Statistical Structure in Indus Sign Sequences

Tanishk Tiwari
DSEU
t.high1945@gmail.com

Abstract

This paper introduces a computational framework for evaluating structural properties of the undeciphered Indus script.The study uses a corpus of 6,579 inscriptions.The analytical approach combines unsupervised visual clustering of sign morphology, entropy-based sequence analysis, Kullback-Leibler divergence comparison, and neural sequence modeling (BiLSTM). The results indicate directional asymmetry and structured combinatorial patterns in sign sequences. We conclude that the Indus sign sequences exhibit statistical properties consistent with structured symbolic systems and not easily explained by random generation.

1 Introduction

The Indus Valley Civilization (c. 2600–1900 BCE) produced a symbolic system primarily found on steatite seals, copper tablets, and pottery sherds. Figure 1 shows an example inscription from the dataset. The short sequences—averaging four signs in length—and the absence of bilingual artifacts have prevented decipherment.

Figure 1: Example Indus seal artifact containing a short sequence of signs above an animal motif.

The inability to decipher the text has generated distinct hypotheses regarding its nature. Some scholarship posits that the sign system exhibits sequential statistical properties consistent with structured symbol systems (Mahadevan, 1977; Parpola, 1994). Conversely, other researchers argue that the sign sequences represent non-linguistic administrative emblems or heraldic symbols (Farmer et al., 2004).

This work does not assume that the Indus inscriptions represent a fully developed writing system. Instead, it investigates whether their statistical properties resemble those observed in known linguistic symbol systems, non-linguistic structured symbol systems, or random sequences.

The central question of this study is whether the statistical structure of Indus sign sequences more closely resembles that of linguistic writing systems or structured non-linguistic symbol systems. Specifically, the study evaluates three questions:

1. Do sign sequences exhibit directional asymmetry?
2. Do signs form stable visual clusters that may correspond to functional categories?
3. Do sequence models detect ordering constraints within inscriptions?

To address these questions, the study applies a multimodal computational framework combining computer vision, statistical sequence analysis, and neural sequence modeling.

2 Related Work

Early efforts by Hunter (1934) pioneered the cataloging of sign variations. Subsequently, Knorozov (1965) and Knorozov et al. (1979) used early computational methods to conduct positional analysis, identifying recurrent sign combinations suggestive of structured sequential composition. Mahadevan (1977) and Parpola (1994) analyzed positional

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 314–319

statistics to suggest the presence of terminal sequence markers.

Recent computational interventions have sought to quantify the script's flexibility. Rao et al. (2009) utilized first-order conditional entropy to suggest that the script's sequential flexibility aligns with natural languages. However, Sproat (2014) demonstrated that entropic measures alone are insufficient diagnostics for language, as they fail to differentiate written language from non-linguistic symbol systems with internal rules. Mukhopadhyay (2023) proposed an administrative model concerning taxation and trade, while Yajnadevam (2024) attempted direct phonetic mapping.

Our work acknowledges the critiques raised by Sproat (2014). We combine sequence modeling with computer vision and differential cross-system baselines to identify statistical regularities in sign ordering and positional constraints, rather than using isolated metrics to claim linguistic status.

3 Methodological Framework

To evaluate the structural constraints of the script, we process the data through a four-stage analytical pipeline. First, we curate a standardized corpus from historical catalogs. Second, to avoid modern human bias in sign classification, we extract visual features directly from glyph images and apply unsupervised clustering to group structurally similar signs independently of sequence statistics. Third, we apply information-theoretic metrics to test for syntax directionality and compare these metrics against linguistic and non-linguistic baselines. Finally, we train a neural sequence model. By feeding the model systematically altered synthetic sequences, we observe which structural manipulations the model assigns lower probability to, thereby analyzing hidden statistical structure.

4 Methodology

4.1 Dataset

The digitized corpus used in this study is derived from publicly available Indus sign lists and inscription catalogs (Mahadevan, 1977; Wells, 2011). Inscriptions containing illegible signs, damaged portions, or non-numeric tokens were filtered. The dataset was partitioned into an 80/20 train/test split, stratified based on archaeological provenance metadata (object type). The vocabulary consists of numeric sign identifiers bounded dynamically by padding (<PAD>), start-of-sequence (<SOS>), and end-of-sequence (<EOS>) tokens.

4.2 Visual Feature Extraction

To establish sign families based on physical traits, we processed binarized glyph images. We extracted Hu Moments to achieve rotational invariance, as signs are frequently carved at varying angles. We additionally extracted Histogram of Oriented Gradients (HOG) features (64×64 pixels, 8×8 cells, 9 orientations) to capture local texture.

4.3 Sign Clustering

We applied feature standardization to the combined visual feature space. To reduce dimensionality while preserving non-linear geometric relationships, we applied Uniform Manifold Approximation and Projection (UMAP). Subsequently, we applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) followed by agglomerative clustering to identify morphological sign communities.

Purpose for Humanities Scholars: The goal of this unsupervised clustering is *not* to reduce the number of signs for modeling, but to discover natural visual families that may reflect functional or scribal categories used by ancient Indus scribes. All subsequent statistical analyses (entropy, KL divergence, sequence modeling) were performed on the **original** fine-grained sign inventory (400 signs), not on the clusters. The 8–20 morphological families serve primarily as an independent validation of structure and help interpret which visual types of signs carry positional or combinatorial importance.

4.4 Entropy Analysis

We calculated unigram (H_1), bigram (H_2), and trigram (H_3) conditional entropies for both the original corpus and a mechanically reversed corpus to test for statistical irreversibility.

Humanities Interpretation: In this context, *conditional entropy* measures the predictability of a sequence. Lower entropy indicates tighter syntactical rules (e.g., if Sign A is almost always followed by Sign B, entropy is low). *Markov orders* refer to the window of context: unigram (1 sign), bigram (2 signs), and trigram (3 signs).

We also calculated *Positional Rigidity* (R_i) for every absolute position i in the sequence, defined as $R_i = 1 - (H_i/H_1)$, where H_i is the positional entropy $H(X|\text{position} = i)$. Positional rigidity measures how strictly a sign is tied to a specific

physical location in a text (e.g., a sign that only ever appears at the very end of a line has high positional rigidity).

4.5 Cross-System Validation

To evaluate whether positional rigidity provides meaningful structural information, we apply the metric to three systems: a known writing system (Linear B), a synthetic heraldic sequence system, and the Indus corpus. Linear B was chosen because its linguistic status is firmly established and it exhibits short administrative inscriptions similar in length to Indus texts. The heraldic system serves as our structured non-linguistic baseline; it mimics European Blazonry (coats of arms), possessing strict positional rules (e.g., crests must be terminal, charges central) but mapping to no spoken language.

4.6 KL Divergence Analysis

To quantify the directional dependence of the joint probability distributions, we computed the Kullback-Leibler (KL) Divergence between the original forward transition probability matrix (P_{fwd}) and the mechanically reversed matrix (P_{bwd}):

$$D_{KL}(P_{fwd}||P_{bwd}) = \sum_{x \in \mathcal{X}} P_{fwd}(x) \log_2 \left(\frac{P_{fwd}(x)}{P_{bwd}(x)} \right) \quad (1)$$

Humanities Interpretation: KL Divergence measures *directional syntax*. A high divergence indicates that reading the text left-to-right follows entirely different structural rules than reading it right-to-left, strongly implying a fixed, intended reading direction.

4.7 Representation Learning of Sign Sequences

We modeled the syntax using a Categorical Hidden Markov Model (HMM), evaluating state distributions using the Bayesian Information Criterion (BIC). We then trained a Bidirectional Long Short-Term Memory (BiLSTM) network to predict the next token to evaluate structural constraints via *perplexity* (a measure of how "confused" or uncertain the model is by a sequence; lower perplexity means the sequence appears more natural to the model's learned grammar).

We procedurally generated three datasets of synthetic inscriptions to test against the model:

1. **Concatenation A+B:** Two inscriptions with identical artifact types and similar lengths were joined.

2. **Prefix-Core-Suffix Swap:** The core of Inscription B was injected between the terminal signs of Inscription A.

3. **Suffix Replacement:** The final sign of A was replaced with the final sign of B.

4.8 Implementation Details

Visual features were extracted after binarization using Otsu's thresholding method and base/affix splitting (components occupying $> 30\%$ of the total glyph area were classified as base). All features were standardized with z-score normalization. Dimensionality reduction was performed with UMAP (10 components, 15 neighbors, min dist 0.1). Clustering used agglomerative clustering with complete linkage. The optimal number of clusters ($k = 8$) was selected by maximizing the silhouette score over $k = 2$ to 20, yielding a global silhouette score of 0.414. The BiLSTM model used an embedding dimension of 128 and hidden size of 256, trained for 10 epochs.

5 Results

Results reported in this section are computed on the Indus inscription corpus unless otherwise stated. Synthetic datasets and cross-system corpora are used only for baseline comparison.

5.1 Clustering Results

The application of HDBSCAN on the UMAP-reduced feature space yielded stable partitions. Figure 2 visualizes the resulting clusters. Figure 3 illustrates the network connectivity (cosine similarity) between these morphological families, and Figure 4 displays algorithmically generated cluster representatives. Several clusters show coherent visual themes (e.g., complex composite signs vs. simple linear strokes), suggesting that ancient scribes may have treated them as related categories despite surface variation.

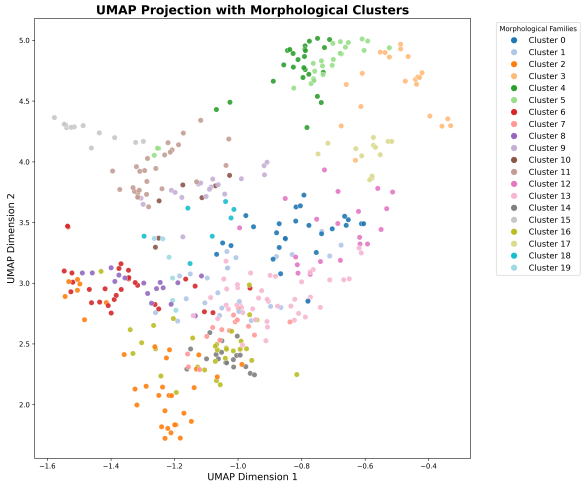

Figure 2: UMAP projection of visual glyph embeddings revealing morphological clusters discovered via HDBSCAN.

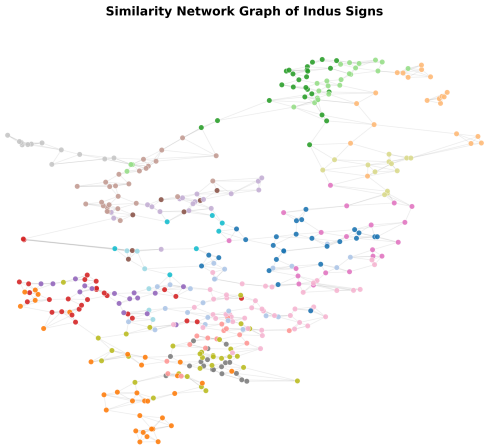

Figure 3: Similarity network graph of Indus script signs based on visual features.Edge thickness represents cosine similarity between morphological families.

5.2 Entropy and Cross-System Validation

The entropy calculations for the original and reversed Indus sequences, alongside comparative baselines, are summarized in Table 1. The original Indus sequence direction yields lower conditional entropy at higher Markov orders compared to its reversed counterpart, indicating a clear directional rule-set.

Metric	Indus (Orig)	Indus (Rev)	Linear B	Heraldic
Unigram (H_1)	6.58	6.58	5.59	3.78
Bigram (H_2)	3.60	4.11	4.43	3.73
Trigram (H_3)	1.42	1.61	2.12	3.45

Table 1: Comparison of conditional entropy decay across systems (measured in bits).

Table 2 presents the cross-system validation of Positional Rigidity (R_i). Linear B shows near-zero rigidity across positions, while the heraldic system exhibits strong rigidity at constrained boundary positions. Indus inscriptions show intermediate behavior. These positional constraints are visualized in Figure 5. While the average Indus sequence is 4-5 signs long, the rightward spike on the x-axis demonstrates that terminal signs remain highly constrained regardless of total sequence length.

System	R_0	R_1	R_2	R_3	R_4	R_5
Linear B	0.023	-0.008	0.009	0.012	0.011	0.014
Heraldic	1.779	-0.126	-0.016	0.071	0.301	1.780
Indus	0.082	0.110	0.050	0.140	0.220	0.740

Table 2: Cross-system comparison of Positional Rigidity (R_i) tracking constraint strengths across specific sequence indices.

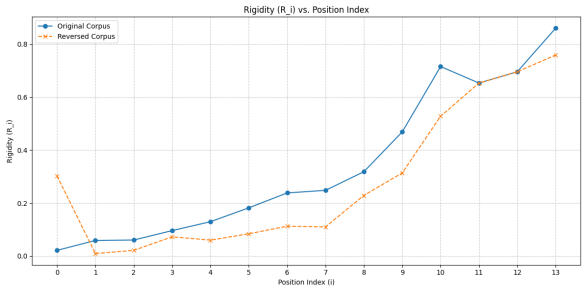

Figure 5: Positional Rigidity (R_i) plotted against absolute position index. The original Indus corpus exhibits terminal boundary constraints that distinguish it from random distribution.

5.3 KL Divergence Results

The KL divergence results comparing the forward and backward joint probability distributions are presented in Table 3.

Metric	Value (bits)
KL Divergence (Bigram)	16.45
KL Divergence (Trigram)	21.43

Table 3: Kullback-Leibler divergence between original and reversed joint distributions measuring statistical irreversibility.

5.4 Sequence Model Predictability

Based on BIC penalty functions, a 4-state HMM was optimal for the dataset. Figure 6 displays the proportional dominance of each hidden state by

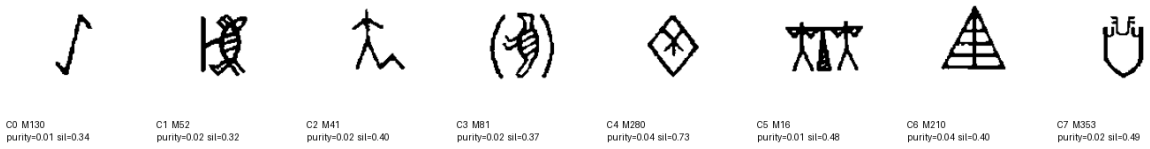

Figure 4: Cluster representatives extracted from the HDBSCAN partition. Labels indicate the standard Mahadevan ID alongside individual silhouette scores for each morphological prototype.

absolute sequence position. These latent states represent hidden functional categories (e.g., "start" and "stop" structural slots) within the script's syntax.

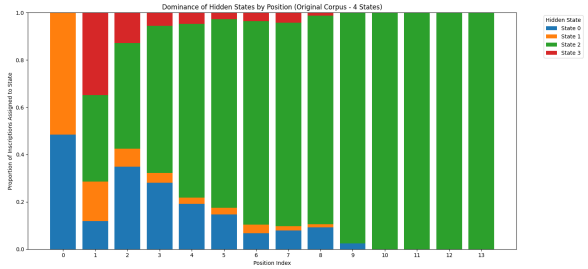

Figure 6: HMM state dominance by sequence position. State 1 dictates sequence beginnings, while State 4 dictates sequence terminals, confirming structured functional transitions.

The BiLSTM model results for synthetic sequence evaluation are presented in Table 4. The ΔPPL metric represents the ratio of synthetic mean perplexity to the real test set mean perplexity (2.60 Abs. PPL). The extremely small p-values ($< 10^{-12}$) indicate that the observed differences between real and synthetic sequence distributions are statistically robust. Figure 7 visualizes these perplexity distributions, and Figure 8 shows the gradient saliency map indicating where the model applies its "attention" during next-token prediction.

Dataset	Abs. PPL	ΔPPL	p-value	Cliff's δ
Real Test Set	2.60	1.00	–	–
Suffix Replace	1.10	0.42	6.53×10^{-13}	0.176
Concat A+B	1.39	0.53	2.65×10^{-177}	-0.698
Prefix-Core Swap	2.42	0.92	2.13×10^{-166}	-0.676

Table 4: Absolute and ratio perplexity scores alongside non-parametric statistical significance evaluations for synthetic sequence modifications.

6 Discussion

The results demonstrate directional structural constraints within the script. The lower conditional entropies for the original sequence orientation (Table 1), combined with the increasing KL divergence

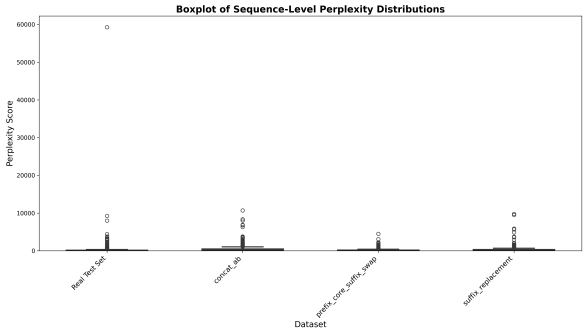

Figure 7: Boxplots of sequence-level perplexity distributions. The BiLSTM smoothly parses suffix replacements (tight low variance) while assigning higher variance to structural concatenations and core swaps.

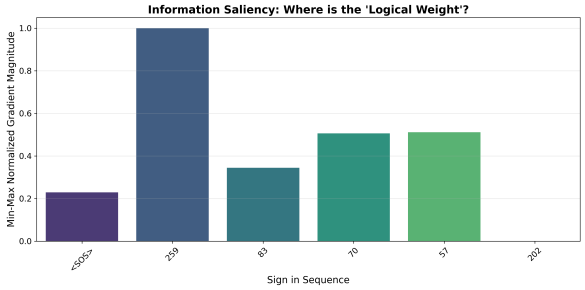

Figure 8: Gradient saliency map indicating specific token weights during sequence prediction. High min-max normalized gradient magnitude on terminal boundaries confirms the model's heavy reliance on positional constraints.

at higher Markov orders (Table 3), indicate statistical irreversibility. Following Sproat (2014), this study avoids interpreting entropy values alone as evidence of language and instead compares multiple symbolic systems.

The cross-system validation (Table 2) clarifies the nature of these constraints. Linear B shows near-zero rigidity, typical of natural languages where words can appear in multiple syntactic slots. The synthetic heraldic system exhibits strict rigidity at boundary positions. Indus inscriptions show intermediate behavior. The elevated rigidity at the final position in the Indus corpus ($R_5 = 0.740$) suggests that terminal signs may function as structural delimiters or positional markers.

The sequence representation model assigned lower probabilities (higher perplexity variance) to concatenated sequences and core-swapped sequences. However, the model assigned higher relative probabilities to sequences where terminal signs were replaced (Figure 7). This response pattern indicates combinatorial constraints, where terminal elements exhibit structural modularity while core elements maintain sequential rigidity.

7 Conclusion

This study evaluated the structural properties of the Indus script using a multimodal computational framework. The central question was whether the statistical structure of Indus sign sequences more closely resembles that of linguistic writing systems or structured non-linguistic symbol systems. Regarding the specific research questions: (1) the script exhibits directional asymmetry, supported by conditional entropy and KL divergence metrics; (2) visual clustering of glyphs yields stable categories independent of human cataloging; and (3) neural sequence modeling detects internal ordering constraints, accepting terminal variations while assigning lower probabilities to arbitrary structural concatenations. These findings indicate that the Indus sign sequences exhibit statistical properties consistent with structured symbolic systems and not easily explained by random generation.

8 Limitations

This analysis cannot determine whether the Indus inscriptions encode language. Instead, it characterizes their statistical structure relative to other symbolic systems. A limitation of this study is the brevity of the source artifacts. Short sequence length limits higher-order syntactic inference, but information-theoretic metrics remain informative for detecting directional constraints. Additionally, because the script remains undeciphered, the functional categories identified by statistical clustering cannot currently be verified against semantic or phonetic meanings.

Data Availability

The digitized sign corpus used in this study is derived from publicly available Indus sign lists and previously published corpora. The processed dataset and analysis scripts will be made available in an open repository upon publication.

References

Steve Farmer, Richard Sproat, and Michael Witzel. 2004. The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies*, 11(2):19–57.

G. R. Hunter. 1934. *The Script of Harappa and Mohenjodaro and Its Connection with Other Scripts*. Kegan Paul, Trench, Trubner & Co.

Yuri V. Knorozov. 1965. *Predvaritel'noe soobshchenie ob issledovanii protoindiyskikh tekstov*. Nauka.

Yuri V. Knorozov, M. F. Albedil, and B. Y. Volchok. 1979. *Proto-Indica: 1979. Report on the Investigation of the Proto-Indian Texts*. Nauka.

Iravatham Mahadevan. 1977. *The Indus Script: Texts, Concordance and Tables*. Archaeological Survey of India.

Bahata Ansumali Mukhopadhyay. 2023. Semantic scope of Indus inscriptions comprising taxation, trade and craft licensing. *Humanities and Social Sciences Communications*, 10.

Asko Parpola. 1994. *Deciphering the Indus Script*. Cambridge University Press.

Rajesh P. N. Rao, Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009. Entropic evidence for linguistic structure in the Indus script. *Science*, 324(5931):1165–1167.

Richard Sproat. 2014. Statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2):457–481.

Bryan K. Wells. 2011. *Epigraphic Approaches to Indus Writing*. Oxbow Books.

Yajnadevam. 2024. A cryptanalytic decipherment of the Indus script. *Preprint*.