Commentary and Discussion

Entropy, the Indus Script, and Language: A Reply to R. Sproat

Rajesh P. N. Rao*
University of Washington

Nisha Yadav**
Tata Institute of Fundamental Research

Mayank N. Vahia**
Tata Institute of Fundamental Research

Hrishikesh Joglekar†

Ronojoy Adhikari‡
The Institute of Mathematical Sciences

Iravatham Mahadevan§
Indus Research Centre

1. Introduction

In a recent Last Words column (Sproat 2010), Richard Sproat laments the reviewing practices of "general science journals" after dismissing our work and that of Lee, Jonathan, and Ziman (2010) as "useless" and "trivially and demonstrably wrong." Although we expect such categorical statements to have already raised some red flags in the minds of readers, we take this opportunity to present a more accurate description of our work, point out the straw man argument used in Sproat (2010), and provide a more complete characterization of the Indus script debate. A separate response by Lee and colleagues in this issue provides clarification of issues not covered here.

2. The Indus Script Debate

The Indus script refers to the approximately 4,000 inscriptions on seals, miniature tablets, pottery, stoneware, copper plates, tools, weapons, and wood left behind by

* Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA. E-mail: rao@cs.washington.edu.
** Department of Astronomy and Astrophysics, Tata Institute of Fundamental Research, Mumbai 400005, India.
† 14 Dhus Wadi, Thakurdwar, Mumbai 400002, India.
‡ The Institute of Mathematical Sciences, Chennai 600113, India.
§ Indus Research Centre, Roja Muthiah Research Library, Chennai 600113, India.

© 2010 Association for Computational Linguistics

the Indus civilization, which flourished ca. 2600–1900 BCE in South Asia. The existing inscriptions (see Figure 1(a) for examples) are relatively short, the average length being 5 signs and the longest inscription on a single surface being 17 signs. The number of different symbols in the script is estimated to be about 400. This large number of symbols, coupled with archaeological evidence indicating extensive use of the script for a variety of purposes, led scholars to suggest that the script was probably a logosyllabic form of writing, each sign representing a word or syllable (Parpola 1994; Possehl 1996).

In 2004, Sproat and colleagues published in the *Electronic Journal of Vedic Studies* an article whose title makes the unconditional pronouncement "The Collapse of the Indus script thesis: The myth of a literate Harappan civilization" (Farmer, Sproat, and Witzel 2004). The article goes on to list arguments for why the authors believe the Indus script is nonlinguistic (the arguments are said to amount to a "proof" [Farmer 2005]). They propose that the script is a collection of religious or political symbols.

Sproat (2010) states that their arguments "have been accepted by many archaeologists and linguists," without actually citing who these "many archaeologists and linguists" are. In fact, a number of respected scholars, not just those who have "spent most of their careers trying to decipher the symbols" (Sproat 2010), have voiced strong disagreement (Kenoyer 2004; Possehl 2004; Mahadevan 2009). Several have published point-by-point rebuttals (Parpola 2005; Vidale 2007; McIntosh 2008). Parpola, who is widely regarded as the leading authority on the Indus script, writes that the arguments of Sproat and co-workers "can be easily controverted" and goes on to expose the inadequacies of each of these arguments (Parpola 2005, 2008). McIntosh, in a recent book on the ancient Indus valley, also discusses the evidence against the arguments of Sproat and colleagues (McIntosh 2008, pages 372–374). Vidale, a well-known archaeologist, notes that the paper (Farmer, Sproat, and Witzel 2004) "is constructed by repeatedly advancing hypotheses and sometimes wild speculation presented as serious scientific evidence" and concludes by saying: "I see no acceptable scientific demonstration of the non-scriptural nature of the Indus sign system; therefore, I see no collapse of such 'thesis'" (Vidale 2007, page 362).

3. Fallacies Resolved

Under a section entitled "The Fallacies," Sproat (2010) describes a result from our article in *Science* (Rao et al. 2009a) which presents evidence against the thesis of Farmer, Sproat, and Witzel (2004). In our article, we show that the conditional entropy of the Indus script is similar to various linguistic sequences. The impression conveyed by Sproat (2010) is that we are claiming such similarity by itself is sufficient to prove that the Indus script, or indeed any symbol system, is linguistic. We do not make such a claim; instead, we only note in Rao et al. (2009a) that our result increases the evidence for the linguistic hypothesis, when one takes into account other language-like properties of the script (see detailed explanation in Section 4 herein).

To set up his criticism of our work, Sproat (2010) presents Figure 1A from our *Science* paper but *never mentions* the results presented in Figure 1B in the same paper. Nor does he describe our more recent block entropy result (Rao 2010b), even though he cites this paper (this new result extends the conditional entropy work). Both of these results include data from demonstrably nonlinguistic sequences, namely, DNA, protein sequences, and Fortran code. To present our work as "simple experiments involving randomly generated texts" is, to say the least, a gross misrepresentation of our work.

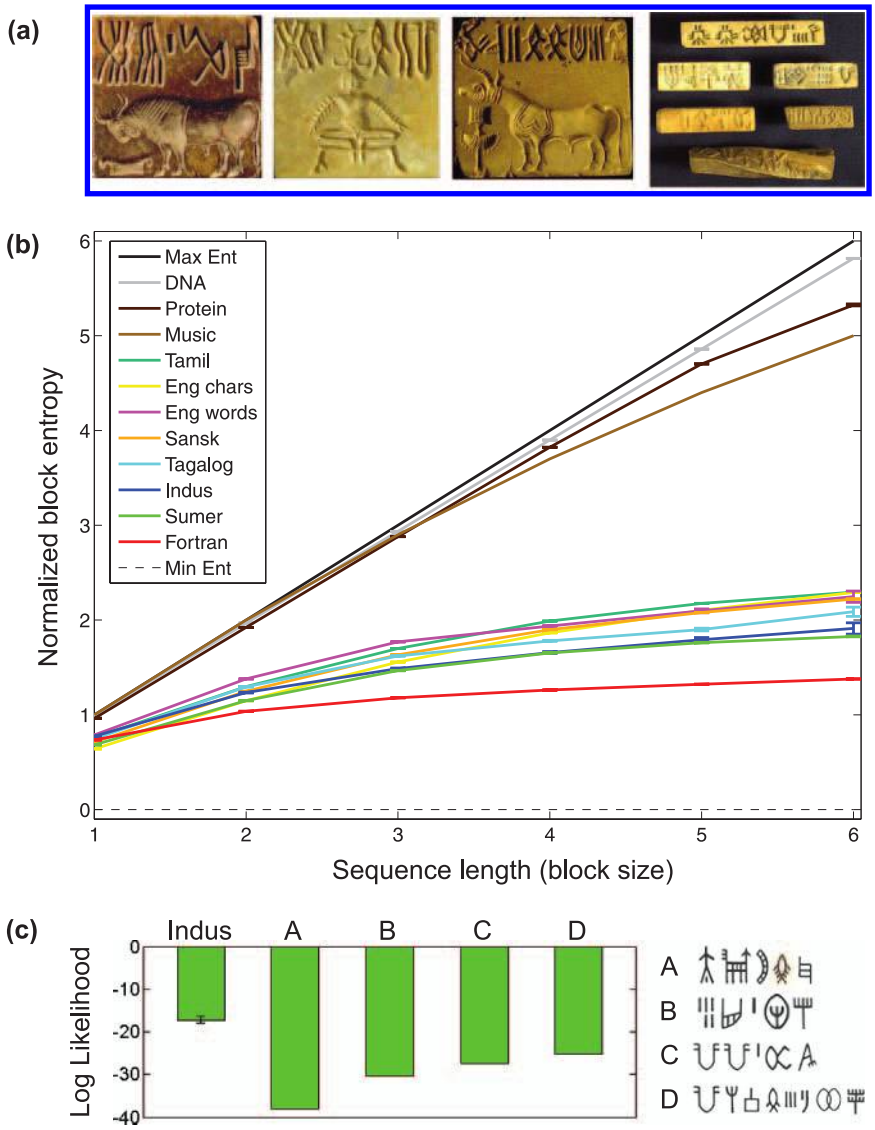

Figure 1
(a) Examples of the Indus script. Three square stamp seals, each with an Indus text at the top. Last image: three rectangular seals and three miniature tablets with inscriptions (image credit: J. M. Kenoyer/Harappa.com). (b) Block entropy scaling of the Indus script compared to natural languages and other sequences. Symbols were signs for the Indus script; bases for DNA; amino acids for proteins; change in pitch for music; characters for English; words for English, Tagalog, and Fortran; symbols in abugida (alphasyllabic) scripts for Tamil and Sanskrit; and symbols in the cuneiform script for Sumerian (see Rao et al. 2009a; Rao 2010a for details). The values for music are from Schmitt and Herzel (1997). To compare sequences over different alphabet sizes L, the logarithm in the entropy calculation was taken to base L (417 for Indus, 4 for DNA, etc.). The resulting normalized block entropy is plotted as a function of block size. Error bars denote one standard deviation above/below mean entropy and are negligibly small except for block size 6. (c) Log likelihood under a first-order Markov model for the Indus corpus for four texts (A through D) found in foreign lands compared to average log likelihood for a random set of 50 Indus region texts not included in the training data (error bar denotes ± 1 standard error of mean). The unusual sequencing of signs in the foreign texts, noted earlier by Parpola (1994), is reflected here in their significantly lower likelihood values.

To correct this misrepresentation, we present in Figure 1(b) the block entropy result (adapted from Rao 2010b). Block entropy generalizes Shannon entropy (Shannon 1948, 1951) and the measure of bigram conditional entropy used in Rao et al. (2009a) to blocks of N symbols. Block entropy for block size N is defined as:

$$H_N = -\sum_i p_i^{(N)} \log p_i^{(N)} \quad (1)$$

where $p_i^{(N)}$ are the probabilities of sequences (blocks) of N symbols. Thus, for $N = 1$, block entropy is simply the standard unigram entropy and for $N = 2$, it is the entropy of bigrams. Block entropy is useful because it provides a measure of the amount of flexibility allowed by the syntactic rules generating the analyzed sequences (Schmitt and Herzel 1997): The more restrictive the rules, the smaller the number of syntactically correct combinations of symbols and lower the entropy. Correlations between symbols are reflected in a sub-linear growth of H_N with N (e.g., $H_2 < 2H_1$).

Figure 1(b) plots the block entropies of various types of symbol sequences as the block size is increased from $N = 1$ to $N = 6$ symbols. To counter the problems posed by the small sample size of the Indus corpus (about 1,550 lines of text and 7,000 sign occurrences), we employed a Bayesian entropy estimation technique known as the NSB estimator (Nemenman, Shafee, and Bialek 2002), which has been shown to provide good estimates of entropy for undersampled discrete data. Details regarding the NSB parameter settings and the data sets used for Figure 1(b) can be found in Rao (2010a).

As seen in Figure 1(b), the block entropies of the Indus texts remain close to those of a variety of natural languages and far from the entropies for unordered and rigidly ordered sequences (Max Ent and Min Ent, respectively). Also shown in the plot for comparison are the entropies for a computer program written in the formal language Fortran, a music sequence (Beethoven's *Sonata no. 32*; data from Schmitt and Herzel [1997]), and two sample biological sequences (DNA and proteins). The biological sequences and music have noticeably higher block entropies than the Indus script and natural languages; the Fortran code has lower block entropies.

Does the similarity in block entropies with linguistic systems in Figure 1(b) *prove* that the Indus script is linguistic? We do not believe so. In fact, we contend that barring a full decipherment, one cannot prove either the linguistic or nonlinguistic thesis, unlike Sproat and colleagues who have previously claimed to have "proof" for the nonlinguistic hypothesis (Farmer, Sproat, and Witzel 2004, pages 34 and 37; Farmer 2005). What we do claim, as we state in our *Science* paper and as explained in more detail subsequently, is that results such as the similarity in entropy in Figure 1(b) increase the evidence for the linguistic hypothesis, given other language-like properties of the Indus script.

However, Sproat, Liberman, and Shalizi (in a blog by Liberman [2009]) and Sproat at EMNLP'09 undertake the exercise of knocking down the straw man ("similarity in conditional entropy by itself implies language") and present artificial counterexamples (e.g., having Zipfian distribution) with conditional independence for bigrams (Sproat 2010). First, such an exercise misses the point: as stated earlier, we do not claim that entropic similarity by itself is a sufficient condition for language. Second, these "counterexamples" ignore the fact that the unigram and bigram entropies are markedly different for both the Indus script and the linguistic systems, as is obvious from comparing Figures 1 and S1 in our *Science* paper (Rao et al. 2009a). More importantly, these artificial examples fail to exhibit the scaling of block entropies beyond unigrams and bigrams exhibited by the Indus script and linguistic systems in Figure 1(b).

Sproat (2010) criticizes our classification of "Type 1" and "Type 2" nonlinguistic systems (corresponding to systems near Max Ent and Min Ent, respectively, in Figure 1(b)), saying these do not characterize any natural nonlinguistic systems. It is clear from Figure 1(b) that there do exist natural "Type 1" nonlinguistic sequences (DNA, protein sequences). The analogous result for conditional entropy was given in Figure 1B in Rao (2010b), which was omitted in Sproat (2010). As for "Type 2" systems, Vidale (2007) provides a number of examples of ancient nonlinguistic systems from Central and South Asia whose properties are in line with such systems. Section 6 herein discusses these systems as well as the specific cases of Vinča and *kudurru* sequences mentioned in Sproat (2010). Sproat, Farmer, and colleagues have objected to the use of artificial data sets in Rao et al. (2009a) to demarcate the Max Ent and Min Ent limits: This objection is a red herring and does not change the result that the Indus script is entropically similar to linguistic scripts. Finally, the allusion in Sproat (2010) that we may be "confused" about the difference between "random" and "equiprobable" is unwarranted and not worthy of comment here. The related issue of artificially generated examples with quasi-Zipfian distributions has already been discussed in this response.

We conclude by noting here that the extension of our original conditional entropy result to block entropies directly addresses the objections of Pereira (2009), who stressed the need to go beyond bigram statistics, which Figure 1(b) does for N up to 6. Beyond $N = 6$, the entropy estimates become less reliable due to the small sample size of the Indus corpus.

4. Inductive Inference

The correct way of interpreting the block entropy result in Figure 1(b) (and likewise the conditional entropy result) is to view it within an inductive framework (rather than in a deductive sense as Sproat and others do in Liberman [2009]). Given that we cannot answer the ontological question "Does the Indus script represent language?" without a true decipherment, we formulate the question as an epistemological problem, namely, one of estimating the posterior probability of the hypothesis H_L that an unknown symbol sequence represents language, given various properties $P_1, P_2, P_3, \ldots$ of the unknown sequence. Using a Bayesian formalism, the posterior $P(H_L|P_1, P_2, P_3, \ldots)$ is proportional to $P(P_1, P_2, P_3, \ldots|H_L)P(H_L)$.

Building on prior work (Hunter 1934; Knorozov, Volchok, and Gurov 1968; Mahadevan 1977; Parpola 1994), we have sought to quantitatively characterize various properties $P_1, P_2, P_3, \ldots$ of the Indus script (Yadav et al. 2008a, 2008b, 2010; Rao et al. 2009a, 2009b; Rao 2010b). In each case, we compare these properties with those of linguistic systems to ascertain whether the property tilts the evidence towards or away from the linguistic hypothesis H_L.

We find these properties to be as follows: **(1) Linearity**: The Indus texts are linearly written, like the vast majority of linguistic scripts (and unlike nonlinguistic systems such as medieval heraldry, Boy Scout merit badges, or highway/airport signs, systems frequently mentioned by Sproat and colleagues). **(2) Directionality**: There is clear evidence for directionality in the script: Texts were usually written from right to left, a fact that can be inferred, for example, from a sign being overwritten by another on its left on pottery (Lal 1966). Directionality is a universal characteristic of linguistic systems but not necessarily of nonlinguistic systems (e.g., heraldry, Boy Scout badges). **(3) Use of Diacritical Marks**: Indus symbols are often modified by the addition of specific sets of marks over, around, or inside a symbol. Multiple symbols are sometimes combined

("ligatured") to form a single glyph. This is similar to linguistic scripts, including later Indian scripts which use such ligatures and diacritical marks above, below, or around a symbol to modify the sound of a root consonant or vowel symbol. **(4) Zipf–Mandelbrot Law**: 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 (Yadav et al. 2010). **(5) Syntactic Structure**: The script exhibits distinct language-like syntactic structure including equivalence classes of symbols with respect to positional preference, classes of symbols that function as beginners and enders, symbol clusters that prefer particular positions within texts, etc. (Hunter 1934; Parpola 1994; Yadav et al. 2008a, 2008b). This structure is evident in both short texts as well as longer texts that are up to 17 symbols long. **(6) Diverse usage**: The script was used on a wide range of media (from seals, tablets, and pottery to copper plates, tools, clay tags, and at least one large wooden board), suggesting a diverse usage similar to linguistic scripts, and unlike nonlinguistic systems such as pottery markings, deity symbols on boundary stones, and so on, whose use is typically limited to one type of medium. **(7) Use in Foreign Lands**: Indus texts have been discovered as far west as Mesopotamia and the Persian Gulf. These texts typically use the same signs as texts found in the Indus region but *alter their ordering*. As shown in Figure 1(c), these "foreign" texts have low likelihood values compared to Indus region texts, *even after taking into account regional variation across the Indus region* (see error bar in Figure 1(c)) (Rao et al. 2009b; Rao 2010b). This suggests that, like other linguistic scripts, the Indus script may have been versatile enough to represent different subject matter or a different language in foreign regions.

Note that although one may find a nonlinguistic system that exhibits one of these properties (e.g., Zipfian distribution) and another that exhibits a different property (e.g., ligaturing), it would be highly unusual for a nonlinguistic system to exhibit a confluence of all of these properties.

To these properties, we add the property in Figure 1(b) that the Indus script shows the same type of entropic scaling as linguistic systems. To estimate the prior probability $P(H_L)$, one could take into account, as a number of scholars have (Vidale 2007; Parpola 2008; Mahadevan 2009), the archaeological evidence regarding the cultural sophistication of the Indus civilization, contact with other literate societies, and the extensive use of the script for trade and other purposes. These factors suggest that $P(H_L)$ is higher than chance. Considering the properties discussed previously and our estimate of $P(H_L)$, the product $P(P_1, P_2, P_3, \ldots | H_L)P(H_L)$ suggests a *higher posterior probability* for the linguistic hypothesis than the nonlinguistic alternative. Given our current data and knowledge about the script, we believe this is the kind of statement one can make about the Indus script, rather than statements about the "collapse" of one thesis or another (Farmer, Sproat, and Witzel 2004).

To claim to have "proof" of the nonlinguistic thesis (Farmer, Sproat, and Witzel 2004, pages 34 and 37; Farmer 2005) would amount to showing a posterior probability of *zero* for the linguistic hypothesis. This is clearly not possible given our current state of knowledge about the script and the lack of an accepted decipherment.

Could the result in Figure 1(b) be an artifact of our particular entropy estimation method? We do not think so. A similar block entropy result was obtained independently by Schmitt and Herzel (1997) using an entirely different entropy estimation method (see Figure 8 in their paper). The overall result is also confirmed by other methods, as discussed by Schmitt and Herzel: "This order—DNA, music, human language, computer language—when ordered by decreasing entropy, is confirmed by the calculation of the Lempel-Ziv complexity (Lempel and Ziv 1976) which also serves as an estimation of the entropy of the source" (Schmitt and Herzel 1997, page 376).

5. Comparison with Ancient Nonlinguistic Systems

Sproat contends that results such as the similarity in entropy scaling in Figure 1(b) are "useless" without analyzing a sizeable number of "ancient nonlinguistic systems" (Sproat 2010). As mentioned earlier, Sproat ignores the fact that the results already include nonlinguistic systems: DNA and protein sequences (perhaps the two "most ancient" nonlinguistic systems!) as well as man-made sequences (Fortran code and music in Figure 1(b)).

We believe entropic results such as Figure 1(b) to be both interesting and useful. An analogy may be apt here: If, in the dim surroundings of a jungle, you notice something moving and then spot some stripes, your belief that what is lurking is a tiger will likely go up, even though it could also be a zebra, a man wearing a tiger costume, or any of a number of possibilities. The observation you made that the object under consideration has stripes is certainly not "useless" in this case, just because you haven't ascertained whether antelopes or elephants in the jungle also have stripes. In other words, we now know that various types of symbol sequences, from natural sequences such as DNA and proteins to man-made systems such as music and Fortran, occupy quite different entropic ranges compared to linguistic systems (Figure 1(b); Figure 8 in Schmitt and Herzel [1997]). Given this knowledge, the finding that Indus sequences occupy the same entropic range as linguistic sequences, although not proving that the Indus script is linguistic, certainly increases the posterior probability of the linguistic hypothesis, just as the observation of stripes increases the posterior probability of the "tiger" hypothesis in our earlier example.[1] As to where ancient nonlinguistic systems may lie among the entropic ranges in Figure 1(b), we discuss this in the next section.

6. Countless Nonlinguistic Sign Systems?

Sproat and colleagues have stated that the properties observed in the Indus script are also seen in "countless nonlinguistic sign systems" (Farmer, Sproat, and Witzel 2004, page 21). Let us consider some of these nonlinguistic systems (Sproat 2010; Farmer, Sproat, and Witzel 2004). Medieval European heraldry, Boy Scout merit badges, and airport/highway signs are not linear juxtapositions of symbols that can be up to 17 symbols long, as we find in the case of the Indus script, nor do they exhibit a confluence of script-like properties as enumerated herein. We invite the reader to compare examples of heraldry (Parker 1894), Boy Scout badges (Boy Scouts of America 2010), and airport/highway signs with the Indus script sequences in Figure 1(a) and judge for themselves whether such a comparison bears merit.

Another nonlinguistic system mentioned in Sproat (2010) is the Vinča sign system, which refers to the markings on pottery and other artifacts from the Vinča culture of southeastern Europe of ca. 6000–4000 BCE. Sproat believes there is order in the Vinča system and states that we "mis-cite" Winn. To set the record straight, here is what Winn has to say in his article in a section on Sign Groups (Winn 1990, page 269):

> Neither the order nor the direction of the signs in these (sign) groups is generally determinable: judging by the frequent lack of arrangement, precision in the order probably was unimportant . . . Miniature vessels also possess sign-like clusters (Figure 12.2j), which are characteristically disarranged.

1 Under certain assumptions, one can derive a quantitative estimate of the increase in posterior probability from a result such as Figure 1(b). We refer the reader to Siddharthan (2009) for details.

This contradicts Sproat (2010) and suggests that the Vinča system, if it indeed lacks precision in the order of signs, would be closer to the maximum entropy (Max Ent) range than to the linguistic scripts in Figure 1(b). The actual amount of lack of precision unfortunately cannot be quantified in entropic terms because a large enough data set of Vinča sequences does not exist.

Sproat also draws attention to the carvings of deities on Mesopotamian boundary stones known as *kudurrus*. He declares that our statement regarding *kudurru* deity sequences obeying rigid rules of ordering compared to linguistic scripts is "clearly false." To shed more light on this issue, we cite here the work of several scholars in this field. Slanski, in a recent in-depth study of the form and function of *kudurrus*, states (Slanski 2003, page 163):

> Hierarchical deployment of the divine symbols. Seidl (1989) observed that, to a certain extent, the divine symbols were deployed upon the Entitlement narûs (*kudurrus*) according to the deities' relative positions in the pantheon. The symbols for the higher gods of the pantheon . . . are generally found upon or toward the top and most prominent part of the monument. Deities lower in the pantheon are deployed toward the bottom of the relief field.

A similar statement on the hierarchical ordering of symbols on *kudurrus* can be found in Black and Green (1992, page 114). The reader will probably agree that a system with even a rough hierarchical ordering among its symbols is more rigid than most linguistic systems. Linguistic systems have no such hierarchy imposed on characters or words, and there is considerable flexibility in where such symbols may be placed within a sequence. Therefore, as originally suggested in Rao et al. (2009a), we expect the entropy of the *kudurru* sequences to be lower than linguistic systems and perhaps slightly above the minimum entropy (Min Ent) range in Figure 1(b). Again, the actual entropy values cannot be estimated because, as admitted in Sproat (2010), a large enough data set of *kudurru* sequences does not exist.

Sproat (2010) says that no one has done the "legwork" of putting together a large data set of ancient nonlinguistic systems. This ignores the work of Vidale (2007), who did put together a set of ten such systems. Vidale questions the relevance of the nonlinguistic systems suggested by Sproat and colleagues because they are neither of the same time period nor from the same geographical region as the Indus script. To rectify this oversight, Vidale lists ten nonlinguistic systems from Central and South Asia that were roughly contemporaneous with the Indus script (Table 1 in Vidale 2007). For this set of much more relevant nonlinguistic systems, Vidale demonstrates that the average number of different signs is only about 44, a far cry from the 400 or so signs in the Indus script.

Are the kind of positional regularities found in the Indus script also found in *countless nonlinguistic sign systems* (Farmer, Sproat, and Witzel 2004, page 21)? Vidale states that the archaeological data lead us to question this "superficial claim" (Vidale 2007, page 344). In the ten nonlinguistic systems roughly contemporary with the Indus script, positional regularities can either be "largely ruled out" (e.g., in potters' markings where signs occur mostly in isolation and rarely in couples) or the regularities take the form of "systematic, large-scale redundancy" (e.g., constant repetition of the same symbols). Such systems would fall roughly in the "Type 2" category of nonlinguistic systems suggested in our *Science* paper (Rao et al. 2009a), lying closer to the minimum entropy (Min Ent) range in Figure 1(b) than to the Indus script and linguistic systems.

7. Implications of the Linguistic versus Nonlinguistic Hypotheses

If the Indus script does encode language, what might the content of the inscriptions be? A majority of the Indus texts are found on stamp seals (Figure 1(a)), which were typically used in Bronze Age cultures for regulating trade. Seals were pressed onto clay used to seal packages of goods. Indeed, a number of such clay tags have been found at various sites in the Indus civilization, bearing seal impressions on one side and impressions of woven cloth, reed matting or other packing material on the other. These archaeological observations suggest that the short Indus texts on seals (Figure 1(a)), like their other Bronze age counterparts, probably represent the contents, the origin or destination, the type or amount of goods being traded, name and title of the owner, or some combination of these. Similar linguistic explanations can be found for the inscriptions on other media.

If, on the other hand, as Sproat and colleagues propose, the script merely represents religious or political symbols, one is hard pressed to explain: (1) how and why were sequences of such symbols, with syntactic rules entropically similar to linguistic scripts (Figure 1(b)), used in trade in a manner strikingly similar to other *literate* Bronze age cultures? and (2) why did the Indus people use these symbols in consistent sequences in their native region and alter their ordering when in a foreign land (Figure 1(c))? As pointed out by other authors (Vidale 2007; Parpola 2008; Mahadevan 2009), such incongruities are more the norm than the exception if one accepts the nonlinguistic thesis espoused by Sproat and colleagues. The principle of Occam's razor then suggests that we reject the nonlinguistic hypothesis in favor of the simpler linguistic one.

8. Conclusion

A large number of identification problems are amenable to statistical tests, and represent perhaps the only way to solve these problems. Practical examples include separating e-mail from spam and recognizing faces in digital camera images. Even though we may not have a perfect test for any of these problems, the statistical methods that are used can be quite useful, even if they are fallible (we all rely on spam detectors for e-mail even if they occasionally let a spam e-mail through; we do not discard these detectors as "useless"). An important goal of our work (Rao et al. 2009a, 2009b; Rao 2010b; Yadav et al. 2010) has been to develop better statistical tests for linguistic systems. As with other statistical tests, it would be foolhardy to expect that a single such test is infallible, as assumed by Sproat and others in their quest to find "counterexamples" (Sproat 2010). The observation that a single statistical test by itself is insufficient was the primary motivation for the inductive framework adopted in our research, where we apply a range of tests and estimate the posterior probability that an unknown sequence represents language (Section 4).

In the concluding remarks of his Last Words column, Sproat says it is not clear if editors of prominent science journals "even know that there are people who spend their lives doing statistical and computational analyses of text" (Sproat 2010). We find such a statement surprising because it fails to acknowledge both the impressive achievements of the field of computational linguistics in recent years and the wide coverage of these accomplishments in the popular press (Fletcher [2010] and Lohr and Markoff [2010], to give two recent examples).

Computational linguistics is playing an important role in our understanding of ancient scripts (Koskenniemi 1981; Knight and Yamada 1999; Rao et al. 2009a, 2009b; Snyder, Barzilay, and Knight 2010; Yadav et al. 2010). Rather than representing a "misuse

of the methods of the field of computational linguistics" (Sproat 2010), techniques from the field are providing new insights into the structure and function of undeciphered scripts such as the Indus script. For example, we now know that the kind of regularities found in the Indus script can be exploited by statistical models to fill in missing or illegible inscriptions with most likely predictions (Rao et al. 2009b; Yadav et al. 2010). Recent *n*-gram analysis of the Indus script has revealed that there are interesting dependencies between signs that go beyond just pairs of signs (Yadav et al. 2010). Additionally, these statistical models have allowed us to quantify the differences between Indus inscriptions found in West Asia and those found in the Indus region (Figure 1(c)), suggesting the script may have been flexible enough to represent different content or even a different language in foreign regions.

Indus script research has benefited immensely from the application of ideas and methods from computational linguistics for almost four decades (Koskenniemi, Parpola, and Parpola 1970; Koskenniemi 1981; Parpola 1994). We believe computational linguistics will continue to make important contributions to Indus script research in the years to come.

References

Black, Jeremy and Anthony Green. 1992. *Gods, Demons and Symbols of Ancient Mesopotamia*. British Museum Press, London.

Boy Scouts of America. 2010. Introduction to merit badges. `http://www.scouting.org/scoutsource/BoyScouts/AdvancementandAwards/MeritBadges.aspx`.

Farmer, Steve. 2005. Simple proof against the 'Indus script'. `http://www.safarmer.com/indus/simpleproof.html`.

Farmer, Steve, Richard Sproat, and Michael Witzel. 2004. The collapse of the Indus script thesis: The myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies*, 11(2):19–57.

Fletcher, Owen. 2010. Microsoft mines Web to hone language tool. *Wall Street Journal*, August 3. `http://online.wsj.com/article/SB10001424052748703545604575406771145298614.html`.

Hunter, Gerald. 1934. *The Script of Harappa and Mohenjodaro and Its Connection with Other Scripts*. Kegan Paul, London.

Kenoyer, Mark. 2004. Quoted in Lawler, Andrew. 2004. The Indus script: Write or wrong? Science, 306:2026–2029; page 2026.

Knight, Kevin and Kenji Yamada. 1999. A computational approach to deciphering unknown scripts. *Proc. of ACL Workshop on Unsup. Learning in Natural Lang. Processing*.

Knorozov, Yuri, Volchok, B. Y., and Gurov, N. V. 1968. *Proto-Indica: Brief Report on the Investigation of the Proto-Indian Texts*. Academy of Sciences of the USSR, Moscow.

Koskenniemi, Kimmo. 1981. Syntactic methods in the study of the Indus script. *Studia Orientalia*, 50:125–136.

Koskenniemi, Seppo, Asko Parpola, and Simo Parpola. 1970. A method to classify characters of unknown ancient scripts. *Linguistics*, 61:65–91.

Lal, B. B. 1966. The direction of writing in the Harappan script. *Antiquity*, XL:52–55.

Lee, Rob, Philip Jonathan, and Pauline Ziman. 2010. Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A*, 466:2545–2560.

Lempel, Abraham and Jacob Ziv. 1976. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22:75–81.

Liberman, Mark. 2009. Conditional entropy and the Indus script. `http://languagelog.ldc.upenn.edu/nll/?p=1374`.

Lohr, Steve and John Markoff. 2010. Computers learn to listen, and some talk back. *New York Times*, June 24. `http://www.nytimes.com/2010/06/25/science/25voice.html`.

Mahadevan, Iravatham. 1977. *The Indus Script: Texts, Concordance and Tables*. Archaeological Survey of India, Calcutta.

Mahadevan, Iravatham. 2009. The Indus non-script is a non-issue. *The Hindu*, May 3. `http://www.hindu.com/mag/2009/05/03/stories/2009050350010100.htm`.

McIntosh, Jane. 2008. *The Ancient Indus Valley: New Perspectives*. ABC CLIO, Santa Barbara, CA.

Nemenman, Ilya, Fariel Shafee, and William Bialek. 2002. Entropy and inference,

revisited. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA, pages 471–478.

Parker, James. 1894. A glossary of terms used in heraldry. http://www.heraldsnet.org/saitou/parker/gallery/Page.html.

Parpola, Asko. 1994. *Deciphering the Indus script*. Cambridge University Press, New York.

Parpola, Asko. 2005. Study of the Indus script. *Proceedings of the International Conference of Eastern Studies*, 50:28–66.

Parpola, Asko. 2008. Is the Indus script indeed not a writing system? In *Airavati: Felicitation Volume in honor of Iravatham Mahadevan*. Varalaaru.com publishers, India, pages 111–131. http://www.harappa.com/script/indus-writing.pdf.

Pereira, Fernando. 2009. Falling for the magic formula. http://earningmyturns.blogspot.com/2009/04/falling-for-magic-formula.html.

Possehl, Gregory. 1996. *The Indus Age: The Writing System*. University of Pennsylvania Press, Philadelphia.

Possehl, Gregory. 2004. Quoted in Lawler, Andrew. 2004. The Indus script: Write or wrong? Science, 306:2026–2029.

Rao, Rajesh. 2010a. Block entropy analysis of the Indus script and natural languages. http://www.cs.washington.edu/homes/rao/BlockEntropy.html.

Rao, Rajesh. 2010b. Probabilistic analysis of an ancient undeciphered script. *IEEE Computer*, 43(4):76–80.

Rao, Rajesh, Nisha Yadav, Mayank Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009a. Entropic evidence for linguistic structure in the Indus script. *Science*, 324:1165.

Rao, Rajesh, Nisha Yadav, Mayank Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009b. A Markov model of the Indus script. *Proceedings of the National Academy of Sciences (PNAS)*, 106:13685–13690.

Schmitt, Armin and Hanspeter Herzel. 1997. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology*, 1888:369–377.

Seidl, Ursula. 1989. *Die babylonischen Kudurru-Reliefs. Symbole mesopotamischer Gottheiten*. Universitätsverlag Freiburg, Freiburg.

Shannon, Claude. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Shannon, Claude. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64.

Siddharthan, Rahul. 2009. More Indus thoughts and links. http://horadecubitus.blogspot.com/2009/05/more-indus-thoughts-and-links.html.

Slanski, Kathryn. 2003. *The Babylonian Entitlement Narûs (*kudurrus*): A Study in Their Form and Function*. American Schools of Oriental Research, Boston, MA.

Snyder, Benjamin, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1048–1057, Uppsala.

Sproat, Richard. 2010. Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics*, 36(3):585–594.

Vidale, Massimo. 2007. The collapse melts down: A reply to Farmer, Sproat and Witzel. *East and West*, 57:333–366.

Winn, Shan. 1990. A Neolithic sign system in southeastern Europe. In M. L. Foster and L. J. Botscharow, editors, *The Life of Symbols*. Westview Press, Boulder, CO, pages 269–271.

Yadav, Nisha, Hrishikesh Joglekar, Rajesh Rao, Mayank Vahia, Ronojoy Adhikari, and Iravatham Mahadevan. 2010. Statistical analysis of the Indus script using n-grams. *PLoS One*, 5(3):e9506. doi:10.1371/journal.pone.0009506.

Yadav, Nisha, Mayank Vahia, Iravatham Mahadevan, and Hrishikesh Joglekar. 2008a. Segmentation of Indus texts. *International Journal of Dravidian Linguistics*, 37(1):53–72.

Yadav, Nisha, Mayank Vahia, Iravatham Mahadevan, and Hrishikesh Joglekar. 2008b. A statistical approach for pattern search in Indus writing. *International Journal of Dravidian Linguistics*, 37(1):39–52.

This article has been cited by:

1. Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, Ramon Ferrer-i-Cancho. 2017. The Entropy of Words?Learnability and Expressivity across More than 1000 Languages. *Entropy* **19**:6, 275. [CrossRef]
2. Aladhahalli Shivegowda Kavitha, Palaiahnakote Shivakumara, Govindaraj Hemantha Kumar, Tong Lu. 2017. A new watershed model based system for character segmentation in degraded text lines. *AEU - International Journal of Electronics and Communications* **71**, 45-52. [CrossRef]
3. Ahmad M. Manasrah, Basim Najim Al-Din. 2016. Mapping private keys into one public key using binary matrices and masonic cipher: Caesar cipher as a case study. *Security and Communication Networks* **9**:11, 1450-1461. [CrossRef]
4. Antoni Hernández-Fernández, Ramon Ferrer-i-Cancho. 2016. The Infochemical Core. *Journal of Quantitative Linguistics* **23**:2, 133-153. [CrossRef]
5. Michael P. OakesLiterary Detective Work on the Computer **12**, . [CrossRef]