A Computational Analysis of the Indus Script: Identifying Sign Functions in Logo-Syllabic Writing Systems

Ruhan Khanna
Sidwell Friends School
Washington, DC

Louie Merriam
Sidwell Friends School
Washington, DC

ABSTRACT
A key challenge in deciphering any logo-syllabic writing system is to distinguish signs that represent complete lexical units (logograms) from signs that represent syllables. The Indus script is widely assumed to be logo-syllabic, an intermediate system in which some signs encode whole words or concepts, others encode syllables, and some function polyvalently depending on context. This ambiguity complicates decipherment: a single grapheme may serve as a word in one inscription and as a syllable in another. This study proposes methods for separating these classes. To begin, one basic premise is assumed: signs that appear alone in inscriptions must be capable of expressing a complete semantic value and hence are the strongest candidates for logograms. Building on this "singleton" premise, the candidate logogram set was expanded by contextual co-occurrence, associations were validated statistically, and distributional relationships were then mapped among signs and sign-pairs. A second component develops an exclusivity-based procedure for identifying likely syllabic pairs. The results demonstrate several key categories of Indus signs.

General Terms
Pattern Recognition, Natural Language Processing, Computational Epigraphy, Statistical Analysis, Script Decipherment.

Keywords
Indus script, logograms, syllabic signs, bigrams, distributional similarity, exclusivity, undeciphered writing systems.

1. INTRODUCTION
The Indus script has eluded decipherment for over a century. The key to deciphering the Indus script is differentiating the different sub-categories of signs that make up the script. The Indus script has been heavily suggested to be logo-syllabic [7][4][6]. A logo-syllabic writing system contains signs representing both logograms and syllables. Multiple lines of internal evidence point to a logo-syllabic system [6]. The sign inventory is too large for an alphabet yet too small and too repetitive for a purely logo-graphic script [4]. Many single-sign inscriptions are unlikely to serve as syllables and rather function as complete words (logograms), while recurrent sign pairs and affix-like elements suggest bound phonological units [1]. Strong positional rules—stable initial/medial/terminal clustering—along with clear numeral series and formulaic collocations are consistent with mixed systems known from the ancient Near East, where logograms coexist with syllabic spellings [6]. Together, these features make a purely alphabetic or purely logographic account unlikely.

However, logo-syllabic writing presents a unique challenge to decipherment. Any given sign in a logo-syllabary could serve as a logogram, syllable, or both. Furthermore, many frequent pairs of signs could represent two related logograms or jointly combine to form one new word. Many previous researchers have applied segmentation analysis to the script [2], [5]. While useful, the segmentation methods do not allow us to further subdivide the script's signs into logograms, syllables, bigrams, etc. Without making any external linguistic assumptions, this paper seeks to computationally analyze the script to identify which signs are semantically complete by themselves, which pairs of signs represent words, and which pairs of signs are potentially phonetic.

This approach is deliberately corpus-driven and language-neutral. First, the study analyzes the distribution of frequent sign pairs across positions and surrounding contexts to distinguish pairs that behave like lexical (logographic) units from those whose even, combinatorial behavior is expected for syllabic material. Second, semantically similar signs were grouped by comparing their full context profiles using a normalized distance metric (adapted from prior matrix-based methods), and the procedure was validated on well-recognized families such as numerals and the "fish" signs. This workflow yields a small, high-confidence set of candidate logograms and syllabic pairs that organizes the signary into functional classes for downstream analysis.

2. IDENTIFYING LOGOGRAMS
Logograms likely make up the majority of the signs in the Indus script, and by extension, likely make up much of the backbone of patterned Indus texts [6]. The goal of this section is to provide a method of identifying logograms based on the premise that any sign that occurs independently represents a semantically complete logogram [6]. An isolated syllable is unlikely to constitute an inscription without simultaneously denoting a complete word. Therefore inscriptions consisting of a single character (hereafter, "singletons") are treated as the most conservative evidence for logograms. Analogous inferences are commonplace in other undeciphered or partially deciphered logo-syllabaries, where lone signs on dedicatory or ceremonial objects are interpreted as titles, names, or ritual terms. Based on the collection of these singletons, contexts shared by other signs and the list of singletons were examined to expand the list of logograms, and then these additional shared-context signs were classified as logograms as well.

In order to build a seed set of definite logograms, all singleton inscriptions from a cleaned corpus were extracted to form a foundational set of "seed" signs. The seed set is intentionally non-exhaustive; rather, it serves as a high-precision anchor for subsequent contextual analysis. For example, a sign such as 070 that repeatedly occurs as a singleton is marked as a probable logogram. These seeds provide fixed points for evaluating multi-sign contexts in longer inscriptions.

From the singleton anchors, a cautious process of contextual expansion was implemented. When two or more seeds share an

identical context within a non-seed sign, the non-seed sign is appended to the list of seed signs. In such contexts, any additional sign that consistently appears in similar contexts to seeds is promoted as a candidate logogram. For instance, if sign 595 is interchangeable with both seed 070 and seed 231 in different inscriptions, then 595 is provisionally admitted as a logographic candidate. Admission requires repeated occurrence in seed-rich contexts to minimize false positives.

This method provided a robust set of around 200 signs determined to be logograms. Interestingly, the majority of these signs are highly frequent in the corpus and make up the overwhelming majority of inscriptions. This implies that patterned Indus texts were primarily logographic. Alternatively, it seems highly probable that many of these signs were polyvalent and could serve as both syllables and logograms in different contexts.

The following signs are identified as semantically complete logograms:

001, 002, 003, 004, 005, 006, 007, 013, 016, 017, 020, 031, 032, 033, 034, 035, 037, 039, 042, 043, 047, 049, 055, 060, 061, 064, 090, 091, 098, 100, 110, 117, 127, 136, 137, 140, 142, 144, 145, 147, 151, 153, 154, 155, 156, 158, 161, 165, 169, 176, 215, 220, 222, 226, 230, 231, 233, 234, 235, 236, 237, 240, 241, 242, 250, 255, 281, 341, 350, 365, 368, 384, 386, 387, 388, 390, 400, 402, 405, 407, 408, 411, 413, 415, 416, 430, 440, 452, 455, 460, 462, 480, 482, 511, 515, 519, 520, 526, 527, 530, 539, 540, 550, 556, 565, 575, 585, 586, 590, 592, 615, 617, 630, 642, 643, 647, 679, 685, 690, 692, 697, 698, 699, 700, 702, 705, 707, 740, 741, 742, 749, 753, 760, 772, 773, 776, 777, 780, 781, 782, 790, 798, 803, 817, 820, 822, 824, 832, 836, 839, 840, 841, 843, 850, 861, 869, 890, 892, 898, 900, 904, 909, 923, 927, 930, 942, 943, 945, 946, 956, 957

3. BIGRAM ANALYSIS

Many pairs of signs could also represent complete words. These word-representing pairs would therefore likely share contexts with the logograms identified earlier. In order to make progress in deciphering the Indus script, bigrams that represent complete words need to be treated as their own semantic units. Incorrectly identifying a semantically complete bigram as two separate words would severely hinder any decipherment efforts. First, bigrams that were statistically significant were identified, thereby implying that their co-occurrence was not by coincidence. Previous studies on the script have also made use of z-scores to segment inscriptions [5], [6], [8]. However, these studies haven't fully identified pairs of signs likely to represent complete words on their own. It was then examined whether these bigrams shared contexts with the semantically complete logograms identified earlier to determine whether these bigrams represented complete words themselves.

To test whether co-occurring signs form stable structural associations rather than random conjunctions, adjacent sign-pairs (bigrams) were analyzed. Let c_a and c_b be the marginal frequencies of signs A and B in a corpus of N inscriptions and let O be the observed bigram count of (A,B). Under an independence model, the expected co-occurrence is

$$E = \frac{c_a \times c_b}{N}$$

Deviation is assessed via a z-score,

$$z = \frac{O - E}{\sigma}$$

with variance estimated under a Poisson approximation ($\sigma^2 \approx E$) or, in sensitivity checks, under a binomial/hypergeometric model. High positive z values indicate pairs that co-occur far more often than predicted by chance [8], supporting the hypothesis that the pair forms a meaningful unit (e.g., a compound or a logogram-modifier construction). Pairs whose observed frequencies do not exceed expectation are treated as noise.

Frequency and z-scores alone are insufficient in identifying semantically complete bigrams: formulaic repetition or scribal conventions can inflate co-occurrence. In order to confirm that the bigrams are semantically complete, they were compared to the logogram list created earlier. Therefore, only those bigrams that occur within contexts already shared by at least one seed sign were retained. The resulting set is further filtered qualitatively by examining consistency across sites, artifact classes, and neighboring signs. This layered procedure, statistical evaluation followed by contextual filtering, reduces over-interpretation and confirms that the bigrams identified are semantically complete.

The following are pairs that likely represent semantically complete words:

001-480, 002-817, 002-820, 002-861, 156-003, 430-003, 390-004, 405-004, 407-004, 900-005, 840-013, 220-016, 390-016, 575-017, 585-017, 220-032, 226-032, 032-840, 877-032, 520-033, 700-033, 033-705, 700-034, 035-171, 060-550, 060-820, 142-061, 090-740, 151-097, 100-415, 740-100, 165-900, 220-415, 520-220, 233-803, 240-235, 630-240, 435-255, 255-705, 920-320, 335-484, 368-817, 390-590, 390-844, 400-525, 405-590, 840-416, 760-440, 460-495, 460-503, 460-510, 806-471, 740-482, 615-503, 527-550, 527-555, 740-752, 740-760, 740-772, 740-773, 740-923

4. DISTRIBUTIONAL SIMILARITY AND CLUSTERING

We also sought to measure the semantic similarity between signs and sign-pairs. To map relationships among signs and sign-pairs, distributional similarity was computed between semantically complete signs and bigrams identified earlier. Each high-confidence bigram is represented as a vector of co-occurrence frequencies with all other signs; Euclidean distances among these vectors yield a similarity space [3]. Two matrices: (i) pair–pair distances among top bigrams, and (ii) pair–sign distances between bigrams and seed logograms. Clustering within these spaces identifies groups that share inscriptional environments. For example, if bigram 595–820 clusters near singleton 070, a common semantic domain (e.g., commodities or numerals) is hypothesized. Clusters do not constitute decipherments per se, but they sketch lexical fields and functional classes.

Analyses were executed on a Google Cloud Platform VM (8 vCPUs, 52 GB RAM) with a 100 GB SSD. Data processing and matrix operations were implemented in Python (pandas/NumPy). Infrastructure issues, including cross-environment file paths, NumPy broadcasting shape mismatches (standardized array shapes), and intermittent archival rate limits (retry logic with backoff) were documented and resolved. These operational details are essential for reproducibility and to prevent silent pipeline failures.

The clustering provided several unique results, useful in identifying semantic categories in the script. Much segmentation work has already been done on the Indus script [6][8][2]. These results were in line with previous clustering analyses of Indus signs. Four distinct categories of signs were identified, roughly corresponding to the initial cluster, medial cluster, bonded cluster, and terminal marker identified by Wells

and Fuls [6][2]. However, these results are the first to cluster sign pairs identified as semantically complete earlier as well. The results identified certain bigrams with shared components that behaved similarly. For example, bigrams ending in sign 465 had smaller distances to one another on average. This potentially suggests some type of suffixing function in the script.

Results were collected in the following manner. All data will be made available upon publication:

- Singleton seeds: putative logograms identified via singletons as discussed earlier.
- Context-expanded seeds: candidates promoted from frequent shared contexts with seeds.
- Top bigrams: ranked by z-score with observed and expected counts.
- Qualified pairs: bigrams surviving semantic and contextual filters.
- Distance matrices: pair–pair and pair–sign similarities for clustering.

Collectively, these outputs provide a reproducible framework for hypothesizing structural relations in the script. First, the seed-based method offers a principled route to identifying logograms from minimal assumptions. Second, the bigram analyses provided a means of identifying whether pairs of Indus signs represented a semantically complete word. Lastly, the clustering analyses bridge isolated signs and larger units, allowing us to identify semantic categories in the script.

5. IDENTIFYING LIKELY SYLLABIC PAIRS

Having established a framework for isolating logograms and semantically complete bigrams, the analysis turns to evidence for a phonetic component. In a logo-syllabic system, some portion of the signary must encode phonological units. Whereas logographic status can be inferred from a sign's ability to stand alone, syllabic status is necessarily indirect and must be inferred from distributional behavior [6].

Distributional exclusivity was adopted as the operative diagnostic for identifying these signs. A high-frequency pair that resists internal substitution by contextually similar alternatives behaves as a bound unit and is therefore plausibly syllabic [3]. By contrast, open collocations of independent words readily admit substitutions while preserving well-formedness.

In essence, if a frequent pair (X,Y) does not admit replacement of X or Y by semantically similar signs in the same position, the pair is a likely candidate for a syllabic word [3]. For example, if "tasty naan" and "tasty roti" were frequent sign pairs, it would concluded that the pair was not syllabic, as roti and naan have similar semantic meanings. On the other hand, lexicalized compounds (e.g., "iPhone", "cellphone") resist internal substitution, implying that they are syllabic.

To operationalize exclusivity, a table of bigrams was constructed from the cleaned corpus. Each record contains: (i) normalized three-digit identifiers for the first and second signs; (ii) the pair's corpus frequency; and (iii) a context-similarity measure, SignSimilarity, derived from Euclidean distances (lower values indicate higher contextual similarity). Two filters focus the analysis on robust patterns: a minimum pair frequency of ≥ 5, and neighbor sets defined by SignSimilarity < 0.025. These settings were tuned empirically to balance recall and precision, given the size of the corpus.

The SignSimilarity score was calculated based on the Euclidean distance between two given signs. For a given sign pair, the Euclidean distance of both signs and all other signs in the corpus was calculated. Their distance to one another was taken as a percentile of all their pairwise distances to other signs in the corpus. This percentile was the metric we used for similarity.

Step 1: Candidate selection by frequency. Retain bigrams occurring at least five times in the corpus, thereby excluding chance co-occurrences.

Step 2: Exclusivity test. For each candidate (X, Y), collect the contextually closest neighbors of X and of Y under the SignSimilarity threshold. If any neighbor of Y forms (X, neighbor) in the same structural position with non-trivial frequency—or, symmetrically, any neighbor of X forms (neighbor, Y)—the candidate is treated as an open collocation and excluded. Absence (or near-absence) of such substitutions supports treatment of (X, Y) as a bound unit.

Step 3: Tolerant overlap. Exclusivity is graded rather than absolute. Rare substitutions are tolerated to account for noise, orthographic variation, or marginal formulae; only systematic overlap leads to exclusion.

The following sign-pairs combine relatively high frequency with strong distributional exclusivity and are therefore provisionally interpreted as syllabic units:

('001','031'), ('001','480'), ('001','595'), ('001','820'), ('017','231'),
('061','845'), ('070','255'), ('070','921'), ('097','700'), ('255','832'),
('255','920'), ('321','407'), ('335','575'), ('407','845'), ('460','495'),
('460','510'), ('480','850'), ('595','820'), ('824','892'), ('003','001'),
('004','001'), ('031','001'), ('055','001'), ('760','001'), ('140','003'),
('423','003'), ('550','003'), ('407','004'), ('700','004'), ('575','017'),
('384','031'), ('407','061'), ('920','140'), ('920','320'), ('407','321'),
('711','335'), ('760','335'), ('595','391'), ('850','407'), ('575','413'),
('892','413'), ('806','465'), ('806','467'), ('806','468'), ('806','472'),
('617','550'), ('850','595'), ('760','605'), ('892','617'), ('845','806')

It is emphasized that this list constitutes a provisional syllabary. It isolates pairs whose distributional behavior is most consistent with phonological binding, pending further validation.

Results are sensitive to modeling decisions. A minimum frequency of five balances robustness against data sparsity; higher thresholds improve precision at the cost of coverage. The SignSimilarity cutoff at 0.025 enforces strict contextual proximity; relaxing it admits more candidates but increases false positives. Exclusivity can be further refined by assigning a continuous score, yielding a ranked list by degree of exclusivity, rather than a binary pass/fail outcome.

Identifying distributionally exclusive pairs provides a tractable entry point to phonological reconstruction. Candidate units can be cross-checked against external lexical hypotheses (e.g., Proto-Dravidian) and compared across sites, artifact classes, and neighboring signs. In combination with the logographic

seeds and semantically complete bigrams identified earlier, the exclusivity framework supports a more integrated account of the script, with signs ranging along a continuum from semantic to phonetic. The method is readily portable to other undeciphered corpora and contributes a generalizable tool for computational epigraphy.

6. CONCLUSION

This paper formalizes a two-track workflow for the Indus script. First, a conservative seed-based method isolates logographic candidates from singleton inscriptions and cautiously expands via contextual co-occurrence, validated by bigram statistics and tempered by semantic filtering. Second, an exclusivity-driven procedure identifies likely syllabic pairs from frequent, substitution-resistant bigrams. Distributional clustering relates signs and sign-pairs into putative lexical fields. The entire workflow is implemented with explicit computational parameters and exported in structured outputs (seed lists, qualified pairs, and distance matrices), emphasizing reproducibility.

The results demonstrate that, conclusively, around 180 signs in the corpus are logograms, and likely many more. However, these signs make up the majority of inscriptions, meaning that the majority of patterned Indus texts exclusively consist of logograms. The analysis also identified several hundred bigrams that could potentially represent complete semantic units themselves. These bigrams behaved like semantically complete signs. Interestingly, many bigrams that behaved similarly to one another shared signs, particularly suffixes. This implies that there could be suffixes in the script, possibly quantifiers. Lastly, several potential phonetic pairs were also identified in the script. Though hard to corroborate, these syllabic pairs could provide a path to decipherment since phonetic systems are much easier to decipher and decipherments can be computationally verified.

While no single analysis can claim decipherment, the methods here establish a durable foundation for subsequent linguistic inquiry. One area of future investigation would be applying the methods used to identify the list of logograms to more specific semantic categories in the script. The logogram analysis identified an initial seed list of logograms and identified signs that shared contexts with these initial seeds. The same process could be performed on other semantic categories, like numerals, nouns, measurements, etc. Other future work may incorporate site-specific strata, artifact-type filters, or temporal priors; extend exclusivity scoring to continuous measures; and integrate cross-linguistic testing where Near Eastern manifestations of Indus signs permit independent validation.

7. REFERENCES

[1] A. B. Mukhopadhyay. Semantic scope of Indus inscriptions comprising taxation, trade and craft licensing, commodity control and access control: Archaeological and script-internal evidence. Humanities and Social Sciences Communications, 10(1): 1–38, 2023. DOI: 10.1057/s41599-023-02320-7.

[2] A. Fuls. Ancient Writing and Modern Technologies—Structural Analysis of Numerical Indus Inscriptions. 2020.

[3] L. Merriam. Deciphering the Indus Script: A Novel Computational Analysis of High-Frequency Sign Pairs. Under Review. 2025.

[4] A. Parpola. Deciphering the Indus Script (Reissue ed.). Cambridge University Press, 2009.

[5] S. Sinha, A. M. Izhar, R. K. Pan, and B. K. Wells. Network analysis of a corpus of undeciphered Indus civilization inscriptions indicates syntactic organization. Computer Speech & Language, 25(3): 639–654, 2011. DOI: 10.1016/j.csl.2010.05.007.

[6] B. K. Wells. The Archaeology and Epigraphy of Indus Writing (UK ed. ed.). Archaeopress Archaeology, 2015.

[7] Bonta, S. C. The Indus Valley Script: A New Interpretation. Academia.edu, 2010.

[8] Sinha, S., Izhar, A. M., Pan, R. K., & Wells, B. K. (2011). Network analysis of a corpus of undeciphered Indus civilization inscriptions indicates syntactic organization. Computer Speech & Language, 25(3), 639–654. https://doi.org/10.1016/j.csl.2010.05.007

8. APPENDEX

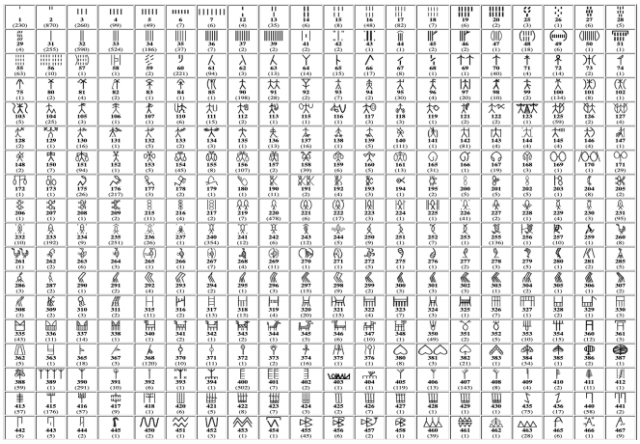

Figure 1: All sign sequences were taken from the Interactive Corpus of Indus Texts (ICIT)

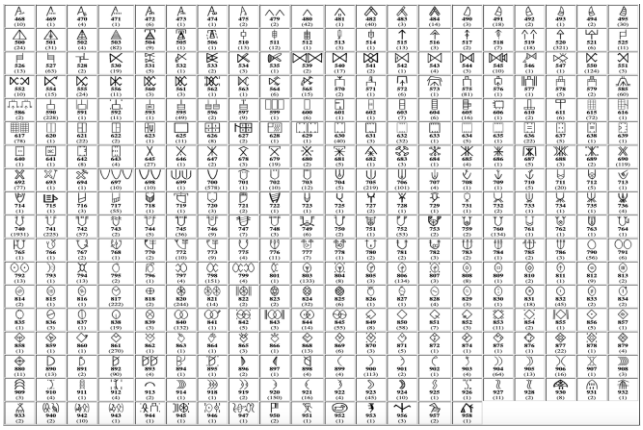

Figure 2: All sign sequences were taken from the Interactive Corpus of Indus Texts (ICIT) - Continued