Quantitative Evidence for Possible Linguistic Structure in the Indus Script

Rajesh P. N. Rao[1*], Nisha Yadav[2], Mayank N. Vahia[2], Hrishikesh Joglekar[3], Ronojoy 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, India
[3] Oracle, Hyderabad 500081, 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

The Indus civilization flourished c. 2550-1900 BC 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 system 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 widely publicized article (*3*,*4*) 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. Here we compare the statistical structure of sequences of signs in the Indus script with those from a representative group of linguistic and nonlinguistic systems, and provide, for the first time, quantitative evidence suggesting that the Indus script may have been a linguistic writing system.

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 (*5*). On the other hand, the sequences of deity signs in Near Eastern inscriptions on boundary stones (*kudurrus*) typically follow a rigid order that is thought to reflect the hierarchical ordering of the deities (*6*).

Linguistic systems tend to fall somewhere between these two extremes: the tokens of a language (such as characters or words) do not randomly follow each other nor are they juxtaposed in a rigid order. There is typically some amount of flexibility in the ordering of the tokens to compose words or sentences, but not too much flexibility. This notion of flexibility in sequential ordering can be quantified statistically using conditional entropy (*7*), which measures the amount of randomness in the choice of a token if the preceding token has been specified (Equation S2 in *Supplementary Information*).

We computed the conditional entropies of four types of known natural linguistic systems (Sumerian logosyllabic system, Old Tamil syllabic system, English words, and English characters), four types of nonlinguistic systems (representative examples of Type 1 and Type 2 nonlinguistic systems as described above, DNA sequence from the human genome, and protein sequences from *E. coli*), and an artificially-created linguistic system (computer program in the 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 the nonlinguistic systems throughout the entire range of token set sizes (Fig. 1*A*; see *Supplementary Information* for details). The conditional entropy of the Indus inscriptions is significantly below those of the two biological nonlinguistic systems (DNA and protein), above that of the computer programming language, and closest to natural linguistic systems (Fig. 1*B*). These results suggest that the sequential structure exhibited by Indus inscriptions is statistically more similar to the kind of sequential structure seen in linguistic systems than in nonlinguistic systems. Moreover, the conditional entropy for Indus inscriptions appears to be most similar to Sumerian (a logosyllabic script roughly contemporaneous with the Indus script) and Old Tamil (a syllabic script), and falls between those for English words and English characters. Both of these observations lend support to previous suggestions (e.g., (*9*)),

made on the basis of the total number of Indus signs, that the Indus script may be logosyllabic. The close relationship to Old Tamil, a Dravidian language, is especially interesting in light of the fact that many of the respected decipherment efforts to date (*9*,*10*) have converged upon a proto-Dravidian hypothesis for the Indus script.

Our results provide the first quantitative evidence for the existence of possible linguistic structure in the Indus script (other arguments in favor of the linguistic hypothesis are implicitly made in (*11*,*12*) and explicitly enumerated in (*13*)). The distinction between nonlinguistic systems (such as DNA) and linguistic systems that we have reported here has also been independently found using a different information theoretic measure (block entropy) (*14*).

A previous study that sought to classify the Indus script as nonlinguistic relied partly on an analysis of the frequencies of isolated Indus signs (*3*). However, the statistics of isolated signs can be shown to be insufficient for distinguishing linguistic from nonlinguistic systems (Fig. S1 and *Supplementary Text*). The fact that linguistic systems appear to cluster together in the space of conditional entropies (Fig. 1) but not in the space of isolated symbol entropies (Fig. S1) suggests that sequential statistics, i.e., the statistics governing which symbols may follow any given symbol, are important in distinguishing linguistic systems from nonlinguistic systems.

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. A. Lawler, *Science* **306**, 2026 (2004).
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. E. Shannon, *The Bell System Technical Journal* **27**, 379 (1948).
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** (1970).
11. K. Koskenniemi, *Studia Orientalia* **50**, 125 (1981).
12. N. Yadav, M. N. Vahia, I. Mahadevan, and H. Joglekar, *International Journal of Dravidian Linguistics* **37**, 39-52 and 53-72 (2008).
13. A. Parpola, in: *Airavati: Felicitation volume in honor of Iravatham Mahadevan*, Chennai, India: Varalaaru.com publishers, pp. 111-131 (2008); A. Parpola, *Transactions of the International Conference of Eastern Studies* **50**, 28 (2005).
14. A. O. Schmitt and H. Herzel, *J. Theor. Biol.* **188**, 369 (1997).
15. This work was supported by the Packard Foundation and Sir Jamsetji Tata Trust.

Figure 1

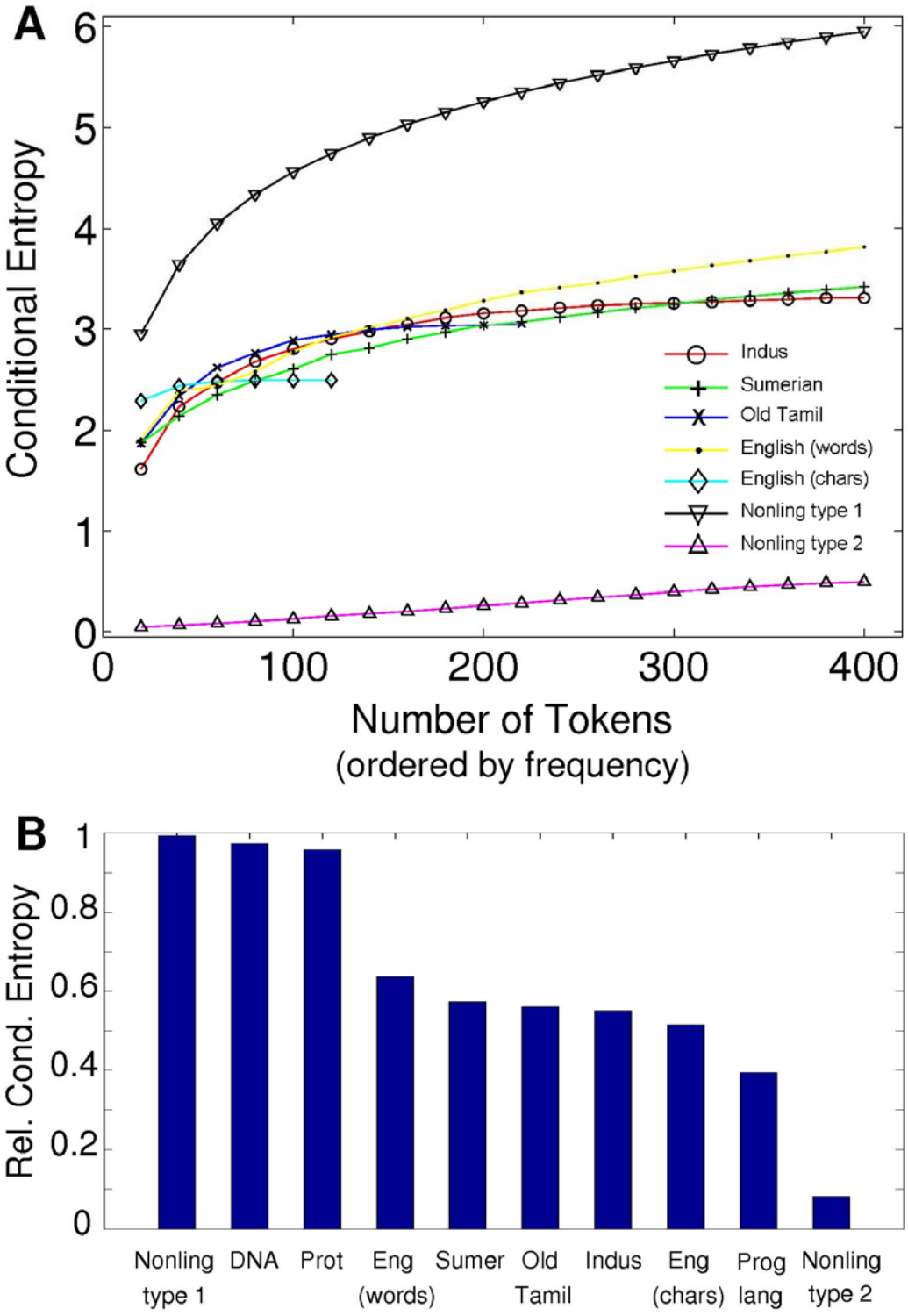
A
Conditional Entropy
0
1
2
3
4
5
6
0
100
200
300
400
Number of Tokens
(ordered by frequency)
Indus
Sumerian
Old Tamil
English (words)
English (chars)
Nonling type 1
Nonling type 2
B
Rel. Cond. Entropy
0
0.2
0.4
0.6
0.8
1
Nonling type 1
DNA
Prot
Eng (words)
Sumer
Old Tamil
Indus
Eng (chars)
Prog lang
Nonling type 2

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/words) ordered according to their frequency in the texts used in this analysis (see *Supplementary Information* for details). **(B)** Relative conditional entropy (conditional entropy relative to a uniformly random sequence with the same number of tokens) for linguistic and nonlinguistic systems. (The abbreviations Prot, Eng, Sumer, and Prog lang stand for Protein sequences, English, Sumerian, and Programming language respectively). Besides the systems in (A), this plot includes two biological nonlinguistic systems (a million-nucleotide DNA sequence from human chromosome 2 and protein sequences from *E. coli*) as well as a programming language (a computer program in Fortran for fluid flow). In both (A) and (B), the conditional entropy of the Indus script is most similar to the conditional entropies of known linguistic systems and remains far from the conditional entropies of the nonlinguistic systems.

Supplementary Information for

Quantitative Evidence for Possible Linguistic Structure in the Indus Script

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

Materials and Methods

Datasets

The following datasets were used for the comparative statistical analysis reported in this 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 (Ref (*8*) in the main text) 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.
- **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* (Kucera & Francis, 1967). This dataset contained 20,000 sentences, with a total of about 1,026,600 words and 5,897,000 characters (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.E. - 300 C.E.). 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 syllables (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 BCE. 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 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 in the Vinča system (Ref (5) in the main text)). 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 corpus 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, e.g., in the Sumerian deity sign system found on *kudurrus* (Ref (6) in the main text), the ordering of deity signs is thought 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 text 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, a syllabic character in Tamil, 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* (Ref (*8*) in the main text).
- **Sumerian texts:** The tokens were the top 417 most frequently occurring Sumerian logosyllabic signs as extracted from the *Electronic Corpus of Sumerian Texts* described above.
- **Old Tamil texts:** The tokens were 244 symbols from the Tamil syllabic alphabet extracted from the Unicode transliteration of the *Eight Sangam-Era Tamil Texts (Ettuthokai)* 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 (Manning & Schütze, 1999). 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 (Manning & Schütze, 1999) for an overview). For the results in this paper, we use a form of smoothing known as "modified Kneser-Ney smoothing" (Chen & Goodman, 1998; based on (Kneser & Ney, 1995)) which has been shown to outperform other smoothing techniques on benchmark datasets. Details of the smoothing procedure can be found in (Chen & Goodman, 1998). 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 (see Ref (*7*) in the main text). 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 of tokens (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 (Ref (*3*) of the main text)). 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 in the middle of those for the linguistic systems. This highlights the fact that the statistics of isolated symbols (quantified by $P(i)$) are insufficient for distinguishing linguistic from nonlinguistic systems. One needs to consider sequential statistics (e.g., the conditional probability $P(j|i)$) to be able to separate linguistic from nonlinguistic systems as demonstrated in Figure 1.

Supplementary Figure

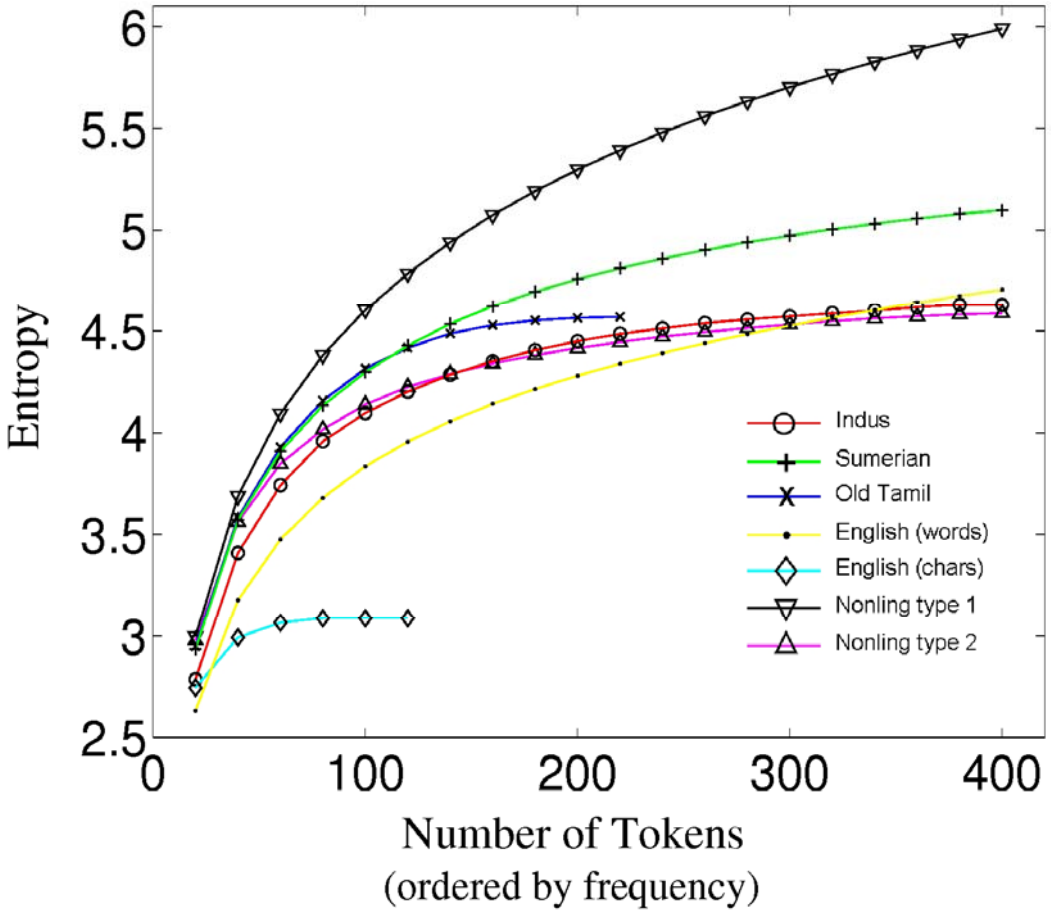

Figure S1: Entropy of isolated signs in the Indus texts compared to entropies of other texts. Entropies (in *nats*) were computed according to Equation S1 for isolated tokens (signs/characters/words) for the same datasets as in Figure 1A in the main text. In contrast to Figure 1A, these single symbol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

Supplementary References

S. F. Chen and J. Goodman (1998), An empirical study of smoothing techniques for language modeling. Harvard University Computer Sci. Technical Report TR-10-98.

R. Kneser and H. Ney (1995), Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pages 181-184.

H. Kucera and W. N. Francis (1967), *Computational analysis of present-day American English. Providence*, Providence, RI: Brown University Press

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