A statistical comparison of written language and nonlinguistic symbol systems

Richard Sproat

Language, Volume 90, Number 2, June 2014, pp. 457-481 (Article)

Published by Linguistic Society of America
DOI: https://doi.org/10.1353/lan.2014.0031

A STATISTICAL COMPARISON OF WRITTEN LANGUAGE AND NONLINGUISTIC SYMBOL SYSTEMS

RICHARD SPROAT

Google, Inc.

Are statistical methods useful in distinguishing written language from nonlinguistic symbol systems? Some recent articles (Rao et al. 2009a, Lee et al. 2010a) have claimed so. Both of these previous articles use measures based at least in part on bigram conditional entropy, and subsequent work by one of the authors (Rao) has used other entropic measures. In both cases the authors have argued that the methods proposed either are useful for discriminating between linguistic and nonlinguistic systems (Lee et al.), or at least count as evidence of a more 'inductive' kind for the status of a system (Rao et al.).

Using a larger set of nonlinguistic and comparison linguistic corpora than were used in these and other studies, I show that none of the previously proposed methods are useful as published. However, one of the measures proposed by Lee and colleagues (2010a) (with a different cut-off value) and 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 Rao and colleagues (2009a) and Lee and colleagues (2010a)—Indus Valley inscriptions and Pictish symbols, respectively—both of these measures classify them as nonlinguistic, contradicting the findings of those previous works.*

Keywords: symbol systems, nonlinguistic symbols, writing systems, statistical models of language

1. INTRODUCTION. One of the defining habits of human beings is the use of visual marks to convey information. People of many if not 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 is intended to be understood without reference to a particular natural language: for example, most traffic signs are intended to be 'readable' in any language. Thus, in such cases the information conveyed is NONLINGUISTIC. One special type of symbol system, a WRITING SYSTEM, represents linguistic information: phonological segments, syllables, morphemes, or in some cases words. The regular use

* A number of people have given input of one kind or another to this work. First and foremost, I would like to thank my research assistants Katherine Wu, Jennifer Solman, and Ruth Linehan, who worked on the development of many of the corpora and the markup scheme. Katherine Wu developed the corpora for totem poles, Mesopotamian deity symbols, Vinča, and Pictish, and Ruth Linehan the heraldry corpus. Jennifer Solman developed the XML markup system used in most of our corpora. An initial report on our joint work was presented in Wu et al. 2012.

I also thank audiences at the 2012 annual meeting of the Linguistic Society of America in Portland, the Center for Spoken Language Understanding, the Oriental Institute at the University of Chicago, King's College London, Johns Hopkins University, the University of Maryland, and Carnegie Mellon University for feedback and suggestions on presentations of this work. Chris Woods gave me useful suggestions for Sumerian texts to use. William Boltz and Ken Takashima were instrumental in pointing me to the corpus of transcribed oracle bones. Thanks to Chris Whelan for discussion of issues surrounding DNA structure. I would like to acknowledge the staff of the Historical Society of Berks County, in particular Ms. Kimberly Richards-Brown, for their help in locating W. Farrell's slide collection of barn stars. I thank Brian Roark, Shalom Lappin, Suyoun Yoon, and especially Steve Farmer for lots of useful discussion. I would also like to thank a referee for *Language* for very detailed and helpful comments on an earlier version of this article.

This material is based upon work supported by the National Science Foundation under grant number BCS-1049308. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

of such a system defines a literate culture (Goody & Watt 1968). 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 (Daniels & Bright 1996, e.g. pp. 33, 73).

Given this definition, the distinction between linguistic and nonlinguistic systems may seem a clear one. But, of course, this definition requires one to know what the symbol system in question denotes, and for many corpora of 'texts' present in the archaeological record, one does NOT know what the symbols mean. Two examples of such cultures are the third-millennium BC Indus Valley culture (Possehl 1996, 1999) and the Pictish culture of Iron Age Scotland (sixth to eighth centuries AD) (Jackson 1990). In both cultures one finds 'texts' consisting of one or more pictographic symbols.[1] The fact that the symbols are pictographic is of no importance to their status, since all ancient writing systems (e.g. Egyptian, Sumerian, Ancient Chinese, Linear B, Mayan) 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.

Recent articles by Rajesh Rao and colleagues (Rao et al. 2009a, Rao 2010) (Indus Valley symbols) and Rob Lee and colleagues (Lee et al. 2010a) (Pictish symbols) have attempted to argue that the statistical distribution of symbols in linguistic systems has 'signatures' that are useful in distinguishing them from nonlinguistic systems. Both of these strands of work make use of the information-theoretic measure of ENTROPY (Shannon 1948): in the case of Rao and colleagues' work, either (bigram) conditional entropy or what Rao terms 'block entropy'; in the case of Lee and colleagues' work, a more complex set of measures that 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 nonlinguistic systems that the authors compared them to. Lee and colleagues interpret their method as DISCRIMINATIVE in that they claim to have found a technique that can classify an unknown system as either nonlinguistic or as one of a set of defined types of linguistic system. Rao and colleagues, by contrast, 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 article I show that none of these approaches work when one compares a larger and more balanced set of nonlinguistic systems drawn from a range of cultures and eras. The nonlinguistic corpora used here, detailed in the online supplementary materials,[2] comprise: Vinča symbols (Winn 1990), Pictish symbols (Jackson 1984, 1990, Royal Commission on the Ancient and Historical Monuments of Scotland 1994, Mack 1997, Sutherland 1997), Mesopotamian deity symbols—KUDURRUS (King 1912, Seidl 1989, Black et al. 1992), totem poles (Newcombe & British Columbia Provincial Museum 1931, Garfield 1940, Barbeau 1950, Gunn 1965, 1966, 1967, Drew 1969, Malin 1986, City of Duncan 1990, Stewart 1990, 1993, Feldman 2003), Pennsylvania German barn stars—'hex signs' (Mahr 1945, Graves 1984, Yoder & Graves 2000), sequences of five-day forecast icons from the weather forecast website Weather Underground (http://www.wunderground.com), and individual Unicode code points in a corpus of Asian emoticons. I also included a small corpus of Indus Valley bar seal texts—a subset of the

[1] I am following common practice in calling a system pictographic if a significant proportion of the symbols in the script appear to be drawings of physical objects; see for example Daniels & Bright 1996:125, 192.

[2] Available at http://muse.jhu.edu/journals/language/v090/90.2.sproat01.pdf. References in this article to tables and figures from the online supplement are indicated with an 'S', for example, Table S5.

texts that Rao and colleagues had at their disposal—developed as part of earlier work. Finally, I discuss analyses of two corpora, mathematical equations and European heraldry, that are still under development. The development of some of these corpora was described in Wu et al. 2012.

These seven systems were compared with fourteen writing systems coded as indicated in parentheses:[3] Amharic (alphasyllabic (C)V syllables), Arabic (abjad letters), Chinese (characters), Ancient Chinese oracle bone texts (characters), Egyptian (glyphs), English (letters), Hindi (alphasyllabic letters), Korean (coded two ways: hangeul syllable blocks, and JAMO letters), Mycenaean Greek (Linear B—syllable glyphs), Malayalam (alphasyllabic letters), Oriya (alphasyllabic letters), Sumerian (glyphs), Tamil (alphasyllabic letters). Again, details of these linguistic corpora can be found in the online supplementary materials.

2. ANALYSES AND RESULTS. A number of statistical measures were computed on these linguistic and nonlinguistic corpora, including measures that have been previously suggested in the literature as well as some novel measures.

In the case of previously proposed measures, the results of using these measures on our corpora were compared with those reported in previous work. One of the results of this comparison is that previously proposed measures are largely uninformative about a symbol system's status. In particular, Rao's entropic measures (Rao et al. 2009a, Rao 2010) are evidently useless when one considers a wider range of examples of real nonlinguistic symbol systems. And Lee's measures (Lee et al. 2010a), with the operating points they propose, misclassify nearly all of our nonlinguistic systems. However, I also show that one of Lee's measures, with different operating points, and another measure developed here do seem useful. But I further demonstrate that they are useful largely because they are both highly correlated with a rather trivial feature: mean text length.

2.1. BIGRAM CONDITIONAL ENTROPY. One popular measure for quantifying the information structure of language is conditional entropy (Shannon 1948), for example, bigram conditional entropy, defined as in 1.

$$(1)\quad H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y)\log p(y|x)$$

Entropy is a measure of information content: in the case of bigram entropy, if for any symbol x there is a unique y that can follow x, then the message carries no information, and the entropy will be minimal. If, by contrast, for any symbol x any other symbol can follow, and with equal probability, then the entropy will be maximal. Most real symbol systems, including language and (as we see below) a whole range of nonlinguistic systems, fall between these two extremes.

Bigram conditional entropy was used in Rao et al. 2009a to argue for the linguistic status of the Indus Valley symbols. One can compute entropy over various-sized units (symbols in the script or symbol system), words, and so forth, and for various portions of the corpus. In the Rao et al. approach, they computed the entropy between the n most frequent units, then the $2n$ most frequent, the $3n$ most frequent, and so forth. This procedure derives an entropy growth curve that starts relatively low and grows until it reaches the full entropy estimate for the corpus.

Figure 1 shows these curves from Rao et al. 2009a, taking $n = 20$, for a variety of actual writing systems, Mahadevan's corpus of Indus inscriptions (Mahadevan 1977), and

[3] For the terms ALPHASYLLABARY and ABJAD, see Daniels & Bright 1996:4.

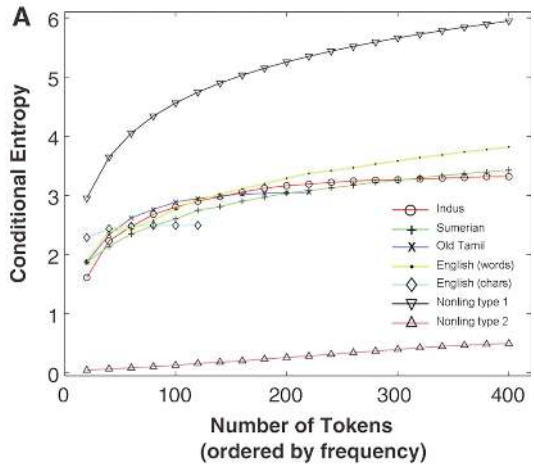

FIGURE 1. Bigram conditional entropy growth curves of various linguistic symbol systems, the Indus Valley symbols, and two artificial models of nonlinguistic systems, from Rao et al. 2009a. See the text for further explanation. Used with permission of the American Association for the Advancement of Science.

two artificial systems with maximum entropy and minimum entropy, which Rao and colleagues label as 'type 1' and 'type 2', respectively. Figure 2 shows the RELATIVE CONDITIONAL ENTROPY—'the conditional entropy for the full corpus relative to a uniformly random sequence with the same number of tokens'—for the various systems represented in Fig. 1, as well as the real nonlinguistic systems, DNA, protein ('prot'), and Fortran ('prog lang').

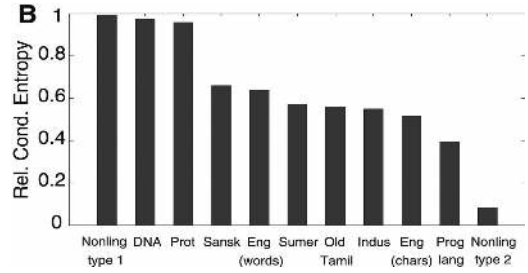

FIGURE 2. Relative conditional entropy (conditional entropy relative to a uniformly random sequence with the same number of tokens) of various systems from Rao et al. 2009a. Used with permission of the American Association for the Advancement of Science.

While the maximum ('type 1') and minimum ('type 2') curves are generated from completely artificial systems, Rao and colleagues argue that they are in fact reasonable models for actual symbol systems: Vinča symbols and kudurrus, respectively. Their beliefs are based, respectively, on Winn's (1990) description of a certain subset of the Vinča corpus as having no discernible order; and on the fact that there is a certain hierarchy observed in the ordering of the symbols on kudurru stones. In Sproat 2010a,b, I showed that this interpretation of how these symbol systems work is incorrect, and this is confirmed by the results presented below.[4]

[4] It is clear from Fig. 2 that the curves of the biological symbol systems (DNA and protein) are very close to the maximum entropy curves. At least in the case of DNA, though, this is because Rao and colleagues take

Unfortunately, Rao and colleagues do not provide sufficient details to perfectly replicate their results. That is, while they report using Kneser-Ney smoothing (Kneser & Ney 1995), a number of details are not clear: for example, it is not clear whether they included estimates for unseen bigrams in their analysis.[5] Discrepancies such as this and others may help explain why the entropy estimates presented below are as a whole lower than those presented in Rao et al. 2009a. Still, while it is not possible to produce plots that are directly comparable to Rao's plots, we can nonetheless produce entropy growth curves for a variety of systems using similar estimation techniques. In Figure 3 conditional entropy growth curves are shown for all of the nonlinguistic systems under investigation here (in blue), all of the linguistic samples (in red), and for a corpus of Indus bar seals from Harappa and Mohejo Daro (in green), described in the online supplementary materials. Entropy statistics were computed using the OpenGrm NGram library (Roark et al. 2012), available from http://www.opengrm.org. Kneser-Ney smoothing was used, including the probability and thus entropy estimates for unseen cases. Start and end symbols are implicit in the computation.

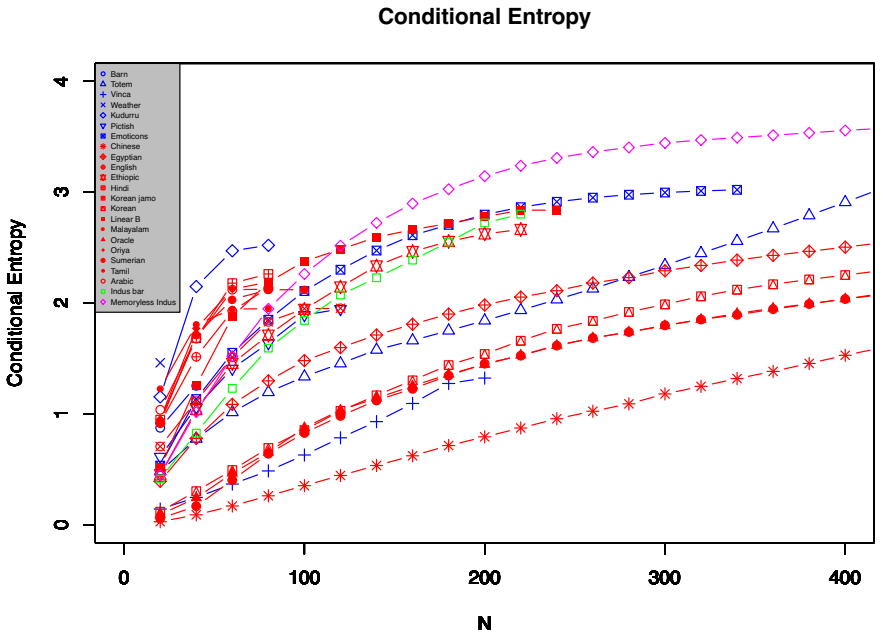

FIGURE 3. Bigram conditional entropy growth curves computed for our linguistic and nonlinguistic corpora, our Indus bar seals, and our memoryless Indus corpus.

the 'alphabet' of DNA to be the bases, an alphabet of size four. But these are surely not the right units of information: the bases are effectively the 'bits' of DNA. The equivalent for natural language would be taking the basic units of English to be the ones and zeroes of the ASCII encoding—and throwing in large amounts of random-looking 'noncoding regions' to mimic the noncoding regions of DNA—surely a meaningless comparison for these purposes. See below for further discussion.

[5] Smoothing and related techniques are much studied in the literature on statistical language modeling and are a way to deal with the issue of data sparsity. Most language models attempt to estimate the probability of seeing a token (e.g. a word) given a previous context (e.g. the last two words). One derives these probability estimates from training data, but even if one has a very large amount of data, it is unlikely that one has seen enough instances of all possible combinations to have accurate estimates for all words in all contexts. Smoothing techniques attempt to provide estimates for unseen combinations by analogizing from combinations that are more common, and for which one has seen sufficient instances. Kneser-Ney smoothing, one of many techniques that have been proposed, has been argued to be a more stable estimate for certain kinds of

Also in the plot (in magenta) is a 'corpus' of 3,373 fake Indus texts with an average text length of five symbols and 16,953 total tokens. This was generated by random sampling from the frequency distribution of the 417 symbol types from Mahadevan 1977. This system is completely MEMORYLESS, in that each symbol is generated with regard only to its unigram frequency, not to its likelihood of following any other symbol. It thus has the same symbol distribution, but none of the structure, of the original system. Note also that the distribution roughly follows a power law (Zipf 1949)—see Sproat 2010a for discussion of this point in connection with nonlinguistic symbol systems.

As can be seen in the plot, the results do not show much structure. The kudurru inscriptions—contrary to what was claimed in Rao et al. 2009a and again in Rao 2010—turn out to have the highest entropy of all the systems. Chinese shows the lowest. Between these two extremes the linguistic and nonlinguistic systems are mixed together. The Indus bar seal corpus, as in the Rao et al. 2009a plot, falls nicely in the middle of this distribution, but that is quite uninformative since the distribution is a mix of both types of system. Even the memoryless system, which shows a higher entropy than most of the other systems, is still close to systems that are unequivocally linguistic.

One concern is that the corpora are of many different sizes, which would definitely affect the probability estimates. How serious is this effect? Could the unexpectedly high entropy estimate (from Rao and colleagues' point of view) for the kudurru corpus be due simply to undersampling? That concern is addressed in Figure 4, which plots the bigram conditional entropy growth curves for three samples from our Arabic newswire headline corpus containing 100 lines, 1,000 lines, and 10,000 lines of text.

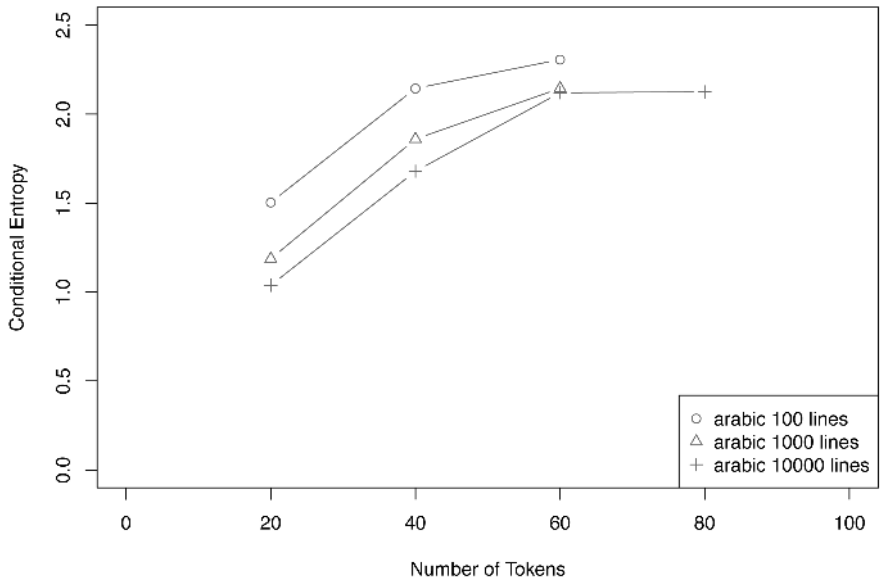

FIGURE 4. Bigram conditional entropy growth curves for three different-sized Arabic corpora.

Note that Arabic has roughly the same number of symbol types as the kudurru corpus, and is about four times larger in terms of numbers of tokens (the full Arabic corpus is over 400 times larger). The entropy does indeed change as the corpus size changes, with

unseen cases that occur in language modeling. For more on smoothing techniques in general, see Manning & Schütze 1999, Roark & Sproat 2007, and Jurafsky & Martin 2008.

the smaller corpora having higher overall entropy than the larger corpora. This is not surprising, since with the smaller corpora, the probability estimates for unseen cases are poorer and hence more uniform, leading in turn to higher entropy estimates. But the difference is not huge, suggesting that one cannot attribute the relatively high entropy of the kudurru corpus wholly to its small size. Given a larger corpus, the estimates would surely be lower. But there is clearly no reason to presume, as Rao does, that a large kudurru corpus would show a conditional entropy near zero.

It is also worth noting that while the kudurru corpus is small, 939 symbols, since there are only sixty-nine distinct symbol types, this is a far better sample of the space than the Mahadevan corpus used by Rao and colleagues. In that case there are about 7,000 symbols, with 417 types. The size of the bigram parameter space for the kudurru corpus is $69^2 = 4{,}761$, and the corpus is about one fifth this size. In the case of the Indus symbols, the bigram parameter space is about 174K, and Mahadevan's Indus corpus is about one twenty-fifth this size. The situation is of course even worse for trigrams and so forth as required for the BLOCK ENTROPY calculations discussed below.

Bigram conditional entropy thus seems not to be very useful as evidence for the linguistic status of symbol systems. This is hardly surprising: an entropy measure in the 'middle' range merely tells us that the system is neither close to completely unconstrained, nor close to completely rigid, something one might expect to hold of most if not all symbol systems.

2.2. BLOCK ENTROPY. We turn now to a second entropic measure proposed by Rao (2010), what he terms 'block entropy', defined as in 2.

$$(2)\quad H_N = -\sum_i p_i^{(N)} \log_{|V|} p_i^{(N)}$$

Here N is the length of an n-gram of tokens, say three for a trigram, so all we are doing is computing the entropy of the set of n-grams of a given length. One can vary the length being considered, say from one to six, and thus get estimates for the entropy of all of the unigrams, bigrams, and so forth, all the way up to hexagrams. In order to compare across symbol systems with different numbers of symbols, Rao uses the log to the base of the number of symbols in each system ($log_{|V|}$ in the equation). Thus for DNA the log is taken to base four. The maximum entropy thus normalized is then just N, the length of the n-grams being considered, for any system.

Figure 5 shows the block entropy estimates from Rao 2010 for a variety of linguistic systems, some nonlinguistic systems, the Indus corpus, and the maximum and minimum entropy systems from Rao et al. 2009a.[6] For this work, Rao used an estimation method proposed in Nemenman et al. 2001, which comes with an associated MATLAB package. In this case, therefore, it is possible to replicate the method used by Rao and produce results that are directly comparable with his results in Fig. 5.

These are presented in Figure 6. As with the conditional entropy, and unlike Rao's clear-looking results, the systems are quite mixed up, with linguistic and nonlinguistic systems interspersed. The Indus system is right in the middle of the range. Note that the curve for our Indus corpus is rather close to the curve for the Mahadevan corpus seen in Fig. 5, suggesting that even though the bar seals constitute a smaller subset of the whole

[6] Here, Rao is wrong: a rigidly ordered system would NOT look like the minimum curve in this plot. For the unigrams, the fact that the next symbol is predetermined from the current symbol is irrelevant to the computation of the block entropy: we are simply considering the probabilities of n-grams of length 1. If the symbols are equally probable, then the value of the block entropy will be 1. In fact, what one would see for a minimum entropy system is a curve that starts at 1, and decreases in value as N grows larger.

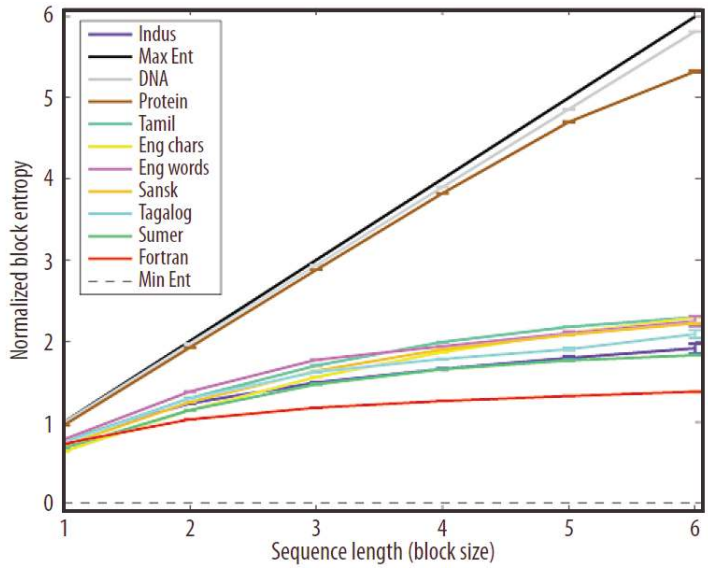

FIGURE 5. Block entropy estimates from Rao 2010 for a variety of linguistic systems, some nonlinguistic systems, and the Indus corpus. Used with permission of the IEEE.

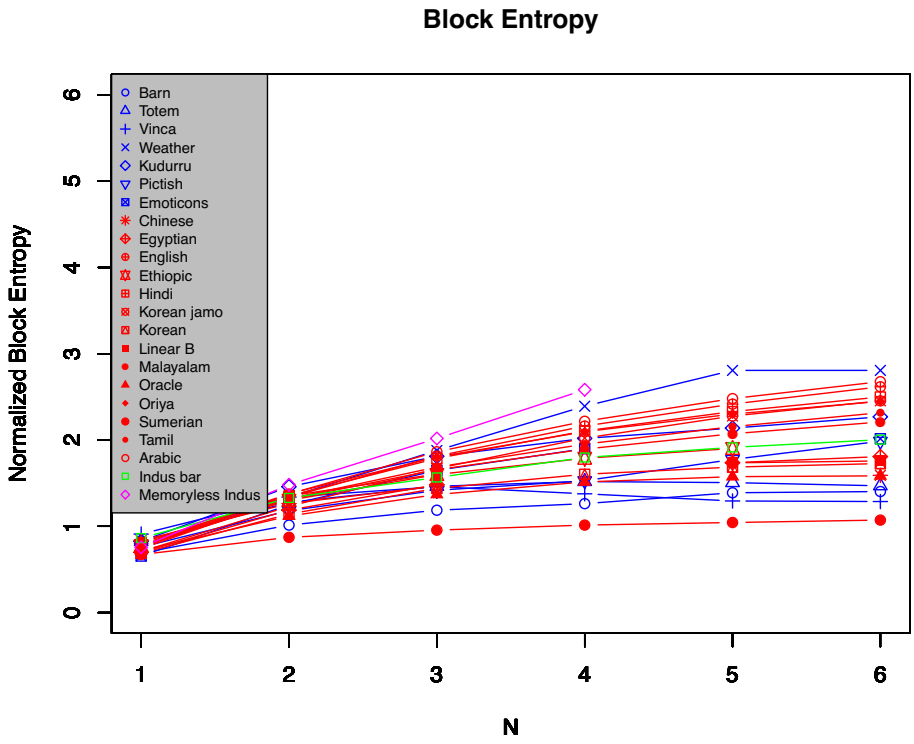

FIGURE 6. Block entropy values for our linguistic and nonlinguistic corpora, the Indus bar seal corpus, and our memoryless Indus corpus. As in Rao's case, we used the method reported in Nemenman et al. 2001. Note that the method became unstable for the two highest n-gram values for the memoryless Indus corpus.

corpus with a rather different average text length, there is some similarity in the distributions of n-grams. By contrast, note that our sample of Sumerian is radically different from Rao's in its behavior, and in fact looks more like Fortran. The Ancient Chinese oracle bone texts are also similar to Fortran, which is perhaps not surprising given that one finds quite a few repeating formulae in this corpus. The memoryless Indus system again shows the highest entropy, especially for the higher n-grams, but nonetheless hugs the curve not far above some of the genuine linguistic systems.[7]

2.3. MEASURES DERIVED IN PART FROM ENTROPY. As noted above, Rao and his colleagues interpret their results 'inductively', though they stated this explicitly for the first time only in Rao et al. 2010: the statistical behavior of the Indus corpus 'increases the likelihood' that the Indus symbols represented language, but does not in and of itself show that it was language.

Unlike Rao, Lee and colleagues (2010a) are forthright in their claim that they have developed a DISCRIMINATIVE METHOD. Their article attempts to use a couple of measures, one of which is derived from bigram conditional entropy, to argue that Pictish symbols were a writing system. As with Rao and colleagues' work, they compare the symbols to a variety of known writing systems, as well as symbol systems like Morse code, European heraldry, and randomly generated texts—by which is meant RANDOM AND EQUIPROBABLE.

Lee and colleagues develop two measures, U_r and C_r, defined as follows. First, U_r is defined as in 3, where F_2 is the bigram entropy, N_d is the number of bigram types, and N_u is the number of unigram types.

$$(3)\quad U_r = \frac{F_2}{\log_2(N_d/N_u)}$$

C_r is defined as in 4, where N_d and N_u are as above, a is a constant (for which, in their experiments, they derive a value of seven, using cross-validation), S_d is the number of bigrams that occur once (HAPAX LEGOMENA), and T_d is the total number of bigram tokens; this latter measure will be familiar as $\frac{n_1}{N}$, the Good-Turing estimate of the probability mass for unseen events (Good 1953, Baayen 2001).

$$(4)\quad C_r = \frac{N_d}{N_u} + a\,\frac{S_d}{T_d}$$

Lee and colleagues use C_r and U_r to train a decision tree to classify symbol systems. If $C_r \geq 4.89$, the system is linguistic. Subsequent refinements use values of U_r to classify the system as segmental ($U_r < 1.09$), syllabic ($U_r < 1.37$), or else logographic. Their decision tree is shown in Figure 7. Note that decision trees, which are also used below in §2.6, are a family of statistical models widely used for classification (for prediction of categorical data) or regression (for continuous-valued data).

What happens if Lee and colleagues' tree is applied 'out of the box' to our data? Table 1 shows the results.[8] Not surprisingly, the Pictish symbols are classified as (logographic) writing, consistent with Lee and colleagues' results. But, with the exception of

[7] For the five- and six-grams, the estimation method was numerically unstable for the memoryless system, returning a NaN result: such instability occurs at certain points for other corpora, including Chinese, Egyptian, Asian emoticons, Amharic, and Linear B, so it probably has no particular significance.

[8] Note that like Lee and colleagues (2010a), in performing these computations, I pad the texts with start and end symbols.

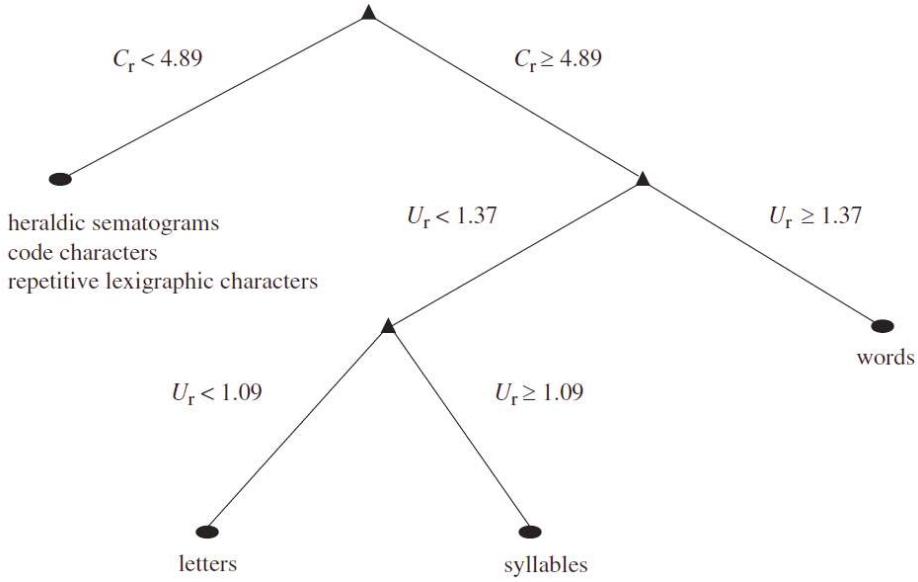

FIGURE 7. Reproduction of figure 6 from Lee et al. 2010a:9.

Vinča symbols, which are appropriately classified as nonlinguistic, the remainder of the known nonlinguistic systems in our set are misclassified as some form of writing.[9]

CORPUS	CLASSIFICATION
Asian emoticons	linguistic: letters
Barn stars	linguistic: letters
Mesopotamian deity symbols	linguistic: syllables
Pictish symbols	linguistic: words
Totem poles	linguistic: words
Vinča symbols	nonlinguistic
Weather icon sequences	linguistic: letters

TABLE 1. The results of applying the Lee et al. 2010a decision tree (Fig. 7) to our data.

2.4. MODELS OF ASSOCIATION. So far we have looked at measures proposed in the literature that, one way or another, involve entropy. But there are other possible measures above and beyond what previous authors have proposed that could be considered. Another property of language is that some symbols tend to be strongly associated with each other in the sense that they occur together more often than one would expect by chance. There is by now a large literature on association measures, but all of the approaches compare how often two symbols (e.g. words) occur together, compared with some measure of their expected cooccurrence. One popular measure is Shannon's (pointwise) mutual information (Shannon 1948, Church & Gale 1991) for two terms t_i and t_j.

$$(5)\quad \mathrm{PMI}(t_i,t_j) = \log \frac{p(t_i,t_j)}{p(t_i)p(t_j)}$$

[9] All of our linguistic corpora are also correctly classified as linguistic, though not always as the right type of linguistic system.

This is known to produce reasonable results for word-association problems, though there are also problems with sensitivity to sample size, as pointed out by Manning and Schütze (1999:182).

In this study, pointwise mutual information between adjacent symbols is used to estimate how strongly associated the most frequent symbols are with each other. For words, at least, it is often the case that the most frequent words—function words such as articles or prepositions—are NOT strongly associated with one another: consider that *the the* and *the is* are not very likely sequences in English.

We therefore look at the mean association of the n most frequent symbols, normalized by the number of association measures computed.

$$(6)\quad \frac{\sum_{j=0}^{n}\sum_{i=0}^{n}\mathrm{PMI}(t_i,t_j)}{n^2}$$

We let n range from 10, 20, … up to the k such that the first k symbols account for 25% of the corpus. As with the entropy computations, we estimated probabilities of bigrams and unigrams using OpenGrm. As with previous measures, there is no clear separation of linguistic and nonlinguistic systems. Both kinds of systems have symbols that are more or less strongly associated with each other. (See Figure S6 in the online supplement for details.)

2.5. MODELS OF REPETITION. To the extent that they involve the statistical properties of the relations between symbols, the measures discussed so far are all local, involving adjacent pairs of symbols. But despite the fact that local entropic models have a distinguished history in information-theoretic analysis of language dating back to Shannon 1948, language also has many nonlocal properties.

One is that symbols tend to repeat in texts. Suppose you had a corpus of Boy Scout merit badge sashes and you considered each badge to be a symbol. The corpus would have many language-like properties. Since some merit badges are awarded far more than others, one would find that the individual symbols follow a roughly Zipfian power-law distribution (Zipf 1949) (the distribution can be seen in Figure S7). As we saw earlier, memoryless systems that have a Zipf-like distribution for symbol types will exhibit 'language-like' entropic behavior, so we expect that a corpus of merit badges would also exhibit that kind of behavior and therefore 'look like' language.

Yet such analyses would fail to capture what is the most salient feature of merit badge 'texts', namely that a merit badge is never earned twice, and therefore no symbol repeats. In this feature, more than any other, a corpus of merit badge 'texts' would differ from linguistic texts.

Symbols in writing systems, whether they represent segments, syllables, morphemes, or other linguistic information, repeat for the simple reason that the linguistic items that they represent tend to repeat. It is hard to utter a sentence in any natural language without at least some segments repeating. Obviously, as the units represented by the symbols in the writing system become larger, the probability of repetition decreases, but we would still expect some repetition.

We therefore argued in Farmer et al. 2004 that it was noteworthy that the Indus inscriptions showed a remarkably low rate of repetition, and that even in 'long' texts one typically finds no symbol repetitions. Indeed, the longest Indus inscription on a single surface, consisting of seventeen glyphs, shows not a single repetition of a symbol. However, it is not just the presence or absence of repetition in a corpus, but also how the repetition manifests itself, that is important. In Farmer et al. 2004 we argued not only that the Indus inscriptions show rather low repetition rates but also that when one does

find repetition of individual symbols within an inscription, it is also often of the form found in Figure 8, with symmetric and other patterns.

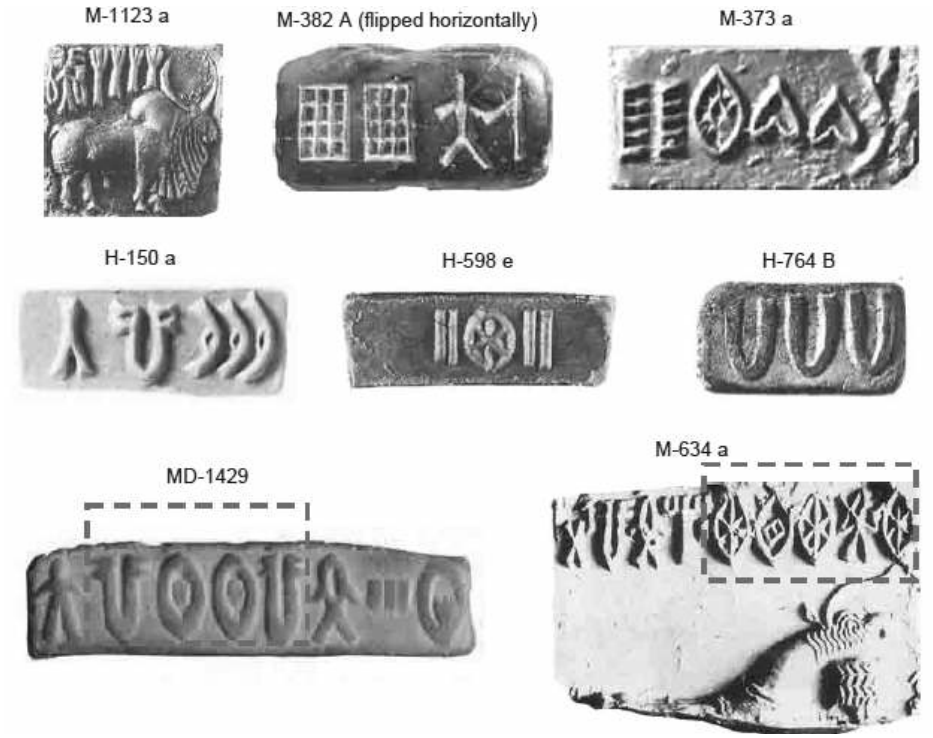

FIGURE 8. Indus repetitions, from Farmer et al. 2004, figure 6. As described there, these are: 'Examples of the most common types of Indus sign repetition. … The most frequent repeating Indus symbol is the doubled sign illustrated in M-382 A, which is sometimes claimed to represent a field or building, based on Near Eastern parallels. The sign is often juxtaposed (as here) with a human or divine figure carrying what appears to be one (or in several other cases) four sticks. M-634 a illustrates a rare type of sign repetition that involves three duplications of the so-called wheel symbol, which other evidence suggests in some cases served as a sun/power symbol; the sign shows up no less than four times on the badly deteriorated Dholavira signboard (not shown), which apparently once hung over (or guarded?) the main gate to the city's inner citadel. The color photo of MD-1429 is reproduced from M. Kenoyer, *Ancient Cities of the Indus Valley Civilization*, Oxford University Press, Oxford 1998, p. 85, exhibition catalog number MD 602. The sign on either side of the oval symbols in the inscription is the most common symbol in the Indus corpus, making up approximately 10% of all symbol cases; despite its high general frequency, repetitions of the symbol in single inscriptions, of the kind seen here, are relatively rare.'

In this study I use a measure of repetition that seeks to quantify the rate of iterated repeats of the kind seen in some cases in Fig. 8, relative to the total number of repetitions found. Specifically, the number of tokens in each text that are repeats of a token that has already occurred in that text are counted, and that number is summed over the entire corpus. Call this number *R*. The number of tokens that are repeats in each text, and that are furthermore adjacent to the token they repeat, are then counted. Sum that over the entire corpus, and call this number *r*. So, for example, for the single text in 7, *R* would be 3 (two As are repeats, and one B), and *r* would be 1 (since only one of the repeated As is adjacent to a previous A).

(7) A A B A C D B

Thus the repetition rate $\frac{r}{R}$ would be 0.33.

Table 2 shows our corpora ranked by $\frac{r}{R}$. This measure is by far the cleanest separator of our data into linguistic versus nonlinguistic. If we set a value of 0.10 as the boundary, only Sumerian and kudurrus are clearly misclassified (while Asian emoticons and Egyptian are ambiguously on the border). Not surprisingly, this measure is the feature most often picked by the decision tree classifier discussed in the next section.

CORPUS	$\frac{r}{R}$
Barn stars	0.86
Weather icons	0.79
Sumerian	0.67
Totem poles	0.63
Vinča	0.59
Indus bar seals	0.58
Pictish	0.26
Asian emoticons	0.10
Egyptian	0.10
Mesopotamian deity symbols	0.099
Linear B	0.055
Oracle bones	0.048
Chinese	0.048
English	0.035
Arabic	0.032
Korean jamo	0.022
Malayalam	0.022
Korean	0.020
Amharic	0.018
Oriya	0.0075
Tamil	0.0060
Hindi	0.0017

TABLE 2. Repetition rate $\frac{r}{R}$ for the various corpora.

But one important caveat is that the repetition measure is also weakly negatively correlated with mean text length: the Pearson's correlation for mean length and $\frac{r}{R}$ is -0.49. This makes sense given that the shorter the text, the less chance there is for repetition, while at the same time, the more chance that if there is repetition, the repetition will involve adjacent symbols. Of course the correlation is not perfect, meaning that $\frac{r}{R}$ probably also reflects intrinsic properties of the symbol systems. How much is length a factor?

One-way analyses of variance with class (linguistic/nonlinguistic) as the independent variable and mean text length or $\frac{r}{R}$ as dependent variables yielded the results in 8.

(8) Mean text length: $F = 5.75$ $p = 0.027$
$\frac{r}{R}$: $F = 15.83$ $p = 0.0008$

Mean text length is thus well predicted by class, with nonlinguistic systems having shorter mean lengths. But the repetition rate is even better predicted, suggesting that our results above cannot be simply reduced to length differences.

To see this in another way, consider the results of computing the same repetition measures over our corpora where the texts have been artificially limited to a length of no more than six (by simply trimming each text to the first six symbols). In order to make the corpora more comparable, the shortened corpus was also limited to the first 500 'texts' from the original corpus. In this analysis, the separation was not as clean as in the case of Table 2. The five corpora with the highest $\frac{r}{R}$ are, nevertheless, nonlinguistic (if one counts the Indus system as nonlinguistic) and the nine corpora with the lowest values are all linguistic. (See Table S5 for the full results.)

2.6. TWO-WAY CLASSIFICATION. We used the above features and the CLASSIFICATION AND REGRESSION TREE (CART) algorithm (Breiman et al. 1984) to train (binary branching) classification trees for a binary classifier. The features used were as follows.

- The PMI association measure over the symbol set comprising 25% of the corpus
- Block entropy calculations for $N = 1$ to $N = 6$ (block 1 … block 6)
- Maximum conditional entropy calculated for Fig. 3
- F_2, the $\log_2$ bigram conditional entropy for the whole corpus, from Lee et al. 2010a
- U_r
- C_r
- Repetition measure $\frac{r}{R}$

Since the set of corpora is small—only twenty-one in total, not counting the Indus bar seals—the system was tested by randomly dividing the corpora into sixteen training and five test corpora, building, pruning, and testing the generated trees, and then iterating this process 100 times. The mean accuracy of the trees on the held-out data sets was 0.8, which is above the baseline accuracy (0.66) of always predicting a system to be linguistic (i.e. picking the most frequent choice).

It is interesting to see how these 100 trees classify the Indus bar seals, which were held out from the training sets. Ninety-eight of the trees classified it as nonlinguistic, and only two of the trees classified it as linguistic. Furthermore, the ninety-eight that classified it as nonlinguistic had a higher mean accuracy—0.81—than the two that classified it as nonlinguistic (0.3). Thus, more and better classifiers tend to classify the Indus symbols as nonlinguistic than those that classify it as linguistic.[10]

Since the various runs involve different divisions into training and test corpora, an obvious question is whether particular corpora are associated with better performance. If a corpus is in the training and thus not in the test set, does this result in better performance, either because its features are more informative for a classifier and thus helpful for training, or because it is easier to classify and thus helpful for testing? Our results show that Asian emoticons and Sumerian result in higher classification rates when they occur in the training and not in the testing, perhaps because both are hard to classify in testing: Sumerian is misclassified by the repetition measure, as we saw above, and emoticons on that measure are close to the border with linguistic systems, as is Egyptian, and its removal from the test data also results in better performance. The next two on the list, weather icons and Pictish symbols, have a repetition value that places them well outside the range of the linguistic corpora, so in these cases it may be that they are useful as part of training to provide better classifiers. Kudurrus are also misclassified by the repetition measure, and thus their removal from the test data could lead to better performance. (See Tables S6 and S7 for fuller results.)

In what was just described, Pictish was included among the nonlinguistic systems. Of course, with the publication of Lee et al. 2010a, this classification has become somewhat controversial. What if the Pictish data was not included? In this case training sets of fifteen corpora were used, and again five corpora were held out for testing. Here the overall mean accuracy is 0.84. What do these classifiers make of Pictish? The results are strongly in favor of the nonlinguistic hypothesis: ninety-seven classified Pictish as nonlinguistic and had a mean accuracy of 0.81; three classified it as linguistic and had a lower mean accuracy of 0.4.

[10] To show that this result is no artifact of the particular setup being used, we ran the same experiment, this time dropping the Indus bar seals entirely, and holding out Oriya from the training. 100% of the trees classified Oriya as linguistic.

It is also interesting to consider the features that are used by the trees. These are presented in Table 3 for the two experimental conditions (with versus without Pictish). In both cases, the most prominent feature is repetition, and the second most common is C_r. The latter is interesting since, if one compares the Lee et al. tree in Fig. 7, it is C_r that was involved in the top-level linguistic/nonlinguistic decision. Anticipating what is argued below, this replicates Lee and colleagues' prior work in that C_r is a good discriminator for this task, but when trained on a more exhaustive set of linguistic and nonlinguistic systems, it achieves results that are the opposite of what they report.

WITH PICTISH		WITHOUT PICTISH	
# TREES	FEATURES	# TREES	FEATURES
84	repetition	71	repetition
13	C_r	26	C_r
1	block 5	1	block 5
1	block 3	1	association
1	block 1	1	

TABLE 3. Features used by trees under the two training conditions. The left-hand column in each case is the number of trees using the particular combination of features in the right-hand column. The final line for the trees trained without Pictish data is for one tree with a single node that predicts 'linguistic' in all cases.

Table 4 shows the results of the Wilcoxon signed rank test for each feature comparing for the two populations of linguistic and nonlinguistic corpora. Only for our repetition measure and C_r are the population means different according to this test (and after a Bonferroni correction, only repetition), suggesting that the other features are largely useless for determining the linguistic status of a symbol system.

MEASURE	W	p
association	29	0.15
block 1	42	0.64
block 2	49	1
block 3	56	0.64
block 4	49	1
block 5	63.5	0.30
block 6	56	0.64
maximum conditional entropy	58	0.54
F_2	63	0.32
U_r	40	0.54
C_r	84	0.0074**
repetition	6	0.00052**

TABLE 4. Wilcoxon signed rank test for each feature with the two populations being linguistic and nonlinguistic corpora. Assuming an alpha level of 0.01, both C_r and repetition are highly significant. Using Bonferroni correction with twelve as the number of tests performed, the adjusted p value is 0.089 for C_r, and 0.0062 for repetition. Thus even with fairly conservative correction, repetition remains significant at the 0.01 level, though C_r is not.

Since our repetition measure is well correlated with mean length of texts in the corpus, and the nonlinguistic corpora in general have shorter mean lengths than the linguistic corpora, it is instructive to consider what happens when we remove that feature and train models as before. As we can see in Table 5, a wider variety of trees is produced, and Lee and colleagues' C_r is the favored feature, being used in eighty-two of them. The results for the held-out Indus bar seal corpus are not as dramatic as when the repetition feature was included, but they still highly favor the nonlinguistic analysis. Eighty-eight of the trees classified the system as nonlinguistic (mean accuracy 0.75), and twelve as linguistic (mean accuracy 0.37).

# TREES	FEATURES
57	C_r
25	C_r, association
6	association
4	maximum conditional entropy
2	block 5
2	
1	U_r
1	block 6
1	association, block 5
1	association, block 4

TABLE 5. Features used by trees when repetition is removed. The sixth row consists of two trees where there is a single leaf node with the decision 'linguistic'.

Similarly lower results were obtained when the repetition rate $\frac{r}{R}$ from Table 2 was replaced with that computed over the artificially shortened corpora discussed above (Table S5). Eighty-five of the trees classified the Indus symbols as nonlinguistic (mean accuracy 0.63), and fifteen classified them as linguistic (mean accuracy 0.36). The features used by these trees are shown in Table 6. Repetition is far less dominant than it is in the original tree set shown in Table 3, but it is still the second most used feature, after C_r, occurring in thirty out of 100 trees. The most useful features thus seem to be our measure of repetition and C_r, and this is further confirmed by the Wilcoxon signed rank test reported in Table 4 above, for which these were the only two features that showed any correlation with corpus type.

# TREES	FEATURES
36	C_r
30	repetition
11	C_r, association
8	
4	maximum conditional entropy
3	association
2	U_r
2	maximum conditional entropy, repetition
2	block 4, repetition
1	block 5

TABLE 6. Features used by trees when repetition is replaced by the repetition rate computed over artificially shortened corpora (see Table S5). The fourth row consists of twelve trees where there is a single leaf node with the decision 'linguistic'.

What do these two features have in common? We know that $\frac{r}{R}$ is correlated with text length: could it be that C_r is also correlated? As Figure 9 shows, this is indeed the case. Pearson's r for C_r and text length is 0.39, and this increases dramatically to 0.71 when the one obvious outlier—Amharic—is not considered.

As argued above, there is a plausible story for why $\frac{r}{R}$ should correlate with text length, but why should C_r correlate? Recall the formula for C_r, repeated in 9, where, again, N_d is the number of bigram types, N_u is the number of unigram types, S_d is the number of bigram hapax legomena, and T_d is the total number of bigram tokens.

$$(9)\quad C_r = \frac{N_d}{N_u} + a\frac{S_d}{T_d}$$

Now recall (n. 8) that Lee and colleagues pad the texts with beginning- and end-of-text delimiters. For corpora consisting of shorter texts, this means that bigrams that include

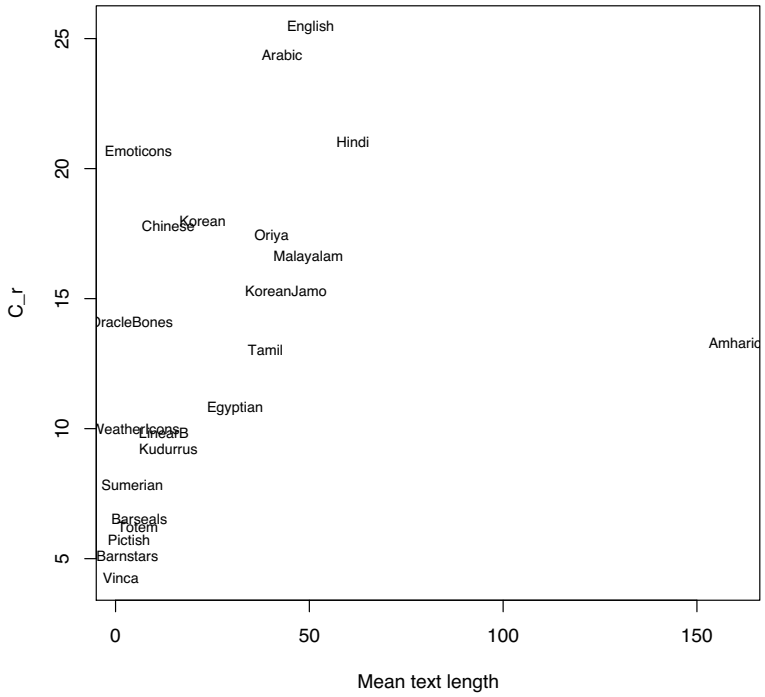

FIGURE 9. Correlation between C_r and text length. Pearson's r is 0.39, but when the obvious outlier Amharic is removed, the correlation jumps to 0.71.

either the beginning or the ending tag will make up a larger portion of the types, and that the type richness will be reduced. The unigrams, by contrast, will be unaffected by this, though they will of course be affected by the overall corpus size. This predicts that the term $\frac{N_d}{N_u}$ alone should correlate well with mean text length, and indeed it does, with $r = 0.4$ ($r = 0.72$ excluding Amharic).

U_r, which is derived from $\frac{N_d}{N_u}$, is also correlated, but negatively: $r = -0.29$. Thus a higher mean text length corresponds to a lower value for U_r. Recall again that it is this feature that is used in the Lee et al. 2010a decision tree (Fig. 7) for classifying the type of linguistic system: the lowest values of U_r are segmental systems, next syllabaries, and next 'logographic' systems. This makes perfect sense in light of the correlation with text length: ceteris paribus, it takes more symbols to convey the same message in a segmental system than in a syllabary, and more symbols in a syllabary than in a logographic system. Thus segmental systems should show a lower U_r, syllabic systems a higher U_r, and logographic systems the highest U_r values.

Returning to C_r, nonlinguistic systems do tend to have shorter text lengths than linguistic systems, so one reason why C_r seems to be such a good measure for distinguishing the two types, apparently, is because it correlates with text length. Of course, where one draws the boundary between linguistic and nonlinguistic on this scale will determine which systems get classified in which ways. For Lee and colleagues, Pictish comes out as linguistic only because they set the value of C_r relatively low. With more data, we were able to determine a more appropriate cut-off point, one that, unfortunately for Lee and colleagues, places Pictish on the other side of the boundary from what they argue.

While there are obviously other factors at play—for one thing, something must explain why Amharic is such an outlier—it does seem that the key insight at work here comes down to a rather simple measure: how long on average are the extant texts in the symbol system? If they are short, then they are probably nonlinguistic; if they are long, they are probably linguistic. We return to the implications of this result in §3.

We have already seen that repetition on its own misclassifies Sumerian as nonlinguistic, due to the short 'texts'—an artifact of the way the corpus was divided. Given what was just discussed, we would therefore expect the decision trees to also misclassify it. Indeed they do: using all features, ninety-two trees classified it as nonlinguistic (accuracy 0.72) and eight trees as linguistic (0.23).

2.7. Two-way classification with additional corpora. Our corpora of mathematical formulae and heraldry are still under development. However, it is of interest to see how the statistical methods hold up when samples of those corpora are added. To that end, 1,000 'texts' from each of these corpora were added. See the online supplementary materials for further discussion. Block entropy curves (as in Fig. 4) with these two new corpora added show that both mathematical formulae and heraldry fall somewhere in the middle of the distribution (see Figure S8).

When the classification experiments reported in §2.6 were replicated with these additional corpora, and using all features, the results were much as before: of the 100 trees trained, ninety-four classified the Indus bar seals as nonlinguistic (mean accuracy 0.73) and six classified it as linguistic (mean accuracy 0.33).

3. Discussion and conclusions. What are we to make of the results reported above? Readers familiar with the Rao et al. and Lee et al. work, and who have more or less accepted their conclusions, must also believe that statistical evidence is relevant to determining the status of an unknown symbol system. If they are fair minded, then they must also accept that a more extensive statistical analysis, of the kind presented here, could result in the opposite conclusions from those drawn in previous work. From that point of view, the above results would seem to lead to the conclusion that the Indus and Pictish symbols were probably not writing, or at least that the results are consistent with a hypothesis that has them as some sort of nonlinguistic symbol system.

To be sure, there are caveats: the two most useful measures were shown to correlate with text length, and texts from our nonlinguistic corpora are on average shorter than our linguistic corpora: more on this point later. But in any case there is clear evidence that the reported methods of Lee and colleagues and Rao and colleagues do not work.

Even this statement requires some qualification. As discriminative methods, the previously reported measures clearly fail. As we saw, Lee and colleagues' method misclassifies all of our nonlinguistic systems but one as some form of writing. And the various entropic measures reported by Rao and colleagues fail equally badly. None of the Rao et al. measures were selected in the CART tree experiments reported in §2.6, meaning that they have no discriminative power, and in any case a simple eyeballing of the plots in Figs. 3 and 6 should be enough to satisfy the reader that these measures are unlikely to be of much use as discriminative measures.

One response to this result would be to broaden one's definition of 'writing' to include other kinds of meaning-bearing systems that do not directly encode language. This was exactly Lee and colleagues' response in Lee et al. 2010b, in reply to the preliminary results for kudurrus presented in Sproat 2010a, where I reported that the Lee et al. system classified it as logographic writing. In their reply, Lee and colleagues noted that they had replicated my result and were comfortable with the classification of the

deity symbols as a form of writing, citing Powell's (2009:13) broader definition of 'writing [as] a system of markings with a conventional reference that communicates information'. There are at least three problems with that response, however.

The first is that if one really wants to adopt such a broad definition, then there is nothing further to discuss: ANY conventional meaning-bearing symbol system is writing. The only distinction that one might consider using statistical methods for is distinguishing meaning-bearing symbols from meaningless decorations, and the Lee et al. 2010a study would thus be pointless: note that nobody ever disputed that the Pictish symbols must have meant something.

The second problem is that most linguists do not accept such a broad and freewheeling definition of writing. As I noted in Sproat 2010b, it seems doubtful that Powell himself really wants such a broad definition, and in any case it is clear that such a broad and almost vacuous definition is clearly not what readers of Lee et al. 2010a would have had in mind when they read the article and learned of its surprising discovery.

The final problem is that not only does the Lee at al. method misclassify most of our nonlinguistic systems as linguistic, but it also does so in ways that cannot be reconciled with even the liberal definition of writing that Powell proposes. Thus barn stars, kudurrus, and weather icons are classified as some sort of phonographic writing system.[11] None of these makes much sense in light of Powell's broader definition of writing, since while one might countenance classifying these systems as logographic under that broader definition, it hardly makes sense to classify them as rather particular forms of phonographic writing. The coup de grace here is the classification of barn stars as 'letters' (segmental writing) since that means the method fails to distinguish even meaningless decorative systems (which barn stars surely are) from real writing.

So much for discriminative interpretations of prior results; but what about Rao and colleagues' INDUCTIVE interpretation of their results—an interpretation that was explicitly stated for the first time only in Rao et al. 2010? As I noted in Sproat 2010b, the 'inductive' approach is familiar to computational linguists as a form of 'generative' modeling, wherein one assumes a hidden set of models and tries to see which of those models is more consistent with the observation. In the case of Rao and colleagues' problem, the two models constitute two hypotheses about the Indus Valley symbols, namely that they were linguistic (call this H_L) or nonlinguistic (H_{NL}). Which hypothesis is more consistent with the facts? Rao and colleagues discuss a range of evidence that they argue is more consistent with H_L. Included in this evidence are the following.

- The linear arrangement of the Indus symbols in texts
- The (apparent) presence of diacritic modifications (ligatures) of symbols (possibly) similar to the kinds of diacritics found in many writing systems
- Evidence for the directionality of the 'writing'
- Apparent evidence, presented in Rao et al. 2009b, for different uses of the symbols in seals unearthed in Mesopotamia, suggesting a difference in the underlying language

If the Indus symbol system was linguistic, the argument goes, then all of the above features are certainly consistent with the hypothesis. And of course one would expect that

[11] This classification of kudurrus differs from what I reported in Sproat 2010a, since the present work uses a larger corpus: 939 tokens, versus 545 tokens in the earlier work. The estimation techniques are also different, meaning that it is not unexpected that one would get different final values that would push results over the boundary.

the entropy of the system would look like that of a language, which is what they demonstrated.

But it is all too easy to be fooled by that line of argumentation. Suppose I had a hypothesis that, counterfactually, the Mesopotamian deity symbols were a form of writing. I could adduce a number of pieces of evidence in support of that idea. For example:

- Deity symbols are frequently written linearly, and in the cases where linearity is less clear, the symbols are written around the top of a stone in an apparent concentric circle pattern; see examples in Seidl 1989. One sees such nonlinear arrangements with scripts too: the Etruscan Magliano disk (Jannot 2005:36–37) and many rune stones have text wrapped around the border of the stone—for example, the Lingsberg Runestones described in Fuglesang 1998.
- There is clear evidence for the directionality of deity symbols: to the extent that 'more important' gods were depicted first, these occur at the left/top of the text.
- Deity symbols are often ligatured together: one symbol may be joined with another.[12]
- The deity symbols obey a power-law distribution.
- Deity symbols are pictographic, like many real scripts—Egyptian, Luwian, Mayan, Hittite hieroglyphs.
- Deity symbols were used largely on standing stones, but were also used in other contexts (see e.g. the 'necklace' depicted in Figure S9), suggesting that there were a variety of media in which one could create 'texts'.

And, following Rao and colleagues' reasoning, we would expect the entropic behavior to fall in the range observed for real languages, which has been demonstrated to be the case. But of course we know that the deity symbols were NOT writing.

But it has also been shown that a memoryless system can mimic language-like entropic behavior, provided the distribution of symbols is nonuniform. So we could turn the above argument on its head and start with the hypothesis that the Indus Valley symbols were some sort of nonlinguistic symbol system that, like many such systems, had a power-law distribution, and even some (nonlinguistic) structure. There are a number of features of the Indus system that support such a hypothesis, as discussed in Farmer et al. 2004. Among these are the following.

- All extant texts are very short, with no evidence of that changing over a 700-year period, as would have been expected if the system were some form of proto-writing evolving into true writing.
- The system was used in a culture where there are no archaeological markers of manuscript production (such as pens, styluses, ink pots, etc.).
- There is no evidence for the development, even over a 700-year period, of the kind of cursive forms that one expects in a true writing system.

Certainly these conclusions have been challenged, for example by Vidale (2007). But assuming one finds these points convincing, it in any case follows that such a system, with a nonuniform distribution of symbols, would be expected to behave entropically like language. Again, this has been shown to be true not only of other genuine nonlin-

[12] We note in passing that it is actually unclear for the Indus symbols that there are diacritics or ligatures of symbols. A given symbol may LOOK as if it is a modified version of another, or ligatured, but unless we know what the symbols mean it is hard to be sure. In the case of the Mesopotamian deity symbols, we actually know what the symbols denoted.

guistic systems, but also even of artificial memoryless systems. In other words, the entropic measures are useless for distinguishing between H_L and H_{NL}.

There is one final point to be made. Rao and colleagues made much of the 'type 1' (maximum entropy) and 'type 2' (minimum entropy) systems, though the form of the argument changed subtly. In the prepublication version of their 2009 article, they presented them as 'representative samples' of nonlinguistic systems, whereas by the time the final archival version of the article was published, these became 'controls', characterized as 'necessary in any scientific investigation, to delineate the limits of what is possible'.[13] The question then is: what in fact IS possible? Rao and colleagues have not found any genuine examples of type 2, though Rao did continue to insist in the face of obvious counterevidence (Rao et al. 2010) that Mesopotamian deity symbols should behave like type 2 systems. For type 1 systems, the set of real systems that approximates type 1 behavior has grown in various work (Rao et al. 2009a, Rao 2010), so those might appear to be genuine. But are they? Consider for example DNA. There is no reason to doubt Rao's results that if one, for example, computes the block entropy over a corpus of DNA sequences at the base level, one will find that the system approaches maximum entropy. But one must recall that maximum entropy means that all possibilities are equiprobable, and that the system is STRUCTURELESS. A moment's thought will reveal that this conclusion, when it comes to DNA, is patently absurd: obviously DNA is highly structured. The problem, as noted in n. 4, is clearly that Rao is sampling at the wrong level: the base pairs are the 'bits' of DNA, but the actual information is carried in sequences of base pairs that are thousands of base pairs long. In addition, while there is much debate on this point, it is at least a viable hypothesis that some portion of DNA is noncoding (Lander et al. 2001, The ENCODE Project Consortium 2012), similar perhaps to nonsensical writing inserted into a corpus of otherwise intelligible prose. Were Rao to sample at a more reasonable level, possibly taking into account (and eliminating) noncoding regions, the entropic behavior would surely be different. What has not been demonstrated is that a real meaning-bearing symbol system, sampled at the right level, has anything like type 1 behavior. Certainly all of the nonlinguistic systems we have examined here fall far away from both entropic extremes.

Thus, while the 'inductive' interpretation that Rao and colleagues propose is more subtle than Lee and colleagues' discriminative interpretation of their results, nonetheless it does not hold up under scrutiny. And thus we can return to our statement above: there is clear evidence that the reported methods of Lee and colleagues and Rao and colleagues DO NOT WORK.

However, as was noted above, if one accepts that statistical argumentation might shed light on the status of an ancient unknown symbol system, then the statistical evidence presented here supports the nonlinguistic hypothesis for both the Indus Valley symbols and the Pictish symbols. But, as we also saw, both of our most discriminative measures, $\frac{r}{R}$ and C_r, also correlate with text length, and so text length would appear also to be an underlying factor. Indeed, nonlinguistic corpora do tend to have shorter texts. 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 nonlinguistic systems.

Text length as a relevant measure for distinguishing systems would seem to be rather trivial, but at its core it seems like common sense, and there is a clear analogy in child

[13] http://homes.cs.washington.edu/~rao/IndusResponse.html

language development. The most basic measure of language development is MEAN LENGTH OF UTTERANCE (MLU), measured in words or in morphemes. There are of course other measures, but often even much more sophisticated measures correlate very strongly with MLU: for example, the INDEX OF PRODUCTIVE SYNTAX (IPSyn) measure (Scarborough 1990), a complex set of syntactic and morphological features used in assessing language development for English-speaking children, has an extremely high correlation with MLU in Scarborough's original data. That MLU should be so basic seems sensible: if an individual only utters four-word sentences, we would not generally think of that individual as having mastered his or her native language. In a similar vein, a symbol system that only allows its users to write short cryptic messages should make one suspicious about its status as a true linguistic script, especially if that situation appeared not to change over seven centuries.

One of the oft-noted problems for the thesis that the Indus symbols were a true writing system is the fact that all extant texts are very short (Parpola 1994, Farmer et al. 2004): the longest text on a single surface is seventeen glyphs long, which is quite a bit shorter than our longest kudurru text (thirty-nine glyphs, and indeed nineteen out of our sixty-nine kudurru texts are longer than the longest Indus text). As has been argued elsewhere (Farmer et al. 2004), the so-called 'lost manuscript' hypothesis, which 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 (e.g. 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 symbols must be taken as a mere act of faith, in the absence of any substantive evidence for it.

And so we end up with some statistical measures that are consistent with the hypothesis that Indus Valley and Pictish symbols were not writing. Both of these measures are correlated with text length, but text length itself seems, on reflection, to be a relevant measure for this purpose. The evidence presented must of course be considered in light of archaeological and cultural evidence concerning the two systems. In the case of Pictish, one relevant data point is that the Picts already had a written language (though scantily attested) based on Ogham (Rhys 1892). In the case of the Indus Valley, various pieces of evidence discussed above, and argued more extensively in Farmer et al. 2004, suggest that the civilization was not literate.

The status of any ancient symbol system as a writing system must be supported by good empirical evidence. As argued in Farmer et al. 2004 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

BAAYEN, HARALD. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

BARBEAU, MARIUS. 1950. *Totem poles*. 2 vols. (Anthropology series 30, National Museum of Canada bulletin 119.) Ottawa: National Museum of Canada.

BLACK, JEREMY; ANTHONY GREEN; and TESSA RICKARDS. 1992. *Gods, demons and symbols of Ancient Mesopotamia*. Austin: University of Texas Press.

BREIMAN, LEO; JEROME FRIEDMAN; RICHARD OLSHEN; and CHARLES STONE. 1984. *Classification and regression trees*. Pacific Grove, CA: Wadsworth and Brooks.

CHURCH, KENNETH WARD, and WILLIAM GALE. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language* 5.1.19–54.
CITY OF DUNCAN. 1990. *Duncan: City of Totems*. Duncan, BC.
DANIELS, PETER, and WILLIAM BRIGHT (eds.) 1996. *The world's writing systems*. New York: Oxford.
DREW, F. W. M. 1969. *Totem poles of Prince Rupert*. Prince Rupert, BC: F. W. M. Drew.
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. Online: http://www.ejvs.laurasianacademy.com/ejvs1102/ejvs1102article.pdf.
FELDMAN, RICHARD. 2003. *Home before the raven caws: The mystery of the totem pole*. Cincinnati: Emmis Books.
FUGLESANG, SIGNE HORN. 1998. Swedish runestones in the eleventh century: Ornament and dating. *Runeninschriften als Quellen Interdisziplinärer Forschung*, ed. by Klaus Düwel and Sean Nowak, 197–218. Berlin: De Gruyter.
GARFIELD, VIOLA. 1940. *The Seattle totem pole*. Seattle: University of Washington Press.
GOOD, IRVING JOHN. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40.3–4.237–64.
GOODY, JACK, and IAN WATT. 1968. The consequences of literacy. *Literacy in traditional societies*, ed. by Jack Goody, 27–68. New York: Cambridge University Press.
GRAVES, THOMAS. 1984. *The Pennsylvania German hex sign: A study in folk process*. Philadelphia: University of Pennsylvania dissertation.
GUNN, SISVAN WILLIAM. 1965. *The totem poles in Stanley Park, Vancouver, B.C*. 2nd edn. Vancouver: Macdonald.
GUNN, SISVAN WILLIAM. 1966. *Kwakiutl House and totem poles at Alert Bay*. Vancouver: Whiterocks.
GUNN, SISVAN WILLIAM. 1967. *Haida totems in wood and argillite*. Vancouver: Whiterocks.
JACKSON, ANTHONY. 1984. *The symbol stones of Scotland: A social anthropological resolution of the problem of the Picts*. Elgin: Orkney.
JACKSON, ANTHONY. 1990. *The Pictish trail*. Elgin: Orkney.
JANNOT, JEAN-RENÉ. 2005. *Religion in Ancient Etruria*. (Wisconsin studies in classics.) Trans. by Jane K. Whitehead. Madison: University of Wisconsin Press.
JURAFSKY, DANIEL, and JAMES MARTIN. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd edn. Upper Saddle River, NJ: Prentice Hall.
KING, LEONARD W. 1912. *Babylonian boundary stones and memorial tablets in the British Museum*. London: British Museum.
KNESER, REINHARD, and HERMANN NEY. 1995. Improved backing-off for m-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95)*, 181–84.
LANDER, ERIC S., ET AL. 2001. Initial sequencing and analysis of the human genome. *Nature* 409.860–921.
LEE, ROB; PHILIP JONATHAN; and PAULINE ZIMAN. 2010a. Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences* 466.2121.2545–60. Online: http://rspa.royalsocietypublishing.org/content/466/2121/2545.
LEE, ROB; PHILIP JONATHAN; and PAULINE ZIMAN. 2010b. A response to Richard Sproat on random systems, writing, and entropy. *Computational Linguistics* 36.4.791–94.
MACK, ALASTAIR. 1997. *Field guide to the Pictish symbol stones*. Balgavies: The Pinkfoot Press. [Updated 2006.]
MAHADEVAN, IRAVATHAM. 1977. *The Indus script: Texts, concordance and tables*. Delhi: Archaeological Survey of India.
MAHR, AUGUST. 1945. Origin and significance of Pennsylvania Dutch barn symbols. *Ohio History: The Scholarly Journal of the Ohio Historical Society* 54.1.1–32.
MALIN, EDWARD. 1986. *Totem poles of the Pacific Northwest coast*. Portland: Timber.
MANNING, CHRISTOPHER, and HINRICH SCHÜTZE. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Nemenman, Ilya; Fariel Shafee; and William Bialek. 2001. Entropy and inference, revisited. *Advances in Neural Information Processing Systems* 14. Online: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-14-2001.
Newcombe, Warren Alfred, and British Columbia Provincial Museum. 1931. *British Columbia totem poles*. Victoria: Charles F. Banfield printer to the King's most excellent majesty.
Parpola, Asko. 1994. *Deciphering the Indus script*. New York: Cambridge University Press.
Possehl, Gregory. 1996. *The Indus age: The writing system*. Philadelphia: University of Pennsylvania Press.
Possehl, Gregory. 1999. *The Indus age: The beginnings*. Philadelphia: University of Pennsylvania Press.
Powell, Barry. 2009. *Writing: Theory and history of the technology of civilization*. Chichester: Wiley-Blackwell.
Rao, Rajesh. 2010. Probabilistic analysis of an ancient undeciphered script. *IEEE Computer* 43.3.76–80.
Rao, Rajesh; Nisha Yadav; Mayank Vahia; Hrishikesh Joglekar; R. Adhikari; and Iravatham Mahadevan. 2009a. Entropic evidence for linguistic structure in the Indus script. *Science* 324.5931.1165.
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* 106.33.13685–90.
Rao, Rajesh; Nisha Yadav; Mayank Vahia; Hrishikesh Joglekar; R. Adhikari; and Iravatham Mahadevan. 2010. Entropy, the Indus script, and language: A reply to R. Sproat. *Computational Linguistics* 36.4.795–805.
Rhys, John. 1892. The inscriptions and language of the northern Picts. *Proceedings of the Society of Antiquaries of Scotland* 26.263–351.
Roark, Brian, and Richard Sproat. 2007. *Computational approaches to morphology and syntax*. Oxford: Oxford University Press.
Roark, Brian; Richard Sproat; Cyril Allauzen; Michael Riley; Jeffrey Sorensen; and Terry Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. *Proceedings of the 50th annual meeting of the Association for Computational Linguistics, System Demonstrations*, 61–66.
Royal Commission on the Ancient and Historical Monuments of Scotland. 1994. *Pictish symbol stones: A handlist*. Edinburgh: Royal Commission on the Ancient and Historical Monuments of Scotland.
Scarborough, Hollis. 1990. The index of productive syntax. *Applied Psycholinguistics* 11.1–22.
Seidl, Ursula. 1989. *Die babylonischen Kudurru-Reliefs: Symbole mesopotamischer Gottheiten*. Freiburg: Universitätsverlag Freiburg.
Shannon, Claude. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27.379–423.
Sproat, Richard. 2010a. Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics* 36.3.585–94.
Sproat, Richard. 2010b. Reply to Rao et al. and Lee et al. *Computational Linguistics* 36.4.807–16.
Stewart, Hilary. 1990. *Totem poles*. Seattle: University of Washington Press.
Stewart, Hilary. 1993. *Looking at totem poles*. Seattle: University of Washington Press.
Sutherland, Elizabeth. 1997. *The Pictish guide*. Edinburgh: Birlinn.
The Encode Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489.57–74.
Vidale, Massimo. 2007. The collapse melts down: A reply to Farmer, Sproat and Witzel. *East and West* 57.333–66.
Winn, Shan M. 1990. A Neolithic sign system in southeastern Europe. *The life of symbols*, ed. by Mary Lecron Foster and Lucy Jayne Botscharow, 269–71. Boulder: Westview.
Wu, Katherine; Jennifer Solman; Ruth Linehan; and Richard Sproat. 2012. Corpora of non-linguistic symbol systems. Paper presented at the annual meeting of the Linguistic Society of America, Portland. Extended abstract online: http://elanguage.net/journals/lsameeting/article/view/2845/pdf.

YODER, DON, and THOMAS GRAVES. 2000. *Hex signs: Pennsylvania Dutch barn symbols and their meaning*. Mechanicsburg, PA: Stackpole.
ZIPF, GEORGE KINGSLEY. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Google, Inc.
76 9th Avenue
New York, NY 10011
[rws@google.com]

[Received 26 June 2013;
revision invited 1 October 2013;
revision received 9 October 2013;
accepted 6 November 2013]