

Probabilistic Analysis of an Ancient Undeciphered Script

➔ **Rajesh P.N. Rao**, *University of Washington*



Probabilistic methods for analyzing sequences are providing new insights into the 4,000-year-old undeciphered script of the Indus civilization.

In the latter half of the 19th century, railway workers in British India found an almost inexhaustible supply of precisely cut baked bricks at Harappa, a small town located in present-day Pakistan. They proceeded to use the bricks as ballast for laying down 100 miles of railroad track. Little did they know that these bricks were more than 4,000 years old, products of South Asia's oldest urban civilization.

The Indus civilization, so named because its first important sites were discovered along the Indus River, flourished from 2600 to 1900 BC. At its peak, it encompassed more than 1 million square kilometers and was larger than the roughly contemporaneous Egyptian and Mesopotamian civilizations. Its cities were laid out in a grid-like pattern with a sophisticated water management and drainage system that would be the envy of many towns today. Citizens of the Indus civilization were highly enterprising, traveling to lands as far away as the Persian Gulf and Mesopotamia (present-day Iraq) to trade.

Remarkably, there is no evidence that powerful kings or elites ruled the

Indus cities, as in other Bronze Age civilizations. No extravagant royal palaces, pyramids, or ziggurats have been found. What archaeologists have unearthed in large numbers are tiny seals like those shown in Figure 1a, most around 1" × 1" in size. Each typically depicts an expertly crafted animal, with a short text of signs at the top. These texts, which also appear on miniature tablets, copper plates, tools, weapons, and pottery, constitute the Indus script, one of the last remaining undeciphered scripts of the ancient world.

THE INDUS SCRIPT

Figure 1b shows a small subset of the approximately 400 signs in the Indus script. The number of signs is more than in purely alphabetic or syllabic scripts, which typically contain a few dozen signs, but less than in logographic scripts such as Chinese, which contain large numbers of signs representing entire words. Researchers have therefore suggested that, like other ancient scripts such as Sumerian and Mayan, the Indus script was logosyllabic in nature, each sign

representing either a word or a syllable.

What the Indus signs actually mean remains a mystery, although the number of books claiming to have deciphered the script could occupy several bookshelves. None of these claims have been widely accepted. The major impediments to decipherment include

- the brevity of existing Indus texts—the average text length is about five signs while the longest text consists of 17 signs;
- our almost complete lack of knowledge of the language spoken by the Indus people; and
- the lack of a bilingual document such as the Rosetta Stone, which was instrumental in deciphering the Egyptian hieroglyphic script.

Given such formidable obstacles, efforts to decipher the script have ranged from inspired guesswork to ideology-driven speculation.

An alternate, more objective approach is to first analyze the script's syntactic structure, in the hope that

such an analysis could eventually lead to decipherment. Are the symbols in Indus texts randomly ordered or do they follow specific rules? Do particular symbols have particular positions within texts? How much flexibility does the script allow when composing a string of symbols? How do the Indus script's syntactical properties compare with those of other ancient and modern languages and scripts? Researchers are investigating such questions using statistics, probabilistic reasoning, and machine learning.

EARLY STATISTICAL ANALYSIS

G.R. Hunter conducted the first rudimentary statistical analysis of the Indus script in the early 1930s. In the absence of computers, Hunter hand-enumerated frequently occurring clusters of signs, segmenting Indus texts into short “words” of two or more signs. This enabled him to infer important syntactic characteristics of the script such as the tendency of certain symbols and words to occur at specific positions within texts. For example, Hunter was among the first to note that the “jar” sign \mathcal{U} , which is the most frequently occurring sign in the texts, acts as a “word ender,” and that the “fish” signs frequently occur in pairs (such as $\mathcal{A}\mathcal{X}$ and $\mathcal{X}\mathcal{A}$), occupying the same relative position within texts.

In the 1960s, the fact that sign clusters have particular positions within Indus texts was confirmed independently with the help of computers by a Finnish team led by Asko Parpola and a Soviet team led by Yuri Knorozov (who played a key role in deciphering the Mayan script). More recent work has demonstrated that the frequency of certain two-, three-, and four-sign combinations is much higher than would be expected by chance, and that a majority of the texts longer than five signs can be segmented into these smaller, frequently occurring sign combinations (N. Yadav et al., “Segmentation of Indus Texts,” *Int'l*

J. Dravidian Linguistics, vol. 37, no. 1, 2008, pp. 53-72). Such regularities point to the existence of distinctive syntactic rules underlying the Indus texts.

MARKOV AND N-GRAM MODELS

The presence of statistically significant clusters of symbols with positional preferences suggests that there is sequential order in the Indus script. One way to capture such sequential order is to learn a Markov model for the script from available texts.

The simplest (first-order) model estimates the transition probabilities $P(s_i|s_j)$ that sign i follows sign j . The obvious way of estimating $P(s_i|s_j)$ is to count the number of times sign i follows sign j , an approach equivalent to maximum likelihood estimation. However, given that there are approximately 400 signs and only a few thousand texts, a large number of sign pairs will have a frequency of 0

even though their actual probability may not necessarily be 0. This is a common problem in statistical language modeling and can be addressed using *smoothing* techniques.

A prominent smoothing technique, the modified Kneser-Ney algorithm, was used to learn a first-order Markov model of the Indus script (R.P.N. Rao et al., “A Markov Model of the Indus Script,” *Proc. National Academy of Sciences*, vol. 106, no. 33, 2009, pp. 13685-13690). The data for training the model came from Iravatham Mahadevan's *The Indus Script: Texts, Concordance and Tables* (Archaeological Survey of India, 1977). Once trained, the Markov model can be used to generate new samples of Indus texts. This can reveal interesting subunits of grammatical structure and recurring patterns, as Figure 2a shows.

There exist a large number of damaged Indus seals, tablets, and other artifacts that contain texts with one or more missing or illegible signs. A Markov model of the Indus texts can

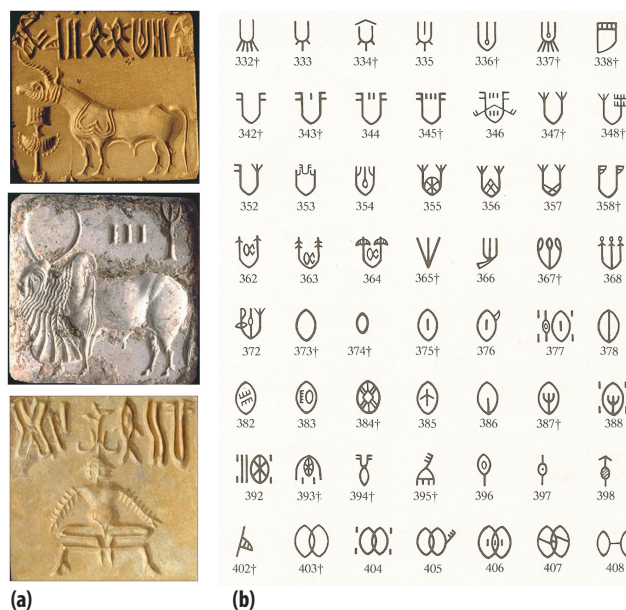


Figure 1. Indus script. (a) Three examples of square stamp seals, each with an Indus text at the top (image credit: J.M. Kenoyer/Harappa.com). Texts were usually written from right to left (inferred, for example, from writing on pottery where a sign is overwritten by another on its left) but this direction was reversed in seals (that is, left to right as in these images) to form correctly oriented impressions. (b) A small subset of the 400 or so signs in the Indus script (selected from Mahadevan's concordance).

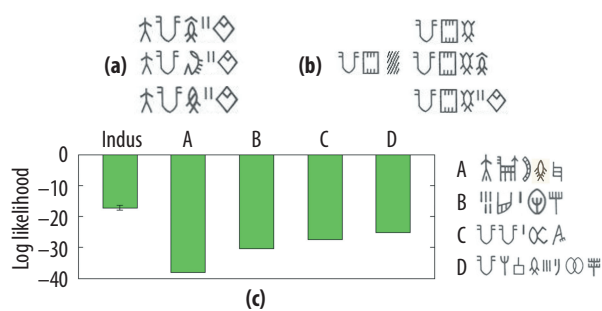


Figure 2. Markov model of the Indus script. (a) (Top) A new Indus text generated by the Markov model. (Below) Two closest matching texts in the training corpus. (b) (Left) Text from a damaged seal containing one or more missing signs (indicated by the shaded box). (Right) Three possible restorations predicted by the Markov model. The first and third texts actually exist in the corpus. (c) Log likelihood under the Markov model for four texts (A through D) found in foreign lands compared to average log likelihood for a random set of 50 Indus region texts not included in the training data (error bar denotes ± 1 standard error of mean). The 50 Indus region texts had the same average length as the foreign texts.

be used to predict these missing or illegible signs. The first-order Markov model was found to be surprisingly good at predicting signs deliberately obliterated for testing purposes, performing at a 75 percent accuracy level in a fivefold cross-validation study. Figure 2b shows an example of restoration of an actual damaged Indus inscription from Mahadevan's concordance, as suggested by the first-order Markov model.

Several seals with Indus signs have been discovered outside the Indus region, as far away as Mesopotamia and the Persian Gulf. One can compute the likelihood of these "foreign" texts with respect to a Markov model trained only on texts from the Indus region. As Figure 2c shows, the likelihood values for many of these foreign texts are several orders of magnitude lower than those for Indus region texts, indicating their low probability of belonging to the same language.

Indeed, an examination of these foreign texts reveals that although they contain commonly used Indus signs, the sequential order of the signs differs dramatically from that in texts originating in the Indus region—for example, the sequence U^fU in the foreign text C in Figure 2c never occurs on an Indus seal. This suggests that

Indus traders in foreign lands may have used the script to represent different content, such as foreign names or goods, or an altogether different language.

More recent work examined the utility of higher-order N -gram models. An N -gram model is essentially an $(N - 1)$ th-order Markov chain where the transition probability depends on the previous $N - 1$ symbols instead of just the previous symbol. The results suggest that a bigram model ($N = 2$) captures a significant portion of the syntax, with trigrams and quadrigrams making more modest contributions (N. Yadav et al., "Statistical Analysis of the Indus Script Using N -Grams," *PLoS One*, to appear in 2010).

THE LANGUAGE QUESTION AND ENTROPIC ANALYSIS

The brevity of existing Indus inscriptions and other attributes, such as the low frequency of many Indus signs, has prompted some to propose that the Indus script is not a script at all but instead is a collection of religious or political symbols. Adherents of the "non-script" thesis have likened the Indus script to nonlinguistic systems such as traffic signs, medieval heraldry,

markings on pottery in the Vinča culture of southeastern Europe, and carvings of deities on boundary stones in Mesopotamia.

Interestingly, this is not the first time that a script of a major ancient civilization has been deemed to be nonlinguistic. The Mayan script was long considered not to be a writing system at all until Knorozov and others finally worked out the rich phonetic underpinnings of the script in the 1950s and 1960s, revealing it to be a fully functional writing system.

Several key features of the Indus script suggest that it represents language:

- the texts are usually linear, like the vast majority of linguistic scripts and unlike nonlinguistic systems such as heraldry or traffic signs;
- symbols are modified by the addition of specific sets of marks over, around, or inside a symbol, much like later Indian scripts that use such marks to modify the sound of a root consonant or vowel symbol;
- the script possesses rich syntactic structure, with particular signs or clusters of signs preferring particular positions within texts, similar to linguistic sequences;
- the script obeys the Zipf-Mandelbrot law, a power-law distribution on ranked data, which is often considered a necessary (though not sufficient) condition for language; and
- texts found in Mesopotamia and the Persian Gulf use the same signs as texts found in the Indus region but alter their ordering, suggesting that the script was versatile enough to represent different subject matter or a different language.

Such attributes are hard to reconcile with the thesis that the script

merely represents religious or political symbols.

Further evidence for the Indus script's linguistic nature comes from quantitative studies comparing the entropy of the Indus texts with that of various languages. In some non-linguistic systems, such as the Vinča system, the signs do not seem to follow any order and appear to be juxtaposed randomly. Other non-linguistic systems, such as deities carved on Mesopotamian boundary stones, exhibit a rigid order reflecting, for example, the hierarchical order of the deities.

In languages, on the other hand, sequences of words and characters exhibit a degree of order intermediate between random and rigid. This intermediate degree of randomness arises from the grammatical rules and morphological structure of languages. The degree of randomness in a sequence can be measured quantitatively using entropy.

The smoothed first-order Markov model can be used to compute conditional entropy, which measures the average flexibility allowed in choosing the next sign given a preceding sign. The conditional entropy of Indus texts has been shown to fall within the range of natural languages (R.P.N. Rao et al., "Entropic Evidence for Linguistic Structure in the Indus Script," *Science*, vol. 324, no. 5931, 2009, p. 1165).

A potential shortcoming of the conditional entropy result is that it only captures pairwise dependencies. Figure 3 shows new results on higher-order entropies for blocks of up to six symbols. These block entropies were calculated using the state-of-the-art NSB estimator (I. Nemenman, F. Shafee, and W. Bialek, "Entropy and Inference, Revisited," *Advances in Neural Information Processing Systems 14*, MIT Press, 2002, pp. 471-478), which has been shown to provide good estimates of entropy for undersampled discrete data.

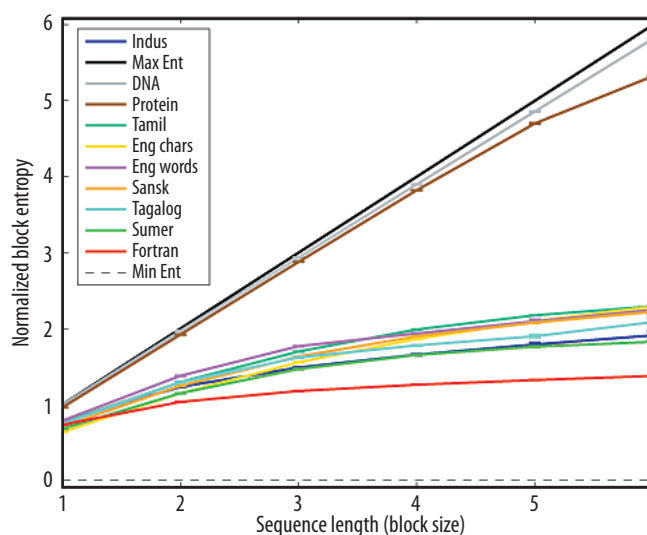


Figure 3. Entropy of the Indus script compared to natural languages and other sequences. Symbols were signs for the Indus script; bases for DNA; amino acids for proteins; characters for English; words for English, Tagalog, and Fortran; symbols in abugida (alphasyllabic) scripts for Tamil and Sanskrit; and symbols in the cuneiform script for Sumerian. To compare sequences over different alphabet sizes L , the logarithm in the entropy calculation was taken to base L : 417 for Indus, 4 for DNA, and so on. The resulting normalized block entropy is plotted as a function of block size. Error bars denote one standard deviation above/below mean entropy and are negligibly small except for block size 6.

The new results in Figure 3 extend the conditional entropy result to sequences of length up to six: The block entropies of the Indus texts remain close to those of a wide range of natural languages and far from the entropies for randomly and rigidly ordered sequences (Max Ent and Min Ent, respectively). Also shown in the plot for comparison are the entropies for a computer program written in Fortran and two sample biological sequences (DNA and proteins). The Fortran program and the biological sequences have noticeably lower and higher block entropies, respectively, than the Indus script and natural languages.

Entropic similarity to natural languages by itself is not sufficient to prove that the Indus script is linguistic. However, given that it exhibits other key features of linguistic scripts as enumerated above, this similarity increases the probability in a Bayesian sense that the Indus script represents language.

PROSPECTS FOR DECIPHERMENT

Can the Indus script be deciphered without a bilingual artifact such as the Rosetta Stone? History suggests it could be: The Linear B script used in ancient Greece was deciphered in the 1950s without a bilingual artifact. The decipherment relied on several factors such as being able to identify common roots and suffixes, hypothesizing that the script was syllabic, and guessing the pronunciation of some symbols, which revealed the script to be a form of Greek. In the case of the Indus script, the short length of the available texts makes such an approach difficult. It may be possible, however, to obtain results by focusing on particular types of Indus texts and the contexts in which they are found.

Most of the Indus texts found are on stamp seals, which were typically used in Bronze Age cultures for regulating trade. Seals were pressed onto clay tags affixed to packaged goods. The tags often listed the contents,

origin or destination, type or amount of goods being traded, name and title of the owner, or some combination of these. Numerous such clay tags have been found at various sites in the Indus civilization, bearing seal impressions on one side and impressions of woven cloth, reed matting, or other packing material on the other.


If the Indus script was used for trade, as the evidence suggests, then we would expect to find signs representing numerical quantities and units of measure. Progress in this direction has recently been reported by Bryan Wells, who estimated the volumes of two pots, one bearing the inscription UIII and the other the inscription IIIIU. By showing that the estimated volume of the second pot was in fact twice that of the first, Wells was able to conclude that strokes such as III and IIIII represent numbers, and the sign U probably represents a unit of volume.

Other efforts by Parpola and Mahadevan have assumed that at least some of the texts probably represent names. Phonetic values for specific signs can then be sug-

gested by assuming an underlying language—for example, proto-Dravidian—and using the rebus principle to guess the pronunciation of pictorial signs such as “fish,” “jar,” and “arrow.” In English, for example, the rebus principle could be used to represent an abstract word such as “belief” with the picture of a bee followed by a picture of a leaf. Ancient scripts often used the rebus principle to represent language.

Probabilistic models could help in this decipherment process in several ways. Recently proposed algorithms for probabilistic grammar induction could allow construction of a partial grammar for the Indus texts, facilitating the identification of root words, suffixes, prefixes, and other modifiers. This may facilitate the use of deciphering techniques similar to those applied to Linear B. Reconstructing a grammar would also allow comparison with the grammars of other languages, helping narrow down the set of candidate language families to consider when using the rebus principle.

A site-by-site analysis of the Indus texts using probabilistic models could indicate whether different languages or dialects were spoken in different regions of the Indus civilization. Similarly, training probabilistic models on texts found on specific types of artifacts, such as seals versus tablets, could ascertain whether the content of the texts varies according to artifact type.

In summary, the study of the Indus script has emerged as an exciting area of interdisciplinary research, offering a unique opportunity for probabilistic models to shed new light on one of the world's oldest civilizations. 

Rajesh P.N. Rao is an associate professor in the Department of Computer Science & Engineering at the University of Washington. Contact him at rao@cs.washington.edu.

Editor: Naren Ramakrishnan, Dept. of Computer Science, Virginia Tech, Blacksburg, VA; naren@cs.vt.edu

Silver Bullet Security Podcast

In-depth interviews with security gurus. Hosted by Gary McGraw.

www.computer.org/security/podcasts

Sponsored by  