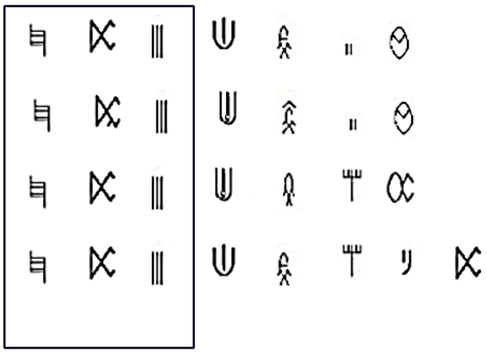
The consistent occurrence of a jar symbol preceding a man symbol, followed by another symbol , suggests a systematic pattern within the Indus script. This pattern implies a structured syntactic or semantic relationship between these symbols, potentially indicating a grammatical or contextual rule governing their arrangement.



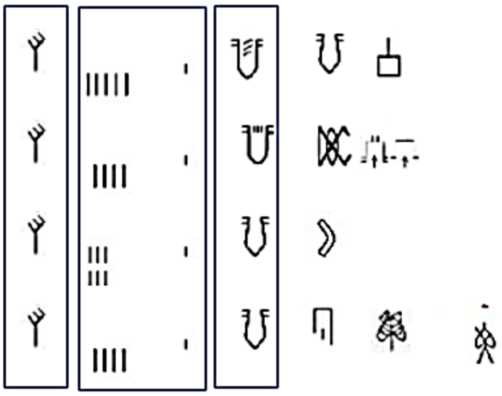
The consistent occurrence of a symbol preceding a jar symbol, followed by another symbol  , suggests a systematic pattern within the Indus script. This pattern implies a structured syntactic or semantic relationship between these symbols, potentially indicating a grammatical or contextual rule governing their arrangement.



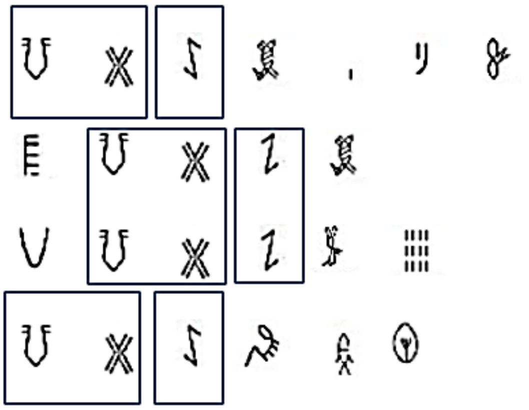
The occurrence of a numeral sign at the third position from the last, followed by symbols possibly representing time at the second position from the last, when another symbol like appears in the last position, suggests a structured pattern within the Indus script. This arrangement implies a systematic relationship between numerical, temporal, and other symbolic elements, hinting at a grammatical or contextual rule guiding their arrangement.



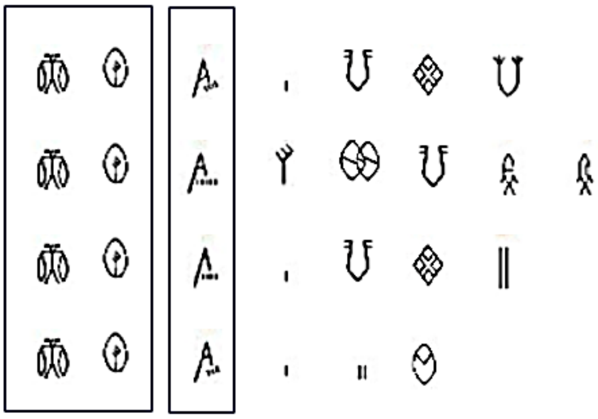
The consistent occurrence of following two stroke symbols, especially when preceded by jar symbols, suggests a structured pattern within the Indus script. This recurring motif implies a potential grammatical or contextual association between stroke symbols, jar symbols, and the subsequent appearance of .



When appears after the symbol , it frequently follows with the usage of the connecting symbol or to join it with the subsequent elements in the sequence. This consistent pattern of employing or after A and before suggests a structural rule within the Indus script, indicating a grammatical or syntactical relationship among these symbols.

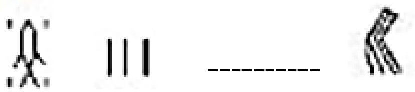


When is preceded by the symbol , it is commonly preceded by symbols resembling A in form or function. This consistent pattern suggests a structural relationship between and the subsequent occurrence of A-like symbols before in the Indus script.

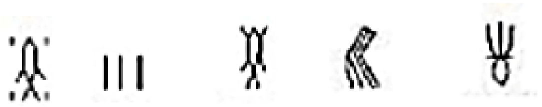


4.2.2 Prediction of Missing Values using N-Grams

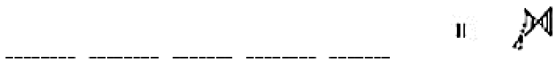
Partial Sequence :



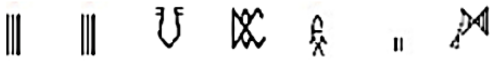
Filled Sequence:



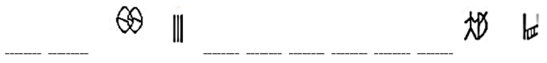
Partial Sequence:



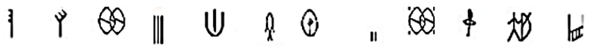
Filled Sequence:



Partial Sequence:



Filled Sequence:



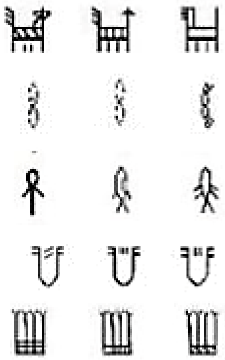
To fill missing symbols within symbol sequences, an algorithm was devised based on contextual information from adjacent symbols and insights derived from n-gram analysis. This algorithm systematically traversed the sequence, identifying missing symbols and predicting their likely symbols by examining neighboring symbols and frequent symbol patterns observed in the dataset. It iteratively inferred missing symbols until no further predictions could be made, ensuring a comprehensive filling of missing symbols.

The efficacy of this approach was validated through comparisons with known symbol sequences and manual inspection. For a visual representation of the filled symbol sequences, please refer to the associated figure. The algorithm operates as follows: First, it locates the first non-null value in the sequence to be filled. Then, it searches for that symbol in the corresponding position in the n-gram analysis results and stores all possible n-grams that can be used to fill the missing symbol. Each possible n-gram is then applied to the sequence to be filled one by one.

Once the first n-gram is filled, the updated sequence is used to find the next non-null value, starting the search from the maximum of the current index + 1 or the index to which it is filled. This process continues until it is no longer possible to fill, utilizing a backtracking approach. This process is repeated for the second possible n-gram and continues until all possibilities of the missing values are exhausted. This algorithm ensures thorough exploration of potential symbol combinations, offering a comprehensive approach to filling missing symbols in the sequence. Additionally, the results were verified using a next symbol predicting LSTM model, further validating the accuracy and reliability of the filled symbol sequences.

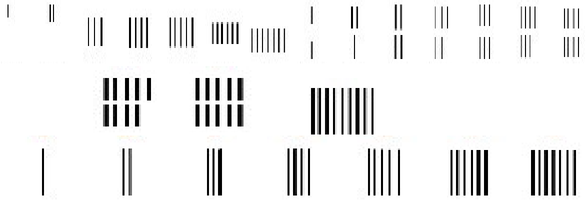
4.3 Clustering Analysis

Moreover, the analysis identified primary, secondary, and tertiary symbols within the script corpus, shedding light on their hierarchical semantic structure. Symbols categorized as primary likely hold fundamental meanings, while secondary and tertiary symbols may represent variations or elaborations. This hierarchical classification enhances the understanding of the script's semantic organization and usage patterns.

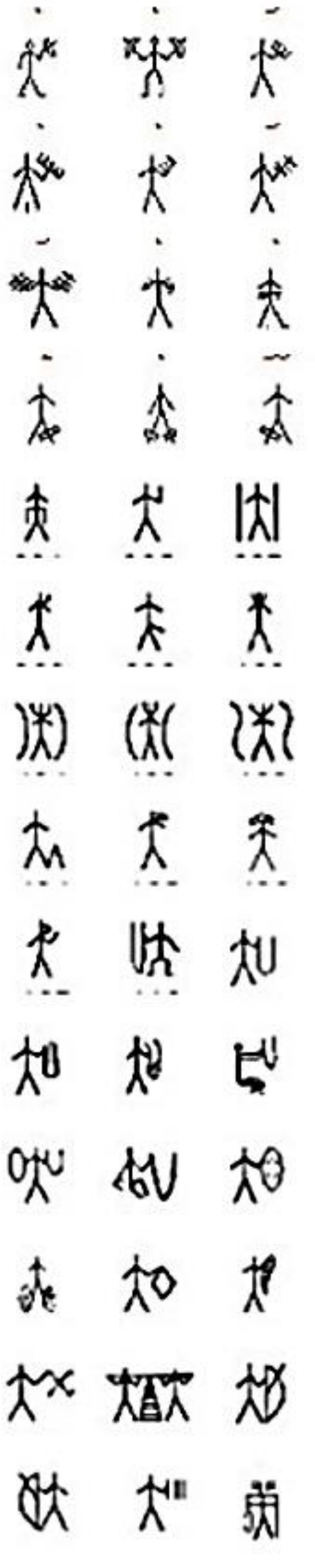


Furthermore, symbols clustered together displayed notable contextual coherence, indicating shared semantic or functional significance. This Clustering analysis offers insights into the script's contextual usage, suggesting thematic or functional associations among clustered symbols. Understanding these contextual nuances enriches the interpretation of the script's societal, religious, or administrative contexts.

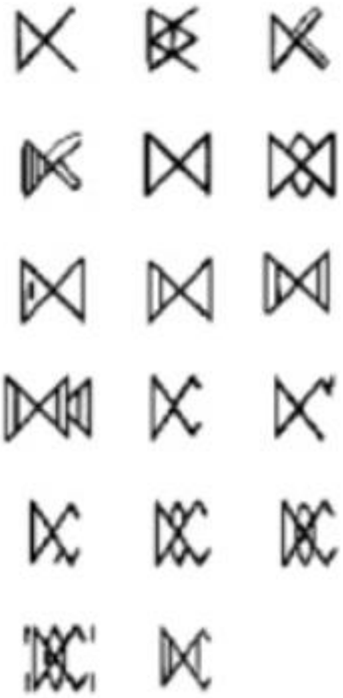
Additionally, it's noteworthy that symbols potentially representing numbers consistently fell into the same cluster across all Clustering methods employed. This observation suggests a consistent numerical semantic association among these symbols, enhancing the understanding of the script's numerical notation system.



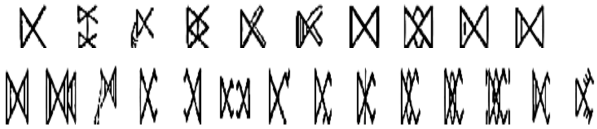
A notable finding emerged concerning symbols depicting figures of men with objects in their hands or legs. Across all four Clustering algorithms employed, these symbols consistently fell within the same clusters. This consistent Clustering pattern suggests a shared semantic association among these symbols, indicating potential thematic or functional similarities. It is conceivable that such depictions may represent activities, roles, or attributes associated with the depicted individuals. For instance, a man holding a staff or engaged in a specific posture may signify leadership, authority, or ceremonial roles within the societal context of the Indus civilization.



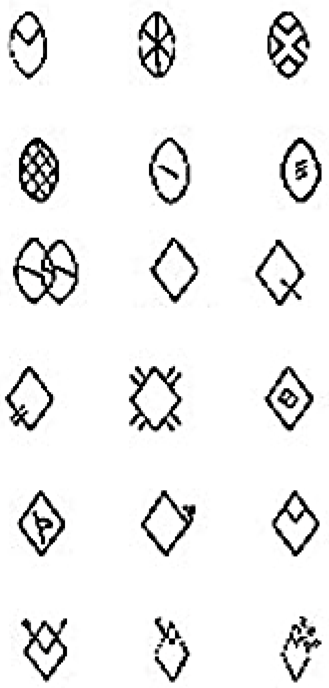
Symbols potentially representing the concept of time consistently clustered together across all four Clustering algorithms employed. This cohesive Clustering pattern suggests a shared semantic association among these symbols, indicating their likely representation of temporal concepts within the context of the Indus civilization. The Clustering of these symbols provides compelling evidence of the systematic organization and conceptual coherence of the Indus script, offering valuable insights into how time may have been conceptualized and represented in the ancient Indus Valley civilization.



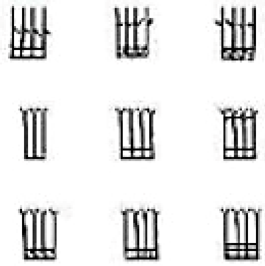
When utilizing a smaller number of clusters, a distinct pattern emerged: twenty-five symbols associated with time consistently clustered together. Across all four algorithms, this cohesive grouping suggests a shared semantic association among these symbols, likely representing temporal concepts within the ancient Indus civilization. Such clustering not only underscores the systematic organization and conceptual coherence of the Indus script but also provides valuable insights into how time may have been conceptualized and represented in this ancient culture.



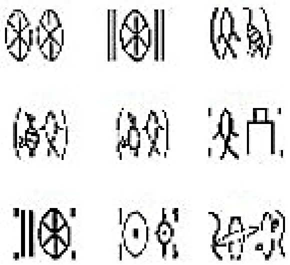
An intriguing pattern emerged as symbols resembling diamond shapes and oval shapes consistently clustered together across multiple Clustering algorithms. This consistent Clustering suggests a potential semantic connection between these geometric forms, indicating that they may represent similar concepts or objects within the context of the Indus script. While it's tempting to speculate that these shapes could indeed refer to analogous or related ideas, such as valuable items or symbolic representations, further investigation and contextual analysis are necessary to ascertain their precise meanings.



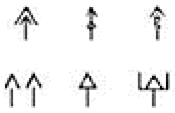
In Spectral Clustering, a distinct cluster consistently emerged, comprising symbols associated with commodities. This separation indicates a cohesive grouping with shared semantic associations among these symbols, likely representing commodities within the ancient Indus civilization. Such clustering not only underscores the systematic organization and conceptual coherence of the Indus script but also provides valuable insights into the economic activities and trade networks of the ancient civilization.



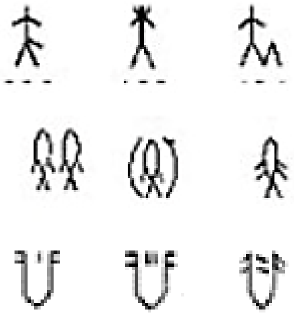
In BIRCH, with the utilization of 10 clusters, a significant observation arose: symbols that appeared to contain two or more sub-symbols within them consistently clustered together. This clustering suggests a cohesive association among these symbols, potentially representing complex concepts or composite symbols within the ancient Indus civilization. Such grouping underscores the systematic organization and conceptual coherence of the Indus script, offering valuable insights into its symbolic structure and potential linguistic conventions.



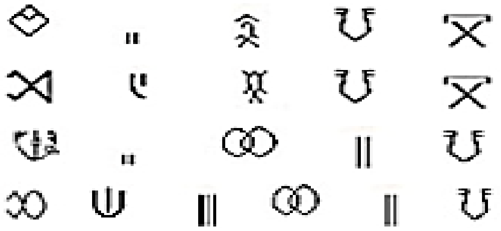
In Agglomerative Clustering, a noteworthy discovery occurred: symbols representing directions consistently fell into the same cluster. This clustering suggests a cohesive association among these symbols, likely indicating their shared semantic concept of direction within the ancient Indus civilization. Such grouping emphasizes the systematic organization and conceptual coherence of the Indus script, offering valuable insights into how spatial concepts were represented and utilized in the ancient script.



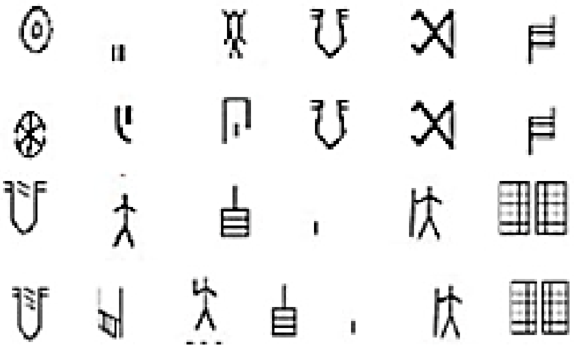
Moreover, an intriguing finding emerged as all primary symbols and those derived or evolved from them consistently fell under the same cluster across all Clustering methods. For instance, symbols resembling fish, humans, or jars, which are considered primary symbols, demonstrated cohesive Clustering patterns with their derived variants. This consistency across Clustering methods reinforces the hierarchical semantic structure of the script and provides valuable insights into its evolutionary dynamics.



In the analysis of Sequence Clustering using K-Means, a notable observation emerged wherein sequences falling within the same cluster exhibited a high degree of similarity, with only minimal symbol replacements across the sequences. For instance, when comparing two different sequences within the same cluster, it was found that the majority of symbols remained unchanged, with only a single position exhibiting a variation in symbols. This consistent pattern suggests a strong structural coherence within clustered sequences, indicating potential syntactic or semantic regularities in the arrangement of symbols. The minimal symbol replacements imply subtle alterations in meaning or emphasis, while preserving the overall thematic or contextual integrity of the sequences.

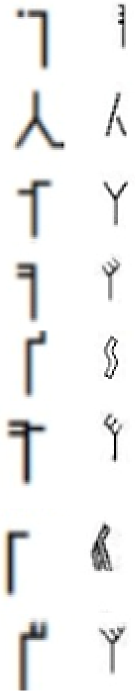


Additionally, sequences of symbols within clusters exhibited patterns suggestive of rephrased sentences or syntactic structures with similar meanings. This observation underscores the script's syntactic and semantic richness, providing clues about its communicative capabilities and linguistic conventions. Exploring symbol sequences within clusters advances the understanding of the script's linguistic complexity and expressive range.



4.4 Brahmi vs Indus Similarity Comparison

The Clustering analysis of the Indus script symbols revealed significant findings regarding their linguistic connections, hierarchical usage, and contextual coherence. Notably, certain symbols exhibited striking similarities with Tamil Brahmi symbols, suggesting potential linguistic connections and aiding in determining pronunciation. This observation underscores the script's linguistic richness and its potential ties to broader linguistic traditions.



This comparative analysis provides valuable contextualization for understanding the Indus script within the broader framework of ancient South Asian writing systems, encompassing Tamil and other regional scripts. It offers avenues for further exploration of historical interactions and cultural exchanges.

5. CONCLUSION

In conclusion, this interdisciplinary study has made significant strides towards deciphering the elusive Indus script, utilizing a multifaceted approach that combines computational techniques with historical inquiry. Through the application of N-gram analysis, predictive modeling, and Clustering analysis, valuable insights into the linguistic, structural, and cultural dimensions of the script have been gained. The N-gram analysis revealed recurring symbol patterns, providing a deeper understanding of the script's linguistic structures and potential semantic meanings. By predicting missing values within symbol sequences and validating the findings through advanced LSTM models, coherent texts were reconstructed. Furthermore, the Clustering analysis highlighted linguistic connections, hierarchical usage, and contextual coherence within the script, offering insights into its historical and cultural significance. Employing four Clustering algorithms—K-means, hierarchical Clustering, BIRCH, and Spectral Clustering—enhanced the robustness of the analysis, ensuring a comprehensive examination of symbol clusters. Contextual usage of clustered symbols enriched the understanding of their intended meanings and symbolic representations, shedding light on cultural, religious, and administrative contexts. Additionally, the grammar analysis, employing frequency analysis, pattern recognition, and contextual analysis, provided further insights into the grammatical structures and syntactical nuances of the Indus script. The identification of recurring grammatical patterns and their contextual interpretations contributes significantly to the ongoing efforts to decipher and understand the ancient Indus civilization's language and communication systems. Though this study represents a significant milestone, future research should aim to expand the dataset and explore new avenues for analysis, leveraging collaboration with experts in related fields. By continuing to innovate and collaborate, optimism remains about further discoveries in the field of ancient linguistics and archaeology.

6. REFERENCES

[1] Community Detection in Networks: A Comprehensive Survey (2016) by Fortunato, S. & Castellano, C. (Published in: Physics Reports, 586, 74-174. https://arxiv.org/list/stat.ME/recent

[2] Gong, H., & Zhang, Y. (2020). Analyzing the Indus script using a combination of convolutional neural networks and n-grams. Pattern Recognition Letters, 130, 272-278.

[3] Shan, X., Yao, J., & Wang, J. (2019). Rethinking Indus script complexity through information theory. Entropy, 21(12), 1222.

[4] Shu-Ming Hsieh and Chiun-Chieh Hsu. 2008. Graph-based representation for similarity retrieval of symbolic images. Data Knowl. Eng. 65, 3 (June, 2008), 401–418. https://doi.org/10.1016/j.datak.2007.12.004.

[5] Guanglin Huang, Wan Zhang, and Liu Wenyin. 2008. A Discriminative Representation for Symbolic Image Similarity Evaluation. Graphics Recognition. Recent Advances and New Opportunities: 7th International Workshop, GREC 2007, Curitiba, Brazil, September 20-21, 2007. Selected Papers. Springer-Verlag, Berlin, Heidelberg, 71–79. https://doi.org/10.1007/978-3-540-88188-9_8

[6] Sarat Sasank Barla, Sai Surya Sanjay Alamuru, and Peter Zsolt Revesz. 2022. Feature Analysis of Indus Valley and Dravidian Language Scripts with Similarity Matrices. In Proceedings of the 26th International Database Engineered Applications Symposium (IDEAS '22). Association for Computing Machinery, New York, NY, USA, 63–69. https://doi.org/10.1145/3548785.3548801.

[7] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability

[8] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.

[9] Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. In *ACM*

[10] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (pp. 849-856).

[11] Ansumali Mukhopadhyay, B. Interrogating Indus inscriptions to unravel their mechanisms of meaning conveyance. *Palgrave Commun* **5**, 73 (2019). https://doi.org/10.1057/s41599-019-0274-1