Technical Report no. 08-11-02, Computer Science and Eng., U. of Washington, November 2008

A Markov Model of the 4500-year-old Indus Script

Rajesh P. N. Rao[a], Nisha Yadav[b], Mayank N. Vahia[b], Hrishikesh Joglekar[c], R. Adhikari[d], Iravatham Mahadevan[e]

[a] Department of Computer Science & Engineering
University of Washington, Seattle, WA 98195, USA

[b] Department of Astronomy & Astrophysics
Tata Institute of Fundamental Research, Mumbai 400005 and
Centre for Excellence in Basic Sciences, Mumbai 400098, India

[c] 14, Dhus Wadi, Laxminiketan, Thakurdwar, Mumbai 400002, India

[d] The Institute of Mathematical Sciences, Chennai 600113, India

[e] Indus Research Centre, Roja Muthiah Research Library, Chennai 600113, India

Although no historical information exists about the Indus civilization (fl. c. 2600-1900 BC), archaeologists have uncovered about 3800 short samples of a script that was used throughout the civilization. The script remains undeciphered, despite a large number of attempts and claimed decipherments over the past 80 years. Here, we propose the use of probabilistic models to analyze the structure of the Indus script. The goal is to reveal, through probabilistic analysis, syntactic patterns that could point the way to eventual decipherment. We illustrate the approach using a simple Markov chain model to capture sequential dependencies between signs in the Indus script. The trained model allows new sample texts to be generated, revealing recurring patterns of signs that could potentially form functional sub-units of an underlying language. The model also provides a quantitative way of testing whether a particular string belongs to the putative language as captured by the Markov model. Applying this test to Indus seals found in Mesopotamia and other sites in West Asia reveals that the script may have been used to express different content in these regions. Finally, we show how missing, ambiguous, or unreadable signs on damaged objects can be filled in with most likely predictions from the model. Taken together, our results indicate that the Indus inscriptions exhibit sequential structure and regularities that are suggestive of a linguistic rather than a nonlinguistic writing system.

Ancient scripts | Markov models | Statistical inference | Linguistic systems | Machine learning

From circa 2600-1900 BC, in a region spanning what is now Pakistan and northwestern India, flourished a vast civilization known as the Indus (or Harappan) civilization whose trade networks stretched all the way to the Persian Gulf and the Middle East. The civilization emerged from the forgotten depths of antiquity in the late 19th century when General Alexander Cunningham (1814-93) visited Harappa and published a description [1] of the site, including an illustration of a tiny seal with characters in an unknown script. Since then, much has been learned about the Indus civilization through the painstaking work of archaeologists (see [2,3] for reviews), but the script still remains an enigma.

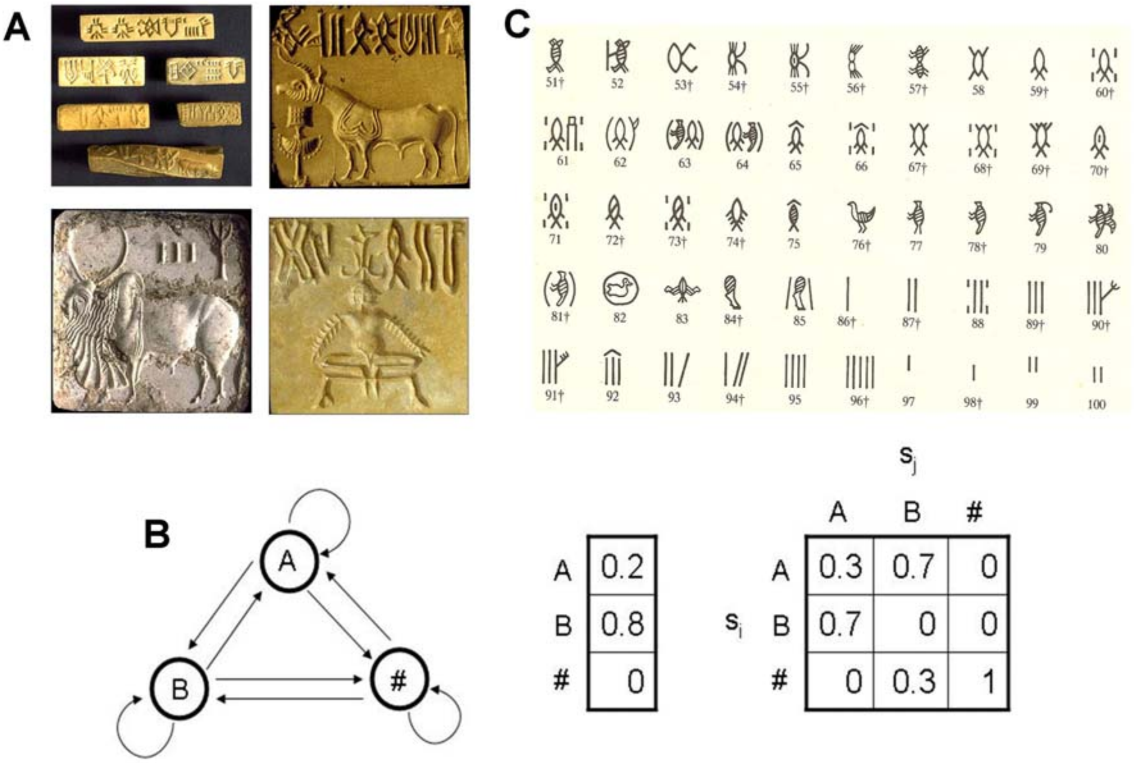
More than 3800 inscriptions in the Indus script have been unearthed on stamp seals, sealings, amulets, small tablets, and ceramics (see Fig. 1*A* for examples). A presumed monumental inscription has also been discovered [4]. Although there have been over 60 claimed decipherments (see [5] for a review), none of these decipherments has passed the rigor of scholarly scrutiny and none has been widely accepted by the community. Several obstacles to decipherment have been identified [5] including the lack of any bilinguals, the brevity of the inscriptions (the average inscription is about 5 signs long), and our almost complete lack of knowledge of the language(s) used in the civilization.

Given these formidable obstacles to direct decipherment, we propose instead to analyze the script's syntactical structure using techniques from the fields of statistical pattern analysis and machine learning [6]. It is our belief that such an approach could provide valuable insights into the grammatical structure of the script, paving the way for a possible eventual decipherment. As a first step in this endeavor, we present here results obtained from analyzing the sequential structure of the Indus script using a simple type of probabilistic graphical model known as a Markov model [6]. Markov models assume that the probability distribution of the current "state" (e.g., symbol in a text) depends only on the previous state. Although a simplification in many circumstances, the Markov assumption renders learning and inference over sequences tractable. Markov models have been successfully used in the analysis of time-series data, for example, in speech [7] and natural language processing [8]. Here, we apply them for the first time (to our knowledge) to analyzing an undeciphered ancient script.

Markov Models for Analyzing the Indus Script

Perhaps the simplest form of analysis one can perform on strings of symbols from an undeciphered script is to calculate the set of probabilities that a given symbol follows another. This idea is formalized in the concept of Markov models (also called Markov chains) [9,10]. A Markov model consists of a finite set of N "states" s_1, s_2, …, s_N (e.g., the states could be the signs in the script) and a set of conditional probabilities $P(s_i|s_j)$ that determine how likely it is that state s_i follows state s_j. There is also a set of prior probabilities $P(s_i)$ that denote the probability that state s_i starts a sequence. Figure 1*B* shows an example of a "state diagram" for a Markov model with 3 states labeled A, B, and #.

The circles denote the states and the arrows denote possible transitions between states. The column in the center gives the prior probabilities for each state and the table on the right provides the probabilities $P(s_i|s_j)$ for transition between states, picked arbitrarily here for the purposes of illustration. Suppose that A and B denote signs or letters in a language, and let # be the terminal sign that denotes the end of a text. Some example sequences generated by this Markov model are BAAB, ABAB, B, etc. (the terminal sign # is not shown). Texts that are *not* generated by this Markov model include all texts that contain a repetition of B (…BB…) and all texts that end in A, since these are precluded by the transition probability table. A more complex example would be a Markov model for English texts involving the 26 letters of the alphabet plus space. In this case, the transition probability table (or matrix) would be of size 27 x 27. In the matrix, we would expect, for example, higher probabilities for the letter "s" to be immediately followed by letters such as "e", "o", or "u" than letters such as "x" or "z" due to the morphological structure of words in English. The Markov model would thus capture such statistical regularities inherent in the English language.



$$\begin{array}{c|c|}
A & 0.2 \\
B & 0.8 \\
\# & 0
\end{array}
\qquad
\begin{array}{cc|c|c|c|}
 & & \multicolumn{3}{c}{s_j} \\
 & & A & B & \# \\
 & A & 0.3 & 0.7 & 0 \\
s_i & B & 0.7 & 0 & 0 \\
 & \# & 0 & 0.3 & 1
\end{array}$$

Fig. 1. The Indus Script and Markov Models. (A) Examples of Indus inscriptions on seals and tablets (top left) and three square stamp seals (images from harappa.com, J.M. Kenoyer, Courtesy Dept. of Archaeology and Museums, Govt . of Pakistan). Note that inscriptions appear reversed on seals. (B) Example of a simple Markov chain model with 3 states. (C) Subset of Indus signs from Mahadevan's list of 417 signs [11].

Markov models make the simplifying assumption that the current state depends only on the preceding state and is independent of states from other previous time steps given this preceding state. While this assumption may not hold true in many cases, Markov models (and their variants hidden Markov models or HMMs) have nevertheless been successfully used in speech recognition [7] and natural language modeling [8]. Markov models and HMMs are examples of a more general class of probabilistic models known as graphical models [12] in which random variables are modeled as nodes in a graph and dependencies between random variables are modeled using edges. Graphical models can be used to model complex relationships between states, including higher-order dependencies such as the dependence of a symbol on the past N-1 symbols (equivalent to N-gram models in language modeling).

In this article, we restrict our attention to first-order Markov models. In other work [13], we have compared higher-order (N-gram) models for the Indus texts using the information-theoretic measure of "perplexity" [8] and found that the bulk of the perplexity in the Indus corpus can be captured by a bigram (N = 2) or equivalently, a 1st-order Markov model.

To model strings using a Markov model, one needs to estimate the transition probabilities $P(s_i|s_j)$ between signs and the prior probabilities $P(s_i)$ from the corpus of data available. While $P(s_i)$ can be computed from frequencies of signs, the estimation of $P(s_i|s_j)$ requires "smoothing" i.e., transitions not observed in the data still get a small share of the probabilities rather than zeros (see, e.g., Chap. 6 in [14]). We use modified Kneser-Ney smoothing [15] (based on [16]), which has been shown to outperform other types of smoothing on benchmark datasets.

We focus in this article on three applications of Markov models to analyzing undeciphered scripts such as the Indus script: (1) *Sampling*: We show how new sample texts can be generated by: (i) randomly sampling from the prior probability distribution $P(s_i)$ to obtain a starting sign (say S_1) and then (ii) sampling from the transition probability distribution $P(s_i \mid S_1)$ to obtain the next sign S_2, and so on. The string generation process can be terminated by assuming an end of text (EOT) sign that transitions to itself with probability one. (2) *Likelihood computation*: If **x** is a string of length L and is of the form $x_1x_2\ldots x_L$, where each x_i is a sign from the list of signs, the likelihood that **x** was generated by a Markov model *M* can be computed as: $P(\mathbf{x}|M) = P(x_1x_2\ldots x_L|M) = P(x_1)P(x_2|x_1)P(x_3|x_2)\ldots P(x_L|x_{L-1})$. The likelihood of a string under a particular model tells us how closely the statistical properties of the given string match those of the original strings used to learn the model. Thus, if the original strings were generated according to the statistical properties of a particular language, the likelihood is useful in ascertaining whether a given string might have been generated by the same language. (3) *Filling in missing signs*: Given a learned Markov model *M* and an input string **x** with some known signs and some missing signs, one can estimate the most likely complete string **x*** by calculating the most probable explanation (MPE) for the unknown parts of the string (see *Methods*).

The Indus Script

A prerequisite for understanding the probabilistic structure of a script is to identify and isolate its basic signs. This task is particularly difficult in the case of an undeciphered script such as the Indus script because (a) there is considerable variability in the rendering of the signs by different scribes, making it difficult to ascertain whether two signs that look different are stylistic variants of the same sign or two independent signs, and (b) many signs appear to be ligatures, composed of a simpler sign modified by one or more auxiliary marks. After a painstaking analysis of the positional statistics of variant signs in the corpus of known inscriptions (circa 1977), Mahadevan arrived at a list of 417 independent signs in his concordance [11] (Parpola used similar methods to estimate a slightly shorter list of 386 signs [17]). We utilized this list of 417 signs as the basis for our study. Figure 1*C* shows a subset of signs from this list.

Although the script remains undeciphered, there is widespread consensus on the direction of writing in the script. Barring a few exceptions (see p. 14 in [11]), the writing direction is predominantly right to left (i.e., left to right in seals and right to left in the impressions). There exists convincing external and internal evidence supporting this claim (e.g., [5,11,17]). We consequently assumed a right-to-left direction for the writing and learned the sequential structure of the Indus texts based on a right-to-left reading of their signs, although a Markov model could equally well be learned for the opposite direction.

Results

Markov Model of Indus Texts

The learned Markov model provides several interesting insights into the nature of the Indus script. First, examining the learned prior probabilities $P(s_i)$ provides valuable information about how likely it is that a particular sign s_i starts a text. Figure 2*B* shows this probability for the ten most frequently occurring signs (Fig. 2*A*) in the corpus of Indus inscriptions.

An examination of Figure 2*B* reveals that certain frequently occurring signs such as and (signs numbered 3 and 10 in the figure) are much more likely to start a text than the others. On the other hand, certain signs such as (the most frequent sign in the corpus) and are highly unlikely to start a text (they are in fact highly likely to end a text – see Fig. 3*C* and Table 1). These observations are consistent with previous analyses by Mahadevan, Parpola, and others [11,17,18,19].

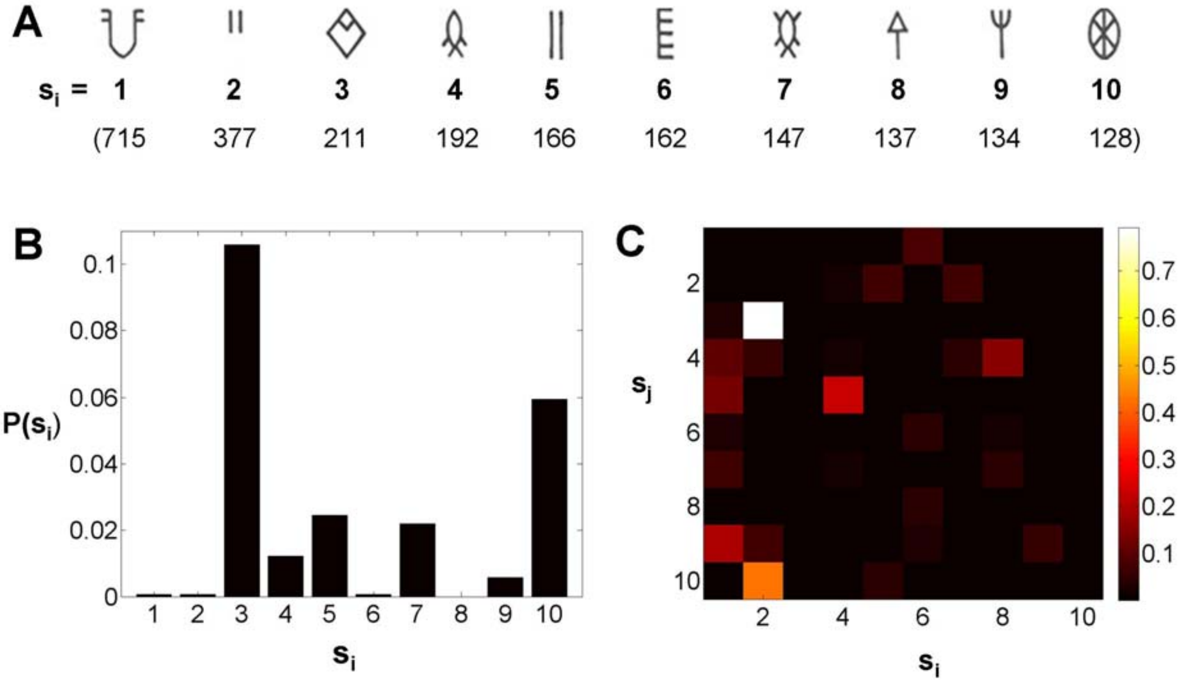


Fig. 2. Learned Markov Model for the Indus Script. (A) The ten most frequently occurring Indus signs in the dataset. The numbers within parenthesis in the bottom row are the frequencies of occurrence of each sign depicted above in the dataset used to train the model. (B) Prior (starting) probabilities P(s_i) learned from data, shown here for the ten most frequently occurring Indus signs depicted in (A). (B) Matrix of transition probabilities P($s_i|s_j$) learned from data. A 418x418 matrix was learned but to aid visualization, only the 10x10 portion corresponding to the ten most frequent signs is shown.

From the learned values for P(s_i), one can also extract the ten most likely signs to start a text and the ten least likely signs to do so (among signs occurring at least twenty times in the dataset) (Fig. 3*A*). These results suggest that some signs that look similar, such as and , subserve different functions within the script and thus cannot be considered variants of the same sign.

The matrix of transition probabilities P($s_i|s_j$) learned from data is shown in Figure 2*C* (only the portion corresponding to the ten most frequent signs is shown in the figure but all 418 x 418 probabilities were learned). The learned transition probabilities for Indus sign pairs can be shown to be significant, in the sense that the null hypothesis for independence can be rejected (see [13] for details). To interpret the transition probability matrix, consider the entry with the highest value (colored white). This value (approximately 0.8) corresponds to P(s_i = 2| s_j = 3) and indicates that the sign ◇ is followed by the sign || with a very high probability of about 80%. Other highly probable pairs and longer sequences of signs are shown in Figure 3*B*.

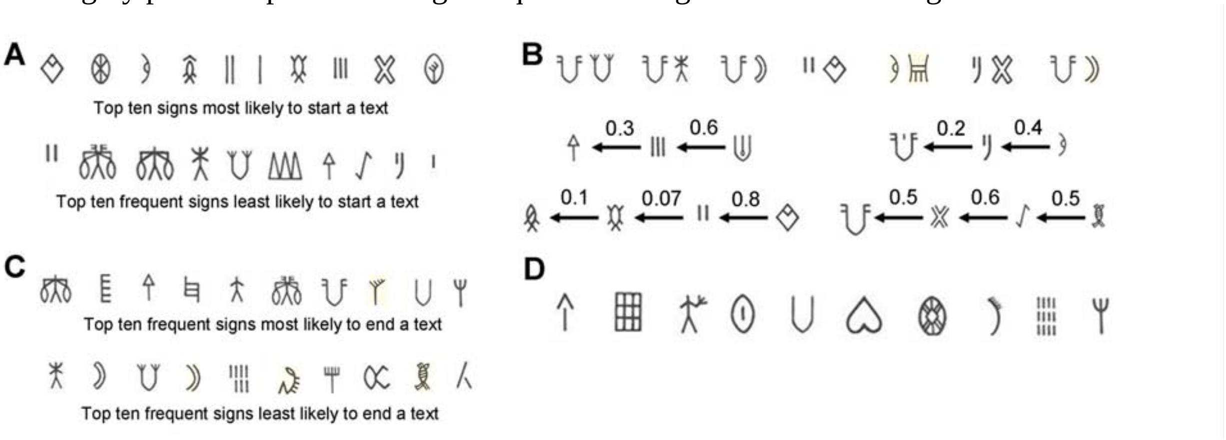

A
Top ten signs most likely to start a text
Top ten frequent signs least likely to start a text
B
0.3
0.6
0.2
0.4
0.1
0.07
0.8
0.5
0.6
0.5
C
Top ten frequent signs most likely to end a text
Top ten frequent signs least likely to end a text
D

Fig. 3. Some characteristics of the Indus script extracted from the learned Markov model. (A) Signs most likely and least likely to start an Indus text. (B) Highly probable pairs (top row) and longer sequences of Indus signs predicted by the learned transition matrix. The numbers above the arrows are the learned probabilities for the transitions indicated by the arrows. (C) Signs most likely and least likely to end an Indus text. (D) Top ten 10 signs with the highest probability of repeating (sign following itself).

The transition matrix also tells us which signs are most likely to end a text. This can be done by reading off the signs that have a high probability of being followed by the end of text symbol. Figure 3*C* (top row) shows ten signs that occur at least twenty times in the dataset which have high probabilities (in the range 0.87 to 0.39) of terminating a text. The bottom row in the figure shows ten frequently occurring signs (occurring at least twenty times in the dataset) that have the least probability of terminating a text.

The diagonal of the transition matrix, which represents the probability $P(s_i|s_i)$ of self-transitions, tells us how likely it is that a given sign follows itself (i.e., repeats). Figure 3*D* shows 10 signs with the highest probability of repeating. The sign 人 occurs in only one inscription where it occurs as a pair, while the sign ↑ occurs in 8 inscriptions, 6 times as a pair. More interestingly, the frequently occurring sign ⊞ occurs as a pair in 33 out of the 58 inscriptions in which it occurs. The presence of such repeating symbols in the Indus texts puts strong constraints on their semantic interpretation since the interpretation has to remain intelligible when repeated.

The dark regions in the transition matrix in Figure 2*C* indicate a transition probability that is near zero. Many entries can be expected to be near zero because of the small sample size of the training dataset compared to the size of the matrix (our use of modified Kneser-Ney smoothing ameliorates some of these effects). However, when the transition matrix is restricted to the top 20 to 100 most frequently occurring signs in the corpus, the number of near zero entries in transition matrix still ranges from 34% to 62% (Table S1), which is an intermediate range between random and rigid symbol order. This intermediate degree of flexibility in choosing the next sign can be quantified using conditional entropy and falls within the range of natural languages [20] (see also [21]). Such a structure in the transition matrix is suggestive of specific grammar-like rules governing the sequencing of signs in the Indus inscriptions. The existence of such rules would lend support to the hypothesis that the Indus script encodes linguistic information, rather than random or rigid juxtapositions of emblems or religious, heraldic, or political symbols [22].

Table 1 provides additional support for the hypothesis that sequences of signs in the Indus inscriptions may be governed by specific rules. The table shows, for the top ten most frequently occurring signs in the dataset, a list of signs that tend to be followed by or follow a given sign with some probability (given in parenthesis below). For each of the ten selected signs, only a small set of signs has a high probability of being followed by or following the given sign, hinting at specific rules governing sequential order in the texts.

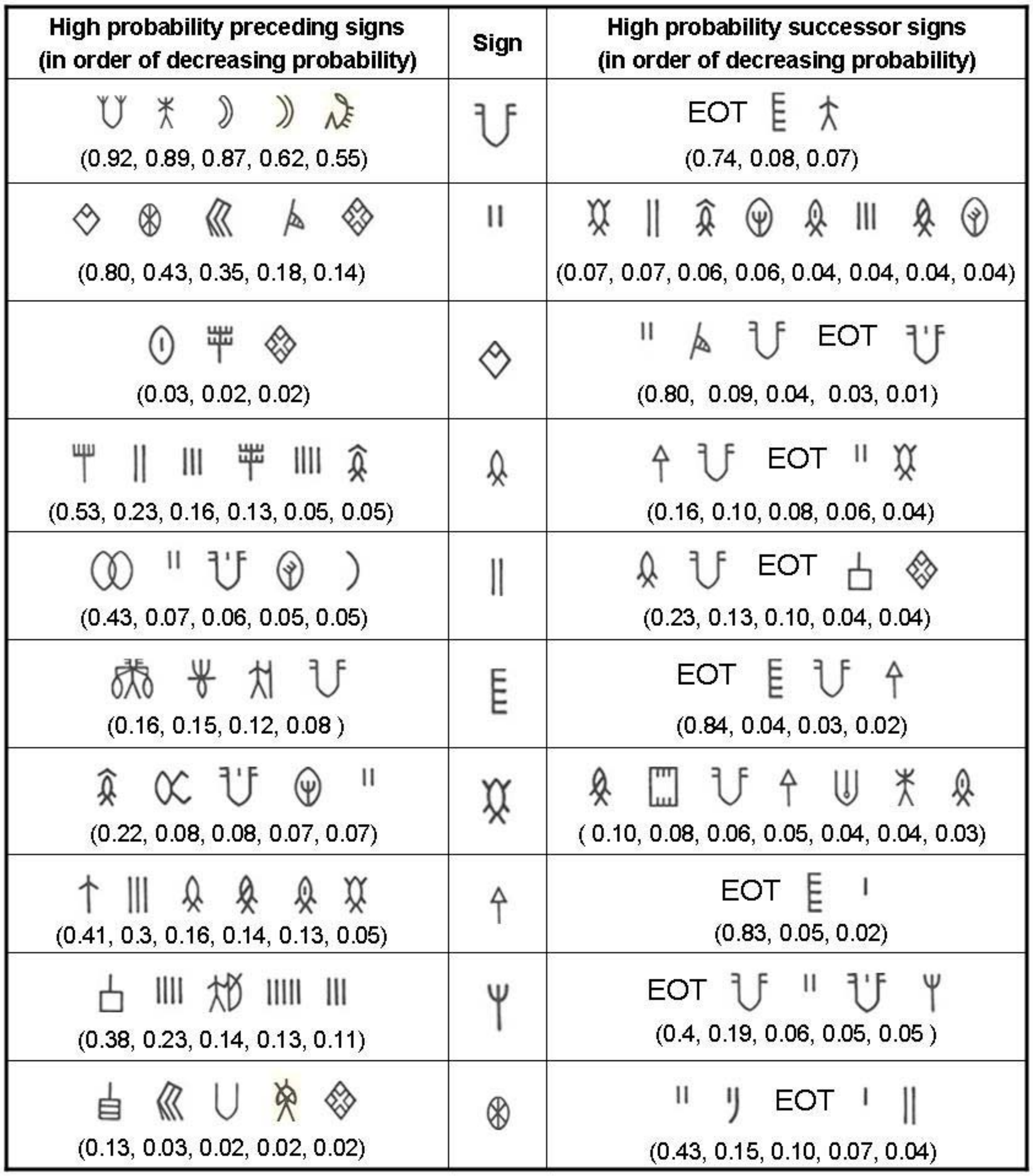
Generating New Sample Indus Texts from the Learned Model

The Markov model learned from the Indus inscriptions acts as a "generative" model in that one can sample from it and obtain new sequences of signs that conform to the sequential statistics of the original inscriptions (albeit limited to pairwise statistics). Figure 4*A* provides an example of a new text obtained by sampling and the closest matching text in the original corpus. The closest match was computed using the "string edit distance" between strings, which measures the number of additions, deletions, and replacements needed to go from one string to the other. The closest matching Indus text in Figure 4*A* is not identical to the generated sample but differs from it in two ways. First, the symbol occurs as the starting symbol instead of . An examination of the transition matrix reveals that both and have a high probability of being followed by the sign . Second, the sample text contains the sign instead of in the same position. This suggests that and perhaps have similar functional roles, given that they occur within similar contexts.

Figure 4*B* gives another example of a new generated Indus text (top) and two closest matching texts from the Indus dataset of inscriptions. Once again, based on their interchangeability in these texts, one may infer that the signs , , and share similar functional characteristics.

Filling-In Incomplete Indus Inscriptions

Many of the inscribed objects excavated at the various Indus sites are damaged, resulting in inscriptions that contain missing or illegible signs. To ascertain whether the model trained on complete texts could be used to fill-in the missing portion of these incomplete inscriptions, we first generated an artificial dataset of "damaged" inscriptions by taking complete inscriptions from the Indus dataset and obliterating one or more signs. Figure 4*C* (top row) shows an example of one such inscription. The complete inscription (middle row) predicted by the Markov model using the "most probable explanation" (MPE) method matched a pre-existing Indus inscription (bottom row). Results from a detailed cross-validation study of filling-in performance are given in [13].



High probability preceding signs (in order of decreasing probability)	Sign	High probability successor signs (in order of decreasing probability)
(0.92, 0.89, 0.87, 0.62, 0.55)		EOT (0.74, 0.08, 0.07)
(0.80, 0.43, 0.35, 0.18, 0.14)		(0.07, 0.07, 0.06, 0.06, 0.04, 0.04, 0.04, 0.04)
(0.03, 0.02, 0.02)		EOT (0.80, 0.09, 0.04, 0.03, 0.01)
(0.53, 0.23, 0.16, 0.13, 0.05, 0.05)		EOT (0.16, 0.10, 0.08, 0.06, 0.04)
(0.43, 0.07, 0.06, 0.05, 0.05)		EOT (0.23, 0.13, 0.10, 0.04, 0.04)
(0.16, 0.15, 0.12, 0.08)		EOT (0.84, 0.04, 0.03, 0.02)
(0.22, 0.08, 0.08, 0.07, 0.07)		(0.10, 0.08, 0.06, 0.05, 0.04, 0.04, 0.03)
(0.41, 0.3, 0.16, 0.14, 0.13, 0.05)		EOT (0.83, 0.05, 0.02)
(0.38, 0.23, 0.14, 0.13, 0.11)		EOT (0.4, 0.19, 0.06, 0.05, 0.05)
(0.13, 0.03, 0.02, 0.02, 0.02)		EOT (0.43, 0.15, 0.10, 0.07, 0.04)

Table 1. Signs that tend to be followed by or follow each of the top ten frequently occurring signs. (Note: Followed by and follow here assumes a right to left reading of the texts). The numbers in the parenthesis below each list of signs is the value from the transition probability matrix for the corresponding sign preceding (left column) or following (right column) a given sign (center column). EOT denotes “End of Text” meaning that the text or inscription ends at that juncture. Each list of signs was extracted from the learned transition matrix. Only signs that occur 20 or more times in the dataset are included in these lists.

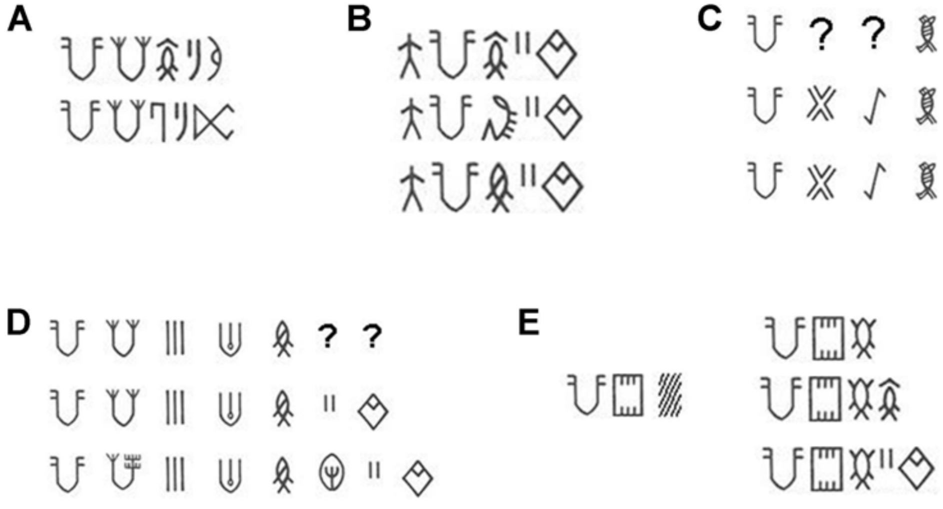
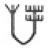
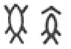


Fig. 4. Generating new Indus texts and filling in missing signs. (A) The top row shows a new sequence of signs (read right to left) that was generated by sampling from the learned Markov model. The lower row is the closest matching actual Indus inscription from the corpus used to train the model. (B) Inferring functionally-related signs. A sample from the Markov model (top) is compared to two closest matching inscriptions in the training dataset, highlighting signs that function similarly within inscriptions. (C) Filling-in missing signs in a known altered text. The inscription in the top row was produced by replacing two signs in a complete inscription with blanks (denoted by ?). The middle row shows the "most probable explanation" (MPE) output. The last row shows the closest matching text in the dataset. (D) Filling-in an actual incomplete Indus inscription. The inscription at the top is an actual Indus inscription (from [17]; Fig. 4.8) with two signs missing. The text in the middle is the MPE output and the text at the bottom is the closest matching complete Indus text in the corpus. (E) Filling in of another actual incomplete inscription from [11]. The text on the left has an unknown number of missing signs (hashed box). The right side shows three complete texts of increasing length predicted by the model. The first and third texts actually exist in the corpus.

Figure 4*D* (top row) shows an actual Indus text with missing signs (from [17]). The middle row shows the completed text generated by the MPE method, with the closest matching Indus text at the bottom. The generated text differs from the known matching text in two ways: the "modifier" is omitted and the sign is replaced by a visually related sign . The text shown at the left of Figure 4*E* is another actual Indus inscription with an unknown number of signs missing (from [11]). The three texts shown at the right are MPE outputs assuming one, two, or three signs are missing. The first and third MPE texts actually occur in the Indus corpus while the middle text contains the frequent pair . Additional examples of filling-in of damaged texts are given in *Supplementary Information* (Table S2).

Testing the Likelihood of Indus Inscriptions

We also computed the likelihood of particular sequences of Indus signs with respect to the learned Markov model. The computed likelihood tells us how likely it is that a particular sequence of signs belongs to the putative language model encoded by the Markov model. Altering the order of signs in an existing Indus text causes the likelihood of the text to drop dramatically (see Fig. S1 for an example), supporting the hypothesis that Indus texts may be subject to specific syntactic rules for sequencing of signs.

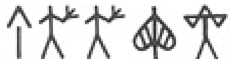
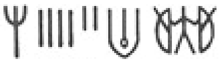
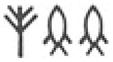
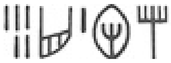
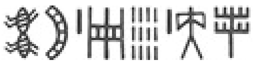
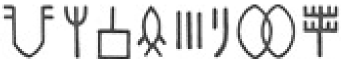
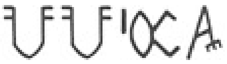
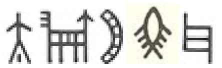
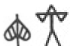
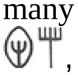
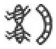
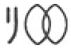
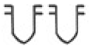
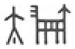
West Asian Text (from [11])	**Likelihood**
	0
	2.71 x 10^{-10}
	6.32 x 10^{-8}
	4.66 x 10^{-14}
	0
	8.82 x 10^{-12}
	1.20 x 10^{-12}
	2.22 x 10^{-17}
Indus valley held-out texts (median)	6.85 x 10^{-8}

Table 2. Likelihood of West Asian Texts compared to Indus valley texts. Only complete and unambiguous West Asian texts from [11] are included in this analysis. Two texts have a likelihood of zero because they each contain a symbol not occurring in our training dataset. The last row shows for comparison the median likelihood value for a randomly selected set of eight texts originating from within the Indus valley which were held out and not used for training the Markov model.

Applying this analysis to Indus texts discovered outside of the Indian subcontinent, for example, in Mesopotamia and other sites in West Asia, we find that the likelihoods of most of these inscriptions are extremely low compared to their counterparts found in the Indus valley (Table 2). Indeed, the median value of likelihoods for the West Asian texts is 6.22 x 10^{-13}, about 100,000 times less than the median value of 6.85 x 10^{-8} obtained for a random set of texts of Indus valley origin that were excluded from the training set for comparison purposes. These findings suggest the intriguing possibility that the Indus script may have been used to represent a different language or subject matter by Indus traders living or conducting business in West Asia. Such a possibility has already been suggested by Parpola, who noted that the West Asian texts often contain unusual sign combinations [17]. In fact, in Table 2, many of the West Asian texts with the low likelihoods contain sign combinations such as , , , , , and  that never appear in any texts found in the Indus valley.

Discussion

A number of researchers have made observations regarding sequential structure in the Indus script, focusing on frequently occurring pairs, triplets, and other groups of signs [11,17,18,19,23]. Koskenniemi suggested the use of pairwise frequencies of signs to construct syntax trees and segment texts, with the goal of eventually deriving a formal grammar. More recently, Yadav, Vahia and colleagues [18,19] have performed statistical analyses of the Indus texts, including explicit segmentation of texts based on most frequent pairs, triplets and quadruplets.

In this article, we provide to our knowledge the first investigation of sequential structure in the Indus script based on Markov models. An analysis of the transition matrix learned from a corpus of 1548 Indus texts provided important insights into which signs tend to follow particular signs and which signs do not, revealing interesting patterns in the script that are unlikely to occur if the script was merely used to represent religious, political, or heraldic symbols in random or rigid linear order [22]. The transition matrix also provides a quantitative probabilistic basis for extracting common sequences and subsequences of signs in the Indus texts. We demonstrated how the learned Markov model can be used to generate new sample texts, revealing groups of signs that tend to function similarly within a text.

The approach we have proposed can also be used to fill-in missing portions of illegible and incomplete Indus inscriptions by computing the "most probable explanation" for the data based on the corpus of complete inscriptions. Finally, a comparison of the likelihood of Indus inscriptions discovered in West Asian sites with those from the Indus valley suggests that many of the West Asian inscriptions may represent subject matter different from Indus valley inscriptions.

Our results favor the hypothesis that the Indus script represents a linguistic writing system. Our Markov analysis of sign sequences makes it clear that the signs do not occur in a random manner within inscriptions but appear to follow certain rules: (1) some signs have a high probability of occurring at the beginning of inscriptions while others almost never occur at the beginning, and (2) for any particular sign, there are signs that have a high probability of occurring after that sign and other signs that have negligible probability of occurring after the same sign. Furthermore, signs appear to fall into functional classes in terms of their position within an Indus text, where a particular sign can be replaced by another sign in its equivalence class. Similar conclusions have been arrived at by others using different methods [17,18,19]. For example, Yadav, Vahia and colleagues have shown [19] that larger Indus texts can be effectively split into smaller segments that do not necessarily form complete standalone texts but need to be enclosed within specific signs or sign sets to make them complete. A range of other arguments in favor of the linguistic hypothesis for the Indus script are provided by Parpola [24].

Our study suffers from some shortcomings that could be addressed in future work. First, our 1st-order Markov model only captures pairwise dependencies between signs, ignoring important longer-range dependencies. Although much of the "perplexity" [8] in the Indus corpus can be captured by a 1st-order Markov model, indicating that the most important correlations in Indus sign sequences come from the immediately preceding sign, additional reduction in perplexity can be obtained by considering triplets and quadruplets of signs [13]. Thus, higher-order Markov models represent a promising direction for future work.

A second potential shortcoming is our use of an Indus corpus of texts from 1977 [11] which does not include many new texts and signs that have since been discovered. Also, new sign lists have been suggested with up to 650 signs [25]. We believe these additions and variations will have only a minor perturbative effect on the structural analysis presented in the paper for the following reasons: (1) The types of new material that have been discovered are in the same categories as the texts in the 1977 corpus, namely, more seals, tablets, etc. The new material thus exhibits syntactic structure that is similar to the material we have analyzed in this paper. (2) The new signs that have been discovered are still far outnumbered by the most commonly occurring signs in the 1977 corpus, most of which also occur frequently in the newly discovered material. Thus, variations in sign frequencies due to the new material will only slightly change the conditional probabilities in the Markov model. Nevertheless, a more complete analysis with all known texts and new sign lists remains an important objective for the future. Additionally, our analysis combined data from different geographical locations. A more detailed site-by-site analysis could shed light on the interesting question of whether there are differences in the sequential patterning of signs across regions.

Finally, it should be noted that the corpus of Indus texts currently available likely represents only part of the total corpus of Indus writing. Evidence for one text on perishable material (a wooden signboard) has been found archaeologically [4]. This opens up the possibility that there are more. The texts in our dataset are almost entirely from special purpose artifacts such as seals and tablets. They therefore may not fully capture the breadth of Indus writing, but rather represent only a small sample consisting of terse statements such as names, titles, occupations etc. related to seals and tablets. The large number of Indus sites yet to be fully excavated leave open the possibility that new texts will be discovered that expand the breadth of subject matter represented by the script.

In summary, the results we have presented strongly suggest the existence of specific rules governing the sequencing of Indus signs in a manner that is indicative of a grammar. A formidable but perhaps not insurmountable challenge for the future (also articulated in [17,18,23]) is to apply statistical and machine learning techniques to infer a grammar directly from the corpus of available Indus inscriptions.

Methods

Dataset

We applied our Markov model analysis techniques to a subset of Indus texts extracted from Mahadevan's 1977 concordance [11]. This dataset, called EBUDS [18], excludes all texts from Mahadevan's concordance containing ambiguous or missing signs and all texts having multiple lines on a single side of an object. In the case of duplicates of a text, only one copy is kept in the dataset. This resulted in a dataset containing 1548 lines of text, with 7000 sign occurrences. We used Mahadevan's list of 417 signs plus an additional "End of Text" (EOT) sign to denote the end of each text. Signs were fed to the model from right to left in each line of text, ending in the EOT sign.

Learning a Markov model from data

The parameters of a Markov model include the prior probabilities $P(s_i)$ and the transition probabilities $P(s_i|s_j)$. A simple method for computing these probabilities is counting frequencies, e.g. $P(s_i)$ is set equal to the number of times sign s_i occurs in the dataset divided by the total number of occurrences of all signs. This can be shown to be equivalent to maximum likelihood estimation of the parameters [8]. However, such an estimate, especially for $P(s_i|s_j)$, can yield poor estimates when the dataset is small, as is the case with the Indus script, since many pairs of signs may not occur in the small dataset at all, even though their actual probability may be nonzero. There has been extensive research on “smoothing” techniques which assign small probabilities to unseen pairs based on various heuristic principles (see Chap. 6 in [8] for an overview). For the results in this paper, we used modified Kneser-Ney smoothing, a technique that has been shown to outperform other smoothing techniques on a number of benchmark datasets [15].

Filling in missing signs

Let $\mathbf{x} = x_1x_2\ldots x_L$ be a string of length L where each x_i is a random variable whose value can be any sign from the list of signs. Let X denote the set of x_i for which the values are given and Y the set of x_i with values missing. For a general graphical model M, the most probable explanation (MPE) for the missing variables Y given the “evidence” X is computed as $Y^* = \arg\max_Y P(Y|X,M)$. For the learned Markov model, $M = (\alpha, \pi)$, where $\alpha_i = P(s_i)$ and $\pi_{ij} = P(s_i|s_j)$. For a Markov model, we can compute the most probable explanation (MPE) $Y^* = \arg\max_Y P(Y|X, \alpha, \pi)$ using a version of the “Viterbi algorithm” [26], which is itself based on the broader technique of dynamic programming (see [7,10] for algorithmic details).

ACKNOWLEDGMENTS. This work was supported by the Packard Foundation, the Simpson Center for the Humanities at the University of Washington (UW), the UW College of Arts and Sciences, the UW College of Engineering and CSE Dept., Sir Jamsetji Tata Trust, and the Indus Research Centre at the Roja Muthiah Research Library, Chennai.

References

1. Cunningham A (1875) *Archaeological Survey of India Report for the Year 1872-73* (Archaeological Survey of India, Calcutta).
2. Kenoyer JM (1998). *Ancient cities of the Indus Valley Civilisation* (Oxford University Press).
3. Possehl GL (2002). *The Indus Civilisation* (Alta Mira Press, Walnut Creek).
4. Bisht RS (1990) Dholavira: New horizons of the Indus Civilization. *Puratattva,* 20:71–82.
5. Possehl GL (1996) *The Indus Age: The Writing System* (University of Pennsylvania Press, Philadelphia, PA).
6. Bishop C (2008) *Pattern Recognition and Machine Learning* (Springer Verlag).
7. Jelenik F (1997) *Statistical* Methods for *Speech* Recognition (MIT Press, Cambridge, MA).
8. Manning C, Schütze H (1999) *Foundations of* Statistical Natural *Language Processing* (MIT Press. Cambridge, MA).
9. Drake AW (1967) *Fundamentals of Applied Probability Theory* (McGraw-Hill, New York).
10. Rabiner LR (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* 77(2):257–286.
11. Mahadevan I (1977) *The Indus Script: Texts, Concordance, and Tables* (Memoirs of Archaeological Survey of India, New Delhi).
12. Jordan MI (2004) Graphical models*. Statistical Science (Special Issue on Bayesian Statistics)* 19:140-155.

13 Yadav N, Joglekar H, Rao RPN, Vahia MN, Mahadevan I, Adhikari R, Statistical analysis of the Indus script using n-grams. arxiv.0901.3017 (2009).

14 Manning C and Schütze H, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press (1999).

15 Chen SF and Goodman J, Harvard University Computer Sci. Technical Report TR-10-98 (1998).

16 Kneser R and Ney H, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pages 181-184 (1995).

17. Parpola A (1994) *Deciphering the Indus script*. (Cambridge University Press, Cambridge).
18. Yadav N, Vahia MN, Mahadevan I, Joglekar H (2008) A statistical approach for pattern search in Indus writing. *International Journal of Dravidian Linguistics* 37(1):39-52.
19. Yadav N, Vahia MN, Mahadevan I, Joglekar H (2008) Segmentation of Indus Texts. *International Journal of Dravidian Linguistics* 37(1):53-72.

20 Rao RPN, Yadav N, Joglekar H, Adhikari R, Vahia MN, Mahadevan I, Entropic evidence for linguistic structure in the Indus script. *Science*, April 2009.

21 Schmitt AO and Herzel H, Estimating the Entropy of DNA Sequences. *J. Theor. Biol.* **188**, 369 (1997).

22 Farmer S, Sproat R, and Witzel M, The Collapse of the Indus-Script Thesis: The Myth of a Literate Harappan Civilization. *Electronic Journal of Vedic Studies* **11**, 19 (2004).

23. Koskenniemi K (1981) Syntactic methods in the study of the Indus script. *Studia Orientalia* 50:125-136.
24. Parpola A (2008) Is the Indus script indeed not a writing system? in *Airavati: Felicitation volume in honor of Iravatham Mahadevan* (Varalaaru.com publishers, Chennai, India) pp. 111-131.

25 Wells BK, *Epigraphic Approaches to Indus Writing* (PhD Dissertation, Harvard University). Cambridge, MA: ASPR Monograph Series (2009).

26. Viterbi AJ (1967) Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory* 13:260-269.

Supplementary Information

Sparseness analysis of the transition matrix

We define sparseness as the percentage of probability values in a transition matrix that are near zero (below a threshold of 0.0024 for the results below). We chose 0.0024 based on the fact that uniform probability of succession given the set of 417 Indus signs is 1/417 = 0.0024. For random strings where any sign has a non-zero probability of following any other, the sparseness value will be relatively low. On the other hand, if the sequencing is rigid and a sign can only be followed by another unique sign, the matrix will contain many values near zero and have a high sparseness value. One caveat is that the full transition matrix (417x417 = 173,889 entries) can be expected to be quite sparse because of the small size of the dataset (1548 lines of text with about 7000 sign occurrences for 417 signs) -- this implies that many pairs of signs will not occur in the dataset at all (our use of modified Kneser-Ney smoothing ameliorates this to some extent). However, the matrix need not be sparse when restricted to the most frequently occurring signs in the dataset. We therefore computed the sparseness of the transition matrix for Indus texts when only the top 20, 40, 60, 80, or 100 most frequently occurring signs are considered. The results, shown in Table S1, indicate that the sparseness value for the Indus transition matrix falls in the intermediate range between random and rigid symbol order, i.e., there is some flexibility in choosing the next sign but not too much flexibility. This is also the case for natural languages (Refs [20] and [21] in the main text) and hints at possible grammatical rules underlying the sequencing of signs in the Indus texts.

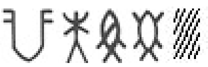
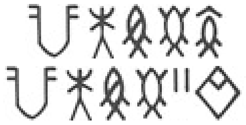
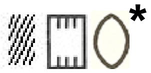
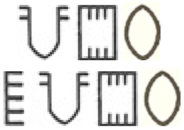
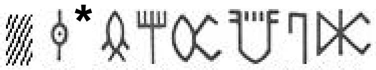
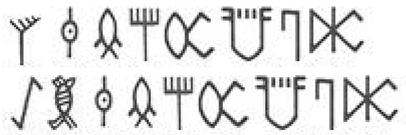
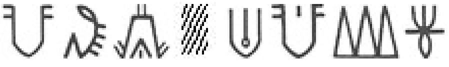
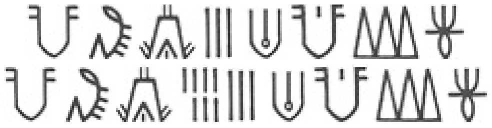
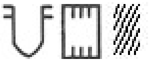
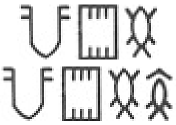
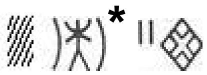
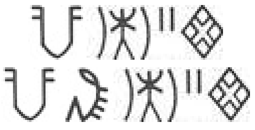
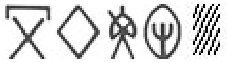
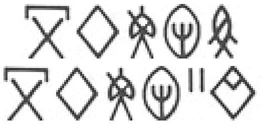
Number of Frequent Signs	Sparseness of Transition Matrix
20	34%
40	44%
60	49%
80	55%
100	62%

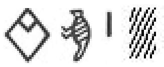
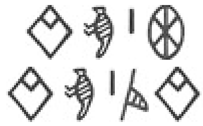
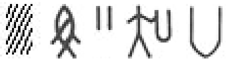
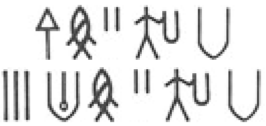
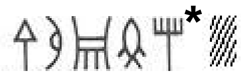
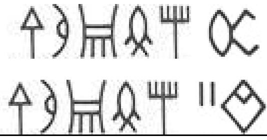
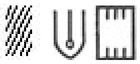
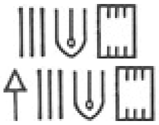
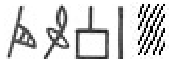
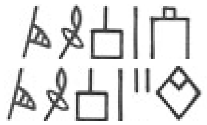
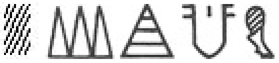
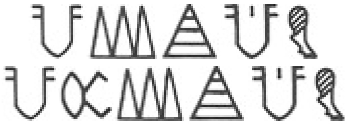
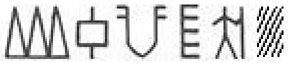
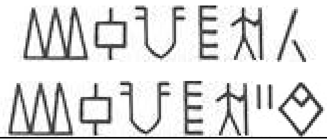
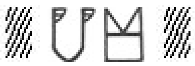
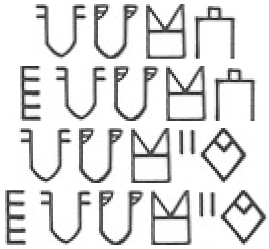
Table S1. Sparseness of transition matrix as a function of number of most frequent signs.

Examples of damaged texts from Mahadevan's concordance filled in using MPE

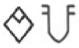
The table below shows complete texts predicted using MPE for a number of damaged texts given in Mahadevan's concordance (Ref [11] in the main text). The Text Numbers below are from the concordance. ▨ denotes one or more missing signs. * denotes a doubtfully read sign. Since the exact number of missing signs can be obscured by the damage to the object, filled-in texts are shown for the cases of 1 and 2 signs missing for each ▨.

Table S2. Filling in of damaged texts from Mahadevan's concordance.

Text No.	Text with Missing Signs	Filled-In Texts
1001		
1002		
1017		
1020		
1034		
1059		
1111		

1144		
1165		
1176		
1240		
1265		
1289		
1343		
1408		

Likelihood of Altered and West Asian texts

To test whether the model could be used to ascertain whether a text belongs to the putative language encoded by the Markov model, we generated several example texts by altering existing Indus texts – alterations involved randomly switching the positions of signs within a text. Figure S1 shows an example of an Indus valley text, the same text altered (last sign moved to the beginning), and a West Asian text. The likelihoods for each according to the learned model are given on the right. The altered text has a very small probability (4.4 x 10^{-10}), consistent with the observations that rarely begins a text and the sign pair  almost never occurs.

Interestingly, the West Asian text is about a 100 times even less likely than the altered text under the learned model, suggesting that the script was used to possibly encode different subject matter in West Asia. Additional examples of West Asian texts and their likelihoods are given in Table 2 in the article.

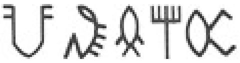
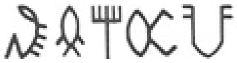
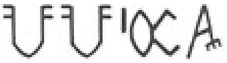
		Likelihood
Indus text		7.3×10^{-6}
Altered text		4.4×10^{-10}
West Asian seal		1.20×10^{-12}

Fig. S1. Comparison of Likelihoods of Example Indus, Altered, and West Asian Texts.