Rebuttal of Sproat, Farmer, et al.'s supposed "refutation"

In a 2004 paper, Farmer, Sproat, and Witzel claimed that the Indus civilization was illiterate and that Indus writing was a collection of political or religious symbols. The publication of our paper in *Science* elicited hostile reactions from them, ranging from off-the-cuff dismissive remarks such as "garbage in, garbage out" (Witzel) to ad-hominem attacks (labeling us "Dravidian nationalists") and a vicious campaign on internet discussion groups and blogs to discredit our work. Their first knee-jerk reaction was to call our two artificial control datasets in our study "invented data sets" (Farmer). This was followed by Sproat and others claiming to have constructed "counterexamples" to our result.

Here, we respond to their arguments in a point-by-point fashion. First, their arguments:

(1) Two datasets, used as controls in our work, are artificial.

(2) Counterexamples can be given, of non-linguistic systems, which produce conditional entropy plots like those presented in our *Science* paper.

(3) Conditional entropy cannot even differentiate between language families.

(4) The absence of writing material and long texts is "proof" that the Indus people were illiterate.

We view arguments (1)-(3) as arising from a misunderstanding of our approach and an overinterpretation of the conditional entropy result.

Our responses to the above arguments are as follows:

(1) As stated in our *Science* paper, the two artificial data sets (which Farmer et al. call "invented data sets") simply represent controls, necessary in any scientific investigation, to delineate the limits of what is possible. The two controls in our work represent sequences with maximum and minimum flexibility, for a given number of tokens. Though this can be computed analytically, the data sets were generated to subject them to the same parameter estimation process as the other data sets. Our conclusions do not depend on the controls, but are based on comparisons with real world data: DNA and protein sequences, various natural languages, and FORTRAN computer code. All our real world examples are bounded by the maximum and the minimum provided by the controls, which thus serve as a check on the computation.

(2) Counterexamples matter only if we claim that conditional entropy by itself is a sufficient criterion to distinguish between language and non-language. We do not make this claim in our *Science* paper. As clearly stated in the last sentence of the paper, our results provide

evidence which, given the rich syntactic structure in the script, *increases the probability that the script represents language.* The methodology, which is Bayesian in nature, can be summarized as follows. We begin with the fact that the Indus script exhibits the following properties:

- The Indus texts are linearly written, like the vast majority of linguistic scripts (and unlike nonlinguistic systems such as medieval heraldry or traffic signs),
- Indus symbols are often modified by the addition of specific sets of marks over, around, or inside a symbol. Multiple symbols are sometime combined ("ligatured") to form a single glyph. This is similar to later Indian scripts which use such ligatures and marks above, below, or around a symbol to modify the sound of a root consonant or vowel symbol;
- 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;
- The script exhibits rich syntactic structure such as the clear presence of beginners and enders, preferences of symbol clusters for particular positions within texts etc. (see *References*), similar to linguistic sequences;
- Indus texts that have been discovered 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 in foreign regions.

Given that the Indus script shares the above properties with linguistic scripts, we claim that the entropic similarity of the Indus script to other natural languages provides additional evidence in favor of the linguistic hypothesis.

As mentioned above, the entire exercise of finding "counterexamples," as pursued by Sproat and others on a blog, is irrelevant to our conclusion because it is based on an overinterpretation of the conditional entropy result. It assumes that similarity in conditional entropy by itself is a sufficient condition for language, a claim not made in our paper. It is nevertheless interesting to analyze these "counterexamples." First, they have no correlations between symbols, making the conditional and unconditional entropies identical. In the Indus script, there is a large difference between these two quantities (cf. Figs. 1 and S1 in our *Science* paper). To produce a counter example in which conditional and unconditional entropies differ would require, as Sproat admits in a blog, tweaking of several parameters. This exercise gets even more difficult and convoluted when one considers higher-order block entropies. Clearly, the parsimonious explanation would be that the Indus script represents language.

(3) Sproat has endeavored to produce a plot where languages belonging to different language families have similar conditional entropies, thereby claiming that the conditional entropy result "proves nothing." This claim is once again based on an overinterpretation of the result in our *Science* paper. We specifically note on page 10 in the supplementary information 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." In other words, conditional entropy provides a quantitative measure of the amount of flexibility allowed in choosing the next symbol given a previous symbol. It is useful for characterizing the average amount of flexibility in sequences of different kinds. We do not make the claim that it can be used to distinguish between language families – this requires a more sophisticated measure.

(4) With regard to the length of texts, several West Asian writing systems such as Proto-Cuneiform, Proto-Sumerian, and the Uruk script have statistical regularities in sign frequencies and text lengths which are remarkably similar to the Indus script (Details can be found in http://indusresearch.wikidot.com/script). These writing systems are by all accounts linguistic. Furthermore, the lack of archaeological evidence for long texts in the Indus civilization does not automatically imply that they did not exist (*absence of evidence is not evidence of absence*). There is a long history of writing on perishable materials like cotton, palm leaves, and bark in the Indian subcontinent using equally perishable writing implements (see Parpola's paper below). Writing on such material is unlikely to have survived the hostile environment of the Indus valley. Thus, long texts may have been written, but no archaeological remains are to be found.

As regards the cultural sophistication of the Indus people and literacy, we believe Iravatham Mahadevan has addressed this adequately in his op-ed piece:
http://www.hindu.com/mag/2009/05/03/stories/2009050350010100.htm

---

References

- Final version of the *Science* paper (including Supplementary Information):
  - http://www.cs.washington.edu/homes/rao/ScienceIndus.pdf
- Parpola's point-by-point rebuttal of the nonlinguistic claim:
  - 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. http://www.harappa.com/script/indus-writing.pdf
- Massimo Vidale's "The collapse melts down: a reply to Farmer, Sproat and Witzel":
  - http://www.docstoc.com/docs/document-preview.aspx?doc_id=9163376
- Syntactic structure in the Indus script:
  - Koskenniemi K (1981) Syntactic methods in the study of the Indus script. *Studia Orientalia* 50:125-136.
  - Parpola A (1994) *Deciphering the Indus script*. (Cambridge University Press), Chaps. 5 & 6.
  - 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. http://www.harappa.com/script/tata-writing/indus-script-paper.pdf
  - Yadav N, Vahia MN, Mahadevan I, Joglekar H (2008) Segmentation of Indus texts. *International Journal of Dravidian Linguistics* 37(1):53-72. http://www.harappa.com/script/tata-writing/indus-texts.pdf