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

Supplementary Materials

Richard Sproat

Google, Inc.

76 9th Avenue

New York, NY 10011, USA

1 Introduction

In this supplement, I describe the methodology used to arrive at the results discussed in the main paper. This includes a description of our corpora, as well as the statistical techniques and the results of applying them to the corpora. I compare our results to those of previous work, in particular (*1, 2*) and (*3*).

Since the notion of a non-linguistic system will be unfamiliar to many readers — even though everyone uses such systems every day — I start with a brief taxonomy of such systems

as a segue into the more technical discussion to follow.

2 Taxonomy

Linguistic symbol systems — writing — have been classified in various ways by scholars of writing systems for the past half century (*4–9*). While the proposed classifications differ, they all share the property that they differentiate systems on the basis of the kind of information encoded in the system, whether it be phonological segments, syllables, morphemes or words.

Systematic taxonomies of *non-linguistic* systems have certainly not been as prevalent and in fact may not exist. Where such systems are discussed — e.g. (*7, 10*) — it is usually just a few systems that are presented in juxtaposition to writing systems. In a forthcoming publication (*11*), I outline and justify a preliminary taxonomy of non-linguistic systems. Here I merely summarize the taxonomy, with examples, to give some context for the ensuing discussion.

But before we delve into the taxonomy, I should perhaps start by asking: Is the distinction between linguistic and non-linguistic systems that I laid out in the main paper really that clear? Lee and his colleagues (*12*) in their response to my critique (*13*) of their claim that Pictish symbos were writing, claimed that they adhere to a much looser notion of what it means to be a writing system, accepting Powell's (*14*) definition, namely that "writing is a system of markings with a conventional reference that communicates information". Notice that Powell's definition makes no reference to (natural) language. Indeed from Powell's point of view practically every symbol system we are discussing here would be considered writing. If Lee and his colleagues

really accept such a broad definition, then their paper, published in the *Proceedings of the Royal Society*, would seem to have simply derived the obvious fact that Pictish symbols were a "system of markings" that had a "conventional reference that communicates information", something that nobody would have doubted: Presumably the Picts would not have expended energy carving symbols on stones if they were not intended to communicate something, and the fact that the symbols repeat on different stones suggests that they were conventional. In contrast, by claiming that their method "reveals" Pictish symbols to be language, it seems that Lee and his colleagues intended that readers accept the conclusion that the symbols served a similar function to the marks that you are currently reading. Indeed, Powell aside, most students of writing systems would accept the conclusion that writing is a special form of symbol system, one that encodes linguistic information that is mostly or completely lacking in non-linguistic systems. Without accepting such a distinction, work such as that of (*1–3*) would have no purpose. With this distinction in mind I turn to a taxonomy.

Non-linguistic systems may be classified into at least the following types:

- **Simple informative systems**. In simple informative systems, the symbols by and large convey a single piece of information. The helical red, white and blue barber pole, for example, indicates the presence of barber shop. In weather reports, icons are used to represent various states of the weather, such as whether it is sunny, partly cloudy, raining, etc. Further examples: other symbols of guild such as three balls for a pawnbroker; institutional logos; traffic information signs; ownership signs, such as brands, e.g. Mongolian horse brands (*15*) or house marks (`http://en.wikipedia.org/wiki/House_`

`mark`).

- **Emblematic systems**. Closely related to simple informative systems are emblematic systems where the symbols represent some special distinction earned by the bearer. Examples: symbols of military rank and distinction, scouting merit badges, Phi Beta Kappa keys, and symbols of other scholarly fraternities; letter grades on academic assignments or in courses.

- **Religious Iconography**. Religious iconography could also be characterized as a simple informative system, except that here though here the notion of "single piece of information" is much less clear, since such symbols are intentionally often highly multivocal in the meanings they evoke. Examples: Christian cross, star of David, star and crescent, dharmachakra, swastika (in its original Buddhist usage) .

- **Heraldic systems**. Heraldic systems are similar to emblematic system in that they usually represent a particular set of features of the bearer, including possibly marks of distinction. They differ, however, in that heraldic systems are frequently highly combinatoric, involving "texts" built of many symbols, often with a quite rigid syntax. Examples: European heraldry, *kudurrus*, some functions of Totem poles (*16*) .

- **Formal systems**. In a formal system, the individual symbols have well-defined meanings, and there are generally strict rules on how the symbols may be combined. Examples: mathematical symbols, alchemical symbols, chemical notation, Feynman diagrams (*17*), programming flowcharts and Systems Biology Graphical Notation (*18*).

- **Performative systems**. Performative systems indicate a sequence of actions to be taken to perform a particular task. Perhaps the most familiar to many people today is are the wordless assembly instructions that come with furniture from Ikea. Other examples: Silas John's system for notating Apache prayers (*19*), musical notation (*20*), dance notation and other movement notation systems (*21*), chess notation, systems that can be used to indicate the sequence of plays in a game, knitting patterns (*10*).

- **Narrative systems or "prompt" texts**. Narrative systems are used to recount stories and as such are the most language-like of the non-linguistic systems. In narrative systems, the symbols typically represent actors or events in the story in an iconic way. Examples: Dakota winter counts (*22*), (probably) Naxi symbology (*23*).

- **Purely decorative systems**. Some systems that involve what are commonly thought of as symbols seem nonetheless to be purely decorative. In such systems, the symbols may derive historically from symbols that had meanings or ranges of meanings, but where those meanings are quite irrelevant in their current use. A clear modern example is the use of Chinese characters in body tattoos, worn by people who may be completely unaware of the character's original meaning, and use them solely because they look "cool". Other examples are Pennsylvania German barn stars (*24*), and Asian emoticons (e.g. (*25*)) where in the latter case symbols from various scripts are combined into a "text" that represents an image, usually a face. The face itself may convey some sort of emotion (e.g. sadness, via depiction of crying), but other than that has no real meaning and mostly functions to

decorate the surrounding text.

For the purposes of the discussion here, I am particularly interested in systems where the symbols may be combined into texts. As we have seen already in the discussion above, some systems — narrative systems, performative systems, heraldic systems, formal systems, for example — are particularly to be found in texts, but in fact systems of almost any kind may be used combinatorically, depending upon what kinds of things the system denotes. Clearly a mathematical system that only allowed single symbols, or a performative system for furniture assembly that was similarly simplex, would not be of much utility. On the other hand, in religious symbology, a single symbol on its own may suffice to communicate the intended message: to those devoted to the faith, the Christian cross can be a highly evocative symbol.

3 Materials and Data Preparation

For this work I selected seven non-linguistic symbol systems and fourteen linguistic symbol systems for analysis. Note that in addition to the non-linguistic symbol systems described here, a further five are under development and will be released along with the corpora described here.

3.1 Non-linguistic corpora

Two kinds of non-linguistic corpora were collected. The first is ancient or traditional systems which serve as exemplars of what a complex ancient non-linguistic system can be like. If a set of "texts" in a previously unknown symbol system is discovered among the ruins of an ancient

civilization, these systems could serve as models of what that symbol system might have been, assuming it was not a form of written language. Our current set of such systems comprises Mesopotamian diety symbols (*kudurrus*) (*26*), Vinča symbols (*27*), Pictish symbols (*28–32*), Totem poles (*16, 33–43*), and Pennsylvania German barn stars (*24, 44, 45*). The development of these ancient or traditional corpora is described in (*46*). Two of these systems are of interest for other reasons: Mesopotamian deity symbols and Vinča systems were mentioned by (*1*) as models for low- and high-entropy systems, respectively. As we shall see, Rao and his colleagues were completely wrong about this, a point that could only have been known for certain by actually doing the work of collecting the corpora. A third — Pictish symbols — were, of course, already discussed by (*3*), and argued to be writing: their inclusion here as non-linguistic accords with what is probably still the most widely accepted view, but in addition to that, the tests presented below — including a retrained version of one of Lee and colleagues' own measures — are at least consistent with it being non-linguistic.

The second kind are modern systems, which can easily be collected electronically. These, presumably, are poorer models of ancient symbol systems, but have the advantage that they are easy to collect from online sources, which was mostly not the case for the ancient or traditional systems, which had to be transcribed from print sources.

After discussing the various corpora, we present in Section 3.1.8 statistics on the size of the various corpora. **S2** contains examples of the first five symbol systems, along with the corresponding transcription in the XML markup scheme that was introduced in (*46*). One feature of our XML markup is that it allows one some flexibility in how one extracts the elements of

the text. For example, if there are clear lines in the text, then we preserve line information. If the text is circular (so that it is not defined where the beginning or end of the text occurs) that information is also marked. Furthermore, one of the questions that arises in the analysis of a symbol system is whether to treat elements that seem to be composed of more atomic symbols as single symbols or as a composite of the individual symbols. The `symbolUnit` preserves the grouping structure so that one can make the choice of level of analysis later on.

At the head of each section I propose a classification for the symbol system in question in terms of the taxonomy developed in Section 2. The breakdown of the types is as follows. Note that totem poles account for two of the types — heraldic and narrative, since poles can have either of these functions:

Type	# of systems represented
Heraldic	3
Narrative	1
Decorative	2
Unknown	2
Simple informative	1
Formal	1

3.1.1 Vinča symbols

Classification: Unknown, possibly religious

The Vinča were a late Neolithic people of Southeastern Europe between the sixth to the third

millenium BCE who left behind pottery inscribed with symbols. Often the symbols occurred singly, but on about 120 items the symbols occur in "texts" of two or more symbols; see **S2** for a sample Vinča text. In such texts, the symbols are sometimes, but not always arranged linearly as in a typical script. The most authoritative compilation of the Vinča materials is (*27*), who classifies the symbols into about 200 types, and provides a corpus of the multisymbol texts from 123 sources, comprising more than 800 tokens. In addition to their sometimes linear arrangement, the Vinča symbols had other script-like properties: in Winn's analysis, some signs seemed to be comprised of two signs ligatured together. Though Winn characterizes the system as "pre-writing", it must be stressed that there is no reason to think this was a linguistic writing system. Very likely, the system had religious significance. As Winn states (p. 255): "In the final analysis, the religious system remains the principle source of motivation for the use of signs." However the meaning of the symbols remains unknown.

Developing Winn's materials into an electronic corpus was relatively straightforward, since he very nicely lays out the texts in his catalog, giving both the form and the type number for each sign. We followed his linearization of the texts, but since he also provides line drawings of the artifacts themselves, it was possible in some cases to indicate the true spatial relationship between the symbols.

3.1.2 Mesopotamian deity symbols on *kudurrus*

Classification: Heraldic

S2 gives an example of a *kudurru* boundary stone with deity symbols. *Kudurrus* were

legal documents that specified property rights. The stones often also included actual cuneiform Babylonian writing. The deity symbols had an essentially heraldic function, to indicate favored deities of the owner of the property. Our source for the deity symbol corpus was Seidl (*26*). We picked texts from all stones where the depictions in (*26*) were clear enough to read. Fortunately (*26*) includes a chart that lists, for each stone, the symbols that appear on that stone (though not in the order they appear); that chart proved useful for checking that the reading of the symbols was correct.

3.1.3 Pictish stones

Classification: Unknown, possibly heraldic

The Picts were an Iron Age people (or possibly several peoples) of Scotland who, among other things, left a few hundred standing stones inscribed with symbols, with "texts" ranging from one to a few symbols in length. The meaning of the symbols is unknown, but until recently it would never have occurred to anyone to assume that they are some form of writing. If nothing else, the Picts evidently had at least some literacy in the (segmental) Ogham script (*47*). An example of Pictish symbols can be found in **S2**.

Fortunately for our work, a corpus of images, comprising 340 stones, along with information on the symbols on each stone is available from the University of Strathclyde `http://www.stams.strath.ac.uk/research/pictish/database.php`. The stones are cross-referenced to standard texts such as (*28–32*). The main work that was needed was to correct the ordering of symbols in some cases since the ordering of the symbols in the Strathclyde

transcription does not always correspond to the symbols' ordering on the stone.

3.1.4 Totem Poles

Classification: Heraldic, narrative

Totem poles are the product of a wide range of Native American cultures of the Pacific Northwest, dating from the 19th and 20th centuries. The texts, which consist of anthropomorphic and zoomorphic symbols, carved vertically on cedar or other tree trunks, represent a variety of kinds of information from legends, genealogical information, or depictions of important events. Types of poles include memorial poles, house frontal poles, mortuary pole, and heraldic poles. Different tribes had different styles so that, for example, on Haida poles the figures are carved in bas relief (*39*). The "texts" also served different purposes among different groups: Thus totem poles are a good example of how a symbol system can vary across different regions and cultures without that variation implying that the system encoded language; cf. (*48*).

We used a number of published resources in developing our data. In addition to Malin (*39*), we transcribed texts from (*16, 33–38, 40–43*).

There are many recurring themes on totem poles. For example, certain combinations of symbols always allude to the same folk story. A whale with a man and a woman always refers to the Nanasimget story (*42*). In principle then this could be represented by a single symbol (e.g. NANASIMGET), but instead we elected to use the `symbolUnit` tag (*46*) to represent the grouping:

```
<symbolUnit>
```

```
  <symbol><title>Whale</title></symbol>
  <symbol><title>Man</title></symbol>
  <symbol><title>Woman</title></symbol>
</symbolUnit>
```

3.1.5 Pennsylvania German Barn Stars

Classification: Decorative

Barn stars, commonly known as "hex signs", are a traditional decorative art among Pennsylvania German ("Dutch") communities. They are found widely in northeastern Pennsylvania, particularly Berks County, as well as in scattered communities in Ohio and elsewhere. They are mostly used to decorate barns, but can also be found on other structures, such as porches. Traditional barn symbols consist mostly of stars, rosettes, wheels-of-fortune, and swastikas.

There is a common belief that "hex signs" had a magical function, such as to ward off evil spirits, and indeed this belief was sometimes expressed by the German farmers who used the symbols on their barns (*44*). And the barn symbols themselves can be traced back in many cases to symbols that probably had definite sets of meanings at one time (*24, 44, 45*). That said, the consensus of much of the small scholarly literature on this topic is that barn stars probably had no particular meaning at all for the people who used them and were used rather for decoration (*24, 45*).[1]

[1] If barn stars are purely decorative, why include them here? There are two reasons. First of all, the symbols can occur in "texts" of several symbols in length. While the texts look rather unlike written texts — for one thing, they

1 Our barn star corpus is based upon the slide collection of W. Farrell, who toured areas
2 around Berks County in the 1940's and photographed many barns that were decorated with

are almost always symmetric — they are nonetheless interesting from the point of view of developing statistical techniques that might possibly distinguish linguistic from non-linguistic systems.

Second, much of the critique of (*49*) has depended on a fundamental confusion of what is meant by the term "non-linguistic symbol system," the most common misconception being that we were referring to randomly arranged meaningless symbols. A good example of this misconception is expressed by Vidale in his critique of our paper (*50*). Using the obviously decorative pottery designs of Shahr-e Sukhteh and elsewhere as his examples he states (page 344):

> Together with coupling and opposition of selected symbols, systematic, large-scale redundancy (constant repetition of the same designs or symbols) is a distinctive feature shared by the more evolved and formally elaborated non-linguistic symbolic systems considered (highly repetitive patterns on the pottery of Shahr-i Sokhtai, endless repetition of icons such as scorpions, men-scorpions, temple facades, water-like patterns and interwoven snakes at Jiroft, and redundant specular doubling of most major symbols in the Dilmunite seals). While positional regularities might be detected in part of the Jiroft figuration, redundancy in all these systems dismiss one of the basic assumption of Farmer & others, who take the rarity of repeating signs as a proof of the non-linguistic character of the Indus script.

Repetition is indeed a characteristic of decorative art: consider the pineapple motif popular in American Colonial interior decoration and stenciled in repeated patterns on walls. And, not surprisingly, high symbol repetition rates do show up as a strong feature of barn stars. But high rates of repetition are not particularly characteristic of non-linguistic systems, though as we shall see high rates of local *reduplicative* repetition relative to the total repetition rate *do* seem to be characteristic of many non-linguistic systems. Systems that have massively higher than expected symbol repetition, such as barn stars, tend to be decorative. Vidale is simply confused on this point.

barn stars. His slides, in five boxes of about 100 slides each, are now housed in the archives of the Berks County Historical Society in Reading, PA. The most useful boxes, in terms of amount of material, are boxes 1 and 2: in these boxes the photos show the whole barn with associated decorations. Boxes 3 and 4 contain a little more material, but in many cases the photos only show a portion of the barn and its decorations, and there are a number of duplicates of barns already seen in Boxes 1 and 2. Box 5 seems for the most part to be useless from the point of view of collecting material on barn stars.

The designs of barn stars are quite numerous, but break down into a relatively small set of categories, following the classification of Farrell himself, and Graves (*24*). The main ones, in order of descending frequency of occurrence, are: 8-POINT-STAR, 5-POINT-STAR, 4-POINT-STAR, 6-POINT-STAR, ROSETTE, SWIRLING-SWASTIKA, 12-POINT-STAR, WHEEL-OF-FORTUNE, 14-POINT-STAR, 7-POINT-STAR, 10-POINT-STAR, 15-POINT-STAR, 16-POINT-STAR. In many cases the names of the slides in Farrell's system give a clear indication of what category the (main) symbol on the barn belongs to: thus `Box01/F.106St6` depicts a barn with three six-pointed stars (`St6`). Farrell however does not distinguish some cases that Graves does distinguish: the main example of this is rosette, which Farrell classifies under six-pointed star; see Figure 1. In such cases I followed Graves' classification, where I was able to clearly assign the symbol to the rosette category. In one notable case I invented a new category, what I term 4-SLICE-PIE, which Farrell classifies as 4-POINT-STAR; see Figure 2. The "four-slice pie" is, however, so distinct that it seems to deserve its own category.

3.1.6 Weather Icons

Classification: Simple Informative

The Weather Underground (`www.wunderground.com`) provides weather forecasts for many parts of the world. The forecast includes icons that represent the predominant weather expectation for a given day. For example, `http://icons-pe.wxug.com/i/c/a/rain.gif` is the icon for rain during the day. There are about 20 distinct icons.[2] These icons, taken in series, form a "text" that corresponds to the weather predictions for a five-day period, one icon per day. In this case we are dealing with a human-designed symbol system, but one where the distribution of the symbols is determined by natural phenomena (or more properly a computational model thereof).

We stored weather icon "texts" by collecting weekly forecasts from the Weather Underground site for a selection of 161 cities throughout the world for 72 days, giving us a corpus of 50,710 symbols.[3] See Figure 3 for an example of a weather "text".

3.1.7 Asian Emoticons

Classification: Decorative

As part of a project on normalization of Twitter messages, my colleagues and I developed an analysis system to detect and parse Asian emoticons (*kaomoji*) (*25*). Unlike the familiar

[2]There are actually double this number since there are separate icons for day and night: for example alongside the "partly cloudy" icon for daytime, there is a nighttime version. In this work we only included the daytime icons.

[3]Note that on occasion the scraping of the page resulted in no data returned.

90-degree flipped ASCII "smileys" — `:-)`, `;-)`, `:-(`, `8-)` ... — Asian emoticons (so-called because they were popularized by Japanese and other East Asian users) are oriented horizontally, and make use of a much wider range of characters. Some examples can be seen in Figure 4. Traditional ASCII smileys are relatively limited, comprising perhaps a few tens of examples. Asian smileys, in contrast, are productive and open ended: our collection includes thousands of examples.

Of interest from the point of view of this project are the individual characters used in the emoticons. Asian emoticons tend to be somewhat (though often not perfectly) symmetric. However unlike in the symmetric "texts" found with Pennsylvania barn stars, the mate characters found in Asian emoticons are different symbols, chosen because they are visually close mirror images. A statistical analysis of the symbol distributions would easily miss the fact that the texts are symmetric.

3.1.8 Statistics on Corpus Sizes

The sizes comprising the numbers of texts, the number of tokens, the number of types and the mean text length of the above-discussed corpora is given in Table 1.

In addition, we also included in our analysis a corpus of Indus bar seals from Harappa and Mohejo Daro, which we developed as part of the work reported in (*49*). The details on this latter corpus are given below:[4]

[4]The reason for not using the same Indus data as that used in (*1, 2*) is that none of the Indus corpora that have been collected by various groups over the years have been made available to other researchers, which has in turn

Corpus	# Texts	# Tokens	# Types	Mean text length
Indus bar seals	206	1,265	209	6.14

Note that the bar seals are relatively late instances of Indus inscriptions, and that the mean length of the texts, 6.14, is substantially longer than the mean length of all Indus inscriptions taken together, which is 4.5 (*49, 51*).

3.2 Linguistic Corpora for Comparison

For comparison with the non-linguistic systems described above, I gathered a set of linguistic corpora from various existing sources. Just as I have tried to collect non-linguistic corpora of a variety of types, so I have also tried to gather linguistic corpora that are varied along a couple of dimensions. The first dimension is age: we have both extremely ancient systems — Sumerian, Egyptian, Ancient Chinese and Linear B (Mycenaean Greek), as well as various modern systems. The second dimension was type of writing system: we have examples of morphosyllabic writing, syllabaries, alphasyllabic writing, abjads and segmental writing. See, e.g., (*7*) for descriptions of what these various terms mean. In most cases each individual "text" in our corpus corresponded to an individual line in the encoding in the source corpus from which we extracted our sample.

The following sections give brief descriptions of each of the corpora. Summary statistics can be found in Table 2.

made it difficult to verify results.

3.2.1 Amharic

Collection of texts from http://www.waltainfo.com. The text was tokenized at the syllable level, with spaces represented as a separate token ("#").

3.2.2 Modern Standard Arabic

Collection of the first 10,000 headlines from the Linguistic Data Consortium (LDC) Arabic Gigaword corpus http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T12. The text was tokenized at the letter level, with spaces represented as a separate token ("#").

3.2.3 Modern Chinese

Collection of the first 10,000 headlines from the LDC Chinese Gigaword corpus http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T09. The text was tokenized at the character level.

3.2.4 Ancient Chinese

Oracle bone texts from Chenggong University, Taiwan. We kept only lines for each document that contain no omissions. The text was tokenized at the character level.

3.2.5 Egyptian

Collection of texts transcribed using the JSesh hieroglyph editor (`http://jsesh.qenherkhopeshef.org/`) downloaded from `http://webperso.iut.univ-paris8.fr/~rosmord/hieroglyphes`. The text was tokenized at the glyph (individual symbol) level.[5]

3.2.6 English

Collection of the first 10,000 headlines from the LDC English Gigaword corpus `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T05`. The text was tokenized at the letter level, with spaces represented as a separate token ("#").

3.2.7 Hindi

Collection of the first 1,000 lines of text from the Hindi corpus developed at the Central Institute of Indian Languages. The text was tokenized at the letter level, with spaces represented as a separate token ("#").

3.2.8 Korean

Collection of the first 10,000 headlines from the LDC Korean Newswire corpus `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2000T45`.

[5]The following texts were used, all in `http://webperso.iut.univ-paris8.fr/~rosmord/hieroglyphes/`: CT160_S2P.hie, DoomedPrince.hie, HAtra.hie, HetS.hie, L2.hie, LC26.hie, Pacher-ereniset.hie, Prisse.hie, gurob.hie, amenemope/*.gly, ikhernofret.hie, ineni.hie, kagemni.hie, lebensmuede.hie, mery.hie, naufrage.hie, sethnakht.hie, twobro.hie, year400.hie.

Following the standard Unicode encoding of Korean, the text was tokenized at the Hangul syllable level, with spaces represented as a separate token ("#").

3.2.9 Korean Jamo

Same source as above, but here the text was tokenized at the level of the (*jamo*) — the individual letters that make up the syllable composites, with spaces represented as a separate token ("#").

3.2.10 Linear B

Transcription by the author of 300 tablets from (*52*). Omitted lines with uncertainties and/or omissions. Tokenized at the glyph level.

3.2.11 Malayalam

The Malayalam corpus (937 lines) developed at the Central Institute of Indian Languages. The text was tokenized at the letter level, with spaces represented as a separate token ("#").

3.2.12 Oriya

Collection of the first 1,000 lines of text from the Oriya corpus developed at the Central Institute of Indian Languages. The text was tokenized at the letter level, with spaces represented as a separate token ("#").

3.2.13 Sumerian

Tablets from Gudea ruler of Lagah (c. 2100 BCE), extracted from the Cuneiform Digital Library Initiative (`http://cdli.ucla.edu/`) via a transliteration search for *gu3-de2-a*, with "rime 3" in the primary publication field. Omitting lines with "#", "<" or "[", since these mark uncertainty/reconstructions.[6] "Texts" consisted of individual lines in the CDLI transcription, meaning that the Sumerian "texts" are rather short.[7]

3.2.14 Tamil

Collection of the first 1,000 lines of text from the Tamil corpus developed at the Central Institute of Indian Languages. The text was tokenized at the letter level, with spaces represented as a separate token ("#").

[6] Thanks to Chris Woods for suggesting appropriate search terms.

[7] One issue with Sumerian that will need to be resolved in future work is that the symbols are transcribed, following standard Sumerological practice, with romanized spellings of either the pronunciation of the symbol (if in lower case) or of the morpheme denoted by the symbol (if in upper case), followed by a subscript. Thus aya_2 is a transcription of 𒀀𒀾, which is one of the ways of writing the two-syllable sequence *a-ya*. The problem is that the system is not many-to-one, but many to many: 𒀀𒀾 could also be $duru_5$, $eš_{10}$, and several others. Standard Sumerological transcriptions thus give an overestimate of the actual number of glyph types appearing in a document. This issue can only be resolved by knowing which transcribed elements map back to a single glyph.

3.3 Preparation

Non-linguistic corpora were processed by extracting each text from the XML markup, and representing the corpus with one text per line. In cases where there was a `SymbolUnit` (see above) we extracted the individual symbols making up that unit. Linguistic corpora were represented with one text per line, tokenized as described above.

All probability-based entropy statistics discussed below were computed using tools based on the OpenGrm NGram library (*53*), available from `http://www.opengrm.org`; these tools will be made available along with the corpora. We used Kneser-Ney smoothing, including the probability and thus entropy estimates for unseen cases. Start and end symbols to pad the text are implicitly computed by the software.

4 Methods and Results

In the ensuing sections we describe the various statistical measures we applied to our corpora, along with the results we derived.

4.1 Bigram Conditional Entropy

One measure that has been widely used to measure the information structure of language is conditional entropy (*54*), for example, bigram conditional entropy, defined as follows:

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

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 on the other hand, 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 a whole range of non-linguistic systems, fall between these two extremes.

Bigram conditional entropy was used by (*1*) to argue for the linguistic status of the Indus Valley symbols. One can compute entropy over various sized units of granularity — symbols in the script or symbol system, words, etc. — and for various portions of the corpus. In Rao and colleagues' 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, given a particular granularity.

Figure 5 shows these curves, taking $n = 20$, for a variety of actual writing systems, Mahadevan's corpus of Indus inscriptions, and two artificial systems with maximum entropy, and minimum entropy, which Rao and colleagues label as "Type 1" and "Type 2", respectively. Figure 6 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 Figure 5, as well as the real non-linguistic systems, DNA, protein ("prot") and Fortran ("prog lang").

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 Mesopotamian deity symbols, respectively. Their beliefs are based, respectively, on Winn's (*55*) 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 (*13*) and (*56*), I argued that Rao's interpretation of how these symbol systems work is a misinterpretation, and indeed the results we present below support that conclusion. Rao and colleagues also present results for DNA and protein in Figure 6 (though not in Figure 5): it is clear from that figure that the curves of these biological symbol systems are very close to the maximum entropy curves.[8][9]

Turning to our own data, in Figure 7 I show conditional entropy growth curves for all of our non-linguistic systems (in blue), all of our linguistic samples (in red), and for our corpus of Indus bar seals (in green).

As can be seen in the plot, the results are pretty much across the map. The *kudurru* inscriptions turn out, contrary to what Rao and colleagues claim — and what Rao continued to insist in (*2*) — to have the highest entropy of all the systems. Chinese shows the lowest. Between

[8]At least in the case of DNA, though, this is because Rao and colleagues take the "alphabet" of DNA to be the base pairs, an alphabet of size 4. But these are surely not the right units of information: the base pairs 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 "non-coding regions" to mimic the non-coding regions of DNA — surely a meaningless comparison for these purposes.

[9]Note that while Rao reports that they used Kneser-Ney smoothing (*57*), a number of details are not clear: for example, it is not clear whether they included estimates for unseen bigrams in their analysis. Discrepancies such as this and others may help explain why the entropy estimates I present below are on bulk lower than those presented in Rao and colleagues' work.

these two extremes the linguistic and non-linguistic systems are mixed together. The Indus bar seal corpus falls in the middle of this distribution, but that is quite uninformative since the distribution is a mix of both types of system.

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 (from Rao and colleagues' point of view) unexpectedly high entropy estimate for the *kudurru* corpus be due simply to undersampling? In Figure 8 I address that concern. This 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. Note that Arabic has roughly the same number of symbol types as the *kudurru* corpus; also, the smallest sample of 100 lines is about 4 times the size of the *kudurru* corpus, and thus is within the same order of magnitude. The entropy does indeed change as the corpus size changes, with the smaller corpora having higher overall entropy than the larger corpora. This is not surprising, since with the smaller corpora, the estimates for unseen cases are poorer and, one would expect, 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.

4.2 Block Entropy

We turn now to a second entropic measure, what Rao (*2*) terms "block entropy", defined as follows:

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

Here N is the length of an ngram of tokens, say 3 for a trigram, so all we are doing is computing the entropy of the set of ngrams of a given length. One can vary the length being considered, say from 1 to 6, and thus get estimates for the entropy of all the unigrams, bigrams, etc., 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 4. The maximum entropy thus normalized is then just N, the length of the ngrams being considered, for any system.

Figure 1a in the main paper shows the block entropy estimates from (*2*) for a variety of linguistic systems, some non-linguistic systems and the Indus corpus from (*1*). For this work, Rao used an estimation method proposed in (*58*), 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. These are shown in Figure 1b in the main paper. As with the conditional entropy, and unlike Rao's clear-looking results, the systems are quite mixed up, with linguistic and non-linguistic 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 Figure 1a in the main paper, suggesting that even though

the bar seals constitute a smaller subset of the whole corpus with rather different average text length, there is some similarity in the distributions of ngrams. Note on the other hand 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.

4.3 Measures Derived in Part from Entropy

Lee and colleagues (*3*), in their study of Pictish symbols, develop two measures, U_r and and C_r, defined as follows. First, U_r is defined as:

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

where F_2 is the bigram entropy, N_d is the number of bigram types and N_u is the number of unigram types. C_r is defined as:

$$C_r = \frac{N_d}{N_u} + a\frac{S_d}{T_d} \tag{4}$$

where N_d and N_u are as above, a is a constant (for which, in their experiments, they derive a value of 7, using cross-validation), S_d is the number of bigrams that occur once (*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.

Lee et al. use C_r and U_r to train a decision tree to classify symbol systems. The resulting tree is shown in Figure 9. Thus 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. As I note in the main paper, when one applies Lee et al's tree, "out of the box" to our data, all but one of our systems are classified as linguistic; see Table 3.

4.4 Models of Association

All of the measures thus far discussed involve entropy. 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 (*54, 59*) for two terms t_i and t_j:

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

This is known to produce reasonable results for word association problems, though there are also problems with sensitivity to sample size as pointed out in (*60*).

In this study I use pointwise mutual information between adjacent symbols 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* or *the is* are not very likely sequences in English.

I therefore look at the mean association of the n most frequent symbols, normalized by the

number of association measures computed,

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

and 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, I estimated probabilities of bigrams and unigrams using OpenGrm.

As we can see, and as with the entropic measures, there is no clear separation of linguistic and non-linguistic systems. Both kinds of systems have symbols that are more or less strongly associated with each other (as well as symbols pairs that are not so strongly associated). For example, in the case of Mesopotamian deity symbols there is a particularly strong association between the HORNEDCROWN and the SYMBOLBASE symbols.

4.5 Models of Repetition

To the extent that they involve the statistical properties of the relations between symbols, the measures that we have 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 ((*54*)), language also has many non-local 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 distribution, as we see in Figure 11. Furthermore, some badges tend to be earned before others, and while there is no requirement that merit badges be applied to the sash in any particular order (`http://www.scoutinsignia.com/sash.htm`), nonetheless one would expect that the earlier earned badges would tend to come earlier in the text. Thus we would expect a hierarchy rather like what we find in the *kudurru* texts. All of this implies that entropic measures of the kinds we have been considering would find "structure" in these "texts". Yet such measures 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 (e.g. whole morphemes versus phonological segments), the probability of repetition decreases, but we would still expect some repetition.

It is not only the amount of repetition in a corpus, but how the repetition manifests itself, that is important. In (*49*) we argued that the Indus inscriptions not only showed rather low repetition rates but that when one does find repetition of individual symbols within an inscription, the repetition patterns often involve repetitions of adjacent symbols as well as symmetric arrangement (see (*49*), Figure 6).

We can capture part of this oddity with a measure of repetition that seeks to quantify the *rate of iterated repeats*, relative to the *total number of repetitions* Specifically, we count the number of tokens in each text that are repeats of a token that has already occurred in that text, and sum that number over the entire corpus. Call this number R. We then count the number of tokens that are repeats in each text, and which are furthermore adjacent to the token they repeat. Sum that over the entire corpus, and call this number r. So for example for a single text:

A A B A C D B

R would be 3 (two *A*s are repeats, and one *B*), and r would be 1 (since only one of the repeated *A*s is adjacent to a previous *A*). Thus the repetition rate $\frac{r}{R}$ would be 0.33. Table 4 shows our corpora ranked by $\frac{r}{R}$. This measure is by far the cleanest separator of our data into linguistic versus non-linguistic. If one sets a value of 0.10 as the boundary, only Sumerian and Mesopotamian deity symbols are clearly misclassified (while Asian emoticons and Egyptian are ambiguously on the border). Not surprisingly, this is the feature most often picked by the decision tree classifier we will discuss in the next section.

One important caveat is that the repetition measure is also 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, whereas at the same time, the more chance that if there *is* repetition, the repetition will involve adjacent symbols. Of course the cor-

relation 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/non-linguistic) as the independent variable and mean text length or $\frac{r}{R}$ as dependent variables yielded the following results:

$$\begin{array}{lll} \text{Mean text length} & F = 5.75 & p = 0.027 \\ \frac{r}{R} & F = 15.83 & p = 0.0008 \end{array}$$

Mean text length is thus well predicted by class, with non-linguistic systems having shorter mean lengths. But the repetition rate is even better predicted, suggesting that the 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 we have artificially limited the texts to length no more than six (by simply trimming each text to the first six symbols). Also, in order to make the corpora more comparable, we limit the shortened corpus to the first 500 "texts" from the original corpus. As we can see in Table 5, the separation is not as clean as in the case of Table 4. Nevertheless, the five corpora with the highest $\frac{r}{R}$ are non-linguistic (if one counts the Indus system as non-linguistic) and the nine corpora with the lowest values are all linguistic. Again this measure is a far better separator than any of the models that we have discussed previously, which are based ultimately on probabilities.

4.6 Two-Way Classification Using Decision Trees

This study has collected a variety of features that might in principle be useful for determining whether an unknown symbol system is linguistic or non-linguistic. Some of these features have been explored in previous literature, others are novel here. A common technique in computational linguistics is to combine features, each of which may on its own not be a very good predictor, into a discriminative model. To that end I used the above features, along with the Classification and Regression Tree (CART) algorithm (*61*) to train (binary branching) classification trees for a binary classifier (linguistic vs. non-linguistic). The features used were:

- The PMI association measure over the symbol set comprising 25% of each corpus
- Block entropy (*2*) calculations for $N = 1$ to $N = 6$ (block 1 . . . block 6).
- Maximum conditional entropy (*1*) calculated for Figure 7.
- F_2, the $\log_2$ bigram conditional entropy for the whole corpus, from (*3*).
- U_r from (*3*).
- C_r from (*3*).
- The repetition measure $\frac{r}{R}$.

Since the set of corpora is small — only 21 in total, not counting the Indus bar seals — I tested the system by randomly dividing the corpora into 16 training and 5 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, which in every case was to classify the system as linguistic).

It is interesting to see how these 100 trees classify the Indus bar seals, which were held out from the training sets. 98 of the trees classified it as *non-linguistic*, and only 2 of the trees classified it as linguistic. Furthermore, the 98 that classified it as non-linguistic had a higher mean accuracy — 0.81 — than the 2 that classified it as non-linguistic (0.3). Thus, more and better classifiers tend to classify the Indus symbols as non-linguistic 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 corpus X 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 harder to classify and thus helps performance if it is not part of the test set? Tables 6 and 7 provide some results on this. 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, which also results

[10]To show that this result is no artifact of the particular setup we are using, I 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.

in better performance if it is not in the test data. 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. Mesopotamian deity symbols are also misclassified by the repetition measure, and thus could lead to better performance when they are not part of the test data.

In what was just described, I included Pictish among the non-linguistic systems. Of course, with the publication of (*3*), this classification has become somewhat controversial. What if one does not include the Pictish data? In this case I had training sets of 15 corpora, and again held out 5 corpora 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 *non-linguistic* hypothesis: 97 classified Pictish as non-linguistic, with a mean accuracy of 0.81; 3 classified it as linguistic, with a lower mean accuracy of 0.4.

It is also interesting to consider the features that are used by the trees. These are presented in Table 8 for the two experimental conditions (with versus without Pictish). In both cases, the most prominent feature was repetition, and the second most common is C_r. The latter is interesting since if one compares Lee et al.'s tree in Figure 9, it is C_r which was involved in the top-level linguistic/non-linguistic decision. This may seem surprising confirmation of Lee et al's results, given that I showed in Section 4.3 that C_r and U_r, with Lee et al's settings, systematically misclassified nearly all of our non-linguistic corpora. What this means, however is that while their own settings — the ones they used to arrive at the linguistic hypothesis for Pictish symbols — are probably not useful, the measure itself is somewhat useful, if one is

allowed to retrain the boundary between linguistic and non-linguistic for C_r.

To see the importance of the various measures in another way, Table 9 shows the results of the Wilcoxon signed rank test for each feature comparing for the two populations of linguistic and non-linguistic corpora. Only for the repetition measure, and C_r are the population means different according to this test, suggesting that the other features are largely useless for determining the linguistic status of a symbol system.

Since the repetition measure is well correlated with mean length of texts in the corpus, and the non-linguistic 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 10, a wider variety of trees is produced, and Lee and colleagues' C_r is the favored feature, being used in 82 of the trees. The results for the held out Indus bar seal corpus are not as dramatic as when I included the repetition feature, but they still highly favor the non-linguistic analysis. 88 of the trees classified the system as non-linguistic (mean accuracy 0.75), and 12 as linguistic (mean accuracy 0.37).

Similarly lower results were obtained when I replaced the repetition rate $\frac{r}{R}$ from Table 4 with that computed over the artificially shortened corpora seen above in Table 5, 85 of the trees classified the Indus symbols as non-linguistic (mean accuracy 0.63), and 15 classified it as linguistic (mean accuracy 0.36). The features used by these trees are shown in Table 11. Repetition is far less dominant than it is in the original tree set shown in Table 8, but it is still the second most used feature, after C_r, occurring in 35 out of 100 trees.

The most useful features thus seem to be the measure of repetition $\frac{r}{R}$ and C_r; again this is further confirmed by the Wilcoxon signed rank test reported above for which these were the only two features that showed a significant correlation with corpus type.

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 12 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 we 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 here:

$$C_r = \frac{N_d}{N_u} + a\frac{S_d}{T_d} \tag{7}$$

where, again, N_d is the number of bigram types, N_u the number of unigram types, S_d is the number of bigram hapax legomena, and T_d is the total number of bigram tokens. Now, Lee et al. pad their texts with beginning and end of text delimiters (as the OpenGrm software we use does implicitly). For corpora consisting of shorter texts, this means that bigrams that include 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, on the other hand, 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 somewhat correlated, but negatively, and not so strongly: $r = -0.29$. 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 Lee and colleagues' (*3*) decision tree for classifying the type of linguistic system: the lowest values of U_r are segmental systems, next syllabaries, next "logographic" systems. This makes sense in light of the (weak) 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, non-linguistic 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 non-linguistic 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.[11]

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

[11] We have already seen that repetition on its own misclassifies Sumerian as non-linguistic, due to the short "texts" — an artifact of the way we divided the corpus. Given what we have just discussed, we would therefore expect that the decision trees would also misclassify it. Indeed they do: using all features, 92 trees classified it as non-linguistic (accuracy 0.72) and 8 trees as linguistic (0.23).

rather simple measure: how long on average are the extant texts in the symbol system? If they are short, then they are probably non-linguistic; if they are long they are probably linguistic.

5 Conclusions

This work has used a set of corpora of non-linguistic symbol systems that we have developed and compared them using a variety of statistical measures to a large set of corpora of written language. The non-linguistic corpus set was designed to include types of systems that are likely to be relevant to assessing whether ancient systems such as the Indus Valley signs, or Pictish symbols, are linguistic or not. I also included one system (barn stars) that is almost certainly purely decorative (thus addressing a common misconception about the claims of (*49*) to the effect that we were claiming the Indus symbols were decorative); as well as some modern systems that can easily be harvested from the Web. Finally, I picked a wide variety of linguistic symbol systems, ranging across both time and script type.

I have examined a variety of statistical measures and shown that two of them — Lee and colleagues' (*3*) C_r and the repetition measure $\frac{r}{R}$ — are useful for distinguishing linguistic and non-linguistic systems. However, the fact that they are useful seems to be in part because both of them are also correlated with a measure that is far more basic and trivial: mean text length.

In both depth and breadth of coverage, this work arguably goes far beyond previous work in this area, but the apparent conclusion — that a core distinction between linguistic and at least some non-linguistic systems — comes down to text length, seems troublesome. Could the

answer be that trivial?

I believe the answer to that question is yes, for a simple reason: if you have a real writing system, it allows you to write anything you want to say, meaning that you can generate very long texts. Non-linguistic systems can be expected to be much more varied in this regard, depending upon what kind of information they represent. Mathematical formulae can, indeed, be quite long, as can deity symbol strings on *kudurru* stones. However Pictish "texts" are all very short, presumably because whatever kind of information was represented by those "texts" was by its nature concise. Totem pole texts are short for similar reasons (and probably also because of the labor involved in carving whole tree trunks). These characterizations are informal and somewhat circular to be sure, but the point is that with non-linguistic systems, unlike linguistic systems, there is no *a priori* expectation that the texts should be long.

One of the main problems all along for the theory that the Indus Valley symbols were a written language has been the extremely short length of the "texts". While few Indus researchers would perhaps admit this, the brevity of the inscriptions is an embarrassment. Why else would researchers since Marshall (*62*) have appealed to a "lost manuscript" hypothesis if it were not the sense that a symbol system that was only used for very short inscriptions did not look much like a writing system? The "lost manuscript" hypothesis and its ramifications were discussed extensively in (*49*), and those arguments will not be repeated here. But one of the conclusions of that work was that one discovery that would be most needed to support the script hypothesis would be a genuine long text in the symbols. This claim is supported by the statistical results we have obtained here.

At its core this just seems like common sense, and there is a clear analogy in child 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 (*63*), 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.

And so it is with symbol systems. If a symbol system is a true writing system, then it is able to represent anything that can be said in the language of its users. Language is in principle unbounded in that there are no limitations, apart from obvious physical ones, on how long a linguistic message can go on. Linguistic texts can range from short texts such as a couple of names to lengthy stories that can extend into the hundreds of thousands of words or more. Texts in a true writing system should thus show a wide range of lengths *because of what it is they encode*. Indeed, in societies that make extensive use of writing, written texts can often be much longer than any given spoken message, precisely because their creation is not limited by the same factors, such as breathing or fatigue, that limit spoken utterances. If the Indus Valley system was a full writing system as some have claimed, it would be quite anomalous if the Indus Valley scribes had not figured out, over a span of roughly 700 years, that they could do a bit more with the system than write short cryptic texts (*49*).

Needless to say, text length, while a plausible predictor of symbol-system status, is nonetheless a crude feature. Non-linguistic systems *can* still have relatively long texts, as in some of the corpora we have collected. And while our useful features seem to reduce ultimately to text length, it is not out of the question that in future work we will discover other features that are useful in classifying symbol systems, and that are not correlated with length.

For contentious systems — in particular the Indus Valley symbols — we do not expect the debate to end here. For the Indus Valley symbols in particular, too many people have a lot invested in the idea that the symbols were writing to give that idea up. I hope however that I have provided an analysis of statistical approaches to the question of symbol system type that is better informed than what has been prominently displayed in the past few years.

Acknowledgments

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 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 (*46*).

I also thank audiences at the Linguistic Society of America 2012 annual meeting in Port-

land, the Center for Spoken Language Understanding, the Oriental Institute at the University of Chicago, Kings College London, Johns Hopkins University and the University of Maryland, 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.

I would like to acknowledge the staff of the Berks County Historical Society, 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.

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.

References

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

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

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

4. I. Gelb, *A Study of Writing: The Foundations of Grammatology* (University of Chicago, Chicago, 1952).

5. G. Sampson, *Writing Systems* (Stanford University Press, Stanford, CA, 1985).

6. J. DeFrancis, *Visible Speech: The Diverse Oneness of Writing Systems* (University of Hawaii Press, Honolulu, HI, 1989).

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

8. R. Sproat, *A Computational Theory of Writing Systems* (Cambridge University Press, Cambridge, 2000).

9. H. Rogers, *Writing Systems: A Linguistic Approach* (Blackwell, Malden, MA, 2005).

10. R. Harris, *Signs of Writing* (Routledge, London, 1995).

11. R. Sproat, *Non-Linguistic Symbol Systems: A Taxonomy and Analysis of their Statistical Properties* (TBD, 2014).

12. R. Lee, P. Jonathan, P. Ziman, *Computational Linguistics* **36**, 791 (2010).

13. R. Sproat, *Computational Linguistics* **36** (2010).

14. B. Powell, *Writing: Theory and History of the Technology of Civilization* (Wiley-Blackwell, Chichester, 2009).

15. C. Waddington, *American Ethnologist* **1**, 471 (1974).

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

17. D. Kaiser, *American Scientist* **93**, 156 (2005).

18. N. Le Novère, et al, *Nature Biotechnology* **27**, 735 (2009).

19. K. Basso, N. Anderson, *Science* **180** (1973).

20. J. McCawley, *The World's Writing Systems*, P. Daniels, W. Bright, eds. (Oxford, New York, 1996), pp. 847–854.

21. B. Farnell, *The World's Writing Systems*, P. Daniels, W. Bright, eds. (Oxford, New York, 1996), pp. 855–879.

22. G. Mallery, *Pictographs of the North American Indians: A Preliminary Paper*, no. 4 in Annual Report of the Bureau of Ethnology (Smithsonian Institution, Washington, DC, 1883). Available from Google Books.

23. L. Li, *Naxizu xiangxing biaoyin wenzi zidian* (Yunnan Minzu Chubanshe, 2001).

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

25. S. Bedrick, R. Beckley, B. Roark, R. Sproat, *Workshop on Language and Social Media* (Montreal, Canada, 2012).

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

27. S. M. Winn, *Pre-Writing in Southeastern Europe: The Sign System of the Vinča Culture, ca. 4000 B.C* (Western Publishers, 1981).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

46. K. Wu, J. Solman, R. Linehan, R. Sproat, *Linguistic Society of America* (Portland, OR, 2012).

47. J. Rhys, *Proceedings of the Society of Antiquaries of Scotland* **26**, 263 (1892).

48. R. Rao, *et al.*, *Proceedings of the National Academy of Sciences* **106**, 13685 (2009).

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

50. M. Vidale, *East and West* **57**, 333 (2007).

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

52. M. Ventris, J. Chadwick, *Documents in Mycenaean Greek* (Cambridge University Press, Cambridge, 1956).

53. B. Roark, *et al.*, *Proceedings of the Association for Computational Linguistics, System Demonstrations* (Association for Computational Linguistics, Jeju Island, South Korea, 2012), pp. 61–66.

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

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

56. R. Sproat, *Computational Linguistics* **36**, 807 (2010).

57. R. Kneser, H. Ney, *International Conference on Speech and Signal Processing* (1995), pp. 181–184.

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

59. K. W. Church, W. Gale, *Computer Speech and Language* **5**, 19 (1991).

60. C. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA, 1999).

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

62. J. Marshall, *Illustrated London News* pp. 528–32, 548 (1924).

63. H. Scarborough, *Applied Psycholinguistics* **11**, 1 (1990).

Figure 1: 6-pointed star (left) versus rosette.

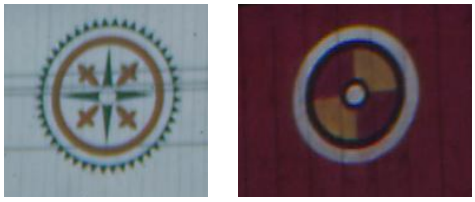

Figure 2: 4-pointed star (left) versus 4-slice pie.

Figure 3: A sample weather icon sequence: forecast for Portland, Oregon, April 29, 2011.

|^_^|	(๑˘•ټ˘•๑)
[○_−]	(˘¦-˘¦ʅ)
\(^v^)/	(*≧m≦*)

Figure 4: Some representative *kaomoji* emoticons

Corpus	# Texts	# Tokens	# Types	Mean text length
Asian emoticons	10,000	59,186	334	5.9
Barn stars	310	963	32	3.1
Mesopotamian deity symbols	69	939	64	13.6
Pictish stones	283	984	104	3.5
Totem poles	325	1,798	477	5.5
Vinča	591	804	185	1.4
Weather icons	10,142	50,710	16	5.0

Table 1: Number of texts, type and token counts, and mean text length for the non-linguistic corpora.

Corpus	# Texts	# Tokens	# Types	Mean text length
Amharic	111	17,747	219	159.9
Arabic	10,000	429,463	62	42.9
Chinese	10,000	136,246	2,738	13.6
Ancient Chinese	22,359	91,010	701	4.1
Egyptian	1,259	38,804	691	30.8
English	10,000	503,309	76	50.3
Hindi	1,000	61,254	77	61.2
Korean	10,000	223,869	984	22.4
Korean (Jamo)	10,000	438,440	102	43.8
Linear B	439	5,465	220	12.4
Malayalam	937	46,497	80	49.6
Oriya	1,000	40,242	75	40.2
Sumerian	11,528	50,318	692	4.4
Tamil	1,000	38,455	61	38.5

Table 2: Linguistic corpora for comparison.

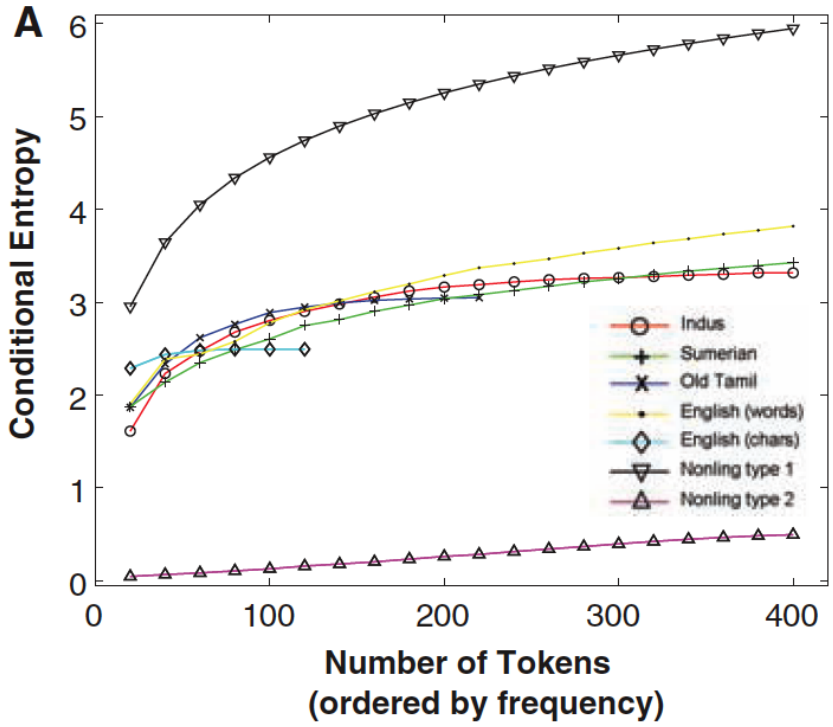

Figure 5: Bigram conditional entropy growth curves of various r linguistic symbol systems, the Indus Valley symbols, and two artificial models of non-linguistic systems, from (*1*). See the text for further explanation. Used with permission of the American Association for the Advancement of Science.

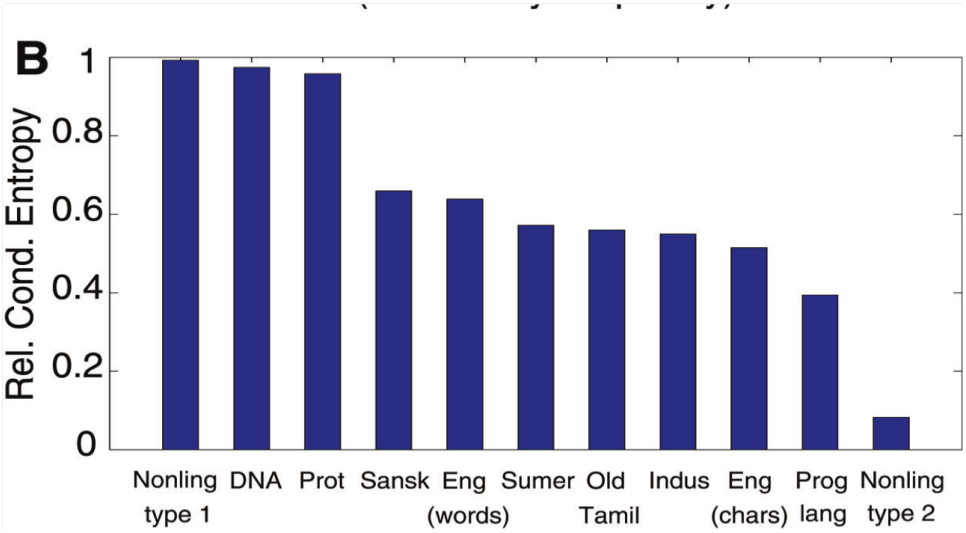

Figure 6: Relative conditional entropy (conditional entropy relative to a uniformly random sequence with the same number of tokens) of various systems from (*1*). Used with permission of the American Association for the Advancement of Science.

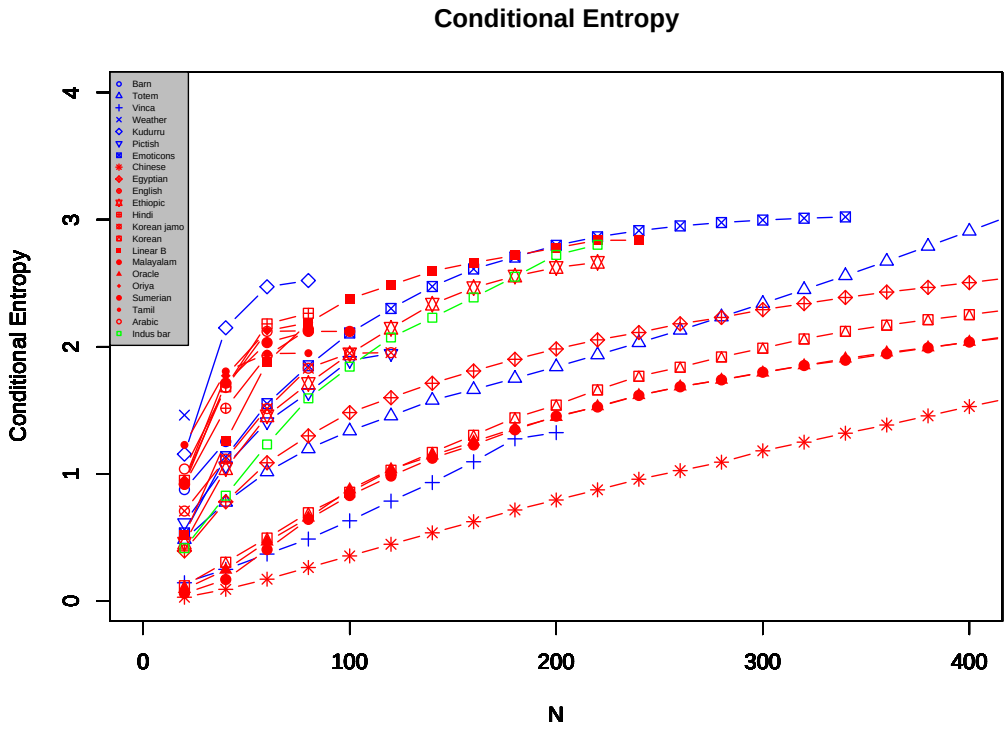

Figure 7: Bigram conditional entropy growth curves computed for our linguistic (red) and non-linguistic (blue) corpora, and for Indus bar seals (green).

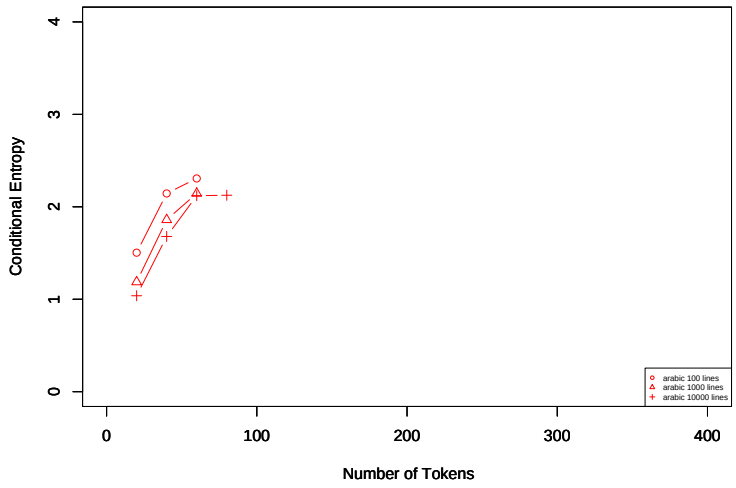

Figure 8: Bigram conditional entropy growth curves for three different sized Arabic corpora.

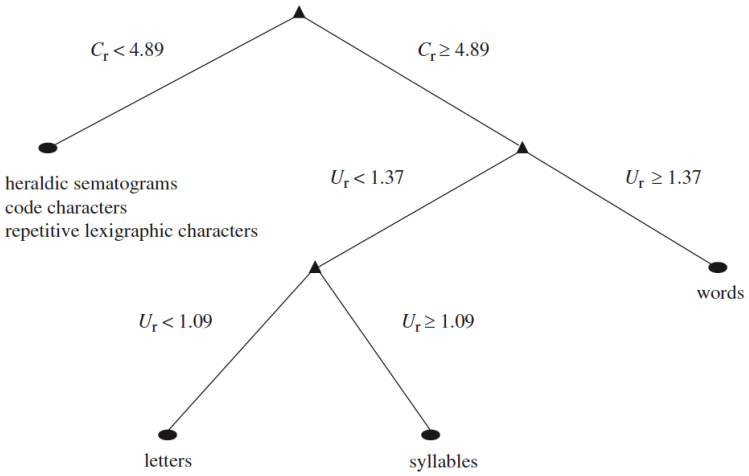

Figure 9: Reproduction of Figure 6, page 9, from: (*3*).

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	non-linguistic
Weather icon sequences	linguistic: letters

Table 3: The results of applying Lee et al.'s (*3*) decision tree from Figure 9 to our data.

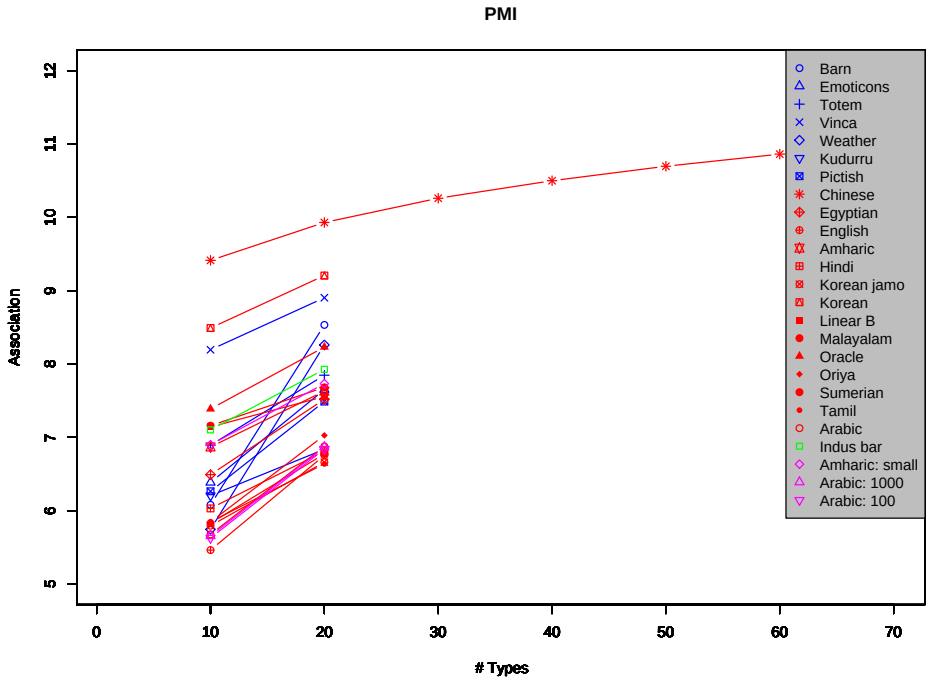

Figure 10: PMI association computations for our corpora, computed over subsets of each corpus starting with the 10 most frequent symbols, the 20 most frequent, and so forth, up to 25% of the corpus. See the text for more detailed explanation.

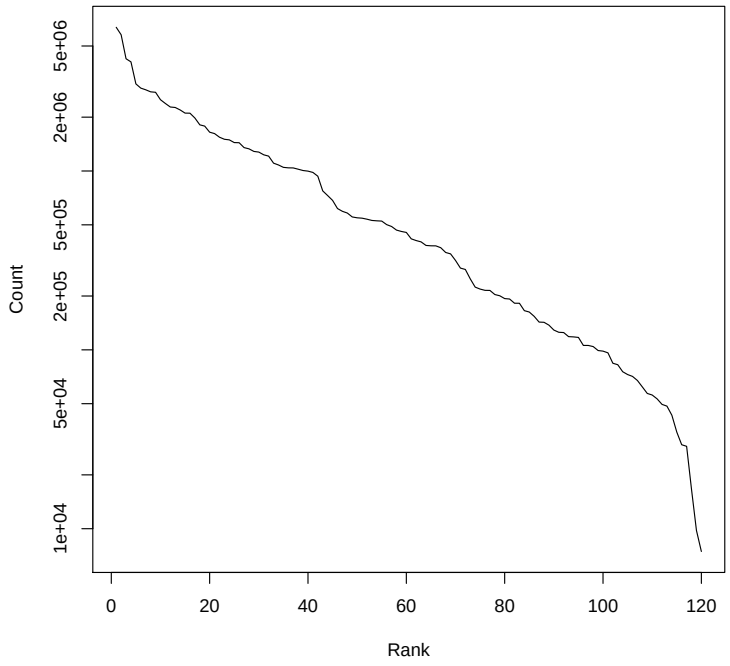

Figure 11: Unigram distribution of awarded boyscout merit badges, from `http://meritbadge.org/wiki/index.php/Merit_Badges_Earned`, showing the standard power-law inverse log-linear relationship between count and rank.

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 4: Repetition rate $\frac{r}{R}$ for the various corpora.

1

Corpus	$\frac{r}{R}$
Barn Stars	0.85
Weather Icons	0.80
Indus bar seals	0.77
Totem Poles	0.71
Vinča	0.63
Egyptian	0.42
Sumerian	0.33
Chinese	0.31
Pictish	0.29
English	0.29
Mesopotamian diety symbols	0.287
Amharic	0.25
Asian emoticons	0.22
Linear B	0.21
Malayalam	0.19
Arabic	0.14
Korean jamo	0.08
Oriya	0.04
Korean	0.015

Hindi	0.015
Tamil	0.0040
Oracle Bones	0.0

Table 5: Repetition rate $\frac{r}{R}$ for versions of the corpora with artificially shortened texts of length 6 or less, and no more than 500 texts.

1

Corpus	Mean accuracy	# training runs
Asian Emoticons	0.85	79
Mesopotamian deity symbols	0.85	71
Sumerian	0.84	76
Egyptian	0.83	72
Weather icons	0.82	80
Pictish	0.82	79
Malayala	0.81	79
Hindi	0.80	78
Oriya	0.80	82
Korean jamo	0.80	77
Arabic	0.80	74
Linear B	0.79	70
English	0.79	76
Tamil	0.79	76
Amharic	0.79	73
Totem poles	0.79	86
Chinese	0.79	75
Korean	0.79	72
Oracle bones	0.78	69

Vinča	0.78	76
Barn stars	0.78	80

Table 6: Mean accuracy of results for each of the corpora when occurring in the training data.

[1]

Corpus	Mean accuracy	# training runs
Barn stars	0.91	20
Totem poles	0.89	14
Vinča	0.88	24
Oracle bones	0.85	31
Chinese	0.85	25
Korean	0.84	28
English	0.84	24
Tamil	0.84	24
Amharic	0.84	27
Linear B	0.83	30
Korean jamo	0.83	23
Arabic	0.82	26
Oriya	0.82	18
Hindi	0.81	22
Malayala	0.80	21
Pictish	0.75	21
Weather icons	0.74	20
Egyptian	0.74	28
Mesopotamian deity symbols	0.70	29

Sumerian	0.69	24
Asian Emoticons	0.64	21

Table 7: Mean accuracy of results for each of the corpora when occurring in the test data.

[1]

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 8: Features used by trees under the two training conditions. The left column in each case is the number of trees using the particular combination of features in the right 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.

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 9: Wilcoxon signed rank test for each feature with the two populations being linguistic and non-linguistic corpora. Only repetition and C_r show significance.

# 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 10: 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”.

# 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
1	block 1, repetition

Table 11: Features used by trees when repetition is replaced by the repetition rate computed over artificially shortened corpora (Table 5. The fourth row consists of twelve trees where there is a single leaf node with the decision “linguistic”.

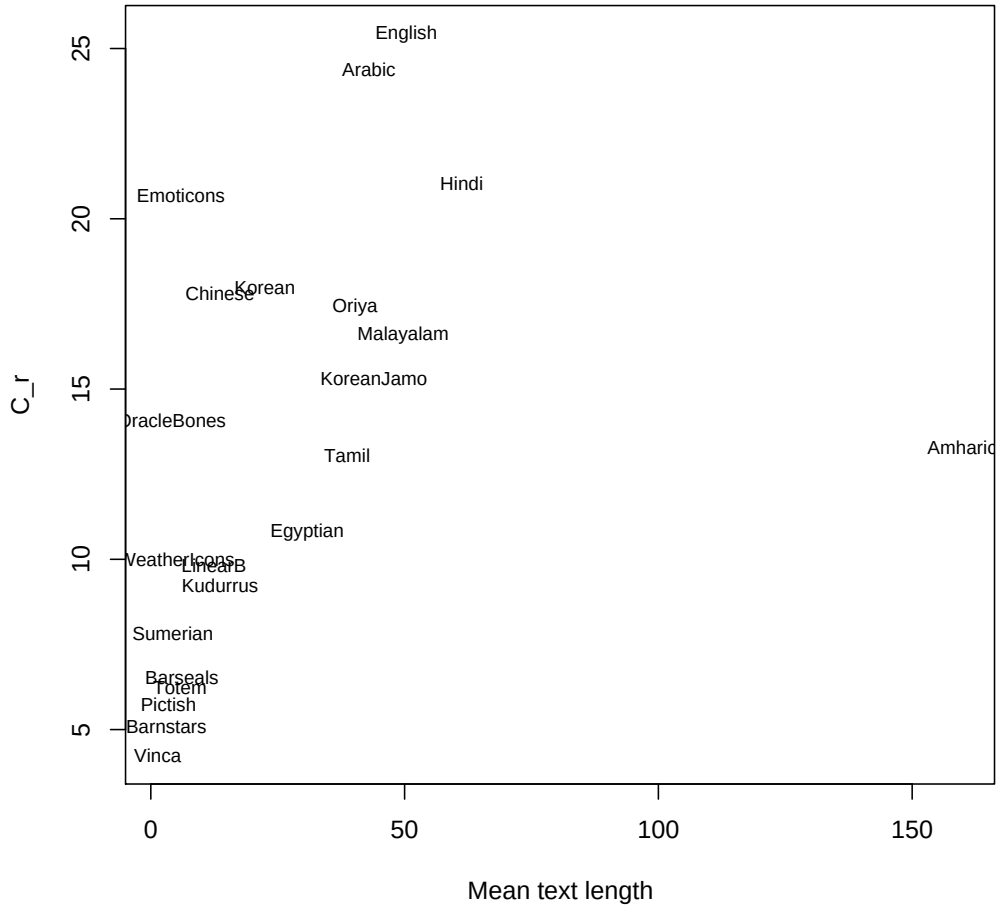

Figure 12: Correlation between C_r and text length. Pearson's r is 0.39, but when the obvious outlier Amharic is removed, the correlation jumps to 0.71.