FINAL VERSION (To appear in *Science*, 2009)

Entropic Evidence for Linguistic Structure in the Indus Script

Rajesh P. N. Rao[1*], Nisha Yadav[2], Mayank N. Vahia[2], Hrishikesh Joglekar[3], R. Adhikari[4], Iravatham Mahadevan[5]

[1] Dept. of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA
[2] Dept. of Astronomy & Astrophysics, Tata Institute of Fundamental Research, Mumbai 400005 and Centre for Excellence in Basic Sciences, Mumbai 400098, India
[3] 14, Dhus Wadi, Laxminiketan, Thakurdwar, Mumbai 400002 India
[4] The Institute of Mathematical Sciences, Chennai 600113, India
[5] Indus Research Centre, Roja Muthiah Research Library, Chennai 600113, India

[*] To whom correspondence should be addressed. E-mail: rao@cs.washington.edu

Online abstract

The script of the ancient Indus civilization remains undeciphered. The hypothesis that the script encodes language has recently been questioned. Here we present evidence for the linguistic hypothesis by showing that the script's conditional entropy is closer to those of natural languages than various types of nonlinguistic systems.

The Indus civilization flourished ~2600 to 1900 before the common era in what is now eastern Pakistan and northwestern India (*1*). No historical information exists about the civilization, but archaeologists have uncovered samples of their writing on stamp seals, sealings, amulets, and small tablets. The script on these objects remains undeciphered, despite a number of attempts and claimed decipherments (*2*). A recent article (*3*) questioned the assumption that the script encoded language, suggesting instead that it might have been a nonlinguistic symbol system akin to the Vinča inscriptions of southeastern Europe and Near Eastern emblem systems. We compared the statistical structure of sequences of signs in the Indus script with those from a representative group of linguistic and nonlinguistic systems.

Two major types of nonlinguistic systems are those that do not exhibit much sequential structure (type 1 systems) and those that follow rigid sequential order (type 2 systems). For example, the sequential order of signs in Vinča inscriptions appears to have been unimportant (*4*).

On the other hand, the sequences of deity signs in Near Eastern inscriptions found on boundary stones (*kudurrus*) typically follow a rigid order that is thought to reflect the hierarchical ordering of the deities (*5*).

Linguistic systems tend to fall somewhere between these two extremes: The tokens of a language (such as characters or words) do not follow each other randomly nor are they juxtaposed in a rigid order. There is typically some amount of flexibility in the ordering of tokens to compose words or sentences. One way of quantifying this flexibility is to use conditional entropy (*6*), which measures the amount of randomness in the choice of a token given a preceding token (*7*).

We computed the conditional entropies of five types of known natural linguistic systems (Sumerian logo-syllabic system, Old Tamil alpha-syllabic system, Rig Vedic Sanskrit alpha-syllabic system, English words, and English characters), four types of nonlinguistic systems (two artificial control datasets representing type 1 and type 2 nonlinguistic systems as described above, human DNA sequences, and bacterial protein sequences), and an artificially-created linguistic system (the computer programming language Fortran). We compared these conditional entropies with the conditional entropy of Indus inscriptions from a well-known concordance of Indus texts (*8*).

We found that the conditional entropy of Indus inscriptions closely matches those of linguistic systems and remains far from nonlinguistic systems throughout the entire range of token set sizes (Fig. 1*A*) (*7*). The conditional entropy of Indus inscriptions is substantially below those of the two biological nonlinguistic systems (DNA and protein) and above that of the computer programming language (Fig. 1*B*). Moreover, this conditional entropy appears to be most similar to those of Sumerian (a logo-syllabic script roughly contemporaneous with the Indus script) and Old Tamil (an alpha-syllabic script) and falls between those for English words and for English characters. These observations are consistent with previous suggestions (e.g., (*9*)), made on the basis of the total number of Indus signs, that the Indus script may be logo-syllabic. The

similarity in conditional entropy to Old Tamil, a Dravidian language, is especially interesting in light of the fact that many of the prominent decipherment efforts to date (*9*,*10*,*11*) have converged upon a proto-Dravidian hypothesis for the Indus script.

Given the prior evidence for syntactic structure in the Indus script (*9*,*12*), our results increase the probability that the script represents language, complementing other arguments that have been made explicitly (*13*, *14*) or implicitly (*15*,*16*) in favor of the linguistic hypothesis.

References and Notes

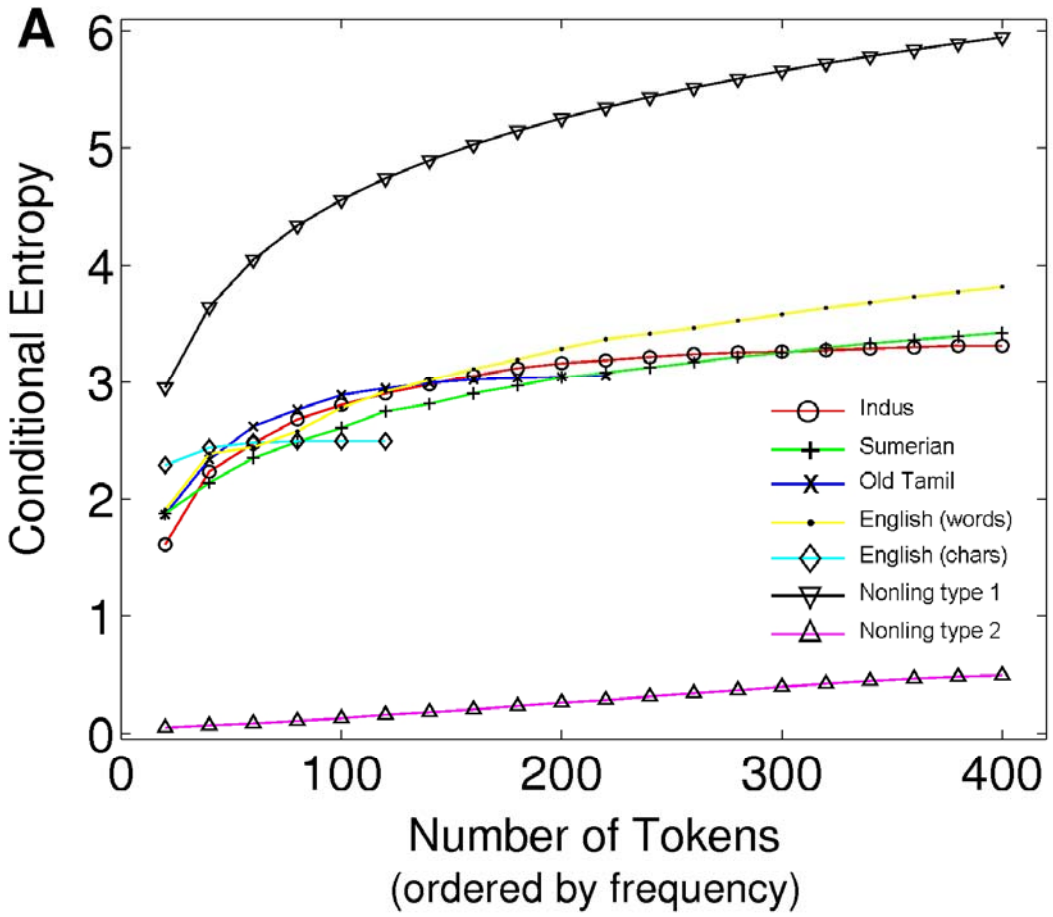
1. A. Lawler, *Science* **320**, 1276 (2008).
2. G. L. Possehl, *Indus Age: The Writing System*. Philadelphia: Univ. of Pennsylvania Press (1996).
3. S. Farmer, R. Sproat, and M. Witzel, *Electronic Journal of Vedic Studies* **11**, 19 (2004).
4. S. M. M. Winn, in *The Life of Symbols*, M. L. Foster and L. J. Botscharow (eds.), pp. 263-83. Boulder, Colorado: Westview Press (1990).
5. J. A. Black and A. Green, *Gods, Demons and Symbols of Ancient Mesopotamia*. London: British Museum Press (1992).
6. C. E. Shannon, *The Bell System Technical Journal* **27**, 379 (1948).
7. Materials and methods are available on *Science Online*.
8. I. Mahadevan, *The Indus Script: Texts, Concordance and Tables*. New Delhi: Archaeological Survey of India (1977).
9. A. Parpola, *Deciphering the Indus Script.* Cambridge, UK: Cambridge Univ. Press (1994).
10. I. Mahadevan, *Journal of Tamil Studies* **2**(1), 157 (1970).

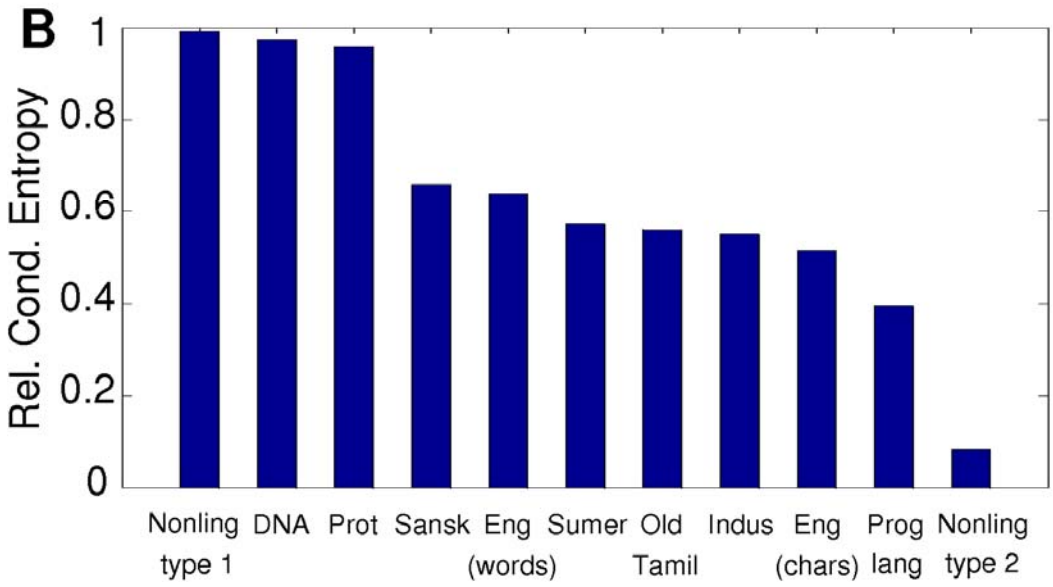
11. Y. V. Knorozov, M. F. Albedil, and B. Y. Volchok, *Proto-Indica: 1979, Report on the Investigations of the Proto-Indian Texts*. Moscow: Nauka Publishing House (1981).
12. K. Koskenniemi, *Studia Orientalia* **50**, 125 (1981).
13. A. Parpola, *Transactions of the International Conference of Eastern Studies* **50**, 28 (2005).
14. A. Parpola, in: *Airavati: Felicitation volume in honor of Iravatham Mahadevan*, Chennai, India: Varalaaru.com publishers, pp. 111-131 (2008)
15. N. Yadav, M. N. Vahia, I. Mahadevan, and H. Joglekar, *International Journal of Dravidian Linguistics* **37**, 39-52 (2008).
16. N. Yadav, M. N. Vahia, I. Mahadevan, and H. Joglekar, *International Journal of Dravidian Linguistics* **37**, 53-72 (2008).
17. This work was supported by the Packard Foundation, the Sir Jamsetji Tata Trust, and the University of Washington. We thank T. Sejnowski, M. Shapiro, and B. Wells for their comments and suggestions.

Figure Caption

Figure 1: Conditional entropy of Indus inscriptions compared to linguistic and nonlinguistic systems. (A) The conditional entropy (in units of nats) is plotted as a function of the number of tokens (signs, characters, or words) ordered according to their frequency in the texts used in this analysis (*7*). **(B)** Relative conditional entropy (conditional entropy relative to a uniformly random sequence with the same number of tokens) for linguistic and nonlinguistic systems. Prot indicates protein sequences; Sansk, Sanskrit; Eng, English; Sumer, Sumerian; and Prog lang, programming language. Besides the systems in (A), this plot includes two biological nonlinguistic systems (a human DNA sequence and bacterial protein sequences) as well as Rig Vedic Sanskrit and a computer program in the programming language Fortran (*7*).

Figure


A
Conditional Entropy
0
1
2
3
4
5
6
0
100
200
300
400
Number of Tokens
(ordered by frequency)
Indus
Sumerian
Old Tamil
English (words)
English (chars)
Nonling type 1
Nonling type 2


B
Rel. Cond. Entropy
0
0.2
0.4
0.6
0.8
1
Nonling type 1
DNA
Prot
Sansk
Eng (words)
Sumer
Old Tamil
Indus
Eng (chars)
Prog lang
Nonling type 2

Supplementary Information for

Entropic Evidence for Linguistic Structure in the Indus Script

Rajesh P. N. Rao, Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar,
R. Adhikari, Iravatham Mahadevan

Materials and Methods

Datasets

The following datasets were used for the comparative statistical analysis reported in the paper. Note that the datasets are of different sizes because they were obtained from different sources – a smoothing technique was used to counter the effects of different sample sizes in estimation (see *Calculation of Conditional Entropy* below).

- **Indus – Corpus of Texts from Mahadevan's *The Indus Script: Texts, Concordance and Tables*:** We used a subset of Indus texts from Mahadevan's concordance (*1*) obtained by excluding all texts 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 was kept in the dataset. This resulted in a dataset containing 1548 lines of text, with 7000 sign occurrences in total – the same dataset has been used in other studies (*2*,*3*).
- **English – The Brown University Standard Corpus of Present-Day American English:** The Brown corpus is a well-known dataset of modern American English. Sentences in the corpus are drawn from a wide variety of texts, including press reports, editorials, books, periodicals, novels, short stories, and scientific articles. The corpus was compiled by Kucera and Francis, and first used in their classic work *Computational Analysis of Present-Day American English* (*4*). This dataset contained 20,000 sentences, with a total of about 1,026,600 words and 5,897,000 characters (including spaces).

• **Sanskrit – Hymns from the Rig Veda:** We used the first 100 hymns (in Sanskrit) from Book 1 of the Rig Veda, which is believed to have been composed circa 1700–1100 B.C. These hymns, represented in Devanagari script, were obtained from http://www.sacred-texts.com/hin/rvsan/index.htm and converted to digital form using Unicode to allow quantitative analysis. The converted dataset contained a total of approximately 65,000 alpha-syllabic tokens (including spaces).

• **Old Tamil – Eight Sangam Era Tamil Texts (Ettuthokai):** This text corpus comprised of Eight Anthologies of Sangam Era poems (Ettuthokai), generally regarded as the earliest known literature in Tamil (dated roughly 300 B.C.–300 A.D.). The texts were obtained from http://www.tamilnation.org/literature/anthologies.htm and converted to digital form using Unicode to allow quantitative analysis. The dataset contained a total of approximately 876,000 alpha-syllabic tokens (including spaces).

• **Sumerian – Electronic Corpus of Sumerian Texts:** This corpus, available at http://www-etcsl.orient.ox.ac.uk/, comprises of a selection of nearly 400 literary compositions from ancient Mesopotamia dating to the late third and early second millennia B.C. The corpus includes narrative, historical, and mythological compositions, royal praise poetry, letters, hymns, and songs. The dataset used consisted of a transliterated subset of this corpus containing about 10,300 logo-syllabic signs (excluding spaces).

• **Nonlinguistic System of Type 1 (e.g., Vinča system):** Type 1 nonlinguistic systems involve signs that may occur in groups but the ordering of signs is not important (as it appears to have been, for example, in the Vinča system (5)). To enable comparison with the Indus texts, we assumed a Type 1 nonlinguistic system with the same number of signs as in the Indus corpus above and created a dataset of 10,000 lines of text, each containing 20 signs, based on the assumption that each sign has an equal probability of following any other.

• **Nonlinguistic System of Type 2 (e.g., Sumerian deity symbol system on kudurrus):** Type 2 nonlinguistic systems exhibit ordering of signs but the order is rigid. For example, in the Sumerian deity sign system found on boundary stones (*kudurrus*) (*6*), the ordering of deity signs appears to follow the established hierarchy among the various deities. As in the case of Type 1 systems above, we assumed a Type 2 nonlinguistic system with the same number of signs as in the Indus corpus above and created a corpus of 10,000 lines of text, each containing 20 signs, based on the assumption that each sign has a unique successor sign (variations of this theme where each sign could be followed by, for example, 2 or 3 other signs produced similar results).

• **DNA – Sequence from human chromosome 2:** We used the first one million nucleotides in human chromosome 2 obtained from the Human Genome Project (http://www.ncbi.nlm.nih.gov/genome/guide/human/), made available as a text file by Project Gutenberg (http://www.gutenberg.org/etext/11776). Roughly similar values for conditional entropy were obtained when sequences from other chromosomes were used.

• **Protein – Sequences from Escherichia coli:** The entire collection of amino acid sequences for the bacteria *E. coli* was extracted from the *E. coli* genome obtained from the NCBI website http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=U00096.2. This yielded a dataset containing a total of 374,986 amino acids comprising the sequences.

• **Programming Language:** We used a representative computer program in the programming language FORTRAN for solving a physics problem (fluid flow) using the finite element method. The program contained 28,594 lines of code (including comments). We removed the comments and used for our analysis the remaining code sequence containing 55,625 occurrences of tokens (examples of tokens include: *if*, *then*, *else*, *integer*, *x*, =, 50, etc.)

Calculation of Conditional Entropy

We describe here the method used to calculate the conditional entropy of the various datasets used in this study. We use the word "token" to denote the fundamental unit of the text being analyzed, such as a character in English, a word in English (for word-level analysis), a symbol in Sumerian, an alpha-syllabic character in Tamil or Sanskrit, or a sign in the Indus script. We consider texts as sequences of tokens: $T_1T_2\ldots T_M$. For example, if English characters are the tokens, the sentence "To be or not to be that is the question" consists of the token sequence T, o, <space>, b, e, <space>, etc., where as if the tokens are words, the token sequence would be: [To], [be], [or], [not], etc. We used the following sets of tokens in our analysis:

- **Indus texts:** The tokens were 417 signs identified by Mahadevan in *The Indus Script: Texts, Concordance and Tables* (*1*).
- **Sumerian texts:** The tokens were the top 417 most frequently occurring Sumerian logo-syllabic signs as extracted from the *Electronic Corpus of Sumerian Texts* described above.
- **Old Tamil texts:** The tokens were 244 symbols from the Tamil alpha-syllabic set extracted from the Unicode transliteration of the *Eight Sangam-Era Tamil Texts (Ettuthokai)* described above.
- **Sanskrit texts:** The tokens were 388 symbols from the Devanagari alpha-syllabic set extracted from the Unicode transliteration of the first hundred hymns in Book One of the Rig Veda as described above.
- **English characters:** The tokens comprised of 128 ASCII characters (letters A-Z, a-z, numbers, punctuation marks, and other miscellaneous characters).
- **English words:** For analysis at the word level, we used the 417 most frequent words in the Brown corpus as tokens.

- **DNA sequence:** The tokens were the 4 bases A, T, G, and C (Adenine, Thymine, Guanine, and Cytosine).
- **Protein sequence:** The tokens were the 20 amino acids: Glycine (G), Proline (P), Alanine (A), Valine (V), Leucine (L), Isoleucine (I), Methionine (M), Cysteine (C), Phenylalanine (F), Tyrosine (Y), Tryptophan (W), Histidine (H), Lysine (K), Arginine (R), Glutamine (Q), Asparagine (N), Glutamic Acid (E), Aspartic Acid (D), Serine (S), and Threonine (T).
- **Programming language:** The tokens were the various programming language constructs (*if*, *then*, *else*, *write*, *call*, etc.), operators (=, +, -, etc.), and user-defined variables and constants (maxnx, maxny, reynld, len, 80, 17, etc.). For the analysis, we used the top 417 most frequently occurring tokens.

The calculation of conditional entropy requires the estimation of conditional probabilities for pairs of tokens. Given a set of tokens (numbered 1,…,N) and a dataset containing sequences such as $T_1T_2\ldots T_M$ of such tokens, we compute, for each pair of tokens i and j, the conditional probability that token j immediately follows token i, i.e., $P(j|i)$. A standard approach to computing these probabilities is to count the number of times token j follows token i in the text sequences in the dataset; this is equivalent to computing the maximum likelihood estimate of the conditional probabilities (*7*). However, this estimate often yields poor estimates when the dataset is small, as is the case with the Indus script, and is susceptible to biases that come from datasets being of different sizes (as in our case). There has been extensive research on "smoothing" techniques which provide better estimates by relying on other sources of information and heuristics (see Chap. 6 in (*7*) for an overview). For the results in this paper, we use a form of smoothing known as "modified Kneser-Ney smoothing" (*8*) (based on (*9*)) which has been shown to outperform other smoothing techniques on benchmark datasets. Details of the smoothing procedure can be found in (*8*). The smoothing procedure ameliorates the effect of dataset sizes in our estimates of

$P(j|i)$. The probability $P(i)$ of token i was calculated based on the frequency of the token in the dataset.

Entropy and conditional entropy are well-established concepts in information theory and were first introduced by Shannon (*10*). The *entropy* of tokens (numbered $i = 1,\ldots,N$) in a particular dataset of texts is defined as:

$$H = -\sum_{i=1}^{N} P(i) \log P(i) \qquad \text{(Equation S1)}$$

The entropy (measured in units of *nats* when the natural logarithm is used as above) quantifies the amount of randomness in the text in the sense that it attains the highest value when all tokens are equally likely and the lowest value when one token has a probability of 1 and all other tokens have a probability of 0 (i.e., the text is made up of a single token that repeats).

The *conditional entropy* of a token j following token i is defined as:

$$C = -\sum_{i=1}^{N} P(i) \sum_{j=1}^{N} P(j \mid i) \log P(j \mid i) \qquad \text{(Equation S2)}$$

The conditional entropy quantifies the amount of flexibility in the choice of a token given a fixed preceding token – it thus captures the flexibility in the pairwise ordering of tokens in a dataset. For example, if a given token can be followed by any other token (as in Type 1 nonlinguistic systems), the conditional entropy is high. If a given token can only be followed by a unique token (as in certain Type 2 nonlinguistic systems), the conditional entropy is low.

Supplementary Text

Entropy of single signs in the Indus texts compared to other texts

Analyzing the frequencies of single tokens (signs/characters/words) by themselves is not sufficient for distinguishing nonlinguistic systems from linguistic systems (this is already evident in Farmer *et al.*'s Fig. 2 (*11*)). Figure S1 demonstrates this fact by comparing the single token entropies (defined in Equation S1) for linguistic and nonlinguistic systems. Unlike Figure 1 in the main text, the plots for linguistic systems are no longer clustered together or separated from the nonlinguistic systems. In fact, the entropy plot for Type 2 nonlinguistic systems falls right in the middle of those for the linguistic systems because the overall frequencies of signs in this particular dataset happened to be similar to those for some linguistic systems. This highlights the fact that the statistics of isolated symbols, as quantified by $P(i)$, are insufficient for distinguishing linguistic from nonlinguistic systems. One needs to consider sequential statistics (e.g., the conditional probability $P(j|i)$ and beyond) to capture the syntactic structure of language and be able to separate linguistic from nonlinguistic systems as in Figure 1 of the main text.

Zipf-Mandelbrot law for linguistic systems

Although the frequencies of isolated signs may not be sufficient to distinguish linguistic from nonlinguistic systems, these frequencies can still provide useful information when expressed in the form of "Zipf's law" (or the "Zipf-Mandelbrot law") (*7*,*12*). Suppose the most frequent sign in a corpus is given rank $r = 1$ and its frequency is denoted f_1, the next most frequent sign is given rank $r = 2$ and its frequency denoted f_2, and so on. For most linguistic corpora, it has been found that the rank-ordered frequency f_r of words can be fit very well to the Zipf-Mandelbrot law: $\log f_r = a - b \log(r + c)$ (*7*). The law essentially states that a small number of tokens contribute to the bulk of the distribution but a large number of rare signs contribute to a long tail. For the Indus dataset used in this paper, we have found an excellent fit to the Zipf-Mandelbrot law with the

values a = 15.39, b = 2.59, and c = 44.47 (see (*13*) for a plot). This is roughly in the same range as the values for other linguistic systems (e.g., for English words, the values are: a = 12.43, b = 1.15, and c = 100 (*7*)).

Using higher order sequences and block entropies

The distinction between nonlinguistic systems (such as DNA) and linguistic systems that we have reported here can also be demonstrated using higher order entropies (block entropy). Block entropies are computed based on the probabilities for sequences of n symbols (also known as "n-grams") for n = 1, 2, 3, …, and using these in Equation S1 instead of *P(i)*. Schmitt and Herzel (*14*) have demonstrated a result similar to our Figure 1 using block entropies (for n = 1,..,15) for yeast DNA, music, English characters, and Fortran code (Fig. 8 in (*14*)). Obtaining a block entropy result for the Indus texts is difficult owing to the small size of the corpus, which makes it hard to accurately estimate the higher order probabilities beyond small values of n. The crucial question is whether taking higher order distributions into account could change the conclusions of this paper. We believe this not to be the case based on our related work on Markov models (*15*) and n-gram analysis of the Indus texts (*13*). In particular, for a given value of the sequence length n, one can compute the "perplexity" (*7,13*) of the Indus corpus as 2^{H_n}, where $H_n = -\sum_{i=1}^{M} 1/M \log_2 P(x_i^n)$. Here, the x_i^n are sequences in the corpus of length n and M is the number of such sequences. Intuitively, the lower the perplexity value, the better the probability model P in approximating the true probabilistic model that generated the data. As Table S1 shows, the bulk of the perplexity in the Indus corpus can be captured by a bigram (n = 2) (or equivalently 1st-order Markov) model as used in this paper.

n	1	2	3	4	5
Perplexity	68.82	26.69	26.09	25.26	25.26

Table S1: Perplexity of Indus probabilistic models as a function of sequence length n.

Relationship to cryptographic techniques for recognizing languages

The problem of distinguishing languages from non-languages also arises in the field of cryptography. For example, to decipher an encrypted text, one can conduct an exhaustive search over possible keys by applying a candidate key to the text and checking to see if the resulting "plaintext" is recognizable as language. Ganesan and Sherman pose the latter "language recognition" problem as statistical hypothesis testing based on *n*-gram frequency counts in the plaintext for a fixed *n* (*16*). Such an approach is especially useful if the underlying language is known since it allows one to test if the given plaintext was generated by a model for the given language. In the case where the underlying language is unknown (as in the case of the Indus script), one can still formulate statistical tests to determine, for example, whether a text was generated by an unknown 0^{th}-order versus an unknown 1^{st}-order Markov model, or whether it was produced by a uniform noise generator (Problems 3 and 4 in (*16*)). However, these questions may also be answered through other means. For example, the results in Table S1 above already suggest that the Indus texts are much better modeled by a 1^{st}-order Markov process than a 0^{th}-order one. Similarly, the results in Figure 1 and Figure S1 strongly argue against the hypothesis that the Indus texts were produced by a uniform noise generator. Crypotography-inspired statistical techniques may nevertheless be worth exploring within the context of the Indus and other undeciphered scripts, with the caveat that analyzing a script encoding natural language can be a very different problem from analyzing deliberately disguised and encrypted texts.

Affinity of Indus texts with Dravidian versus Indo-European Languages

There has been considerable debate regarding whether the Indus texts have affinity with the Dravidian languages (such as Tamil), Indo-European languages (such as Sanskrit), or some other language family. A frequently cited argument in favor of the Dravidian hypothesis (see (*17*,*18*,*19*)) is that the Indus texts appear to be agglutinative in their morphological structure: sign

sequences often have the same initial signs but different final signs (ranging from 1 to 3 signs). Such modification of morphemes by a system of suffixes is found in Dravidian languages but is rare in Indo-European languages (such as Sanskrit) which tend to be inflectional (changing the final sound of a word rather than adding suffixes). Our result that the conditional entropy of the Indus texts is closest to Old Tamil, a Dravidian language, among the datasets we considered is suggestive in this regard. However, this result is also tied to our use of an alpha-syllabic script to represent the Sangam era Tamil texts. Using a similar alpha-syllabic Devanagari representation of Rig Vedic Sanskrit produced a comparatively higher conditional entropy value than Tamil (Figure 1B). Using an alphabetic representation (consonants and vowels are not combined into syllables) lowers the conditional entropy value for Sanskrit to below that for alpha-syllabic Tamil and closer to that for English characters. We believe that answering the question of linguistic affinity of the Indus texts requires a more sophisticated approach, such as statistically inferring an underlying grammar for the Indus texts from available data and comparing the inferred rules with those of various known language families including Dravidian and Indo-European.

Choice of the Indus corpus

The results in this paper were based on Mahadevan's corpus of Indus texts compiled in 1977. Since then, approximately 1000 Indus texts have been unearthed (some that repeat texts already in the 1977 corpus, some new) (*20*). Does this new material affect the conclusions of this paper? We do not think so 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. rather than a new class of written material (except for one so-called monumental inscription consisting of only 10 large-sized signs rather than a large number of signs (see p. 113 in (*19*))). The new material thus exhibits syntactic structure that is similar to the material we have analyzed in this paper and is therefore not likely to change the major conclusions of this study. (2) Some

new signs have been discovered in the new material. However, these new signs 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 estimated values of the conditional probabilities, causing a relatively minor change in the value of the conditional entropy.

Supplementary Figure

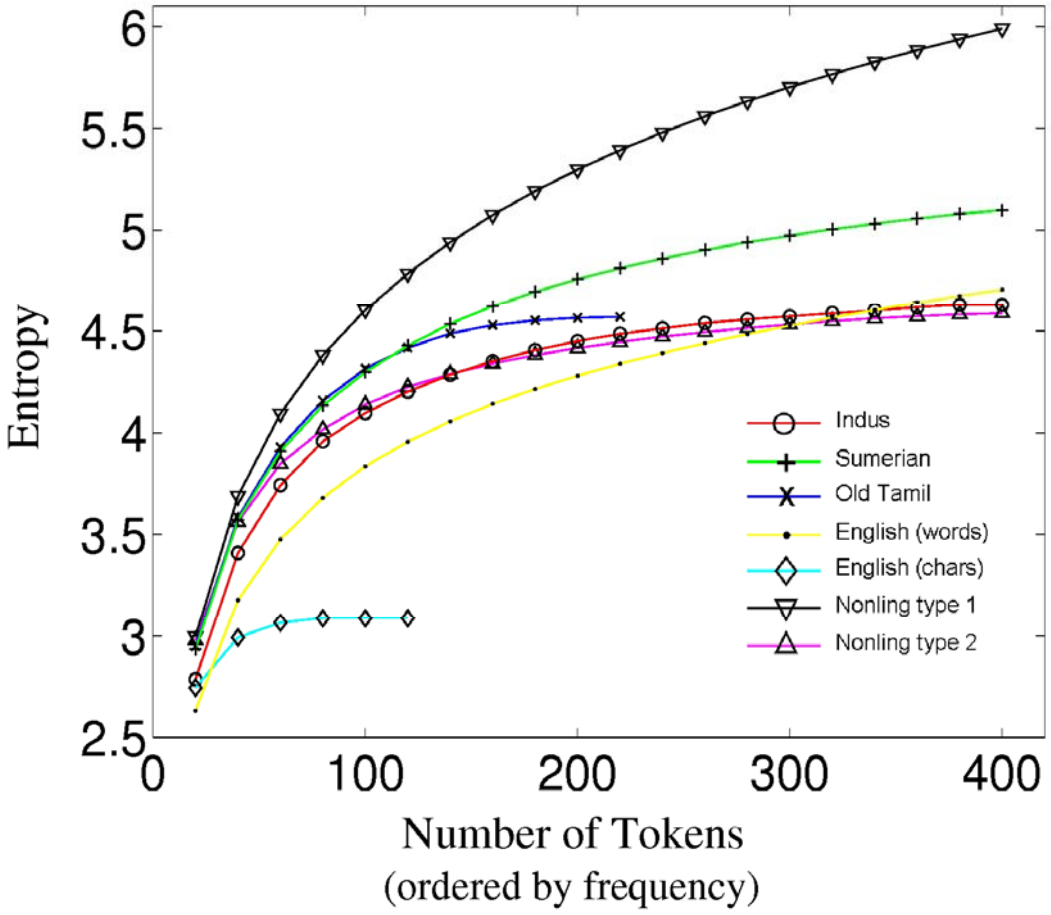


Figure S1: Entropy of isolated signs in the Indus texts compared to entropies of other texts. Entropies (in *nats*) were computed according to Equation S1 for isolated tokens (signs/characters/words) for the same datasets as in Figure 1A in the main text. In contrast to Figure 1A, these single symbol (unigram) entropy plots for linguistic systems do not cluster together and are not well separated from the two types of nonlinguistic systems. In particular, the entropies for the Type 2 nonlinguistic system overlap significantly with those for linguistic systems for this particular Type 2 dataset.

Supplementary References

1. I. Mahadevan, *The Indus Script: Texts, Concordance and Tables*. New Delhi: Archaeological Survey of India (1977).
2. N. Yadav, M. N. Vahia, I. Mahadevan, and H. Joglekar, *International Journal of Dravidian Linguistics* **37**, 39-52 (2008).
3. N. Yadav, M. N. Vahia, I. Mahadevan, and H. Joglekar, *International Journal of Dravidian Linguistics* **37**, 53-72 (2008).
4. H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press (1967).
5. S. M. M. Winn, in *The Life of Symbols*, M. L. Foster and L. J. Botscharow (eds.), pp. 263-83. Colorado: Westview Press (1990).
6. J. A. Black and A. Green, *Gods, Demons and Symbols of Ancient Mesopotamia*. London: British Museum Press (1992).
7. C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press (1999).
8. S. F. Chen and J. Goodman, Harvard University Computer Sci. Technical Report TR-10-98 (1998).
9. R. Kneser and H. Ney, In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pages 181-184 (1995).
10. C. E. Shannon, *The Bell System Technical Journal* **27**, 379 (1948).
11. S. Farmer, R. Sproat, and M. Witzel, *Electronic Journal of Vedic Studies* **11**, 19 (2004).
12. R. H. Baayen, *Word Frequency Distributions*. Dordrecht: Kluwer (2001).
13. N. Yadav, H. Joglekar, R. P. N. Rao, M. N. Vahia, I. Mahadevan, R. Adhikari, arxiv.0901.3017 (2009).

14. A. O. Schmitt and H. Herzel, *J. Theor. Biol.* **188**, 369 (1997).

15. R. P. N. Rao, N. Yadav, H. Joglekar, R. Adhikari, M. N. Vahia, I. Mahadevan, Technical Report no. 08-11-02, Dept. of Computer Sci. and Engineering, Univ. of Washington, Seattle, November, 2008.

16. R. Ganesan and A. T. Sherman, *Cryptologia* **17**(4), 321 (1993).

17. A. M. Kondratov, In G. V. Alekseev *et al.* (eds.), *Preliminary Report on the Investigation of the Proto-Indian Texts*. Moscow: Academy of Sciences U.S.S.R., pages 31-45 (1965).

18. I. Mahadevan, *Journal of Tamil Studies* **2**(1), 157 (1970).

19. A. Parpola, *Deciphering the Indus Script.* Cambridge, UK: Cambridge Univ. Press (1994).

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