

Corpora and Statistical Analysis of Non-Linguistic Symbol Systems

Richard Sproat

January 21, 2013

Abstract

We report on the creation and analysis of a set of corpora of *non-linguistic* symbol systems. The resource, the first of its kind, consists of data from seven systems, both ancient and modern, with two further systems under development, and several others planned. The systems represent a range of types, including heraldic systems, formal systems, and systems that are mostly or purely decorative. We also compare these systems statistically with a large set of linguistic systems, which also range over both time and type.

We show that none of the measures proposed in published work by Rao and colleagues (Rao et al., 2009a; Rao, 2010) or Lee and colleagues (Lee et al., 2010a) works. In particular, Rao’s entropic measures are evidently useless when one considers a wider range of examples of real non-linguistic symbol systems. And Lee’s measures, with the cutoff values they propose, misclassify nearly all of our non-linguistic systems. However, we also show that one of Lee’s measures, with different cutoff values, as well as another measure we develop here, do seem useful. We further demonstrate that they are useful largely because they are both highly correlated with a rather trivial feature: mean text length.

©2012–2013, Richard Sproat

1 Introduction

Humans have been using symbols for many millenia to represent many different kinds of information. In some cases a single symbol represents a single concept, and is not part of any larger symbol system: an example is the red, blue and white helical symbol for a barber shop. In other cases the symbols may be part of a more complex *combinatoric* system that has its own syntax and semantics. Examples are mathematical symbols, European heraldry (Slater, 2002), or the Mesopotamian symbols representing deities (Seidl, 1989), all of which are systems consisting of many symbols, which may be combined together into “texts” of varying degrees of complexity. Writing systems are just one kind of complex symbol system, where the system happens to represent language. In writing systems, the individual symbols of the *script* represent a linguistic unit — often a phoneme or a syllable, but in the case of many (especially ancient) writing systems, also other units such as morphemes, words, or semantic information. Fully developed writing systems — those capable of representing essentially anything that can be spoken — always represent at least some phonological information (DeFrancis, 1984, 1989), so that there are no true “logographic” or “semasiographic” writing systems.

Suppose an archaeologist discovers a clay tablet inscribed with heretofore unknown symbols among the ruins of an ancient civilization. As long as the symbols seem to be arranged more or less linearly (see below) and the inscription meets certain other properties, one’s gut feeling is often to assume that the system must have been writing.

Consider in this light the stone inscription in Figure 1. Here we see a set of symbols, arranged more or less in lines from top to bottom. The symbols repeat in some places. Was this some form of writing? In considering this question one might note that the symbols are highly pictographic: perhaps, then, they are not writing. On the other hand, we know of writing systems that were also highly pictographic: three examples are Egyptian, Luwian and Mayan (Figure 2).

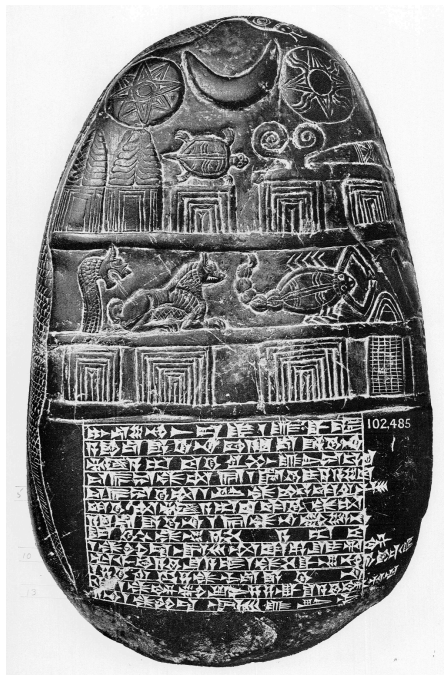


Figure 1: *Kudurru* of Gula-Eresh (King, 1912), Plate I.

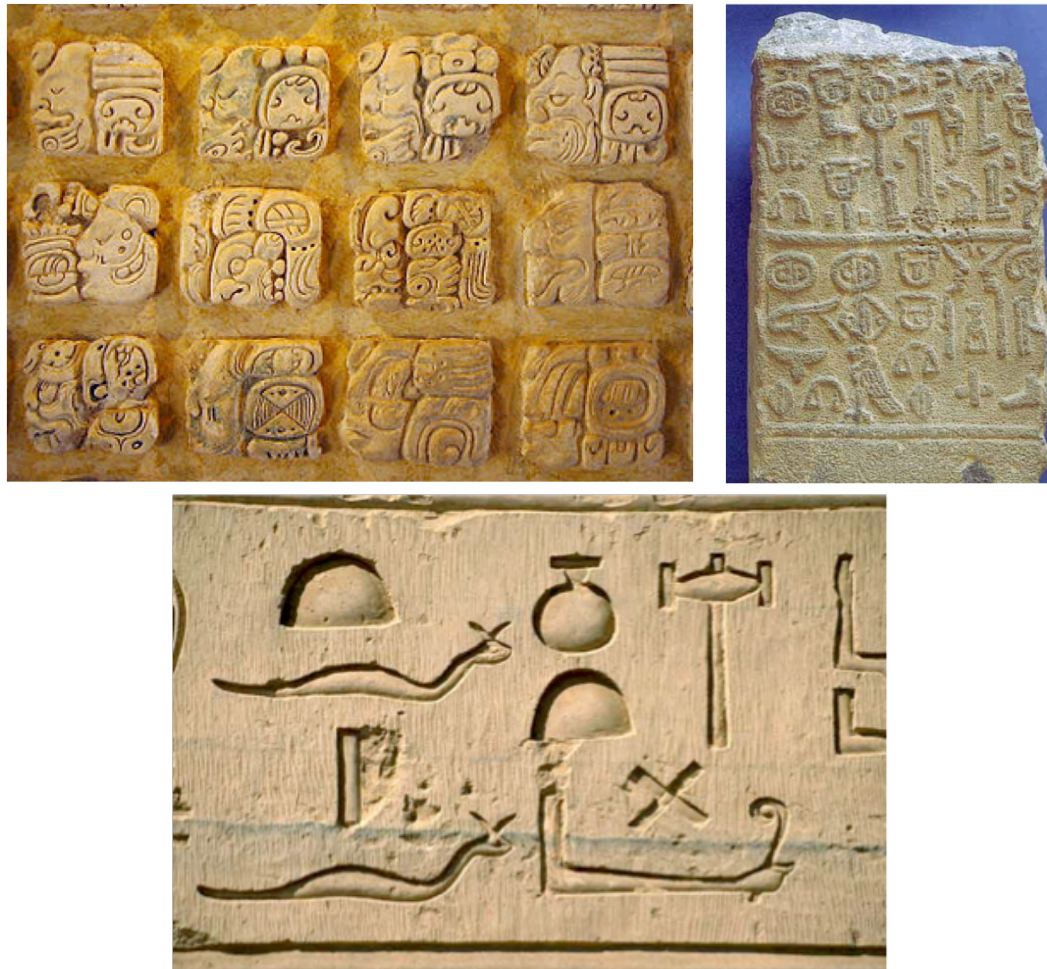


Figure 2: Examples of Mayan (http://en.wikipedia.org/wiki/File:Palenque_glyphs-edit1.jpg), Luwian (Hawkins, 2000), and Egyptian (multiple sources on the Web) pictographic scripts.



Figure 3: Two adjacent instances of the “horned crown” and “symbol base”, two highly associated symbols in the Mesopotamian deity symbol system.

So perhaps being pictographic is not particularly important in considering whether a symbol system is writing. Furthermore, the symbols in Figure 1 have other script-like properties. Note, for example, that some of the symbols are “ligatured” together, as with the HORNEDCROWN and SYMBOLBASE combinations in Figure 3.

In fact, in this case, we know that the symbol system was not writing. Rather, the symbols in question were deity symbols used (among other things) to mark *kudurru* boundary stones, legal documents that specified property rights. The stones, as in this example, also included actual cuneiform Babylonian writing. The deity symbols had an essentially heraldic function, to indicate favored deities of the owner of the property. In this particular case the culture in question was literate in a writing system that we can read, and it was through that written record that we were able to understand the function of the non-linguistic deity symbols. But suppose symbols like those in Figure 1 were the only symbols left behind by the culture? How then would one know whether one is dealing with writing or not?

The question is an important one. The first impulse when dealing with an unknown symbol system is to attempt to “decipher” it — i.e. decode the language underlying it and the way in which the symbols encode that language. But if the symbols do not encode language, attempts at decipherment are an exercise in futility. The most one can hope to do is to try to understand the symbols in some other way than as an encoding of language, though a priori this is a much harder task.

Thus we arrive at the following question: Is there a way to determine, on the basis of the properties of the symbols, and their distribution, whether the system is likely to be writing (and thus worth the effort of attempting decipherment), or not?

Some recent work has proposed that the answer to this question is “yes”. One paper (Rao et al., 2009a) used bigram conditional entropy to argue that the symbols used by the Indus Valley civilization constituted a writing system and not a non-linguistic symbol system as had been recently proposed in (Farmer et al., 2004). Another paper (Lee et al., 2010a) used a more sophisticated method, but one also based in part on conditional entropy, to argue that Pictish symbols, found on a few hundred standing stones in Scotland, were part of a heretofore unrecognized writing system. Both of these papers were widely reported in the popular science press, where there seemed to be a general consensus that new computational techniques had been discovered that can be used to unravel ancient mysteries. The only problem, as we have argued elsewhere (Sproat, 2010a), and as we shall in any case document much more extensively here, is that the techniques reported in those papers do not work.

A large part of the problem, is that whereas there is now a large amount of linguistic data around in the form of text corpora from languages both ancient and modern, there is a dearth of data from *non-linguistic* symbol systems. While (Rao

et al., 2009a) did have some real data from biological sequences and Fortran code, they compared the Indus and linguistic systems with no real ancient non-linguistic data, but instead used simulated data based on (mis)conceptions about how a couple of different ancient symbol systems worked. Lee et al. (2010a) included examples from European Heraldry (as we had previously in (Farmer et al., 2004)), but no other ancient or traditional non-linguistic systems. Both studies were limited in part because a set of corpora of non-linguistic symbol systems was simply not available. Yet without such corpora, one cannot hope to develop convincing statistical methods that might distinguish between linguistic and non-linguistic systems. And without such a basis for developing a rigorous approach to the problem, studies like those of Rao et al. (2009a) and Lee et al. (2010a), despite their obvious appeal to the popular science press, are effectively meaningless.

This study addresses this situation in two ways. First, we have developed electronic corpora from published (mostly print) sources for a variety of unequivocally non-linguistic symbol systems both from the ancient world, from later traditions, and from modern times. Second, we have used these corpora, in combination with a set of readily available linguistic corpora from a variety of ancient and modern languages, to begin an investigation of whether statistical methods might be developed that can distinguish between the two kinds of systems.

The outline of this monograph is as follows. First, in Section 2 we discuss general issues in the study of symbol systems, in particular the field of Semiotics, that are relevant (or not) to our study of non-linguistic systems. Then, in Section 3 we report on an informal survey on factors that are relevant to the judgment that a sign system “looks like” writing.

Section 4 proposes a first cut at a taxonomy of non-linguistic symbol systems, something that will be useful for understanding the types of corpora we have collected, and which will also be useful for future research.

In Section 5 we discuss non-linguistic corpora that we have developed, as well as others that are in progress or are planned. Section 6 describes the XML markup scheme used to encode the corpora. Then, in Section 7 we describe the linguistic corpora that we have used for comparison in our statistical tests.

Section 8 outlines the statistical methods we have investigated, including separate features, and combinations of those features into a decision tree classifier.

Finally, in Section 9 we summarize our conclusions and discuss future directions for this kind of research.

2 Semiotics

In any discussion of symbol systems a natural starting point might seem to be the field of semiotics (Peirce, 1934; Eco, 1976; Sebeok, 1977). Semiotics is, after all, “concerned with everything that can be taken as a sign” (Eco, 1976), and is in principle the theoretical discipline that deals with sign and symbol systems.

There are several reasons however for not adopting much if any semiotic theorizing in this work. First and foremost, the notion of *sign* as it has come to be used in semiotics is far broader than we need. For example, the entry for *sign* in Bouissac’s *Encyclopedia of Semiotics* tells us:

In scholarly writing, the term *sign* might include, for example, words, sentences, marks on paper that represent words or sentences, computer programs . . . pictures, ideograms, graphs, chemical and physical formulas, fingerprints, ideas, concepts, mental images, sensations, money, postures and gestures, manners and customs, costumes, rules and values, the orienting dance of the honeybee, avian display, fishing lures, DNA, objects made of other signs . . . and also nonrepresentational objects (perhaps in music or mathematics) that have types of structure characteristic of other signs. (Bouissac, 1998, page 572)

Insofar as we are interested here in the statistical distribution and denotation of sets of graphical signs, clearly we are discussing a far narrower concept than what Bouissac lays out above.

Second, semiotics provides no formal theory of the combination of signs in text, the *key* interest here. After all, what we are investigating is whether one can tell merely by looking at the distribution of symbols, what kinds of things they represent. Part of the problem, to put it bluntly, is that the field of semiotics was early on hijacked by deconstructionists, who are not particularly interested in formal mathematical models.¹

Finally we have little use for distinctions common in the semiotics between *symbols*, *indices*, *icons*, and so forth and just use the neutral term *symbol*.² Whether, for example, a given symbol is iconic for what it represents is largely orthogonal to the issues that we will be addressing.

¹Even non-deconstructionist semioticians, such as Eco, are not typically well versed in mathematical theories of information. Cf. the following (incorrect) definition by Umberto Eco: “[according to the mathematical theory of information] information is only the measure of the probability of an event within an equi-probable system” (Eco, 1976, page 42).

²Roughly, a *symbol* is an arbitrary sign denoting an object, an *icon* is similar to the object it represents, and indices are somehow physically connected to the object — e.g. a symptom of an underlying disease. Cf. (Eco, 1976, page 178), though it should be noted that Eco largely rejects this trichotomy.

In this work we will thus restrict ourselves to using the term *symbol*, where by symbol we mean *set* of man-made graphical forms that is itself a member of a set, termed a *symbol system*, that has a well-defined cultural function. Thus mathematical symbols are each a member of the mathematical symbol system. Boy scout merit badges are individually members of a set of symbols whose function has a well-defined function in scouting. Mesopotamian deity symbols were individually members of a set of symbols that had a recognized function of representing favored deities. And letters of the Roman script, as used for say English, are elements of the set of symbols used in the English writing system.

For familiar reasons, a symbol cannot in general be defined purely as a graphical form: “A” and “a” are both instances of the same letter <a>, as are multiple font and handwritten variants of these forms. That is why we defined a symbol as a *set* of forms above. This brings up the question of how we decide what is in each set, and thus which graphical forms are just variants of the same symbol. This is of course a very real problem when dealing with an unknown symbol system: which variations in form distinguish among separate symbols, and which do not? In this work we are fortunate to be dealing with symbol systems that in many cases are known. So, in Mesopotamian deity symbols there may be multiple variants of the GOATFISH symbol, but scholars of this system know that these are just variants of the same symbol. Similarly, in known writing systems, we know which variants to classify as being the same symbol. For symbol systems, such as the Vinča system or Pictish symbols, whose meanings are not known, we can at least make use of prior scholarship in determining which distinctions in graphical form are relevant and which not.

As a practical matter, if one knows the meaning of the symbols the equivalence classes can be determined by the denotation — the *signifié* in Saussure’s (1916) terms, or the *object* in Peirce’s (1934) terms. Of course in many symbol systems, including linguistic symbol systems (writing), symbols may be multivocal. Thus, in the largely segmental writing system of English <a> represents far more than just one vowel, and the digraph <th> represents at least two phonemes (/θ/, /ð/). In systems where the meaning is not known — including many ancient non-linguistic systems, and any undeciphered script — distributional evidence is often used to make guesses about equivalence classes.

A special case is symbol systems where the symbols may not actually denote anything: as we shall discuss in Section 5.5, this may be true of Pennsylvania Dutch barn stars, where scholarship as well as local tradition weighs in on the side of the symbols being purely decorative. In such cases, does it even make sense to talk of a symbol system? We do not attempt to decide this issue here: rather, we will adopt the conventions of previous scholarship in deciding what the variants of a given symbol are, and we will simply sidestep the question of what the symbols

may or may not be represented. Ultimately we are just interested in understanding how the symbols, whatever they may be, are arranged in text.

3 What does Writing “Look Like”?

Can one tell just by looking at a text in an unknown symbol system if it was writing or not? What does writing look like?

In a recent (January 5, 2012) discussion on the Ancient Near East (Yahoo!) discussion forum, writing-systems scholar Peter Daniels, in discussing some controversial instances of early writing supposedly excavated at Jiroft, proclaimed: “Well, it looks like writing.” He then goes on to note that “four small samples of this script are now known.”³ Daniels, apparently, is willing to countenance something as writing on the basis of just four small samples because it “looks like” writing.

But what does it mean to say something “looks like” writing? What particular features should a symbol system have to “look like” writing? With only four small samples of a system, there is little hope the system could be deciphered — unless of course many more samples are found.

In order to elucidate these questions I conducted a small informal survey using Survey Monkey (<http://www.surveymonkey.com>). In the survey, participants were asked to rate, on a scale of 1 to 4 — very unimportant, to very important — the following factors in making something “look like” a writing system:

1. How important is the linear arrangement of symbols in assessing whether a symbol system is a writing system?
2. If the symbols are pictographic, how important is it that they be abstract rather than realistic?
3. How important is the length of the “texts” to a judgment of whether a symbol system is writing?
4. In order to judge something to “look like” writing, how important is it that symbols repeat?
5. How important is the number of distinct texts to a judgment that a symbol system is writing?

Respondents were also asked to give a bit of information about themselves, and in particular what writing systems they were familiar with, and which ones they considered themselves an expert on. Thirty eight people responded. Based on their self-reported levels of expertise, these were classified into three bins as follows:

³<http://groups.yahoo.com/group/ANE-2/message/14005>.

Repetition rate	Very important	18	3.29
Linearity	Very important	16	3.24
Number of distinct texts	Somewhat important	18	3.00
Text length	Somewhat important	16	2.92
Abstractness	Somewhat unimportant	17	2.21

Table 1: Results of informal survey: modes, number of respondents (out of 38) choosing the mode, and means for each of the survey questions.

- Group 0: largely unfamiliar with writing systems (5 respondents)
- Group 1: familiar with and/or expert on several writing systems (22 respondents)
- Group 2: familiar with and/or expert on several writing systems from a variety of writing system types (11 respondents)

The results for the main questions were for the most part unsurprising insofar as they accorded with my own judgments. The modes, number of respondents (out of 38) choosing the mode, and the means for each of these questions were as in Table 1.

Thus it was considered to be very important that symbols be arranged largely linearly, and that symbols should repeat in texts. The length and number of texts was considered to be a somewhat important factor. And abstractness of symbols in pictographic systems was considered to be somewhat unimportant. On the latter, presumably enough people are familiar with Egyptian and other highly pictographic scripts to know that a system can look like pretty pictures, yet still be writing.

There was no significant interaction between expertise level and the five main questions in the survey. The largest difference between group 2 (expert) and groups 1 and 0 (minimal expertise), was on question 3, where members of group 2, perhaps counterintuitively, considered length of texts a somewhat *less* important factor than members of the other groups.

In addition, I asked respondents for other suggestions on other features that might be relevant to deciding if a symbol system looks like a writing system. I quote directly some of the suggestions below. First, there were a few suggestions that relate roughly to the statistical distribution of symbols:

- How important is it that symbols are made on an object such that they could not have been made randomly (i.e., as an expression of a thought rather than meaningless scribble)?

- Statistical distribution of signs should accord with statistical distribution of graphemes, syllables, or words in attested languages.
- The existence of random combinations of the symbols. If we are lucky to have a long enough texts, or many texts, the existence of random combinations that repeat themselves.
- The number of variations of symbols. How many are there? Does it correspond at all to the number of sounds in a given language?

Then there were some suggestions of the importance of provenance:

- Provenance, resemblance to known systems.
- Archaeological setting. No provenance will tend to make me highly suspicious.

Finally, one suggestion had to do with the degree to which cursive styles have developed, suggesting a long tradition of use:

- Degree of cursivity; a fully evolved system should show signs of use, and therefore influence of the writing instruments, medium, and human dexterity.

The latter suggestion is of particular interest given that one of the arguments put forward by Farmer, Sproat and Witzel (2004) for the non-linguistic status for the Indus inscriptions was that over 700 years, there was no evidence of the development of a cursive style.

The above survey and discussion is hardly scientific, nor is it intended to be. No serious scholar believes that one can tell just by looking at a system whether it is writing or not and misjudgments can go either way: Ignace Gelb, the “father” of the study of writing systems famously misclassified Mayan writing as a “limited” preliterate system (Gelb, 1952). But it is nonetheless interesting to see what people believe are necessary characteristics of a symbol system in order for it to be considered writing. As we shall see in any case, all of the characteristics that were considered to be at least somewhat important for writing systems to exhibit, also show up in non-linguistic systems.

4 A Brief Taxonomy of Non-Linguistic Symbol Systems

What is writing then? The way most grammatologists define it, writing is a system of symbols for representing linguistic information. That linguistic information is usually some sort of sound information — either segmental phonemes, or syllables, or perhaps some other unit. Some writing systems, notably many ancient systems such as Sumerian, Egyptian, Mayan but also systems such as Chinese which are still used today encode other information in addition to phonological information: in Chinese, for example characters usually give some clue to the pronunciation but also contain semantic information about the morpheme that the character represents. Still, there is a strong case to be made that *all* fully developed writing systems represent sound information to a fairly sizable degree, and *no* full writing system can get by by purely representing some more abstract linguistic representation such as meaning (DeFrancis, 1989). But no matter: what is clear is that all writing systems represent some form of *linguistic* information. Simply put, they are linguistic symbol systems.

What about the many *non-linguistic* systems? What kind of information do they represent? Whereas writing is narrowly circumscribed by the fact that, whatever it represents, it must be some aspect of language, there is no such constraint on non-linguistic systems. In principle the signs in a system could represent anything that the creators of the system want them to represent. It seems that we need a way of classifying non-linguistic systems, and in this section I present a preliminary taxonomy.

To my knowledge, systematic taxonomies of non-linguistic systems do not exist. Certainly such systems are discussed. In (Daniels and Bright, 1996), a few non-linguistic systems are presented at the end of the book: numerical notation, music notation, movement notation systems. Such short shrift in that book is reasonable, since the main theme is writing systems and, as Daniels puts it (page 785), these other systems are on “the sidelines of the pageant of writing”. Harris 1995 also discusses a few cases of what he terms “non-glottic writing”: mathematics, knitting patterns, dance notation and music. But, again, there is no attempt at a systematic classification of non-linguistic systems.

The following discussion is intended to attempt to fill this void. I start out, however, with a discussion multivocality — the fact that symbols can often represent more than one thing, sometimes simultaneously.

4.1 Multivocality of symbols

Part of the problem in developing a taxonomy of symbol systems is the fact that symbols can often be *multivocal* in that they frequently acquire a wide variety of

meanings.

Multivocality is rampant with non-linguistic symbols. For example, Turner (1967) discusses the Ndembu “Mystery of the Three Rivers” thus:

This mystery (*mpang’u*) is exhibited at circumcision and funerary cult association rites. Three trenches are dug in a consecrated site and filled respectively with white, red, and black water. These “rivers” are said to “flow from Nzambi,” the High God. The instructors tell the neophytes, partly in riddling songs and partly in direct terms, what each river signifies. Each “river” is a multivocal symbol with a fan of referents ranging from life values, ethical ideas, and social norms, to grossly physiological processes and phenomena. (Turner, 1967, page 107)

Or as Farmer et al. (2004) point out, writing of possible interpretations of the Indus Valley symbols:

But it would be naive to assume that the same signs were interpreted at all times in the same way, any more than this was true of the Christian cross in medieval Europe — which in different contexts could serve as a political-military symbol, a magical talisman, a sign of a profession, a mystical aid, or a symbol of death. (Farmer et al., 2004, page 43)

But it is important to realize that such multivocality is by no means unique to non-linguistic systems. We already noted above (Section 2), for example, that <a> in English is multivocal, representing several different vowels depending on the context. For Sumerian, as we shall discuss again later on (Section 7.13), multivocality was the norm, with individual cuneiform symbols taking on a variety of different phonological or logographic denotations. Thus the symbol 𒀭 could represent the sounds (or morphemes) *aya*₂, *duru*₅, *eš*₁₀, among a variety of other functions. Such systematic multivocality was typical for ancient writing systems, but one sees extreme cases also in modern writing systems such as Japanese, where a given Chinese character (*kanji*) often has upwards of six pronunciations (Smith, 1996).

And that’s just when one considers the *linguistic* functions of a symbol. Often, though, linguistic symbols can take on other, non-linguistic, functions. Consider just the letter <a>, and its capital form <A>. In addition to its linguistic functions it can also denote, the first in a series, the top grade in an academic scoring system, a musical note, an adulteress in the context of a particular Hawthorne novel, the 8th Avenue Express in the New York subway system, or *acceleration* in mechanics. For many other uses see http://en.wikipedia.org/wiki/A_

(disambiguation). To be sure, many of these uses derive from the linguistic use of <a> (the words *acceleration* and *adulteress* begin with <a>), and others from the purely grammatological fact that <a> is first in the alphabet (thus explaining, the letter grade, or the musical note). But however the extension occurs, once so extended the symbol has become largely separated from its linguistic function. The recipient of an “A” in a class is unlikely to think of the linguistic uses of that symbol to reflect a range of English vowels.

Or consider the description of the Hebrew letter *he* in the fanciful account of kabbalist Helmont (1667). Helmont’s theory is that the Hebrew letters represent the shapes the tongue takes on when making the sounds associated with the letters, but there are other meanings as well:

When, therefore, the mouth opens, it must close again when it seeks rest; and then the tongue rises perceptibly, as is suitable for producing of the next letter. The name *He* is a definite article, meaning “this” or “that.” A certain mystical meaning concerning generation seems to be hidden in this letter, for all animals produce this sound when panting from the heat of lust. And for this reason it is probable that a *He*, but no other letters, was added to the names of Abraham and Sarah because many people were descended from them. (Translation by Coudert and Corse, pages 115–117.)

Thus we have *he* as the symbol for a sound (/h/), but in addition as the symbol for a morpheme (the definite article, written with *he*), as well as a symbol with the deeper meaning related to generation.

The main difference between cases like <A> and cases like the Ndembu Three Rivers mystery or (probably) Helmont’s *he*, is that in the former case, while the symbol can have multiple meanings, it is unusual for the symbol to have those meanings *all at the same time*. Thus again, the academic grade “A” is not contextually ambiguous with some other use, such as its use to represent the vowels /eɪ/ or /ə/; the “A” on a New York subway sign does not simultaneously carry the meaning “adulteress”. But as Turner makes clear, in the Ndembu case, the rivers simultaneously take on a variety of meanings — indeed that simultaneity of multiple meanings is crucial for the function of those signs in the system. For <a>, not only are the other uses of the symbol not crucial: they are completely irrelevant.

Thus when it comes to multivocality, one is tempted to classify symbols along two dimensions. The first is how multivocal they are: a barber pole (see below) is not notably multivocal in modern times, having essentially one basic function (to advertise the presence of a barber shop); a Christian cross, as noted above, is highly multivocal. Depending on the application domain of the symbol — the type of thing(s) it denotes — there may be more or less ambiguity. Restricting

ourselves to its purely linguistic functions, <a> is far less multivocal in a highly phonemic writing system like Finnish than it is in English: writing systems that by design are highly regular impose strict constraints on how symbols are used and thus the opportunity for multivocality is minimized. For non-linguistic systems, highly formal systems will, again, often impose strict constraints. It is strictly defined, for example, what the symbol “ \forall ” denotes in logic or mathematics. On the other hand, in religious symbology, a high degree of multivocality can be desirable.

The second dimension is how readily the symbol can take on its various meanings simultaneously.⁴ Presumably this relates to the particular domain of application of the symbol in question. Most of the functions of the letter <a> are rather mundane, either denoting particular phonemes, or being used as labels or symbols in mathematical equations. Religious or ritualistic symbols, on the other hand, often are used in contexts where the mystery of the rite is enhanced by the simultaneous appeal to multiple meanings.

Symbols of a variety of types can thus be both highly multivocal, or not; and can allow for simultaneous interpretation of the several meanings, or not, depending upon the kinds of things they denote and their function.

With that (ultimately rather obvious) point established, we now turn to a taxonomy of symbol system types, starting with the simplest type where symbols usually occur alone rather than in “texts” (or if they do, where there is no particular relation between the symbols in the “text”), and where the denotation of the system is usually pretty straightforward and unambiguous.

4.2 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. The symbols in such systems may occur in combination. Thus, in five-day weather forecasts such as one finds on sites like weatherunderground.com, one will see a “text” consisting of five weather icons, representing the predicted weather for the next five days. But though the symbols may be concatenated together, there is no real syntax in the “messages”, other than the trivial syntactic operation of

⁴Note that we are discussing *conscious* awareness of the multiple meanings. It is quite possible, indeed likely, that the various neural representations of what <a> represents are triggered simultaneously by a visual presentation of <a>. This is certainly true with ambiguous linguistic terms and it is possible in experimental conditions to show that the “insect” interpretation of *bug* is primed even when the context clearly calls for the “surveillance device” interpretation (Swinney, 1979); but subjects in such experiments are not generally consciously aware of having both meanings active.

concatenation.

Besides barber poles and weather icons, some further simple informative systems include: 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 (Waddington, 1974) or house marks (http://en.wikipedia.org/wiki/House_mark).

Symbols in simple informative systems may be “etymologically” complex in that a number of symbols were combined into what eventually became a single symbol. Thus for the barber pole, the red helix represents blood (barbers used to perform operations as well as cut hair), white represents bandages, the origin of the blue helix being less clear (see http://en.wikipedia.org/wiki/Barber's_pole). Thus one might characterize the barber pole as being multi-vocal, though it is a safe bet that most people who see this symbol today will not be aware of the origin of the symbol and will merely see it as a sign for a barber shop. In similar ways, symbols that have come to be used to mark ownership of horses in Mongolia, have complex historical origins (Waddington, 1974).

4.3 Emblematic systems

Closely related to simple informative systems are emblematic systems where the symbols represent some special distinction earned by the bearer. Examples include symbols of military rank and distinction, and scouting merit badges. As with other simple informative systems, emblematic symbols may occur in “texts”. Thus boy scout merit badges are typically linearly arranged on a sash (<http://www.scoutinsignia.com/sash.htm>). And as with other simple informative systems, there is no syntax in the system, other than the simple syntax of concatenation. The main reason for separating off emblematic systems from other simple informative systems is their sociological function: such systems are an institutionalized form of bragging right, indicating that the bearer has achieved certain marks of distinction. The function is thus rather different from the more mundane function of other simple informative systems.

Some other emblematic systems besides military rank and scouting merit badges: Phi Beta Kappa keys, and symbols of other scholarly fraternities; letter grades on academic assignments or in courses; designations of scholarly status (*BA*, *MA*, *PhD*), or other honors (*OBE*, *Kt* for “knight bachelor”), which use letters but have in effect become almost atomic symbols of status.

4.4 Religious Iconography

Religious iconography could also be characterized as a simple informative system, except that here though the notion of “single piece of information” is much less clear. The Christian crucifix, for example, represents the Christian faith, and is used to mark Christian houses of worship as such, but it also retains its original meaning (a symbol of Christ’s execution), as well as other meanings (e.g. as a symbol of death). It is, by nature, highly multivocal, a necessary result of its central function as a religious symbol, as we discussed above.⁵

Some other obvious symbols in this category: Star of David (Judaism), Star and Crescent (Islam), Dharmachakra (Buddhism), Swastika (Buddhism), the “Om” symbol (Hinduism, though this is itself derived from linguistic scripts).

Related to this category, though possibly deserving their own category, are magic symbols, which hide hidden meanings (Goodman, 1989). Horoscopal signs, insofar as they generally invoke a range of “powers”, and form part of a usually complex system of lore that relates one symbol to another, can also be seen as falling into this category.

4.5 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. The classic example of a heraldic system is European Heraldry (Friar and Ferguson, 1993; Fox-Davies, 2000; Slater, 2002), where a family’s coat of arms can be built up from a large set of symbols and tinctures. Rules specify how the arms may be laid out, what tinctures (colors, metals, furs) may occur superimposed on each other, and where certain emblems of distinction should be placed.

The deity symbols of Mesopotamian boundary stones (kudurrus) Seidl (1989) are another example of a heraldic system, albeit a less elaborate one than is found in European heraldry. In this case the deity symbols chosen are those that represent the favored deities of the person whose property deed is described on the stone, but the symbols also “invoke[d] divine protection upon private property and the rights of private individuals” (King, 1912, page xi). The syntax of kudurru deity symbol sequences — which can include texts of some tens of symbols long — is

⁵A Christian fish as prominently displayed on bumper stickers also has a multiple meanings including representing Christ as the “fisher of men”, as evocative of the Greek word ΙΧΘΥΣ, literally “fish”, but also an acronym for a phrase meaning “Jesus Christ, Son of God, Savior”, and used emblematically by those who subscribe to certain heavily evangelical Christian sects. Thanks to George Kiraz for this example.

simpler than the syntax of European heraldry, but there are some weak syntactic tendencies, such as the tendency for symbols of higher-ranked gods to occur closer to the beginning of the text.

Heraldic systems occur in many cultures. Apart from European heraldry and kudurrus, other examples include some functions of Totem poles (Barbeau, 1950) as well as well as Japanese *Kamon* ([http://en.wikipedia.org/wiki/Mon_\(emblem\)](http://en.wikipedia.org/wiki/Mon_(emblem))), though these are much simpler syntactically than the aforementioned systems.

4.6 Formal systems

The best example of a formal system is the symbology of mathematics. In a formal system, the individual symbols have a meaning, and there are generally strict rules on how the symbols may be combined. Thus if “a” and “b” represent numbers, and “+” is the arithmetic plus operation, in standard mathematical notation one may write “a + b” or “b + a”, but not “+ a b” or “a b +”. Furthermore, in addition to syntactic combinations, formal systems display compositional semantics in that in general the meaning of a complex expression can be derived by very specific rules from the meaning of the individual symbols and how they are combined. In this way, formal systems are very much like natural language, though in natural language the semantics of complex expressions is less thoroughly compositional than is the case with formal systems.

Other examples of formal systems: alchemical symbols, chemical notation, Feynman diagrams (Kaiser, 2005), programming flowcharts and Systems Biology Graphical Notation (Le Novère and et al, 2009).

4.7 Performative systems

Many systems exist that 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.

Silas John’s system for notating a small set of Apache prayers (see (Basso and Anderson, 1973), and Section 5.11.2) is another clear example of a performative system. Others are: musical notation (McCawley, 1996), dance notation — e.g. Labanotation — and other movement notation systems (Farnell, 1996), chess notation, and other systems that can be used to indicate the sequence of plays in a game, and knitting patterns (Harris, 1995).

4.8 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. For example, in Dakota winter counts (see Section 5.11.3), each winter was notated by a picture that represented some critical event that “defined” that particular year. The depiction thus evoked a story relevant to the given year.

A more elaborate narrative system is the “dongba” symbols of the Naxi; see (Li, 2001) and Section 5.11.1. In this system, a few thousand distinct symbols are used to represent important characters, as well as important plants, animals, topographical and astronomical entities, and so forth. While the system does have some linguistic elements (for example, sound cues are often written as part of the text using the Naxi syllabary), the system is basically non-linguistic in that it does not specify the exact set of words to be read. Naxi texts can be quite long, allowing for some quite elaborate stories to be told using this system.

Possibly also fitting into this category are some Mesoamerican systems such as Aztec “writing”, though some scholarship seems to be leaning more towards considering Aztec writing as true writing, with significant amounts of phoneticism (Lacadena, 2008).⁶

Another possible example is *rongorongo* (Fischer, 1997), the “writing system” of Easter Island, which has defied decipherment as a true writing system for over a century despite the best possible conditions for its decipherment that one could imagine.⁷

4.9 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

⁶A similar shift in the view of Maya writing happened several decades earlier after Gelb’s (1952) pronouncement about Maya writing noted above.

⁷Viz.:

- We know the language spoken by the Easter Islanders.
- We have good ethnographic accounts of how the *rongorongo* were used in ceremonies by eyewitnesses.
- While most of the corpus was lost, over 12,000 symbols’ worth of text still remain.
- The texts are quite long, meaning that a fortuitous fit to the text will be far less likely than what is possible when all one has is very short texts.

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 (Graves (1984), and Section 5.5), and Asian emoticons (Section 5.7), 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.

As we shall discuss in detail in Section 5.5, there has been a lot of confusion in the literature that has equated the notion of “non-linguistic” symbols with decorative systems, under the mistaken belief that something that does not encode language must therefore be meaningless.

4.10 Summary

In this section we have presented an initial taxonomy of non-linguistic signs. While the classification is certainly in need of further work and is undoubtedly incomplete, it is nonetheless useful in that it gives us a way of characterizing the corpora that we will describe in the next section. As will be clear we cover some types well, and are lacking in others, suggesting the need for further work.

One thing that should be clear after reading the foregoing section is what sets non-linguistic systems apart from linguistic systems. For all of the systems we have mentioned, it is clear that whereas the “messages” that can be conveyed by the system can be quite varied, nonetheless the scope of the messages is limited to the kinds of things for which the system was designed. Mathematical symbolology can only be used to express mathematical equations and formulae. Kudurrus could express the deities favored by the owner and “invoke” their power. But none of these systems can be used to express arbitrary messages. One could not, for example, write a treatise on mining using the symbols of European heraldry.

A fully developed writing system is capable of expressing *any* information that can be conveyed using speech, including any of the content expressed by the more specialized systems we have been considering here. One can express the meaning of any mathematical formula using written language: the expression will certainly not be as compact or as clear, but it will nonetheless be possible to do it. The reverse — the expression of anything that can be handled in writing using purely the symbolology of mathematics — is, pace Asimov's psychohistorians in the *Foundation* series, not possible.

This is at its base a rather simple and obvious point, but one that is often lost in the discussion of whether some arbitrary symbol system constitutes a writing

system or not.

It must also be borne in mind that some systems will not be purely of one kind or another, and indeed it is possible to find cases where different kinds of systems are mixed together. A rather nice example of this is Syriac, where a cross symbol may be incorporated into the spelling to indicate that the priest should make the sign of the cross while intoning the words (Kiraz, 2012, page 128): thus ܐܘܢ ܕܥܒܪܟܐ ܕܩܕܝܫܐ “and he blessed and sanctified” includes three instances of the cross symbol ܥ written above the words (and displacing the vowel diacritics, which end up underneath the words). In this case, then, a performative non-linguistic system is melded with a linguistic system.

5 The Corpora

Corpora were selected using both the criteria of what kinds of systems were desirable to represent, as well as what was readily available. Falling into the former category are Mesopotamian deity symbols, Vinča symbols, Pictish symbols, totem pole symbols and European heraldry. Mesopotamian deity symbols (*kudurrus*) and Vinča systems were mentioned by Rao et al. (2009a) 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.

Pictish symbols were, of course, already discussed by Lee et al. (2010a), 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, as we shall see, our tests, including Lee and colleagues' own measure, suggest it is non-linguistic.

Pacific Northwest totem poles are included because there are a large number of “texts” in this system, and many sources. A similar point can be made for European heraldry, where it is possible to find tens of thousands of “texts” in the form of *blazon*.

The function and meaning of Vinča symbols and Pictish symbols is unknown. On the other hand, the meanings of totem poles, European heraldic symbols, and Mesopotamian deity symbols are well understood. Their inclusion, in fact, is not purely serendipitous: All of these systems had, at least in part, a heraldic function, and might serve as a model for the kind of system that the Indus signs represented, given that it was not writing.

One other traditional system included here is Pennsylvania Dutch barn stars. This system is different from the rest since, as we noted above and as we shall discuss more fully below, it is highly likely that the system was purely decorative. Where to draw the boundary between decorative art and symbols is, of course, a difficult one, and as we shall argue below, many critiques of Farmer et al. (2004) have been confused on precisely this point.

Three further corpora were collected because they were easy to collect semi- or fully automatically from the Web. Weather icon sequences⁸ can easily be harvested by a script that runs at regular intervals. Asian emoticons had been collected as part of another project I was involved in. Mathematical formulae are available in the hundreds of thousands from L^AT_EX documents at archiv.org. Such corpora, which are already in electronic format, require relatively little processing.

For the other corpora, it was necessary to transcribe them from print sources. The only exception to this was the Pictish corpus, but even here reference to pho-

⁸Suggested by a colleague, Bill Zaumen.

tographs was needed. In transcribing any symbol system the very first concern is the inventory of signs, something that can be hard to ascertain for the non-specialist. Fortunately, in all cases, we could make use of the standard inventories compiled by experts on the systems. Thus, for totem poles, for example, our sources all gave transcriptions of the depicted poles, naming each symbol; for Mesopotamian deity symbols our source (Seidl, 1989), lists the symbols used on each stone; and so forth.

After discussing the various corpora, we present in Section 5.8 statistics on the size of the various corpora. The Appendix contains examples of the first five symbol systems, along with the corresponding transcription in the XML markup scheme we introduce in Section 6.

At the head of each section we propose a classification for the symbol system in question in terms of the taxonomy developed in Section 4. 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

5.1 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 Figure 4. 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 (Winn, 1981), 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 Vinya 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

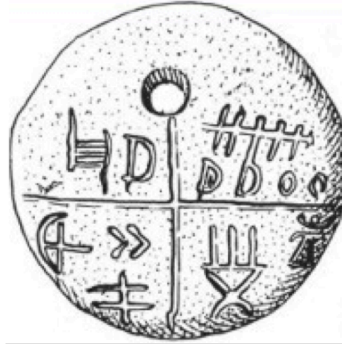


Figure 4: An example of a Vinča “text”, from (Winn, 1981).

significance. As Winn states (p. 255): “In the final analysis, the religious system remains the principle source of motivation for the use of signs.”

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.

5.2 Mesopotamian deity symbols on *kudurrus*

Classification: Heraldic

In Section 1 we introduced deity symbols from Babylonian *kudurru* boundary stones and described their function. Our source for the deity symbol corpus was (Seidl, 1989).⁹ We picked texts from all stones where the depictions in Seidl were clear enough to read. Fortunately Seidl 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.

5.3 Pictish stones

Classification: Unknown, possibly heraldic

⁹Additional sources for deity symbol texts include (King, 1912; Black et al., 1992), which have at least a few additional texts.



Figure 5: The Aberlemno 1 stone from <http://www.stams.strath.ac.uk/research/pictish/database.php>.

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 (Rhys, 1892). An example of Pictish symbols is shown in Figure 5.

As we noted above, a high profile paper by Rob Lee and colleagues (Lee et al., 2010a) has argued on the basis of a statistical measure based on conditional entropy (see Section 8.3), that the symbols comprised a heretofore unrecognized writing system. While this paper generated a lot of press, it is less clear how widely Lee and colleagues’ view is accepted among scholars of ancient Scotland. More to the

point Lee and colleagues in their reply to Sproat (2010a) — (Lee et al., 2010b), revealed that they adhere to a much looser notion of what it means to be a writing system, accepting Powell’s (2009) definition, namely that “writing is a system of markings with a conventional reference that communicates information”. They also state that they are comfortable with the result I reported in (Sproat, 2010a) that, using their methods, the known non-linguistic system of Mesopotamian deity symbols would be classified as logographic writing (like Pictish symbols): indeed, they replicated my result. Clearly using this definition, *any* symbol system where the symbols have a more or less fixed denotation, would be considered writing. If that is the definition one wants then there is really nothing to argue about.

But if one wants a stricter definition (such as the one assumed by many students of writing systems, including (DeFrancis, 1984, 1989; Sproat, 2000) , that writing systems encode specifically *linguistic* (and usually phonological) information, then there seems to be little reason to believe that Pictish symbols were writing.

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 (Jackson, 1984; Royal Commission on the Ancient and Historical Monuments of Scotland, 1994; Jackson, 1990; Mack, 1997; Sutherland, 1997). 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.

5.4 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 (Malin, 1986). 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. (Rao et al., 2009b).

We used a number of published resources in developing our data. In addition to Malin (1986), we transcribed texts from (British Columbia Provincial Museum,



Figure 6: The Grizzly Bear Pole of Yan, from (Drew, 1969, page 16).

1931; Garfield, 1940; Barbeau, 1950; Gunn, 1965, 1966, 1967; Drew, 1969; City of Duncan, 1990; Stewart, 1990, 1993; Feldman, 2003).

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 Nanasinget story (Stewart, 1993). In principle then this could be represented by a single symbol (e.g. NANASIMGET), but instead we elected to use the `symbolUnit` tag (Section 6) to represent the grouping:

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

5.5 Pennsylvania Dutch 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. Contrary to common belief, they are *not* found among the Plain People (Amish and Mennonites), who eschew decoration of any kind. Traditional barn symbols consist mostly of stars, rosettes, wheels-of-fortune, and swastikas. Ironically, none of these forms are readily available from any of the purveyors of “hex signs” for tourists.

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 (Mahr, 1945). And the barn symbols themselves can be traced back in many cases to symbols that probably had definite sets of meanings at one time (Mahr, 1945; Graves, 1984; Yoder and Graves, 2000). 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 (Graves, 1984; Yoder and Graves, 2000).

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 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 (Farmer et al., 2004) 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 (Vidale, 2007). 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.

Our barn star corpus is based upon the slide collection of W. Farrell, who toured areas around Berks County in the 1940's and photographed many barns that were decorated with 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 (1984). 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 7. 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 8. The "four-slice pie" is, however,

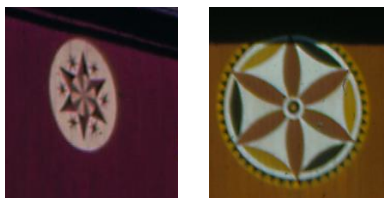


Figure 7: 6-pointed star (left) versus rosette.

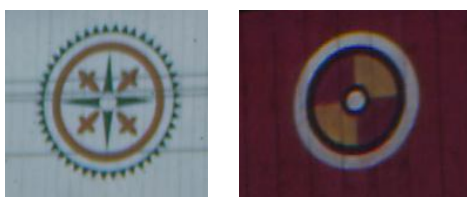


Figure 8: 4-pointed star (left) versus 4-slice pie.

so distinct that it seems to deserve its own category.

5.6 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.¹⁰ 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.¹¹ See Figure 9 for an example of a weather “text”.

¹⁰There 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.

¹¹Note that on occasion the scraping of the page resulted in no data returned.



Figure 9: A sample weather icon sequence: forecast for Portland, Oregon, April 29, 2011.

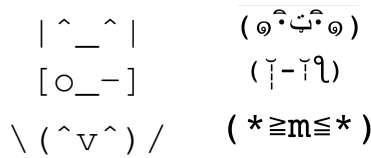


Figure 10: Some representative *kaomoji* emoticons

5.7 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 (Bedrick et al., 2012). Unlike the familiar 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 10. 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.

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 2: Number of texts, type and token counts, and mean text length for the corpora collected so far.

5.8 Statistics on Current 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 2.

5.9 Symbol Systems Currently under Development

The corpora described above are more or less complete. Two corpora are currently under development, and we describe them here since we will make use of them in the statistical analysis later on.

5.9.1 Mathematical Formulae

Classification: Formal

Mathematical symbology is a clear case of a non-linguistic system and it also serves as a good example of the distinction between linguistic and non-linguistic systems. Consider a formula such as:

$$\int e^x dx = e^x + C$$

It is clearly possible to *read* this expression using language: “the integral of e to the x, d x, equals e to the x plus C”. Thus it might seem that one could argue that mathematical symbols are linguistic. But this is missing a crucial point: while it is certainly possible to express mathematics using language, it is clear that the elements in a mathematical formula do not *represent* linguistic information. The symbol \int represents the mathematical concept of integration, not the English

words “integrate”, or “integral”, or “integration”. Similarly e^x represents the mathematical exponentiation of the irrational number e to the power of variable x , not the English expression “e to the x”.

Fortunately mathematical expressions are easy to harvest automatically from online sources. For this project, 35,492 L^AT_EX documents from the ArXiv database (<http://www.arxiv.org>) have been downloaded from <http://www.cs.cornell.edu/projects/kddcup/>. From these we extracted 414,492 equations delimited by `\begin{equation}` and `\end{equation}`.

5.10 European Heraldry

Classification: Heraldic

All of the symbol systems we have discussed so far have the property that the symbols are linearly arranged, or at least mostly linearly arranged: there is no crucial use of a second dimension. European Heraldry, crucially differs in that symbols used in arms are arranged in two dimensions, or perhaps more accurately two and a half dimensions, since while arms appear on a flat surface, one talks of symbols (charges) being placed on top of backgrounds (fields) or other charges.

An obvious question then is how to serialize the symbols. Fortunately there is already the formal language of *blazon* that provides a conventional serialization. Blazon looks a little bit like English in that it uses many English words and observes an English-like syntactic structure. However it differs from English in that the syntax is a great deal more rigid and most terms are unambiguous in their meaning. In addition there is much terminology that is certainly not used in English, or not used with the given sense, such as the terms for colors or metals (e.g. *azure* for “blue”, *gules* for “red”, or *or* for gold). See Figure 11 for some examples of blazon and their corresponding arms.

The dataset we are collecting consists of 10,659 blazons from Burke’s *General Armory* (Burke, 1884) and 2,531 blazons from the Mitchell Rolls from the Heraldry Society of Scotland <http://www.heraldry-scotland.co.uk/mitchell-rolls.html>, consisting of about 115,000 tokens — for a total of 13,190 blazons. The text from Burke was obtained from the OCR’ed version at <http://archive.org/details/generalarmoryofe00burk> and there were a great many OCR errors and other issues. The text for the Mitchell Rolls is mostly much cleaner. The blazons were parsed using *pyBlazon* (<http://web.meson.org/pyBlazon/>), which produces an XML analysis of the blazon, but there were a number of issues with the use of this program. First of all, it is fairly brittle, and large numbers of well-formed blazons are not parsed at all. For example, the system will mostly not handle quarterings — i.e. blazons of the form *quarterly first and fourth blazon₁, second and third blazon₂*, where

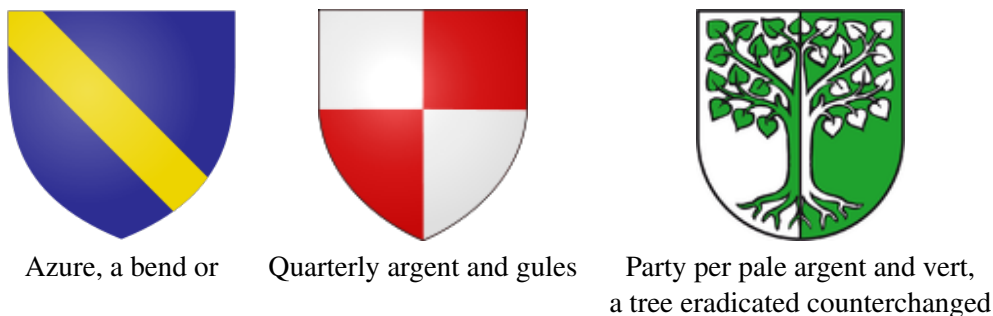


Figure 11: Examples of arms and their corresponding blazon from <http://en.wikipedia.org/wiki/Blazon>.

$blazon_{1,2}$ are full blazons of the specified quarters.¹² Second, it will naturally fail or produce an incorrect analysis if there are OCR or other errors in the text: this feature actually proved useful in that it allowed us to filter out a large number of examples with OCR errors that would have been impractical to clean up by hand, and focus on a still reasonable sized subset that was mostly clean. Third, and most importantly, the XML structures generated by pyBlazon are overly verbose, and fail to mark features that are useful for the kind of analyses we want to perform here. We discuss this issue further in Section 6.

5.11 Further Symbol Systems Planned

In this section we discuss some further systems that are either in the early stages of development or are planned.

5.11.1 Naxi Pictography

Classification: Narrative

¹²The upshot of this is that we only have a small set of quartered arms. This is not really a problem since quartered arms are merely compound arms that include well-formed simple arms, and thus we are not really losing information. However in future work we plan to extend our set to include these kinds of arms by first training a statistical parser on trees derived from the XML corpus we already have, augmented with artificial quartered arms, generated by embedding well-formed simplex arms within quarterings. The trained parser will then be used to parse a wider range of real blazons, and it is expected that it will produce an analysis of a much larger set of quartered arms than pyBlazon is able to handle. In early experiments with the BUBS parser (Bodenstab et al., 2011), we achieved parsing accuracies of at least 99%, a figure that is unheard of in parsing real natural language text — a point that underscores the fact that though blazon may look like English, it is really an artificial language.



Figure 12: The first page of NZA002 in the Library of Congress Collection.

One such system is Naxi pictography. The Naxi, a Tibeto-Burman-speaking minority group in China's Yunnan province, have a pictographic symbol system that is used in the recounting of myths. While the system is often called writing, it is not clearly a real writing system: the symbols do not in general represent words or morphemes of the Naxi language, though many of them do have canonical translations into Naxi words or phrases. A compendium of about 3,000 of these symbols and their translations can be found in (Li, 2001). A complication is that the Naxi also have a true syllabic writing system, and elements from that writing system are often used to add phonetic cues in pictographic texts.

The Library of Congress has a collection consisting of 3,342 manuscripts, some of which are viewable online at <http://rs6.loc.gov/intldl/naxihtml/naxihome.html>. I have collected photographs of 25 of these documents. See Figure 12 for an example. An additional collection can be found at the Harvard Yenching library (http://hcl.harvard.edu/collections/digital_collections/naxi_manuscripts.cfm).¹³ The Naxi system is worth investigating for a number of reasons. As non-linguistic systems go, it is among the most complex that we are aware of in terms of numbers of symbols. It is also clearly on the border between non-linguistic and linguistic systems: while it is not a full writing system and does not have mechanisms for representing all of the Naxi language, it is similar in many ways to semasiographic systems such as Bliss

¹³The main difficulty with the Naxi corpus will be transcription: the system is complex, with thousands of symbols, and it is not clear whether the best available lexicon (Li, 2001), is complete. In principle there are still people who are conversant in this system so it would be possible to find experts. There is also some initial prior work on transcription — C.L.A.U.D.I.A. <http://www.xiulong.it/4.0/Dongba/CLAUDIA/intro.php>, and the author of that project, Stefano Zamblera, has expressed interest in collaborating if resources can be made available.

Symbolics (Bliss, 1965), which do allow users to compose a limited, though still large number of messages. Finally the Naxi culture is an interesting living example of how linguistic and non-linguistic symbol systems can exist side by side, serving different sociolinguistic functions.

5.11.2 Silas John's System.

Classification: Performative

In 1904, Silas John Edwards, a Western Apache Shaman invented a notation system for 62 prayers that he claimed were revealed to him in a vision. While the system has been described by (Basso and Anderson, 1973) as a writing system, and certainly does encode some linguistic features (perhaps most notably the use of a symbol that looks like a script form of the English word *she*, used to represent the Apache first person singular pronoun *šii*), there are a number of ways in which it differs from a true writing system. First, the system is limited to writing exactly the set of 62 prayers and was never intended as a general way of writing Apache. Second, the system includes information on non-linguistic behavior, such as particular actions that are to be performed during the incantation of the prayer. The system is thus more of a performative symbol system, rather like labanotation, rather than a true writing system. However, since it does represent some linguistic information, it will be an interesting case, much like Naxi, for a system that is intermediate between a wholly non-linguistic system and a writing system. Unfortunately it appears that the six texts, comprising a little over a hundred symbols in total, that were published in (Basso and Anderson, 1973) may be the only ones available. These were the only texts which the authors were permitted to see and in which they received instruction as to their interpretation. (Another indication that the system was not a real writing system was that it seems that one had to be instructed in how to interpret each text.) However, Basso and Anderson's description of those few texts is very clear, and it will be straightforward to develop an electronic version of the corpus.

5.11.3 Dakota Winter Counts.

Classification: Narrative


Mallery (1883) compiles the “winter counts” (Lakota *waniyetu wowapi*) for the Dakota covering about 100 years (and thus 100 winter symbols); he also presents a variant of the winter counts from Corbusier. In this system, a year is represented by a symbol that gives a particularly memorable event for that year. For example, the symbol for the year 1786–1787 represents a chief who wore an “iron” shield over his head . Since these symbols represent individual years, each symbol is



Figure 13: Buffalo robe winter counts (http://wintercounts.si.edu/html_version/images/a.jpg)

really a separate “text”, and thus the system is non-combinatoric. However winter-counts were compiled into “texts” such as the famous buffalo robe depicted in Figure 13, which might be mistaken for writing by someone who did not know how the system worked.

Yet further systems. Other systems that various people have suggested include chess notation, game notation (e.g. chess), dance notation (labanotation) and Wikipedia barn stars (<http://en.wikipedia.org/wiki/Wikipedia:Barnstars>). Many of these, including obviously Wikipedia barn stars can easily be mined from online sources.

6 The Markup Scheme

The XML-based markup system, also described in (Wu et al., 2012) is intended to be minimally verbose, and at the same time give enough information to allow for the extraction of symbols from the texts according to different criteria of interest. For example, if the text appears in multiple lines, then the line breaks might be of interest. If the text appears in a circle, e.g. on a disk, then we want to tag the text as circular since in such cases it is not generally clear where the text begins. If symbols are grouped or ligatured together, we would like a way of representing that. If a symbol is uncertain, that information should be marked. Bibliographic information, the source of the particular text, and so forth also needs to be recorded.

An example (see also the Appendix for further examples) is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="mySchema.xsd">

  <entry>
    <bib>
      <author>Drew, F. W. M.</author>
      <year>1969</year>
      <title>Totem poles of Prince Rupert.</title>
      <pubInfo>Prince Rupert, B.C.: F.W.M. Drew.</pubInfo>
    </bib>

    <document type="totemPole" origin="Haida">
      <description>Grizzly Bear Pole of Yan,
        a house frontal pole</description>
      <page> p16 </page>
      <docText>
        <symbol><title>3-Skils</title></symbol>
        <symbol><title>Grizzly-Bear</title></symbol>
        <symbol><title>Bear-Mother</title></symbol>
        <symbol><title>Cub</title></symbol>
        <symbol><title>Cub</title></symbol>
        <symbolUnit>
          <symbol><title>Supernatural-Grizzly-Bear</title></symbol>
          <symbol><title>Frog</title></symbol>
          <description>Supernatural
            Grizzly Bear holding a Frog</description>
        </symbolUnit>
        <symbol><title>Grizzly-Bear</title></symbol>
      </docText>
    </document>
  </entry>
</collection>
```

Each XML file consists of collection of entry elements, consisting of

one or more documents from a single source. The entry starts with a `bib` element, whose contents should be self-explanatory. This is followed by a set of one or more document entries, consisting of an `description` and page information. Attributes are the `type` of document and the `origin`. Then follows the actual `docText`, consisting minimally of a set of `symbol` elements containing the title (the actual symbol). `docText` elements may also contain `side` elements for multisided documents, and `line` elements for multiline texts, each of these containing `symbol` elements. For circular texts we provide a `circle` tag.

Symbols that are grouped together in some way are collected under the `symbolUnit` element, as in the above example. The `symbolUnit` may contain a `description`, which explains the significance of the grouping. 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.

What we have just described covers most of the texts we have encoded to date; further technical details on the XML markup can be had by consulting the XML schema that will be published along with the corpora.¹⁴

One corpus for which the above scheme is not adequate is European heraldry. This is in part because unlike most of the other symbol systems under consideration, heraldry makes meaningful use of two dimensions, and in part because the blazon transcription we are working from includes not only symbols — charges, colors and so forth that are significant elements of the system — but also English words such as *a*, *the* or *and*, which are not part of the underlying symbol system, and should usually be discarded in statistical analysis. We started with `pyBlazon` (<http://web.meson.org/pyBlazon/>), a Python module for parsing blazon descriptions, and as we discussed in Section 5.10, the 13,190 blazons in our set are those blazons (from a much larger set) that could be parsed using `pyBlazon`.

The output of `pyBlazon` is an XML representation, but it is highly verbose and has many levels of embedding that seem largely unnecessary. It also fails to mark terms such as *a* or *the* as grammatical morphemes, which we need. Much of our work on blazon has involved simplifying this output so as to represent all the critical information with a minimum of structure, while still preserving necessary information, and in some cases adding useful information. An example of this modified markup can be seen below for the blazon *azure on a bend ermine three buckles gules*:

```
<blazon>
```

¹⁴Publication details to follow.

```

<treatment>
  <color>azure</color>
</treatment>
<charges>
  <group>
    <spatial>on</spatial>
    <grammar>a</grammar>
    <ordinary>
      <title>bend</title>
      <coloring>
        <fur>ermine</fur>
      </coloring>
    </ordinary>
  </group>
  <group>
    <amount>3</amount>
    <charge>
      <title>buckles</title>
      <coloring>
        <color>gules</color>
      </coloring>
    </charge>
  </group>
</group>
</charges>
</blazon>

```

Following pyBlazon, we encode the field as a `treatment` with a `color`, mark “bend” as a member of the class of so called `ordinaries` with a `coloring` of “ermine”, and set up a `group` consisting of an `amount` of `three` `charges` “buckles gules”. “On” and “a” are marked as, respectively, `spatial` and `grammar`. With the text marked up in this format, it is easy to extract parts of the blazons for statistical analysis. Suppose we want to remove grammatical terms like “a” or “the” since these are not really part of the symbology of heraldry and merely reflect the syntactic requirements of blazon? Or suppose we want merely to look at the statistics of `charges` and `ordinaries` (the actual “symbols” in the system) and ignore the other parts of the blazon? The above markup scheme allows for this.

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 3: Linguistic corpora for comparison.

7 Linguistic Corpora for Comparison

While one aim of this project is to collect corpora of non-linguistic symbol systems, not linguistic systems, yet another aim is to investigate a range of possible statistical methods for distinguishing the two types. To that end, we gathered a set of linguistic corpora from various existing sources.

Just as we have tried to collect non-linguistic corpora of a variety of types, so we 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.

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. This resulted in a wide range of text lengths, which in turn resulted in some interesting and telling statistical outcomes, as we shall detail below.

The following subsections give brief descriptions of each of the corpora. Summary statistics can be found in Table 3.

7.1 Amharic

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

7.2 Modern Standard Arabic

Collection of the first 10,000 headlines from the 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 (“#”).

7.3 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.

7.4 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.

7.5 Egyptian

Collection of texts transcribed using the JSesh hieroglyph editor (<http://jsesh.genherkhopeshef.org/>) downloaded from <http://webperso.iut.univ-paris8.fr/~rosmord/hieroglyphes>. The text was tokenized at the glyph level.¹⁵

7.6 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 (“#”).

¹⁵The 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, Pachereiseniset.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.

7.7 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 (“#”).

7.8 Korean

Collection of the first 10,000 headlines from the LDC Korean Newswire corpus <http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2000T45>. Following the standard Unicode encoding of Korean, the text was tokenized at the Hangul syllable level, with spaces represented as a separate token (“#”).

7.9 Korean Jamo

Same source as above, but here the text was tokenized at the letter (*jamo*) level — the individual components of the syllable composites, with spaces represented as a separate token (“#”).

7.10 Linear B

Transcription by the author of 300 tablets from (Ventris and Chadwick, 1956). Omitted lines with uncertainties and/or omissions. Tokenized at the glyph level.

7.11 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 (“#”).

7.12 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 (“#”).

7.13 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.¹⁶ “Texts” consisted of individual lines in the CDLI transcription, meaning that the Sumerian “texts” are rather short.

One issue with Sumerian that we will need to resolve 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 𒀭𒃶, 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: 𒀭𒃶 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.¹⁷

7.14 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 (“#”).

¹⁶Thanks to Chris Woods for suggesting appropriate search terms.

¹⁷At least some of this information is available in the Pennsylvania Sumerian Dictionary (<http://psd.museum.upenn.edu/epsd/index.html>), though expert manual intervention will almost certainly be needed beyond any automatic method that can be developed.

8 Statistical Models

In this section we explore the use of existing measures, as well as propose new measures and approaches for the task of distinguishing linguistic from non-linguistic symbol systems. One of the byproducts of this investigation is the result 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 non-linguistic symbol systems. And Lee's measures (Lee et al., 2010a), with the cutoff values they propose, misclassify nearly all of our non-linguistic systems. However, we also show that one of Lee's measures, with different cutoff values, as well as another measure we develop here, do seem useful. But we also demonstrate that they are useful largely because they are both highly correlated with a rather trivial feature: mean text length.

We note at the outset the way in which we extract the texts from the XML for each of the corpora: every document is extracted as a single text, ignoring line breaks and sides. We extract symbols rather than symbolUnits and thus extract the finest grained breakdown of the texts. We also ignore uncertainty marking on symbols. These decisions could of course affect our results, though they are unlikely to change the results enough to affect the overall conclusions of our analysis.

Since not all readers will have a background in statistics or information theory, we provide, at the introduction to each of these sections, a non-technical synopsis of the main issues discussed in the section as well as a summary of the basic conclusions.

8.1 Models of Information

Conditional entropy (Shannon, 1948) is a measure of how surprised one should be, given a sequence of symbols, about the value of the next symbol. Suppose a symbol system is completely rigid so that if one sees, say, abc, the symbol d will always follow. Then the entropy will be zero, since once will be completely unsurprised by finding d after a sequence abc. On the other hand if a system is completely free so that no matter what symbol sequence has been seen, any new symbol is equally likely to occur, then the entropy will be maximal — and will increase with the number of symbols in the system, since one would be equally surprised to see any given symbol.

Rao and his colleagues (2009a) argue that the conditional entropy behavior of Indus Valley texts is similar to that of natural language,

which is neither completely rigid nor completely free, and show plots that seem on the face of it to be highly convincing.

But no symbol systems of any kind are completely rigid or completely free, so the entropic behavior of the Indus Valley symbols is unsurprising and uninformative. Indeed analysis of our corpora shows no clear separation of linguistic and non-linguistic systems using the measure of conditional entropy.

One measure that has been used a lot to measure the information structure of language is conditional entropy (Shannon, 1948), for example, bigram conditional entropy, defined as follows:

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

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 (as we shall see) a whole range of non-linguistic systems, fall between these two extremes.

Bigram conditional entropy was used by 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, 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.

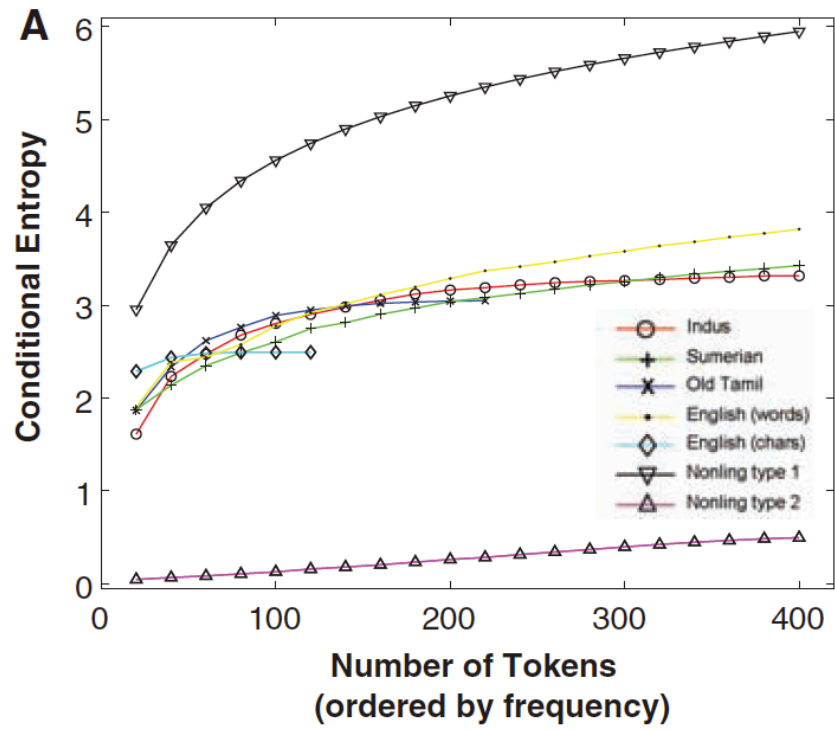


Figure 14: Bigram conditional entropy growth curves of various linguistic symbol systems, the Indus Valley symbols, and two artificial models of non-linguistic systems, from (Rao et al., 2009a). See the text for further explanation. Used with permission of the American Association for the Advancement of Science.

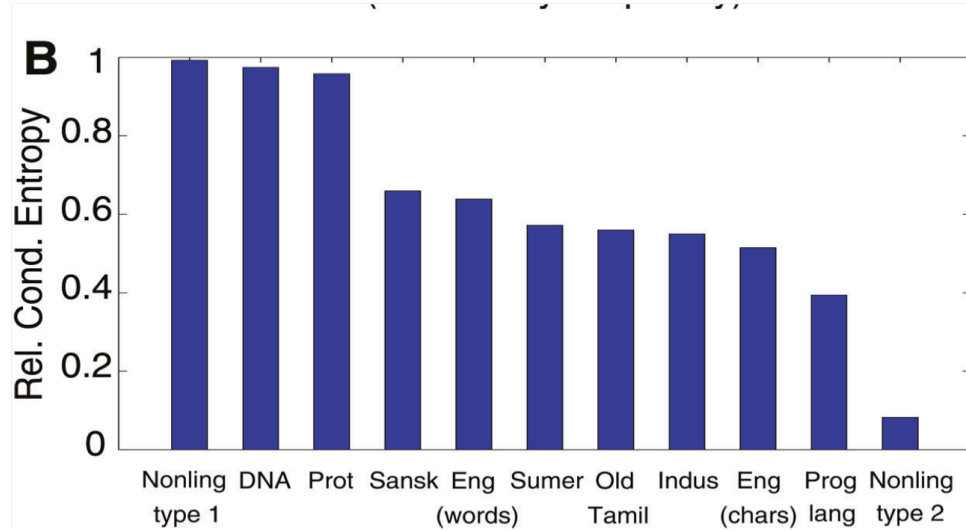


Figure 15: Relative conditional entropy (conditional entropy relative to a uniformly random sequence with the same number of tokens) of various systems from (Rao et al., 2009a). Used with permission of the American Association for the Advancement of Science.

Figure 14 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 15 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 14, 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 (Winn, 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) and (Sproat, 2010b), 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 15 (though not in Figure 14): it is clear from that figure that the curves of these biological symbol systems are very close to the maximum entropy curves. 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.

We are not sure exactly how Rao and colleagues computed their results: while Rao reports that they used Kneser-Ney smoothing (Kneser and 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. Discrepancies such as this and others may help explain why the entropy estimates we present below are on bulk lower than those presented in Rao and colleagues’ work. It is not possible to produce plots that are directly comparable to Rao’s plots. However we can produce entropy growth curves for a variety of systems using similar estimation techniques.

In Figure 16 we show conditional entropy growth curves for all of our non-linguistic systems (in blue), all of our linguistic samples (in red), and for a corpus of Indus bar seals from Harappa and Mohejo Daro (in green), which we developed as part of the work reported in (Farmer et al., 2004). The details on this latter corpus are given below:

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 (Parpola, 1994; Farmer et al., 2004).

Entropy statistics were computed using the OpenGrm NGram library (Roark et al., 2012), available from <http://www.opengrm.org>. We used Kneser-Ney smoothing, including the probability and thus entropy estimates for unseen cases. Start and end symbols are implicit in the computation.

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 (Rao, 2010) — to have the highest entropy of all the systems. Chinese shows the lowest. Between 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

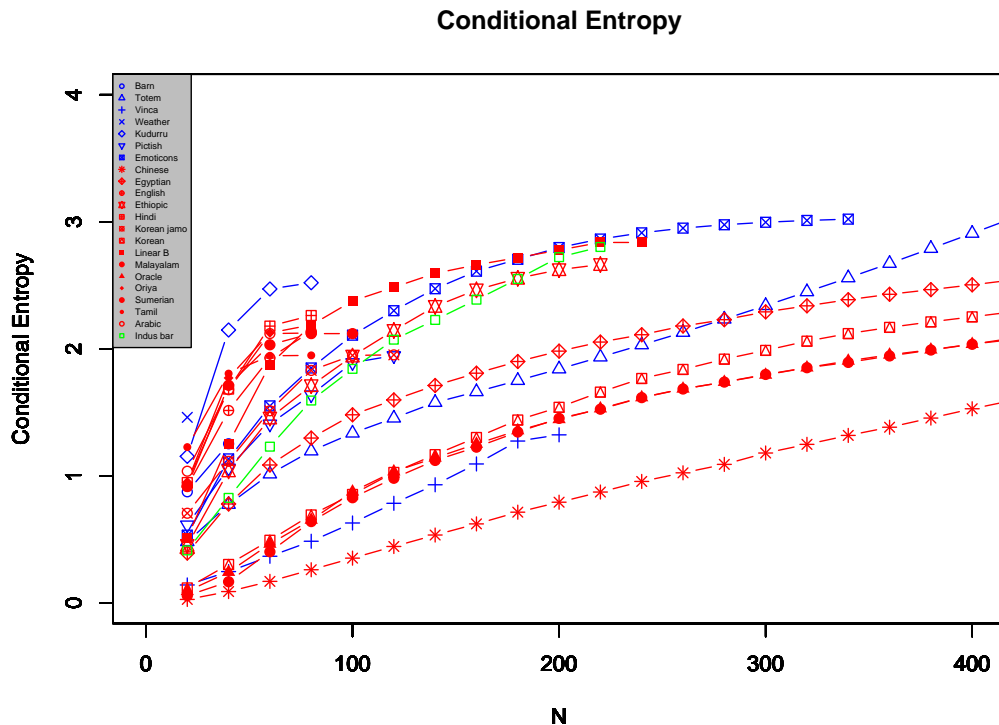


Figure 16: Bigram conditional entropy growth curves computed for our linguistic and non-linguistic corpora, and for Indus bar seals.

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 17 we 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. 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, lead-

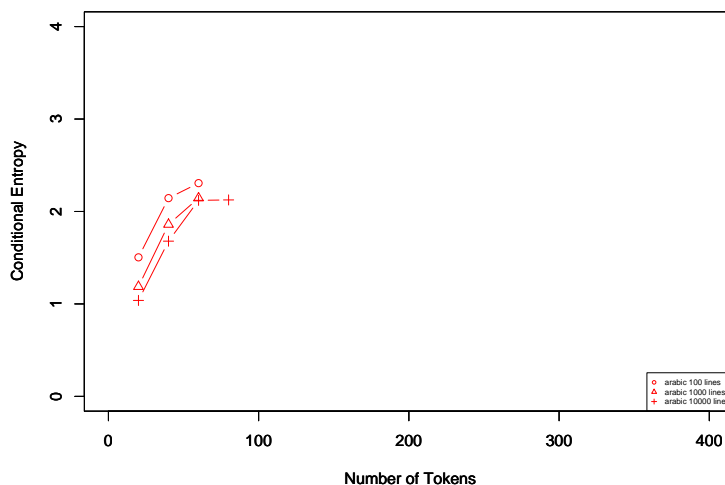


Figure 17: Bigram conditional entropy growth curves for three different sized Arabic corpora.

ing 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.

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 any symbol system. But it does contradict Rao and colleagues’ claims, so a natural question is how to interpret these results vis-à-vis the results presented in Rao et al’s work? In order to answer that question, one needs to understand how Rao and colleagues interpret their own results. When Rao’s original paper came out in *Science* in 2009, it was billed in the popular press as a demonstration that the Indus symbols constituted writing. The most natural interpretation of that would be that the methods proposed by Rao and his colleagues constitute a way of *discriminating* between linguistic and non-linguistic systems. The plot in Figure 14 goes

a long way towards reinforcing that interpretation. However, Rao never actually says this: what the paper says is that the results “increase the probability” that the Indus system was writing. Only when the objections I raised in (Sproat, 2010a) forced a response (Rao et al., 2010), did it become clear that Rao did not interpret their results in a discriminative way. Rather they took a *generative* interpretation: consider two hypotheses H_L which states that the Indus system was linguistic and H_{NL} , which states that it was non-linguistic. Which hypothesis is more consistent with the facts? Rao then goes on to discuss a range of evidence that they argue is more consistent with H_L . Included in this evidence are:

- The linear arrangement of the Indus symbols in texts.
- The (apparent) presence of diacritic modifications 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 one would expect all of the above features, as well as the behavior evidenced by the entropic measures that Rao and colleagues present.

The problem with this line of reasoning is that it is all too easy to come up with a list of features that make a known non-linguistic system look linguistic. Consider the hypothesis that Mesopotamian deity symbols constituted a form of writing. Under that hypothesis, one would expect properties such as the following:

1. Deity symbols are frequently linearly written 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 non-linear arrangements with scripts too: the Etruscan Magliano disk (http://it.wikipedia.org/wiki/Disco_di_Magliano), and many rune stones have text wrapped around the border of the stone (see, e.g., http://en.wikipedia.org/wiki/Lingsberg_Runestone).
2. There is clear evidence for the directionality of deity symbols: to the extent that “more important” gods were depicted first (see below), these occur at the left/top of the text.



Figure 18: The linearly arranged symbols of the major deities of Aššurnasirpal II. From http://upload.wikimedia.org/wikipedia/commons/8/87/Ashurnasirpal_II_stela_british_museum.jpg, released under the GNU Free Documentation License, Version 1.2.

3. Deity symbols are often ligatured together: one symbol may be joined with another.¹⁸
4. The deity symbols obey a power-law distribution.
5. Deity symbols clearly have language-like properties in that certain symbols display positional preference (symbols for the more important gods coming earlier in the text, per Rao et al.), and certain glyphs have an affinity for each other — for example some glyphs such as the “horned crown” seem to like to be joined together with the “symbol base”.
6. Deity symbols are pictographic, like many real scripts — Egyptian, Luwian, Mayan, Hittite hieroglyphs.
7. Deity symbols were used largely on standing stones, but were also used in other contexts such as the “necklace” depicted in Figure 18, suggesting that there were a variety of media in which one could create “texts”.

All of these properties hold of the deity symbols, yet we know that this symbol system did not represent language.

Turning back to the Indus system, one can easily list features of the system that seem more consistent with H_{NL} . Consider features such as:

¹⁸We note in passing that it is actually unclear for the Indus symbols that there are ligaturings or modifications 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.

- All extant texts are very short.
- The system was used in a culture where there are no archaeological markers of manuscript production (such as pens, styluses, ink pots . . .).
- 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.

These features, and others, were discussed in (Farmer et al., 2004). To be sure, that work has not gone without its critics — see, e.g., (Vidale, 2007) — and the interested reader is invited to compare our original arguments with those of the critics to see which seem more convincing.

Returning to the statistical evidence, one thing that is obvious is that if instead of Figure 14 from (Rao et al., 2009a), it had been clear from the outset that the true situation is more like that presented in Figure 16, it would have been significantly less easy to argue that conditional entropy supports the linguistic status of the Indus signs — under *any* interpretation of what “support” means.

8.2 Models of Information: Models using Block Entropy

Conditional entropy, as we have seen, is a measure of one’s surprise at seeing an event, such as the appearance of a particular symbol, given some previous history of events.

“Block entropy” is rather a measure of surprise of a sequence of events, for example a particular symbol sequence abcd. Consider all possible sequences of four symbols over the English alphabet of. If those sequences are all equally probable, then the surprise of seeing any particular sequence will be maximal. If however, the sequence abcd is the only one possible, then entropy will be zero since we will be completely unsurprised when we see it.

Rao (2010) used “block entropy” to further his argument that the Indus symbols are a script, but in this section we show that “block entropy” is just as uninformative as conditional entropy in determining whether a symbol system is linguistic or not.

We turn now to a second entropic measure proposed by Rao (2010), what he terms “block entropy”, defined as follows:

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

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 19 shows the block entropy estimates from (Rao, 2010) for a variety of linguistic systems, some non-linguistic systems, the Indus corpus as well as the maximum and minimum entropy systems from (Rao et al., 2009a).¹⁹ For this work, Rao used an estimation method proposed in (Nemenman et al., 2002), 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 Figure 19.

These we present in Figure 20. 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 19, 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.

Once again, if the true situation represented in Figure 20 had been clear rather than the situation presented in Figure 19, one could not have made much of a case that statistical measures “increase the probability” of the Indus symbols being a writing system.

8.3 Measures Derived in Part from Entropy

In the past couple of sections we considered simple measures such as bigram entropy. But there is no reason why a measure that might be informative about the status of a symbol system should be simple. In

¹⁹Here, Rao is wrong: a rigidly ordered system would *not* look as he has it 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 ngrams 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 one, and decreases in value as N grows larger.

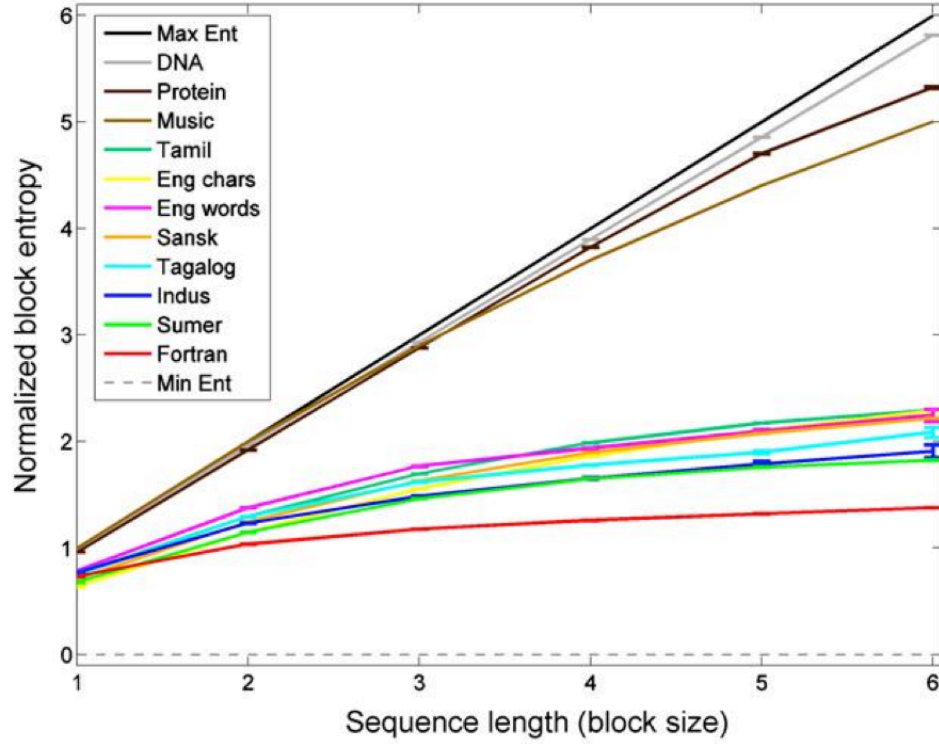


Figure 19: Block entropy estimates from (Rao, 2010) for a variety of linguistic systems, some non-linguistic systems, and the Indus corpus. Used with permission of the IEEE.

this section we discuss a more complex couple of measures proposed by Lee et al. (2010a), which include conditional entropy, but also the number of distinct symbols, the number of distinct pairs of symbols, and the total number of pairs of symbols.

If one follows Lee et al. (2010a)'s proposal to the letter, the system misclassifies nearly all of our non-linguistic corpora as some sort of linguistic system, and thus more or less completely fails as a reliable method for distinguishing between linguistic and non-linguistic systems.

Unlike Rao, Lee and colleagues (2010a) are forthright in their claim that they

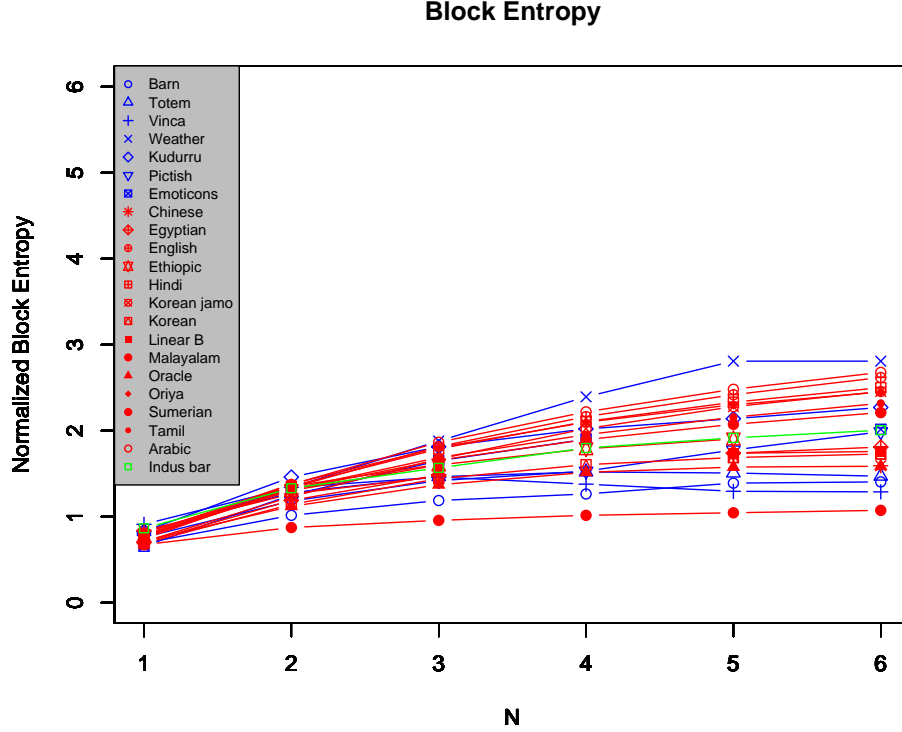


Figure 20: Block entropy values for our linguistic and non-linguistic corpora, and the Indus bar seal corpus. As in Rao’s case, we used the method reported in (Nemenman et al., 2002).

have developed a discriminative method. Lee et al.’s paper 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 et al.’s work, they compare the symbols to a variety of known writing systems, as well as symbol systems like Morse code, and European heraldry, and randomly generated texts — by which, again, 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:

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

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:

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

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. To illustrate the components of C_r , Lee et al. show a plot (their Figure 5.5), reproduced here as Figure 21. According to their description this shows:

[A p]lot of S_d/T_d (degree of digram repetition) versus N_d/N_u (degree of digram lexicon completeness). . . Dashes, sematograms—heraldry; filled diamonds, letters—prose, poetry and inscriptions; grey filled triangles, syllables—prose, poetry, inscriptions; open squares, words—genealogical lists; crosses, code characters; open diamonds, letters—genealogical lists; filled squares, words—prose, poetry and inscriptions. (Lee et al., 2010a, page 8)

Note that the non-linguistic system of heraldry seems to have a much lower number of singleton bigrams than would be expected given the corpus size, clearly separating it from linguistic systems.

Lee et al. 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 22.

What happens If we apply Lee et al’s tree, “out of the box” to our data? Table 4 shows the results of doing this.²⁰ Not surprisingly, the Pictish symbols are classified as (logographic) writing, consistent with Lee et al.’s results. But, with the exception of Vinča symbols, which are appropriately classified as non-linguistic, the remainder of the known non-linguistic systems in our set are misclassified as some form of writing.²¹ I already presented a preliminary result for Mesopotamian deity symbols in (Sproat, 2010a), where at that time I reported that Lee et al’s sys-

²⁰Note that like (Lee et al., 2010a), in performing these computations, we pad the texts with start and end symbols.

²¹Note that all of our linguistic corpora are correctly classified as linguistic, though not always as the right type of linguistic system.

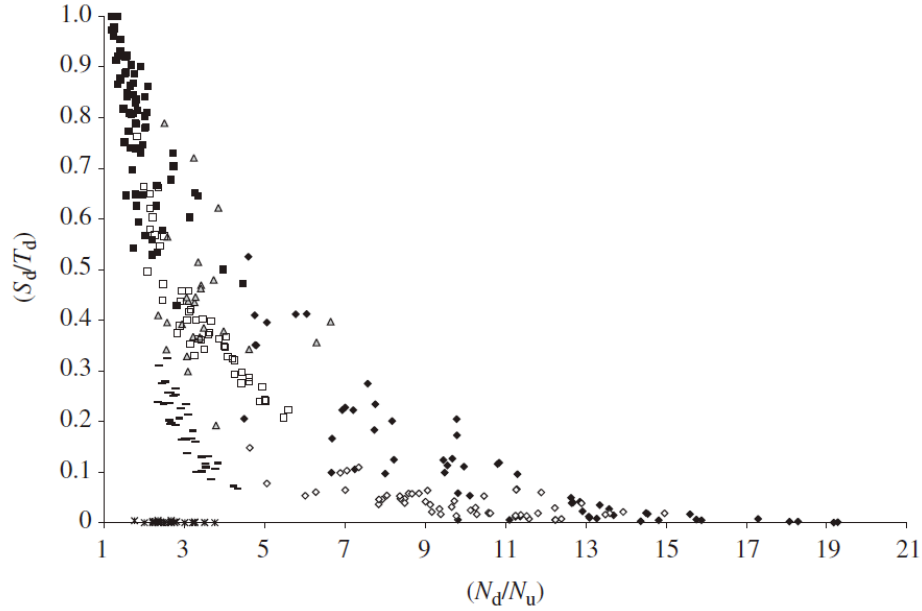


Figure 21: Reproduction of Figure 5.5, page 8, from: Lee, Rob, Philip Jonathan, and Pauline Ziman. “Pictish symbols revealed as a written language through application of Shannon entropy.” *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, pages 1–16, March 31, 2010. Used with permission of the Royal Society. See text for explanation.

tem classified it as logographic writing, not a syllabary as we have it here.²² In their reply to my commentary (Lee et al., 2010b) 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) broader definition of writing discussed above. As we have already noted, if one accepts such a broad definition, the discussion of whether something is writing or not becomes largely vacuous, and there is really nothing further to discuss.

If on the other hand by “writing” one means the kind of system that one can use to write poetry, treatises on iron smelting techniques, or the monograph that you are currently reading, and if one believes that Lee and colleagues’ method

²²This is the result of now working with a larger corpus, as well as (no doubt) some clean up and correction of the data. That in turn points to the sensitivity of any such measure to slight changes in the data, particularly with small sample sizes.

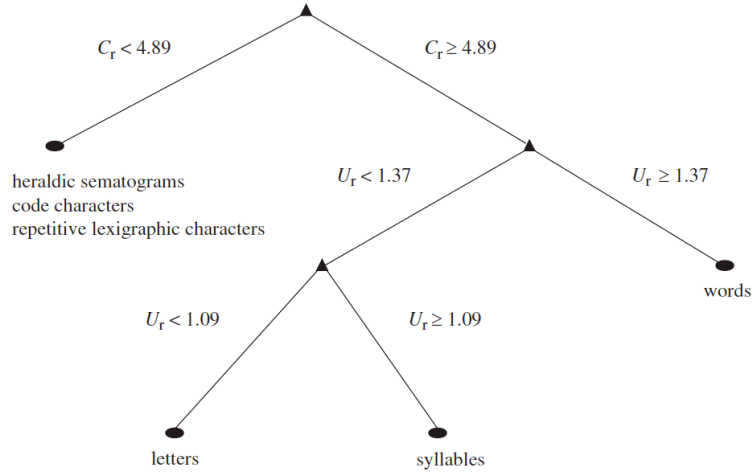


Figure 22: Reproduction of Figure 6, page 9, from: (Lee et al., 2010a).

is supposed to tell you whether or not you are looking at such a system, then it should be somewhat disturbing that, excluding Pictish, it misclassifies five out of six known non-linguistic systems as linguistic.

But, no doubt, Lee and his colleagues would be mostly satisfied with this result: as already noted. Indeed, of the six systems listed in Table 4, four of them certainly convey information of one kind or another, and Vinča (classified as non-linguistic) may have been meaningless (though we do not know). Barn stars, though, are almost surely purely decorative, so it is an embarrassment to their approach that these are classified as a form of “writing”.

8.4 Models of Association

A feature of language is that certain pairs of words tend to be more strongly associated with each other than they are with others. For example, pecan is more associated with pie than it is with appendectomy. Could association measures, computed over an entire corpus, be informative about whether a symbol system is linguistic or not?

Many association measures have been proposed in the computational linguistics literature. Here we use a very simple one, pointwise mutual information (Shannon, 1948; Church and Gale, 1991), which simply measures how common a particular pair of words is relative to

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 4: The results of applying Lee et al.’s (2010a) decision tree from Figure 22 to our data.

how common one would expect that pair to be if there were no particular association between them.

As with entropic measures, this association measure is not useful at distinguishing between linguistic and non-linguistic systems.

So far we have looked at measures that, one way or another, 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 (Shannon, 1948; Church and Gale, 1991) for two terms t_i and t_j :

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

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

In this study we 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.

We therefore look at the mean association of the n most frequent symbols,

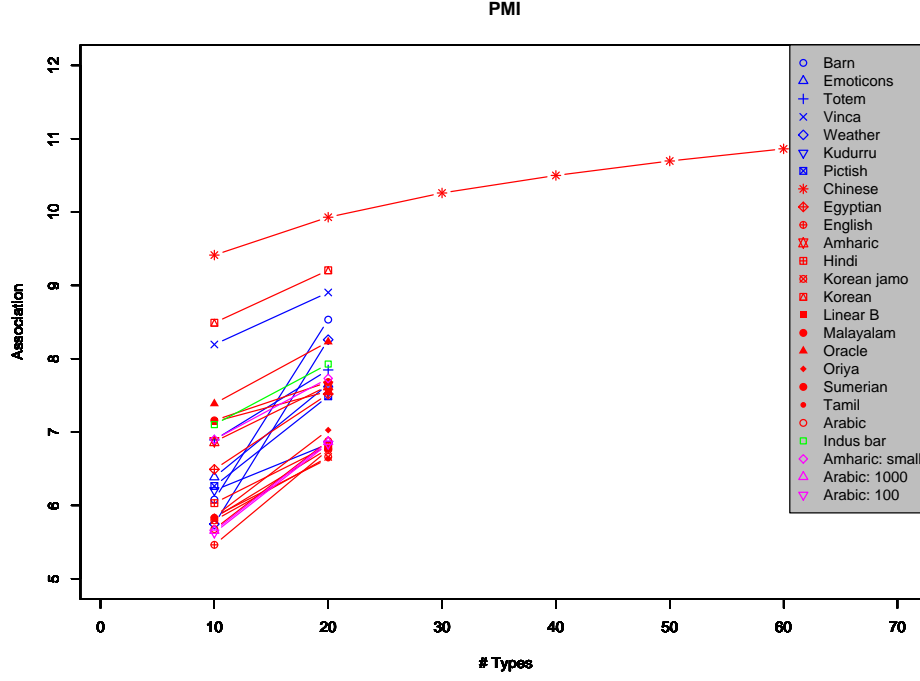


Figure 23: PMI association computations for our corpora, computed over subsets of each corpus starting with the 10 most frequent symbols, the 20 most frequent, and so forth, up to 25% of the corpus. See the text for more detailed explanation.

normalized by the number of association measures computed:

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

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 we can see, and as with previous 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. For example, in the case of Mesopotamian deity symbols there is a particularly strong association between the “horned crown” and the “symbol base”; see again Figure 3.

8.5 Models of Repetition

All of the measures we have discussed so far share one crucial property: they are local, in that they involve statistics computed over adjacent symbols or symbol sequences. But language — and other symbol systems, have many non-local properties too. For example, it is typical that certain words will repeat in a text. If you have read this far in this monograph, for example, you will have seen about 1,000 instances of the word the. With one metalinguistic exception, none of these thes are adjacent to each other and few of them are even very close to each other.

Could a measure of how often symbols repeat in text, whether or not those repetitions are adjacent, be useful in distinguishing between linguistic and non-linguistic symbol systems? We develop a simple measure of repetition and show that of all the measures so far, this one is by far the most informative. However, at least part of its value comes from the fact that the measure correlates with another even simpler feature: mean text length. While repetition rate is still a better predictor than mean text length alone, it is of note that mean text length is so informative. We return to this point in the conclusion.

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 (1948), 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 24.

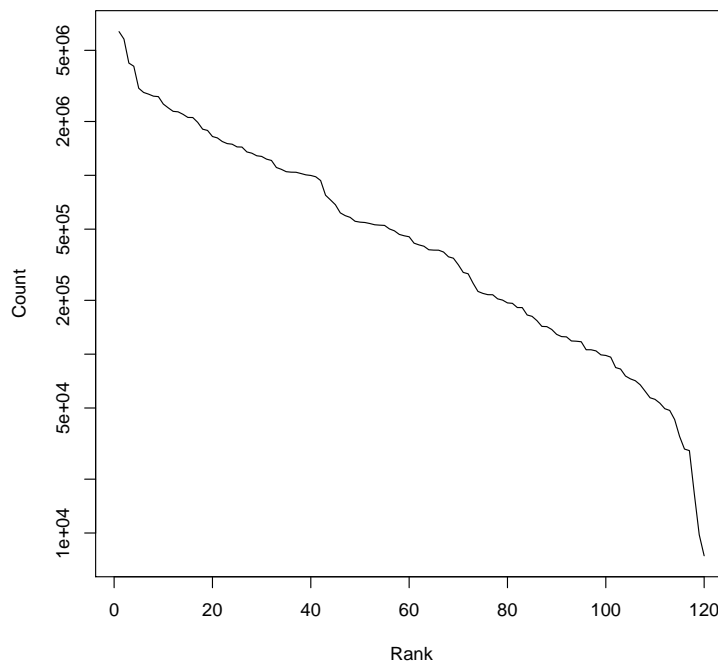


Figure 24: Unigram distribution of awarded boyscout merit badges, from http://meritbadge.org/wiki/index.php/Merit_Badges_Earned, showing the standard power-law inverse log-linear relationship between count and rank.



Figure 25: Indus seal impression M-314, consisting of seventeen symbols with no repeats.

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, 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, show not a single repetition of a symbol (Figure 25).

However it is not only the presence or absence of repetition in a corpus, but how the repetition manifests itself, that is important. In (Farmer et al., 2004) 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, it is often of the form found in Figure 26, with symmetric and other patterns.

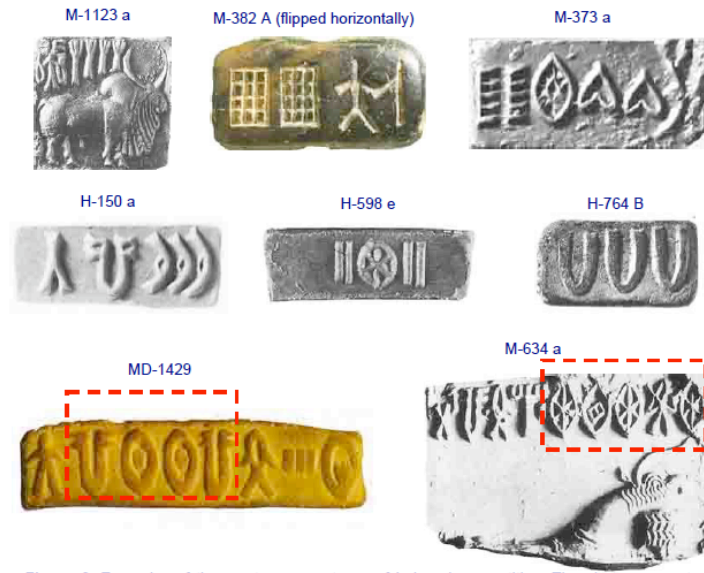


Figure 26: Indus repetitions, from (Farmer et al., 2004), Figure 6. As described there these are: “Examples of the most common types of Indus sign repetition. . . . The most frequent repeating Indus symbol is the doubled sign illustrated in M-382 A, which is sometimes claimed to represent a field or 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 we use a measure of repetition that seeks to quantify the rate of iterated repeats of the kind seen in some cases in Figure 26, relative to the total number of repetitions found. 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 As are repeats, and one B), and r would be 1 (since only one of the repeated As is adjacent to a previous A). Thus the repetition rate $\frac{r}{R}$ would be 0.33. Table 5 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 we set 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.

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

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

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 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 we have artificially limited the texts to length no

Corpus	$\frac{r}{R}$
Barn Stars	0.86
Weather Icons	0.79
<i>Sumerian</i>	0.67
Totem Poles	0.63
Vinča	0.59
Indus bar seals	0.58
Pictish	0.26
Asian emoticons	0.10
Egyptian	0.10
<i>Mesopotamian deity symbols</i>	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 5: Repetition rate $\frac{r}{R}$ for the various corpora.

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 6: 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.

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 6, the separation is not as clean as in the case of Table 5. 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.

8.6 Two-Way Classification

So far in Section 8, we have been doing exploratory data analysis, to see which measures might be useful in distinguishing between linguistic and non-linguistic systems.

In this section we combine these features into a decision tree classifier whose task is to make a two-way classification: linguistic or non-linguistic. A classification decision tree is a simple representation of structured knowledge where each node in the tree represents a question, branches represent answers to those questions, and the leaves the ultimate classification answer. One starts at the root of the tree, answers the question at the root, then depending on the answer takes one or another branch. One then repeats the process until one arrives at a leaf node.

Many algorithms exist for training such trees, given a set of features and a set of training data. The particular algorithm we use here is Classification and Regression Trees (CART) (Breiman et al., 1984).

We perform an exhaustive set of analyses of our data with a variety of different settings for the training sets. One of the upshots of these experiments is that one of Lee et al. (2010a)’s measures — C_r , when appropriately retrained, does prove somewhat useful as a discriminative feature. Sadly though, from the point of view of all analyses performed, both Pictish and Indus symbols look non-linguistic.

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:

- 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 Figure 16
- F_2 , the \log_2 bigram conditional entropy for the whole corpus, from (Lee et al., 2010a).
- U_r
- C_r
- Repetition measure

Since the set of corpora is small — only 21 in total, not counting the Indus bar seals — we 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).

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.²³

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 easier to classify and thus helpful for testing? Tables 7 and 8 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 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 we just described, we included Pictish among the non-linguistic systems. Of course, with the publication of (Lee et al., 2010a), this classification has become somewhat controversial. What if we do not include the Pictish data? In this case we used 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 9 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

²³To show that this result is no artifact of the particular setup we are using, 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, though see below for further discussion on this point.

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

tree in Figure 22, it is C_r which was involved in the the top-level linguistic/non-linguistic decision.

In Table 10 we show the results of the Wilcoxon signed rank test for each feature comparing for the two populations of linguistic and non-linguistic corpora. Only for our 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 our 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 11, 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

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

dramatic as when we 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 we replaced the repetition rate $\frac{r}{R}$ from Table 5 with that computed over the artificially shortened corpora seen above in Table 6 . 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 12. Repetition is far less dominant than it is in the original tree set shown in Table 9, 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 our measure of repetition and C_r , and 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

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 9: 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 10: Wilcoxon signed rank test for each feature with the two populations being linguistic and non-linguistic corpora.

# 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 11: 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 12: Features used by trees when repetition is replaced by the repetition rate computed over artificially shortened corpora (Table 6). The fourth row consists of twelve trees where there is a single leaf node with the decision “linguistic”.

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 27 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:

$$C_r = \frac{N_d}{N_u} + a \frac{S_d}{T_d} \quad (7)$$

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 recall (footnote 20) that Lee et al. pad the texts with beginning and end of text delimiters. 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 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 Lee and colleagues’ (2010a) 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 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 , 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.

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 non-linguistic; if they are long they are probably linguistic. Of course such a trivial computation would hardly get one’s paper published in *Proceedings of the Royal Society* (or *Science*). But it seems that the fancier methods used in Lee and colleagues’ work are in large part proxies for this far more basic measure. We will return to the implications of this result in Section 9.

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).

8.7 Two-way Classification with Additional Corpora

In this section we augment the results of the previous section by incorporating two additional corpora which are still under development: mathematical formulae and heraldic blazon. The results, after incorporating these new corpora, are essentially identical to those of the previous section.

Our corpora of mathematical formulae and heraldry are still under development. However, and putting aside the serious reservations about the methods that we have just discussed, it is of interest to see how the statistical methods hold up when samples of those corpora are added. To that end we added 1,000 “texts” from each of these corpora.

For mathematical formulae, we selected equations that consisted of single lines in the original L^AT_EX. We processed the L^AT_EX code so that commands (a string of alphanumeric characters preceded by a backslash) were kept as single tokens; square brackets, angle brackets and parentheses were split off as separate tokens; subscript (underbar) and superscript (wedge) were also split off as separate tokens; and all other alphanumeric sequences were split into sequences of character tokens.

For heraldry we selected 500 texts from each of Burke and the Mitchell rolls. From each blazon’s XML markup, we extracted all terms, except those marked as “grammatical” (usually articles such as *a* or *the*), or conjunction (*and*). Furthermore, whenever a count was specified (*three roundels*, *four lions*, etc.), we replaced that phrase with the charge, etc., repeated the appropriate number of times. Thus *four lions* is replaced with *lion lion lion lion*. This yields a more accurate representation of the actual symbols used in the arms. Note that the repetition rate $\frac{r}{R}$ of the heraldry corpus is quite high (0.71) whereas for the mathematics corpus it is

Corpus	# Texts	# Tokens	# Types	Mean text length
Mathematical expressions	1,000	23,312	273	23.31
Heraldry	1,000	9,788	410	9.79

Table 13: Number of texts, type and token counts, and mean text length for subsets of the mathematics and heraldry corpora.

low (0.027). For mathematics, it is relatively rare that the same symbol will follow itself, hence the low rate of local repetitions. On the other hand, given our processing of the heraldry corpus so that *four lions* becomes *lion lion lion lion*, a large proportion of the repetitions involve local repetitions, so $\frac{r}{R}$ is high.

Basic statistics for these corpora are given in Table 13. Figure 28 shows the block entropy curves as in Figure 20, with these two new corpora added. As can be seen, both mathematical formulae and heraldry fall somewhere in the middle of the distribution.

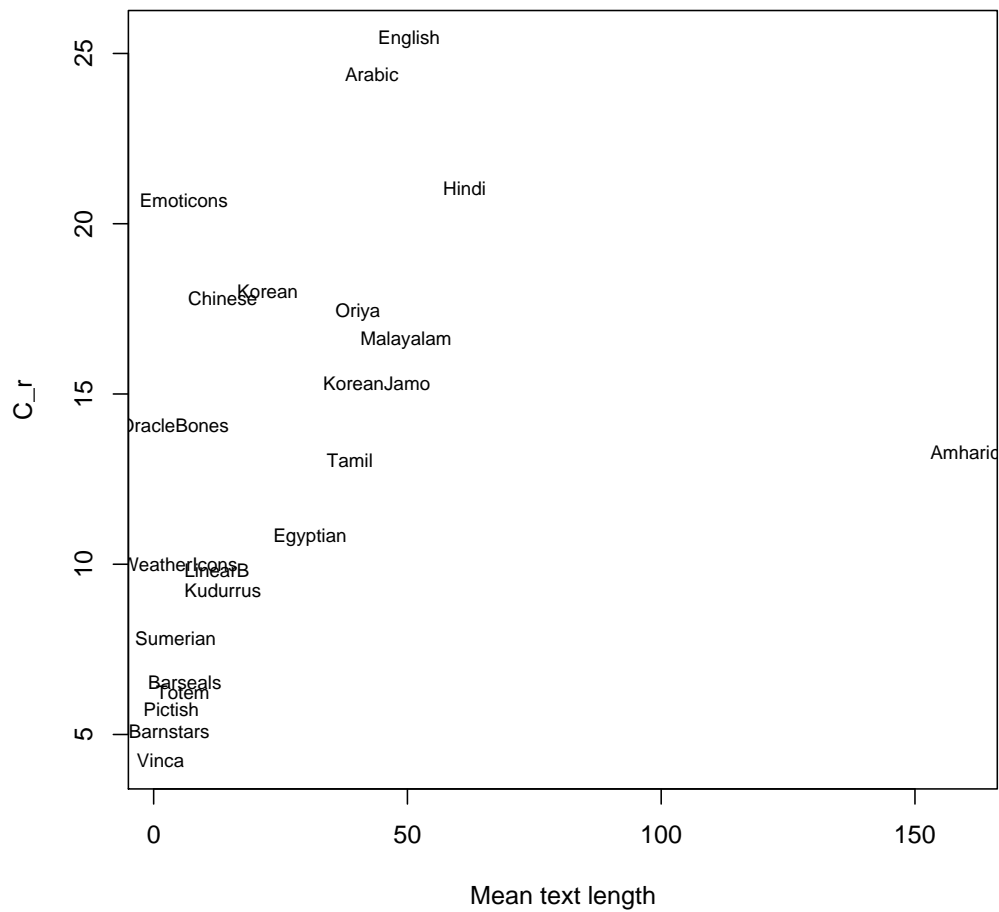


Figure 27: Correlation between C_r and text length. Pearson's r is 0.39, but when the obvious outlier Amharic is removed, the correlation jumps to 0.71.

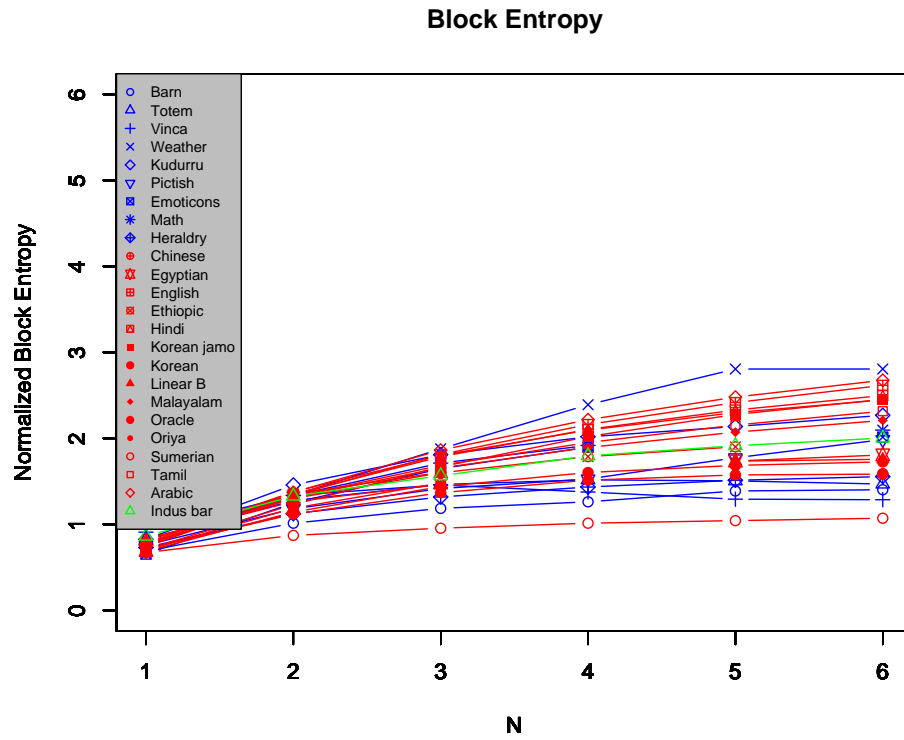


Figure 28: Block entropy values for our linguistic and non-linguistic corpora, including mathematical formulae and heraldry, and the Indus bar seal corpus.

When we replicated the classification experiments reported in Section 8.6 with these additional corpora, and using all features, the results were much as before: of the 100 trees trained, 94 classified the Indus bar seals as non-linguistic (mean accuracy 0.73) and 6 classified it as linguistic (mean accuracy 0.33).

9 Summary and Future Work

In this monograph we have presented a set of corpora of non-linguistic symbol systems that we have developed and compare 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. We also included one system (barn stars) that is almost certainly purely decorative (thus addressing a common misconception about the claims of Farmer et al. (2004)); as well as some modern systems that can easily be harvested from the Web. We also picked a wide variety of linguistic symbol systems, ranging across both time and script type.

We have examined a variety of statistical measures and shown that two of them — Lee and colleagues’ (2010a) C_r and our 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 large measure 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 (1924) 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.²⁴ The “lost manuscript” hypothesis and its ramifications was discussed extensively in (Farmer et al., 2004), 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* (IP-Syn) 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.

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 (Farmer et al., 2004).

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.

²⁴Indeed, while it ranked fourth out of five in our informal survey in Section 3, text length was nonetheless considered to be “somewhat important” as a feature in determining if a system is writing or not.

One possible avenue is automated grammar induction — see, e.g. (Klein, 2005; Clark and Lappin, 2011; Cohen, 2011).²⁵ When one claims that a symbol system is linguistic, one is implicitly claiming that one will find evidence of *linguistic* structure. It is not enough that one can show evidence for structure, as people have argued for the Indus symbols at least since the work of computational linguists like Koskenniemi (1981). After all mathematical equations have structure. What is needed is a demonstration that the structure involves things like segments grouped into syllables, syllables into morphemes or words, and words into noun-phrases, verb phrases, and so forth. This is a tall order to be sure, but without such a demonstration — or without, on the other hand, a convincing decipherment — it is pretty hard to argue that one is dealing with a linguistic system.

Further work would extend the taxonomy of non-linguistic systems that we started in Section 4. In terms of the kinds of things they can denote, there is a much wider range of non-linguistic systems than linguistic systems. We have included here a variety of systems covering several types in our taxonomy. But there are certainly other kinds, and future work will need to develop corpora that are classified into a wider set of groups than what we have developed here.

For contentious systems — the Indus Valley symbols, and to a lesser extent Pictish 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. We do hope however that we 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.

²⁵Some preliminary experiments with a grammar induction system yielded unimpressive results, but there is much further work to be done along those lines.

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 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 (Wu et al., 2012).

I also thank audiences at the Linguistic Society of America 2012 annual meeting in Portland, at the Center for Spoken Language Understanding, and at the Oriental Institute at the University of Chicago 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.

Nate Bodenstab helped out with training the BUBS parser on blazon.

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 also thank the staff of the Asian Reading Room at the Library of Congress for help with the Naxi pictograph collection.

I thank Brian Roark, Shalom Lappin 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

- Barbeau, M., 1950. *Totem Poles*. Vol. 1–2 of *Anthropology Series 30*, *National Museum of Canada Bulletin 119*. National Museum of Canada, Ottawa.
- Basso, K., Anderson, N., 8 June 1973. A Western Apache writing system: The symbols of Silas John. *Science* 180 (4090).
- Bedrick, S., Beckley, R., Roark, B., Sproat, R., June 2012. Robust kaomoji detection in Twitter. In: *Workshop on Language and Social Media*. Montreal, Canada.
- Black, J., Green, A., Rickards, T., 1992. *Gods, Demons and Symbols of Ancient Mesopotamia*. University of Texas Press, Austin, TX.
- Bliss, C., 1965. *Semantography*. Semantography (Blissymbolics) Publications, Sydney.
- Bodenstab, N., Dunlop, A., Hall, K., Roark, B., 2011. Adaptive beam-width prediction for efficient CYK parsing. In: *ACL/HLT 2011*. pp. 440–449.
- Bouissac, P. (Ed.), 1998. *Encyclopedia of Semiotics*. Oxford University Press, New York.
- Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA.
- British Columbia Provincial Museum, 1931. British Columbia totem poles. Printed by Charles F. Banfield printer to the King’s most excellent majesty.
- Burke, B., 1884. *The General Armory of England, Scotland, Ireland, and Wales; Comprising a Registry of Armorial Bearings from the Earliest to the Present Time*. Harrison & Sons, London.
- Church, K. W., Gale, W., 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.
- Clark, A., Lappin, S., 2011. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons, Malden, MA.
- Cohen, S., 2011. Computational learning of probabilistic grammars in the unsupervised setting. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

- Daniels, P., Bright, W. (Eds.), 1996. *The World's Writing Systems*. Oxford, New York.
- DeFrancis, J., 1984. *The Chinese Language: Fact and Fantasy*. University of Hawaii Press, Honolulu, HI.
- DeFrancis, J., 1989. *Visible Speech: The Diverse Oneness of Writing Systems*. University of Hawaii Press, Honolulu, HI.
- Drew, F., 1969. *Totem Poles of Prince Rupert*. F.W.M. Drew, Prince Rupert, BC.
- Eco, U., 1976. *A Theory of Semiotics*. Harvard University Press, Cambridge, MA.
- Farmer, S., Sproat, R., Witzel, M., 2004. The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies* 11 (2).
- Farnell, B., 1996. Movement notation systems. In: Daniels, P., Bright, W. (Eds.), *The World's Writing Systems*. Oxford, New York, pp. 855–879.
- Feldman, R. D., 2003. *Home before the Raven Caws: the Mystery of the Totem Pole*. Cincinnati, OH. Emmis Books.
- Fischer, S. R., 1997. *Rongorongo: The Easter Island Script*. Oxford University Press, Oxford.
- Fox-Davies, A. C., 2000. *A Complete Guide to Heraldry*. Adamant Media Corporation, Boston.
- Friar, S., Ferguson, J., 1993. *Basic Heraldry*. Herbert Press, London.
- Garfield, V. E., 1940. *The Seattle Totem Pole*. University of Washington Press, Seattle.
- Gelb, I., 1952. *A Study of Writing: The Foundations of Grammarology*. University of Chicago, Chicago.
- Goodman, F., 1989. *Magic Symbols*. Brian Trodd Publishing, London.
- Graves, T. E., 1984. The Pennsylvania German hex sign: A study in folk process. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.
- Gunn, S. W., 1965. *The Totem Poles in Stanley Park, Vancouver, B.C.*, 2nd Edition. Macdonald, Vancouver, BC.

- Gunn, S. W., 1966. *Kwakiutl House and Totem Poles at Alert Bay*. White rocks Publications, Vancouver, BC.
- Gunn, S. W., 1967. *Haida Totems in Wood and Argillite*. White rocks Publications, Vancouver, BC.
- Harris, R., 1995. *Signs of Writing*. Routledge, London.
- Hawkins, J. D., 2000. *Corpus of Hieroglyphic Luwian Inscriptions*. Walter de Gruyter, Berlin.
- Helmont, F. M. v., 1667. *Alphabeti vere naturalis hebraici brevissima delineatio*. Sulzbach, translated with an introduction and annotations by Allison P. Coudert & Taylor Corse, Brill, Leiden, 2007.
- Jackson, A., 1984. *The Symbol Stones of Scotland: a Social Anthropological Resolution of the Problem of the Picts*. Orkney Press.
- Jackson, A., 1990. *The Pictish Trail*. Orkney Press.
- Kaiser, D., 2005. Physics and Feynman's diagrams. *American Scientist* 93, 156–165.
- King, L., 1912. *Babylonian Boundary Stones and Memorial Tablets in the British Museum*. British Museum, London.
- Kiraz, G., 2012. *Tūrrāṣ Mamllā: A Grammar of the Syriac Language. Volume 1: Orthography*. Gorgias Press, Piscataway, NJ.
- Klein, D., 2005. The unsupervised learning of natural language. Ph.D. thesis, Stanford University, Stanford, CA.
- Kneser, R., Ney, H., 1995. Improved backing-off for m-gram language modeling. In: *International Conference on Speech and Signal Processing*. pp. 181–184.
- Koskeniemi, K., 1981. Syntactic methods in the study of the Indus script. *Studia Orientalia* 50, 125–136.
- Lacadena, A., 2008. Regional scribal traditions: Methodological implications for the decipherment of Nahuatl writing. *The PARI Journal: A quarterly publication of the Pre-Columbian Art Research Institute* 8 (4), 1–22.
URL <http://www.mesoweb.com/pari/publications/journal/804/PARI0804.pdf>

- Le Novère, N., et al, 2009. The systems biology graphical notation. *Nature Biotechnology* 27, 735–741.
 URL <http://www.nature.com/nbt/journal/v27/n8/full/nbt.1558.html>
- Lee, R., Jonathan, P., Ziman, P., 2010a. Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, 1–16.
- Lee, R., Jonathan, P., Ziman, P., 2010b. A response to Richard Sproat on random systems, writing, and entropy. *Computational Linguistics* 36 (4), 791–794.
- Li, L., 2001. *Naxizu xiangxing biao yin wenzi zidian*. Yunnan Minzu Chubanshe.
- Mack, A., 1997. *Field Guide to the Pictish Symbol Stones*. The Pinkfoot Press, Balgavies, Angus, updated 2006.
- Mahr, A. C., 1945. Origin and significance of Pennsylvania Dutch barn symbols. *Ohio History: The Scholarly Journal of the Ohio Historical Society* 54 (1), 1–32, <http://publications.ohiohistory.org/ohstemplate.cfm?action=toc&vol=54>.
- Malin, E., 1986. *Totem Poles of the Pacific Northwest Coast*. Timber Press, Portland, OR.
- Mallery, G., 1883. *Pictographs of the North American Indians: A Preliminary Paper*. No. 4 in Annual Report of the Bureau of Ethnology. Smithsonian Institution, Washington, DC, available from Google Books.
- Manning, C., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- Marshall, J., September 20 1924. First light on a long forgotten civilization. *Illustrated London News*, 528–32, 548.
- McCawley, J., 1996. Music notation. In: Daniels, P., Bright, W. (Eds.), *The World's Writing Systems*. Oxford, New York, pp. 847–854.
- Nemenman, I., Shafee, F., Bialek, W., 2002. Entropy and inference, revisited. In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press, Cambridge, MA.
- Parpola, A., 1994. *Deciphering the Indus Script*. Cambridge University Press, New York.

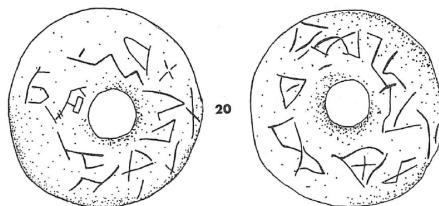
- Peirce, C., 1934. *Collected Papers: Volume V. Pragmatism and Pragmaticism*. Harvard University Press, Cambridge, MA.
- Powell, B., 2009. *Writing: Theory and History of the Technology of Civilization*. Wiley-Blackwell, Chichester.
- Rao, R., 2010. Probabilistic analysis of an ancient undeciphered script. *IEEE Computer* 43 (3), 76–80.
- Rao, R., Yadav, N., Vahia, M., Joglekar, H., Adhikari, R., Mahadevan, I., 2009a. Entropic Evidence for Linguistic Structure in the Indus Script. *Science* 324 (5931), 1165.
- Rao, R., Yadav, N., Vahia, M., Joglekar, H., Adhikari, R., Mahadevan, I., August 5 2009b. A Markov model of the Indus script. *Proceedings of the National Academy of Sciences* 106 (33), 13685–13690.
- Rao, R., Yadav, N., Vahia, M., Joglekar, H., Adhikari, R., Mahadevan, I., 2010. Entropy, the indus script, and language: A reply to R. Sproat. *Computational Linguistics* 36 (4), 795–805.
- Rhys, J., 1892. The inscriptions and language of the northern Picts. *Proceedings of the Society of Antiquaries of Scotland* 26, 263–351.
- Roark, B., Sproat, R., Allauzen, C., Riley, M., Sorensen, J., Tai, T., 2012. The OpenGrm open-source finite-state grammar software libraries. In: *Proceedings of the Association for Computational Linguistics, System Demonstrations*. Association for Computational Linguistics, Jeju Island, South Korea, pp. 61–66.
- Royal Commission on the Ancient and Historical Monuments of Scotland, 1994. Pictish symbol stones: A handlist.
- Saussure, F. d., 1916. *Cours de linguistique générale*. Lausanne, Paris.
- Scarborough, H., 1990. The index of productive syntax. *Applied Psycholinguistics* 11, 1–22.
- Sebeok, T. (Ed.), 1977. *A Perfusion of Signs*. Indiana University Press, Bloomington, IN.
- Seidl, U., 1989. *Die babylonischen Kudurru- Reliefs. Symbole mesopotamischer Gottheiten*. Universitätsverlag Freiburg.
- Shannon, C., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423.

- Slater, S., 2002. *The Complete Book of Heraldry*. Lorenz Books, London.
- Smith, J., 1996. Japanese writing. In: Daniels, P., Bright, W. (Eds.), *The World's Writing Systems*. Oxford, New York, pp. 209–217.
- Sproat, R., 2000. *A Computational Theory of Writing Systems*. Cambridge University Press, Cambridge.
- Sproat, R., 2010a. Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics* 36 (3).
- Sproat, R., 2010b. Reply to Rao et al. and Lee et al. *Computational Linguistics* 36 (4), 807–816.
- Stewart, H., 1990. *Totem Poles*. University of Washington, Seattle, WA.
- Stewart, H., 1993. *Looking at Totem Poles*. University of Washington, Seattle, WA.
- Sutherland, E., 1997. *The Pictish Guide*. Birlinn Ltd, Edinburgh.
- Swinney, D., 1979. Lexical access during sentence comprehension: (re)consideration of contextual effects. *Journal of Verbal Learning and Verbal Behavior* (5), 219–227.
- Turner, V., 1967. *The Forest of Symbols: Aspects of Ndembu Ritual*. Cornell University Press, New York.
- Ventris, M., Chadwick, J., 1956. *Documents in Mycenaean Greek*. Cambridge University Press, Cambridge.
- Vidale, M., 2007. The collapse melts down: a reply to Farmer, Sproat and Witzel. *East and West* 57, 333–366.
- Waddington, C., 1974. Horse brands of the Mongolians: A system of signs in a nomadic culture. *American Ethnologist* 1 (3), 471–488.
- Winn, S., 1990. A Neolithic sign system in southeastern Europe. In: Foster, M. L. (Ed.), *The Life of Symbols*. Westview Press, Boulder, pp. 269–271.
- Winn, S. M. M., 1981. *Pre-Writing in Southeastern Europe: The Sign System of the Vinča Culture, ca. 4000 B.C.* Western Publishers.
- Wu, K., Solman, J., Linehan, R., Sproat, R., January 2012. Corpora of non-linguistic symbol systems. In: *Linguistic Society of America*. Portland, OR.
URL <http://elanguage.net/journals/lsameeting/article/view/2845/pdf>

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

Appendix: Sample Texts and Transcriptions

Tordos Spindle Whorl #20, (Winn, 1981, page 270)



```
<document type="Vinca" region="Tordos" class="Spindle">
  <description>Tor 20</description>
  <docText>
    <side number="A">
      <circle>
        <symbol><title>118</title></symbol>
        <symbol><title>106</title></symbol>
        <symbol><title>66</title></symbol>
        <symbol><title>95</title></symbol>
        <symbol><title>66</title></symbol>
        <symbol><title>105</title></symbol>
        <symbol><title>7</title></symbol>
        <symbol><title>180</title></symbol>
        <symbol><title>12</title></symbol>
        <symbol><title>9</title></symbol>
        <symbol><title>157</title></symbol>
        <symbol><title>178</title></symbol>
      </circle>
    </side>
    <side number="B">
      <circle>
        <symbol><title>155</title></symbol>
        <symbol><title>113</title></symbol>
        <symbol><title>95</title></symbol>
        <symbol><title>97</title></symbol>
        <symbol><title>110</title></symbol>
        <symbol><title>184</title></symbol>
        <symbol><title>108</title></symbol>
        <symbol><title>106</title></symbol>
      </circle>
    </side>
  </docText>
</document>
```

Kudurru stone #67, from (Seidl, 1989), plate 23.



a. Nr. 67 Nabû-kudurri-uṣur I.

```
<document type="Kudurru" group="6">
  <description>67, Nabu-kudurri-usur I. Zeit</description>
  <docText>
    <line number="1">
      <symbol><title>Schlange</title>
        <description>goes along the side of the entire stone,
          all the lines</description></symbol>
      <symbol><title>Stern</title></symbol>
      <symbol><title>Mondsichel</title></symbol>
      <symbol><title>Sonnenscheibe</title></symbol>
    </line>
    <line number="2">
      <symbolUnit>
        <symbol><title>Hoernerkrone</title></symbol>
        <symbol><title>Symbolsockel</title></symbol>
      </symbolUnit>
      <symbolUnit>
        <symbol><title>Hoernerkrone</title></symbol>
        <symbol><title>Symbolsockel</title></symbol>
      </symbolUnit>
      <symbolUnit>
        <symbol><title>Hoernerkrone</title></symbol>
        <symbol><title>Symbolsockel</title></symbol>
      </symbolUnit>
    </line>
    <line number="3">
      <symbolUnit>
        <symbol><title>Spaten</title></symbol>
        <symbol><title>Symbolsockel</title></symbol>
      </symbolUnit>
    </line>
  </docText>
</document>
```

```

    <symbol><title>Schlangendrache</title></symbol>
  </symbolUnit>
  <symbolUnit>
    <symbol><title>Schreibgeraet</title></symbol>
    <symbol><title>Symbolsockel</title></symbol>
  </symbolUnit>
  <symbolUnit>
    <symbol><title>Band</title></symbol>
    <symbol><title>Symbolsockel</title></symbol>
    <symbol><title>Ziegenfisch</title></symbol>
  </symbolUnit>
</line>
<line number="4">
  <symbol><title>Adlerstab</title></symbol>
  <symbolUnit>
    <symbol><title>Widderstab</title></symbol>
    <symbol><title>Loewenstab</title></symbol>
  </symbolUnit>
  <symbol><title>Pferdekopf</title></symbol>
  <symbol><title>Vogel-auf-d.-Stange</title></symbol>
</line>
<line number="5">
  <symbolUnit>
    <symbol><title>Symbolsockel</title></symbol>
    <symbol><title>Hund</title></symbol>
  </symbolUnit>
</line>
<line number="6">
  <symbol><title>Schildkroete</title></symbol>
  <symbol><title>Lampe</title></symbol>
</line>
<line number="7">
  <symbolUnit>
    <symbol><title>Blitzbuendel</title></symbol>
    <symbol><title>Rind</title></symbol>
  </symbolUnit>
  <symbol><title>Skorpion</title></symbol>
</line>
</docText>
</document>

```

The Aberlemno 1 stone from <http://www.stams.strath.ac.uk/research/pictish/database.php>



```
<document type="PictishStone" region="Angus" class="I">
  <description>Aberlemno 1</description>
  <docText>
    <line number="1">
      <symbol><title>Serpent</title></symbol>
    </line>
    <line number="2">
      <symbolUnit>
        <symbol><title>Double-Disc</title></symbol>
        <symbol><title>Z</title></symbol>
      </symbolUnit>
    </line>
    <line number="3">
      <symbol><title>Mirror</title></symbol>
      <symbol><title>Comb</title></symbol>
    </line>
  </docText>
</document>
```

Grizzly Bear Pole of Yan, (Drew, 1969, page 16)



```
<document type="totemPole" origin="Haida">
  <description>Grizzly Bear Pole of Yan,
a house frontal pole</description>
  <page> p16 </page>
  <docText>
    <symbol><title>3-Skils</title></symbol>
    <symbol><title>Grizzly-Bear</title></symbol>
    <symbol><title>Bear-Mother</title></symbol>
    <symbol><title>Cub</title></symbol>
    <symbol><title>Cub</title></symbol>
    <symbolUnit>
  <symbol>
    <title>
      Supernatural-Grizzly-Bear
    </title>
  </symbol>
  <symbol><title>Frog</title></symbol>
  <description>Supernatural Grizzly Bear
holding a Frog</description>
    </symbolUnit>
    <symbol><title>Grizzly-Bear</title></symbol>
  </docText>
</document>
```

Farrell collection AR(2)F



```
<document type="Barn star" group="1">
  <description>Box01/AR(2)F</description>
  <docText>
    <line number="1">
      <symbol><title>WHEEL_OF_FORTUNE</title></symbol>
      <symbol><title>20_POINT_STAR</title></symbol>
      <symbol><title>WHEEL_OF_FORTUNE</title></symbol>
    </line>
  </docText>
</document>
```