Sciencexpress Brevia

Entropic Evidence for Linguistic Structure in the Indus Script

Rajesh P. N. Rao,[1]* Nisha Yadav,[2,3] Mayank N. Vahia,[2,3] Hrishikesh Joglekar,[4] R. Adhikari,[5] Iravatham Mahadevan[6]

[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]Centre for Excellence in Basic Sciences, Mumbai 400098, India.

[4]14, Dhus Wadi, Laxminiketan, Thakurdwar, Mumbai 400002 India.

[5]The Institute of Mathematical Sciences, Chennai 600113, India.

[6]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. 2500-1900 B.C. 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. 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.

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. This flexibility can be quantified statistically using 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 (representative examples of 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 significantly 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 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 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 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 (*9–11*) have converged upon a proto-Dravidian hypothesis for the Indus script.

In summary, our results provide quantitative evidence for the existence of linguistic structure in the Indus script,

complementing other arguments that have been made explicitly (*12*, *13*) or implicitly (*14*–*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. 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. A. Parpola, *Transactions of the International Conference of Eastern Studies* **50**, 28 (2005).
13. A. Parpola, in: *Airavati: Felicitation volume in honor of Iravatham Mahadevan*, Chennai, India: Varalaaru.com publishers, pp. 111-131 (2008)
14. K. Koskenniemi, *Studia Orientalia* **50**, 125 (1981).
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).

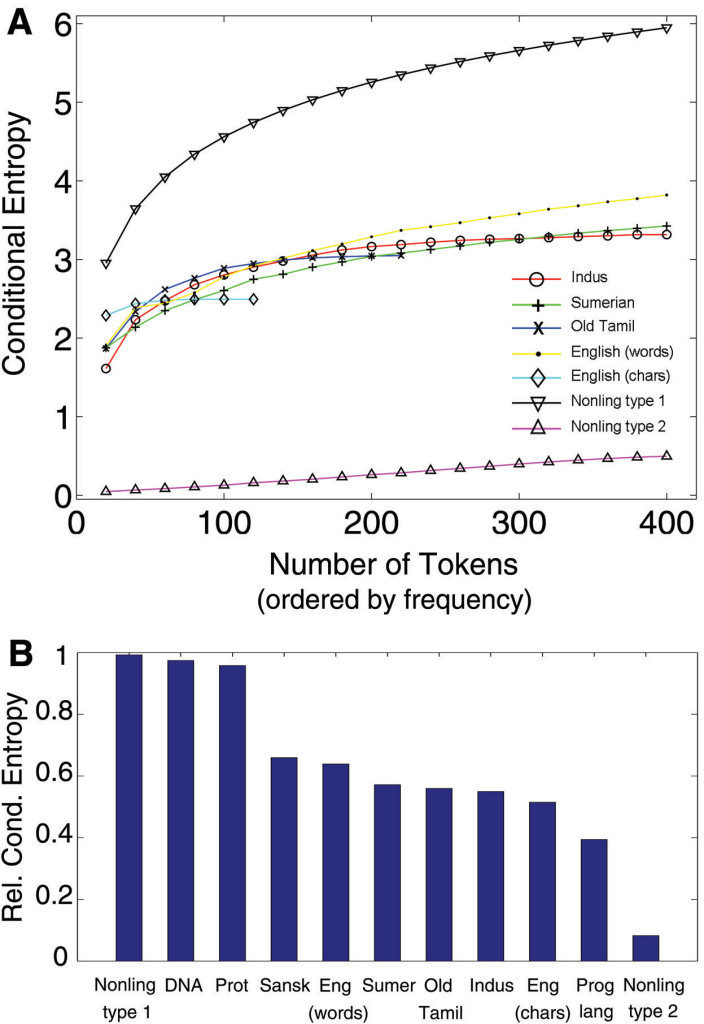
17. This work was supported by the Packard Foundation, the Sir Jamsetji Tata Trust, and the University of Washington. We thank Drs. Terrence Sejnowski, Michael Shapiro, and Bryan Wells for their comments and suggestions.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1170391/DC1
Materials and Methods
Fig. S1
Table S1
References

30 December 2008; accepted 14 April 2009
Published online 23 April 2009; 10.1126/science.1170391
Include this information when citing this paper.

Fig. 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 (*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: Protein sequences, Sansk: Sanskrit, Eng: English, Sumer: Sumerian, 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*).


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