Written Language vs. Non-Linguistic Symbol Systems: A Statistical Comparison

Richard Sproat

Google, Inc.

76 9th Avenue

New York, NY 10011, USA

Are statistical methods useful in distinguishing written language from *non-linguistic* symbol systems? Some recent papers in *Science* (*1*) and the *Proceedings of the Royal Society* (*2*) have claimed so.

Using a larger set of non-linguistic and comparison linguistic corpora than were used in these and other previous papers, I show that none of the previous proposed methods are useful as published. However, one of the measures proposed in (*2*) (with a different cutoff value), as well as a novel measure based on repetition, turn out to be good measures for classifying symbol systems into the two categories. For the two ancient symbol systems of interest to (*1*) and (*2*) —

the Indus Valley "script" and Pictish symbols, respectively, both of these measures classify them as non-linguistic, contradicting the findings of that previous work.

Introduction

One of the defining habits of human beings is the use of visual marks to convey information. People of all cultures have been communicating with symbols etched on stone, pottery, brick, wood, preparations of leaves, and many other materials for thousands of years. Many different types of information can be represented with symbols. Familiar systems such as traffic signs, written music or mathematical symbols represent information relevant to those domains. In cases like these the information conveyed does not depend upon reference to any specific language: for example most traffic signs are intended to be "readable" in any language. Thus in such cases the information conveyed is *non-linguistic*. One special type of symbol system represents linguistic information — phonological segments, syllables, morphemes or in some cases words: this latter type of *linguistic* symbol system we call *writing*, and the regular use of such a system defines a *literate* culture (*3*). The first literate cultures for which we have clear evidence arose about 5,000 years ago in Mesopotamia and at roughly the same time in Egypt (*4*).

Given this definition, the distinction between linguistic and non-linguistic systems may seem a clear one, but of course the definition presumes one *knows* what the symbols denote. In the case of an ancient culture that left behind "texts" consisting of otherwise unknown symbols, the default situation is that one does *not* know what the symbols mean. Two examples of such cultures are the 3rd millenium BCE Indus Valley culture (*5, 6*), and the Pictish culture of Iron Age Scotland (*7*). In both cultures one finds texts consisting of one or more pictographic symbols. The fact that the symbols are pictographic is of no importance to their status, since all ancient writing systems are also pictographic. The traditional assumptions about these systems — that the Indus symbols were writing, and that the Pictish were probably not — are also of no importance.

One might therefore be tempted to ask if the statistical distribution of symbols in linguistic systems has some sort of "signature" that can distinguish these systems from non-linguistic systems. Indeed, just such a claim has been made in recent papers by Rajesh Rao and colleagues (*1, 8*) for the Indus Valley symbols, and Rob Lee and colleagues (*2*) for Pictish symbols. Both of these strands of work make use of the information-theoretic measure of *entropy* (*9*): in the case of Rao et al's work either conditional entropy or what Rao terms "block entropy"; in the case of Lee et al's work a more complex set of measures which include conditional entropy as one of the terms. In both cases the results seem to suggest that the symbol systems of interest behave more like linguistic systems than they do like the handful of non-linguistic systems that the authors compared them to. Lee et al interpret their method as *discriminative* in that they claim to have found a technique that can classify an unknown system as either non-linguistic or as one of a set of defined types of linguistic system. Rao and colleagues (*1, 8*) on the other hand interpret their method as "inductive": the entropic values for the Indus symbols do not demonstrate that the system was writing, but if you assume that the system *was* writing, then it follows that the entropic values should look like those of known writing systems, which is what they claim to have found.

In this paper we show that none of these approaches work when one compares a larger and more balanced set of non-linguistic systems drawn from a range of cultures and eras. Our non-linguistic corpora, detailed in **S1, S2**, comprise: Vinča symbols (*10*), Pictish symbols (*7, 11–14*), Mesopotamian deity symbols (*kudurrus* (*15–17*)), Totem Poles (*18–29*), Pennsylvania German barn stars ("hex signs" (*30–32*)), sequences of five-day forecast icons from the Weather Underground, and individual Unicode code points in a corpus of Asian emoticons. We also included a small corpus of Indus Valley bar seal texts — a subset of the texts that Rao and colleagues had at their disposal — developed as part of earlier work.

We compared these seven systems with fourteen writing systems coded as indicated in parentheses: Amharic (syllables), Arabic (letters), Chinese (characters), Ancient Chinese Oracle Bone texts (characters), Egyptian (glyphs), English (letters), Hindi (letters), Korean (coded two ways: syllables and letters), Mycenaean Greek (Linear B – syllables), Malayalam (letters), Oriya (letters), Sumerian (glyphs), Tamil (letters). Again, details of these linguistic corpora can be found in **S1**.

Results

We computed both Rao et al's (*1*) bigram conditional entropy and Rao's (*8*) block entropy for our corpora (**S1**). For the block entropy, since Rao gave a reference to the exact method he used (*33*), complete with a pointer to the MATLAB software and the parameter settings, it is possible to produce plots that are directly comparable to his. These, side by side with Rao's plot, are given in Figures 1 and 2. Note that unlike Rao's clean separation of linguistic and non-linguistic systems 1, the true situation evidenced in 2 is much more messy. Linguistic (red) and non-linguistic (blue) systems are interleaved, with no clear pattern. A few points are worth noting. First of all note that *kudurrus* have relatively high entropy, which is surprising given that Rao in various publications (*1, 34*) has insisted that the entropy for *kudurrus* must be low. Second, our bar seal corpus (green) matches Rao's Indus corpus (dark blue in his plot) very well; while they are not expected to be identical, given that the bar seals are a subset of Rao's corpus, it would be surprising if they showed radically different behavior. Finally note that Ancient Chinese has a very low entropy, like Rao's estimates for Fortran: this is not surprising given the repetitive nature of Oracle Bone texts, and it underscores the point that measures like entropy are highly sensitive to the *kind* of text being analyzed.

Lee and colleagues develop two measures, U_r and and C_r, defined as follows. First, U_r is defined as $U_r = \frac{F_2}{\log_2(N_d/N_u)}$, where F_2 is the bigram entropy, N_d is the number of bigram types and N_u is the number of unigram types. C_r is defined as $C_r = \frac{N_d}{N_u} + a\frac{S_d}{T_d}$, where N_d and N_u are as above, a is a constant (for which, in their experiments, they derive a value of 7, using cross-validation), S_d is the number of bigrams that occur once and T_d is the total number of bigram tokens. They use these two features to train a decision tree to classify symbol systems into four bins — non-linguistic, and three types of linguistic writing systems: if $C_r \geq 4.89$, the system is linguistic, and then depending on the value of U_r, the system is segmental ($U_r < 1.09$), syllabic ($U_r < 1.37$) or else logographic. On the basis of that tree they classify Pictish as a linguistic system. The problem is that when we apply their system with their parameters to our non-linguistic corpora, it almost uniformly classifies them as linguistic. The classifications are as follows: Asian emoticons (linguistic: letters), Barn stars (linguistic: letters), Mesopotamian deity symbols (linguistic: syllables), Pictish symbols (linguistic: words), Totem poles (linguistic: words), Weather icon sequences (linguistic: letters), Vinča symbols (non-linguistic). Note that the Pictish symbols are classified as words, which replicates Lee et al's conclusion.

We then considered other measures that might potentially distinguish linguistic from non-linguistic systems. One promising measure is the ratio of the number of symbols that repeat in a text and are *adjacent* to the symbol they repeat (r), to the number of total repetitions in the text (R) (**S1**). This measure is by far the cleanest separator of our data into linguistic versus non-linguistic, with higher values for $\frac{r}{R}$ (e.g. 0.85 for barn stars, 0.79 for weather icons, and 0.63 for totem poles), being nearly always associated with non-linguistic systems, and lower values (e.g. 0.048 for Ancient Chinese, 0.018 for Amharic or 0.0075 for Oriya) being associated with linguistic systems. If we set a value of $\frac{r}{R} = 0.10$ as the boundary, only Sumerian and Mesopotamian deity symbols are clearly misclassified, while Asian emoticons and Egyptian are ambiguously on the border. Both Pictish ($\frac{r}{R} = 0.26$) and Indus symbols ($\frac{r}{R} = 0.58$) are solidly non-linguistic by this measure. The latter is particularly noteworthy in light of the discussion in (*35*) that identified odd repetition patterns in the Indus corpus as problematic for the linguistic hypothesis. Note that the repetition measure correlates negatively with mean text length (Pearson's $r = -0.49$). However the correlation is clearly not perfect and $\frac{r}{R}$ still contributes to a reasonable split between linguistic and non-linguistic corpora, even when we artificially truncate the data to make the texts and corpus sizes more comparable as detailed in **S1**.

Finally we trained classification and regression trees (*36*) with the above-discussed and other measures in a series of experiments, holding out our Indus corpus, and then in a second set of experiments the Pictish corpus. In both cases the vast majority of experimental runs — 98% and 97% respectively — classified these corpora as *non-linguistic*. See **S1** for further details. The features that were most often picked by the trees as discriminative were (not surprisingly) our $\frac{r}{R}$ repetition measure — and Lee and colleagues' C_r. As we discussed above, Lee et al's measures perform very poorly on our non-linguistic corpora when used out of the box with their published cutoff values. But C_r, with cutoff values retrained by the CART algorithm, actually turns out to be a reasonable discriminator (though not as good as $\frac{r}{R}$). On the other hand, interestingly, C_r also correlates positively with mean text length (Pearson's $r = 0.4$, or $r = 0.72$ if we exclude the Amharic data, which for some reason are an outlier).

It is clear from the above discussion that while $\frac{r}{R}$ and C_r turn out to be useful all of the methods as proposed in the previous literature fail as *discriminative* methods: for example, Lee et al's published decision tree misclassifies almost all of our non-linguistic systems as writing. But what about Rao et al's "inductive" interpretation? Again, in Rao's view, the fact that the entropy growth curves for the Indus Valley symbols fall into the same narrow band as linguistic systems, serves as sort of "sanity check" on the hypothesis that the Indus signs were writing. But what we have shown is that this narrow band of entropic growth curves is densely populated by all sorts of symbol systems, both linguistic and non-linguistic; and among the non-linguistic systems many that are clearly meaning-bearing. So the middle band of the entropy curves serves equally well as a sanity check that one is dealing with a meaning bearing — but not necessarily linguistic — symbol system. Since nobody has disputed that the Indus signs must have meant something to the people that used them, this tells us nothing new.

The fact that mean text length seems to be a correlate of the two most discriminative measures we have found — $\frac{r}{R}$ and C_r — is noteworthy insofar as our non-linguistic corpora do tend to be shorter. This is not surprising since while a true writing system is expected to have long texts (since language itself is theoretically unbounded in the length of utterances that can be produced), there is no such *a priori* expectation for non-linguistic systems. One of the problems for the thesis that the Indus symbols were a true writing system is the fact that all extant texts are very short (*35, 37*): the longest text on a single surface is 17 glyphs long, which is quite a bit shorter than our longest *kudurru* text (39 glyphs, and indeed 19 out of our 69 *kudurru* texts are longer than the longest Indus text). As has been argued elsewhere (*35*), the so-called "lost manuscript" hypothesis that states that longer texts were written on perishable materials, all of which have been lost, seems questionable given the absence of other "markers" of literate civilization (for example pens, styluses or ink pots). In any case the belief in the existence of a large trove of now lost literary material in the Indus "script" must be taken as a mere act of faith, in the absence of any substantive evidence for it.

The status of any ancient symbol system as a writing system must be supported by good empirical evidence. As argued in (*35*) for the Indus Valley symbols in particular, good arguments for the linguistic status would be a decipherment into one or more languages that succeeds in convincing a wide body of scholars; the discovery of artifacts indicating an active culture of literacy; or the discovery of a long text or a text that is bilingual with a known contemporaneous writing system. These ought to count as minimal requirements for accepting the thesis that *any* unknown ancient symbol system is writing.

References

1. R. Rao, *et al.*, *Science* **324**, 1165 (2009).

2. R. Lee, P. Jonathan, P. Ziman, *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences* pp. 1–16 (2010).

3. J. Goody, I. Watt, *Literacy in Traditional Societies*, J. Goody, ed. (Cambridge University Press, New York, 1968), pp. 27–68.

4. P. Daniels, W. Bright, eds., *The World's Writing Systems* (Oxford, New York, 1996).

5. G. Possehl, *The Indus Age: The Writing System* (University of Philadelphia Press, Philadelphia, 1996).

6. G. Possehl, *The Indus Age: The Beginnings* (University of Philadelphia Press, Philadelphia, 1999).

7. A. Jackson, *The Pictish Trail* (Orkney Press, 1990).

8. R. Rao, *IEEE Computer* **43**, 76 (2010).

9. C. Shannon, *Bell System Technical Journal* **27**, 379 (1948).

10. S. M. Winn, *The Life of Symbols*, M. L. Foster, ed. (Westview Press, Boulder, 1990), pp. 269–271.

11. A. Jackson, *The Symbol Stones of Scotland: a Social Anthropological Resolution of the Problem of the Picts* (Orkney Press, 1984).

12. Royal Commission on the Ancient and Historical Monuments of Scotland, Pictish symbol stones: A handlist (1994).

13. A. Mack, *Field Guide to the Pictish Symbol Stones* (The Pinkfoot Press, Balgavies, Angus, 1997). Updated 2006.

14. E. Sutherland, *The Pictish Guide* (Birlinn Ltd, Edinburgh, 1997).

15. L. King, *Babylonian Boundary Stones and Memorial Tablets in the British Museum* (British Museum, London, 1912).

16. U. Seidl, *Die babylonischen Kudurru- Reliefs. Symbole mesopotamischer Gottheiten* (Universitätsverlag Freiburg, 1989).

17. J. Black, A. Green, T. Rickards, *Gods, Demons and Symbols of Ancient Mesopotamia* (University of Texas Press, Austin, TX, 1992).

18. British Columbia Provincial Museum, British Columbia totem poles, Printed by Charles F. Banfield printer to the King's most excellent majesty (1931).

19. V. E. Garfield, *The Seattle Totem Pole* (University of Washington Press, Seattle, 1940).

20. M. Barbeau, *Totem Poles*, vol. 1–2 of *Anthropology Series 30, National Museum of Canada Bulletin 119* (National Museum of Canada, Ottawa, 1950).

21. S. W. Gunn, *The Totem Poles in Stanley Park, Vancouver, B.C.* (Macdonald, Vancouver, BC, 1965), second edn.

22. S. W. Gunn, *Kwakiutl House and Totem Poles at Alert Bay* (Whiterocks Publications, Vancouver, BC, 1966).

23. S. W. Gunn, *Haida Totems in Wood and Argillite* (Whiterocks Publications, Vancouver, BC, 1967).

24. F. Drew, *Totem Poles of Prince Rupert* (F.W.M. Drew, Prince Rupert, BC, 1969).

25. E. Malin, *Totem Poles of the Pacific Northwest Coast* (Timber Press, Portland, OR, 1986).

26. City of Duncan, Duncan: City of Totems. (1990). Duncan, BC.

27. H. Stewart, *Totem Poles* (University of Washington, Seattle, WA, 1990).

28. H. Stewart, *Looking at Totem Poles* (University of Washington, Seattle, WA, 1993).

29. R. D. Feldman, *Home before the Raven Caws: the Mystery of the Totem Pole*, Cincinnati, OH (Emmis Books, 2003).

30. A. C. Mahr, *Ohio History: The Scholarly Journal of the Ohio Historical Society* **54**, 1 (1945).

31. T. E. Graves, The Pennsylvania German hex sign: A study in folk process, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA (1984).

32. D. Yoder, T. Graves, *Hex Signs: Pennsylvania Dutch Barn Symbols and their Meaning* (Stackpole, Mechanicsburg, PA, 2000).

33. I. Nemenman, F. Shafee, W. Bialek, *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2002), vol. 14.

34. R. Rao, *et al.*, *Computational Linguistics* **36**, 795 (2010).

35. S. Farmer, R. Sproat, M. Witzel, *Electronic Journal of Vedic Studies* **11** (2004).

36. L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees* (Wadsworth and Brooks, Pacific Grove, CA, 1984).

37. A. Parpola, *Deciphering the Indus Script* (Cambridge University Press, New York, 1994).

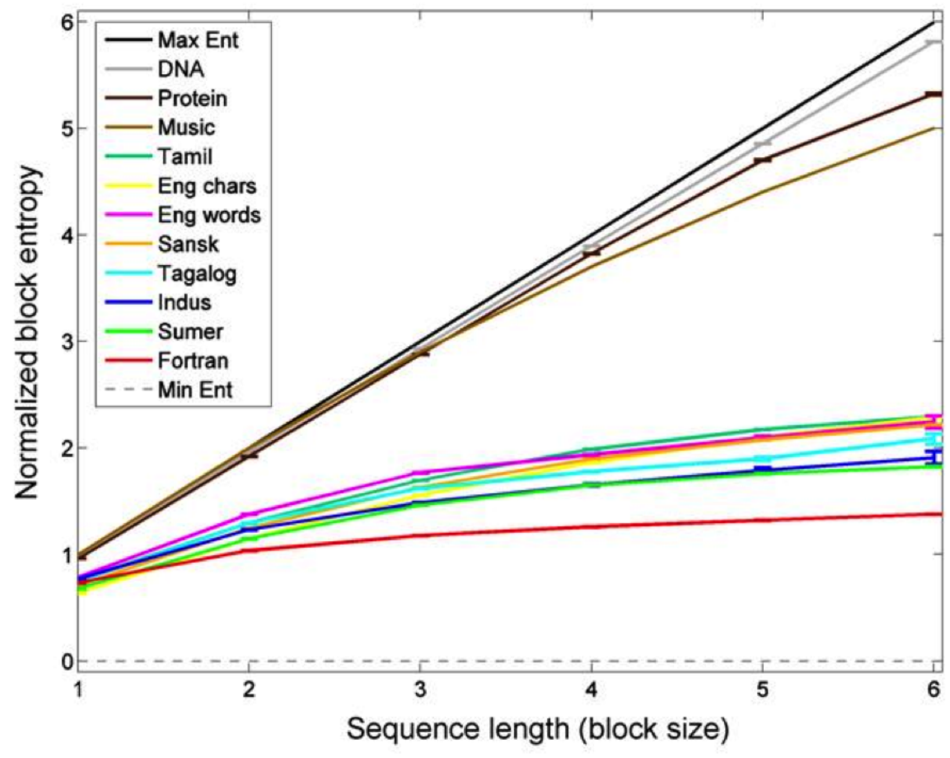

Figure 1: Block entropy estimates from (*8*) for a variety of linguistic systems, some non-linguistic systems, and the Indus corpus. Used with permission of the IEEE.

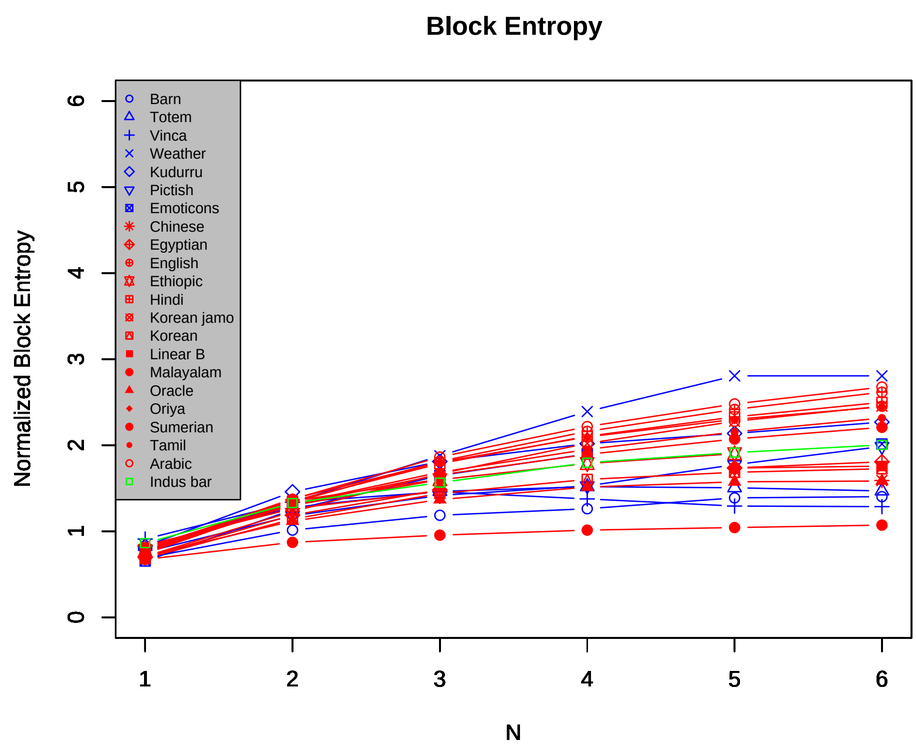

Figure 2: Block entropy values for our linguistic (red) and non-linguistic corpora (blue), and the Indus bar seal corpus (green).