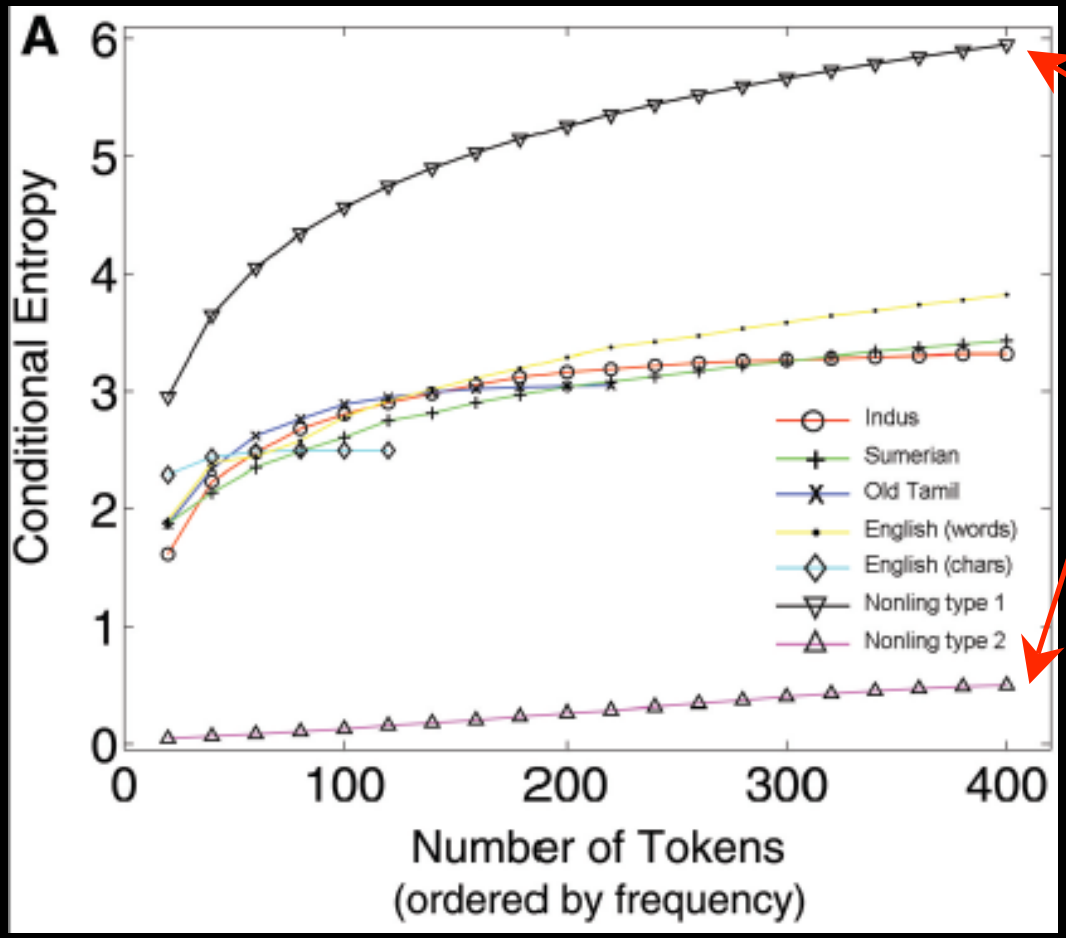
The following two slides are adapted from Steve Farmer, Richard Sproat, and Michael Witzel, "The Collapse of the Indus-Script Thesis, Five Years Later: Massive Nonliterate Urban Civilizations of Ancient Eurasia" (Presentation at the Indus Civilization conference held at the Research Institute for Humanity and Nature [RIHN], Kyoto, Japan, 29-31 May 2009; paper to follow).

Rao *et al.*, "Entropic Evidence for Linguistic Structure in the Indus Script" (*Science*, April 2009)



Invented data unlike any real-world non-linguistic system.

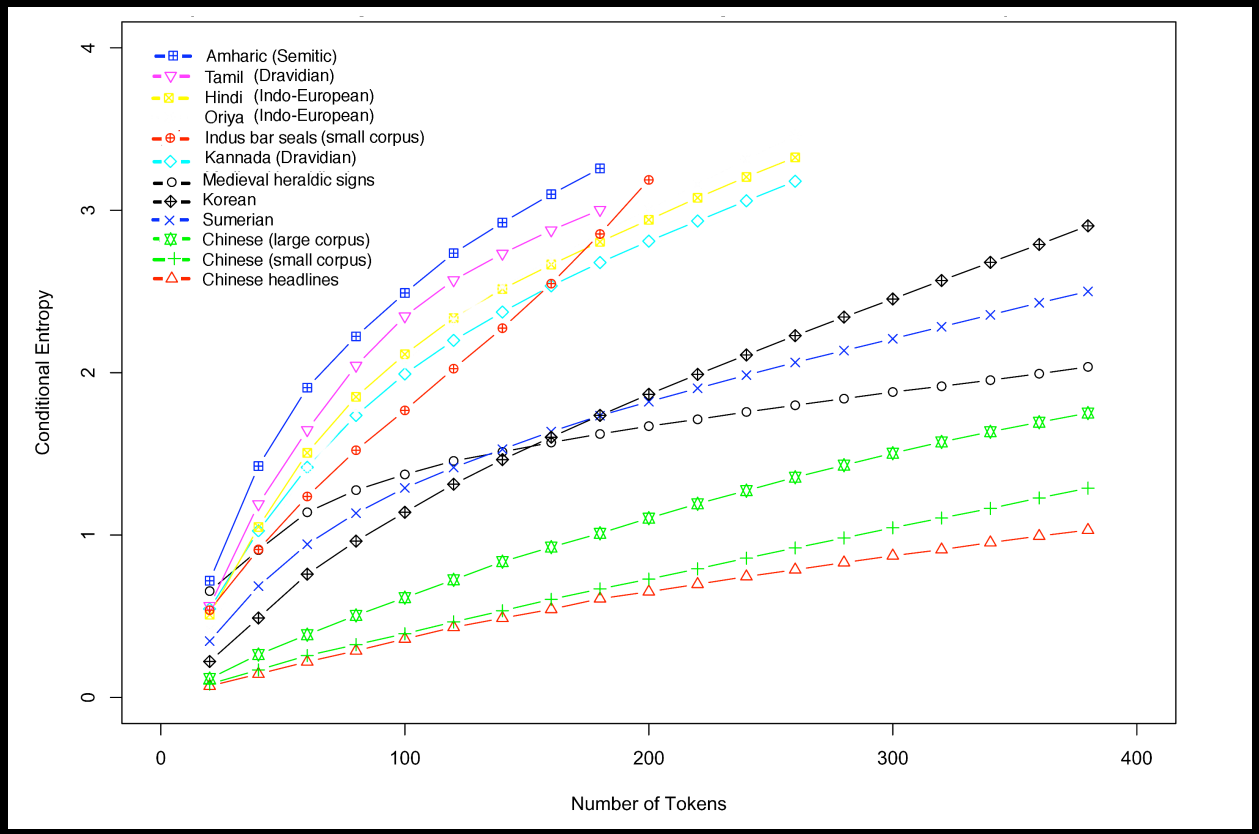
The paper spuriously links these data with Vinca and Mesopotamian sign systems.

(You only find that the data are invented when you go to their online "Supplemental Information".)

This problem is obvious to anyone who carefully reads that paper. *Science* did not properly peer-review this paper. (See our original refutation of the paper at the link below.)

For our initial refutation of Rao *et al.*, see http://www.safarmer.com/Refutation3.pdf, supplemented now by the materials on the next page in this Presentation. For similar harsh judgments of the Rao paper — and of *Science* for publishing it — from well-known computational linguists (Mark Liberman and Fernando Pereira), see http://tinyurl.com/dgwurl and http://tinyurl.com/cfj5wo.

‘Conditional Entropy’ *Cannot* Distinguish Linguistic from Nonlinguistic Symbol Systems



We calculate the conditional entropy of 10 textual corpora representing 10 different languages (and scripts) and two nonlinguistic sign systems (Indus signs and medieval heraldic signs).

'Conditional entropy' is simply a measure of the degree of order or disorder in any sequentially ordered system, man-made or not. The conditional entropy of *all* symbol systems — *not* just linguistic ones — fall somewhere between the two extremes of complete order and disorder. The method cannot tell the difference between literate or non-literate systems nor can it even reliably distinguish bodies of texts that encode materials from totally unrelated language families. (Compare the odd overlaps and divergences in the conditional entropy of the textual corpora calculated above, which among other things suggest impossible structural affinities between various Semitic, Dravidian, and Indo-European languages — as well as the so-called ‘Indus script’.)

In addition, results will vary widely with corpus size, with the genre of texts sampled, with the kinds of encoding used when the symbols *are* linguistic (i.e., the calculations will give radically different results when the same texts are encoded in different scripts), and with the ‘smoothing’ methods used in the calculations (methods used to estimate the probabilities of signs not seen from those already seen; the accuracy of smoothing methods also largely depends on corpus size).

One consequence of these facts: the published results of Rao *et al*. can’t be replicated even in principle unless you use *exactly* the *same* texts used in their plot — a violation of every norm of replicabilty used in legitimate science.