A statistical comparison of written language and nonlinguistic symbol systems: Online Supplementary Materials

LANGUAGE

A JOURNAL OF THE LINGUISTIC SOCIETY OF AMERICA

VOLUME 90, NUMBER 2 JUNE 2014

PUBLISHED BY THE LINGUISTIC SOCIETY OF AMERICA

Richard Sproat

Language, Volume 90, Number 2, June 2014, pp. s1-s27 (Article)

Published by Linguistic Society of America
DOI: https://doi.org/10.1353/lan.2014.0042

A STATISTICAL COMPARISON OF WRITTEN LANGUAGE AND NONLINGUISTIC SYMBOL SYSTEMS: ONLINE SUPPLEMENTARY MATERIALS

RICHARD SPROAT
Google, Inc.

1. INTRODUCTION. In this supplement, I describe the corpora used in deriving the results discussed in the main article.

Since the notion of a nonlinguistic symbol system may be unfamiliar to some readers (even though everyone uses such systems every day), I start with a brief taxonomy of such systems as a segue into the more technical discussion to follow.

2. TAXONOMY. Linguistic symbol systems—writing—have been classified in various ways by scholars of writing systems for the past half century (Gelb 1952, Sampson 1985, DeFrancis 1989, Daniels & Bright 1996, Sproat 2000, Rogers 2005). While the proposed classifications differ, they all share the property that they differentiate systems on the basis of the kind of information encoded in the system, whether it be phonological segments, syllables, morphemes, or words.

Systematic taxonomies of NONLINGUISTIC systems have certainly not been as prevalent and in fact may not exist. Where such systems are discussed—for example, Harris 1995, Daniels & Bright 1996—it is usually just a few systems that are presented in juxtaposition to writing systems. In a forthcoming publication (Sproat 2014), I outline and justify a preliminary taxonomy of nonlinguistic systems. Here I merely summarize the taxonomy, with examples, to give some context for the ensuing discussion.

Nonlinguistic systems may be classified into at least the following types:

- SIMPLE INFORMATIVE SYSTEMS. In simple informative systems, the symbols by and large convey a single piece of information. The helical red, white, and blue barber pole, for example, indicates the presence of barber shop. In weather reports, icons are used to represent various states of the weather, such as whether it is sunny, partly cloudy, raining, and so forth. Further examples: other symbols of guild such as three balls for a pawnbroker; institutional logos; traffic information signs; ownership signs, such as brands, for example, Mongolian horse brands (Waddington 1974), or house marks (http://en.wikipedia. org/wiki/House_mark).
- EMBLEMATIC SYSTEMS. Closely related to simple informative systems are emblematic systems, where the symbols represent some special distinction earned by the bearer. Examples: symbols of military rank and distinction, scouting merit badges, Phi Beta Kappa keys, and symbols of other scholarly fraternities; letter grades on academic assignments or in courses.
- RELIGIOUS ICONOGRAPHY. Religious iconography could also be characterized as a simple informative system, except that here the notion of 'single piece of information' is much less clear, since such symbols are intentionally often highly multivocal in the meanings they evoke. Examples: Christian cross, star of David, star and crescent, dharmachakra, swastika (in its original Hindu/ Buddhist usage).
- HERALDIC SYSTEMS. Heraldic systems are similar to emblematic systems in that they usually represent a particular set of features of the bearer, including possibly marks of distinction. They differ, however, in that heraldic systems are frequently highly combinatoric, involving 'texts' built of

many symbols, often with a quite rigid syntax. Examples: European heraldry, kudurrus, some functions of totem poles (Barbeau 1950).

- FORMAL SYSTEMS. In a formal system, the individual symbols have well-defined meanings, and there are generally strict rules on how the symbols may be combined. Examples: mathematical symbols, alchemical symbols, chemical notation, Feynman diagrams (Kaiser 2005), programming flowcharts, and Systems Biology Graphical Notation (Le Novère et al. 2009).
- PERFORMATIVE SYSTEMS. Performative systems indicate a sequence of actions to be taken to perform a particular task. Perhaps those most familiar to many people today are the wordless assembly instructions that come with furniture from Ikea. Other examples: Silas John's system for notating Apache prayers (Basso & Anderson 1973), musical notation (McCawley 1996), dance notation and other movement notation systems (Farnell 1996), chess notation, systems that can be used to indicate the sequence of plays in a game, knitting patterns (Harris 1995).
- NARRATIVE SYSTEMS OR 'PROMPT' TEXTS. Narrative systems are used to recount stories and as such are the most language-like of the nonlinguistic systems. In narrative systems, the symbols typically represent actors or events in the story in an iconic way. Examples: Dakota winter counts (Mallery 1883), (probably) Naxi symbology (Li 2001).
- 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 in the past, but those meanings are quite irrelevant in their current use. A clear modern example is the use of Chinese characters in body tattoos, worn by people who may be completely unaware of the characters' original meaning and use them solely because they look 'cool'. Other examples are Pennsylvania German barn stars (Graves 1984) and Asian emoticons (e.g. Bedrick et al. 2012), where in the latter case symbols from various scripts and other symbols are combined into a 'text' that represents an image, usually a face. The face itself may convey some sort of emotion (e.g. sadness, via depiction of crying), but other than that has no real meaning and mostly functions to decorate the surrounding text.

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

3. MATERIALS AND DATA PREPARATION. For this work I selected seven nonlinguistic symbol systems, with two additional systems under development, and fourteen linguistic symbol systems for analysis. Note that in addition to the nonlinguistic symbol systems described here, further corpora of nonlinguistic symbol systems are under development. I describe two of these below after presenting the main corpora.

The seven nonlinguistic systems discussed here are available at: http://rws.xoba.com/data/nonlinguistic-symbols.

3.1. NONLINGUISTIC CORPORA. Two kinds of nonlinguistic corpora were collected. The first is ancient or traditional systems that serve as exemplars of what a complex, ancient, nonlinguistic system can be like. If a set of 'texts' in a previously unknown symbol system is discovered among the ruins of an ancient civilization, these systems could serve as models of what that symbol system might have been, assuming it was not a form of written language. Our current set of such systems comprises Mesopotamian deity symbols (kudurrus) (Seidl 1989), Vinča symbols (Winn 1981), Pictish symbols (Jackson 1984, 1990, Royal Commission on the Ancient and Historical Monuments of Scotland 1994, Mack 1997, Sutherland 1997), totem poles (Newcombe & British Columbia Provincial Museum 1931, Garfield 1940, Barbeau 1950, Gunn 1965, 1966, 1967, Drew 1969, Malin 1986, City of Duncan 1990, Stewart 1990, 1993, Feldman 2003), and Pennsylvania German barn stars (Mahr 1945, Graves 1984, Yoder & Graves 2000). The development of these ancient or traditional corpora is described in Wu et al. 2012. Two of these systems are of interest for other reasons: kudurrus and Vinča systems were mentioned in Rao et al. 2009a as models for low- and high-entropy systems, respectively. A third—Pictish symbols—was, of course, already discussed in Lee et al. 2010 and argued to be writing: their inclusion here as nonlinguistic accords with what is probably still the most widely accepted view, but in addition to that, the tests presented in the main article—including a retrained version of one of Lee and colleagues' own measures—are at least consistent with it being nonlinguistic.

The second kind are modern systems, which can easily be collected electronically. The two examples discussed here are Asian emoticons and weather icon sequences. These, presumably, are poorer models of ancient symbol systems, but have the advantage that they are easy to collect from online sources, which was mostly not the case for the ancient or traditional systems, which usually had to be transcribed from print sources.

After discussing the individual corpora, I present in §3.1.8 statistics on the size of the various corpora. The final section of this supplement contains examples of the first five symbol systems, along with the corresponding transcription in the XML markup scheme that was introduced in Wu et al. 2012. One feature of our XML markup is that it allows some flexibility in how one extracts the elements of the text. For example, if there are clear lines in the text, then we preserve line information. If the text is circular (so that it is not defined where the beginning or end of the text occurs), that information is also marked. Furthermore, one of the questions that arises in the analysis of a symbol system is whether to treat elements that seem to be composed of more atomic symbols as single symbols or as a composite of the individual symbols. The symbolUnit (see below for an example) preserves the grouping structure so that one can make the choice of level of analysis later on. I do not discuss the XML markup system in any detail here, but refer the reader to Wu et al. 2012; examples of the XML markup system used can be found in §5 of these supplementary materials.

At the head of each section I propose a classification for the symbol system in question in terms of the taxonomy developed in §2. The breakdown of the types is as in Table S1. 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

TABLE S1. Types of symbol systems.

3.1.1. VINČA SYMBOLS. **Classification: Unknown, possibly religious.**

The Vinča were a late Neolithic people of Southeastern Europe between the sixth and the third millennia BC 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 §5 for a sample Vinča text. In such texts, the symbols are sometimes, but not always, arranged linearly as in a typical script. The most authoritative compilation of the Vinča materials is Winn 1981, which classifies the symbols into about 200 types and provides a corpus of the multisymbol texts from 123 sources, comprising more than 800 tokens. In addition to their sometimes linear arrangement, the Vinča symbols had other script-like properties: in Winn's analysis, some signs seemed to comprise two signs ligatured together. Though Winn characterizes the system as 'prewriting', it must be stressed that there is no reason to think this was a linguistic writing system. Very likely, the system had religious significance. As Winn states: 'In the final analysis, the religious system remains the principle source of motivation for the use of signs' (1981:255). However, the meaning of the symbols remains unknown.[1]

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

3.1.2. MESOPOTAMIAN DEITY SYMBOLS ON KUDURRUS. **Classification: Heraldic.**

Kudurrus were legal documents that specified property rights. The stones often also included actual cuneiform Babylonian writing. The deity symbols had an essentially heraldic function, to indicate favored de-

[1] Recent work—for example, Haarmann 2008, Haarmann & Marler 2008, Winn 2008, and other articles in volume 4 of *The Journal of Archaeomythology*—has revived the idea that the Vinča symbols were part of a broader 'Danube script' dating back to the sixth millennium BC. It is worth pointing out at the outset that even if it turned out that we have misclassified Vinča symbols as nonwriting, that change in the status of one system would not materially affect the results presented here.

But the arguments in support of a 'Danube script' seem tenuous, in any case. Haarmann, for example, presents two arguments in support of the claim. One is a process-of-elimination argument: if the system is not apparently decoration, potter's marks, religious symbology, and so forth, then it must be writing. But in the absence of knowing WHAT the symbols denoted, how can one be sure that it is not one of these things? Consider the deity symbols on kudurru stones: as argued in §3 of the main article, 'texts' in those symbols look a lot like writing—except that we know that they were not. Haarmann seems to be assuming that religious symbols cannot be strung into 'texts' that resemble writing, an assumption we know to be false. Haarmann's second argument is equally tenuous, since it depends on 'identifying properties which the sign system of the Danube civilization shares with other ancient writing systems' (2008:13). Purely formal properties of sign systems are notoriously hard to align with function, and are likely to be very misleading.

As the authors of these articles admit, the chances of 'decipherment' of these symbols are essentially null. We do not know what language or languages the people spoke. The inscriptions are all very short: the mean length of 'texts' for our sample is 1.4 symbols long, far worse even than the situation for Indus inscriptions. Finally, there is no 'Rosetta Stone'. Indeed, how could there be? These artifacts date to thousands of years before the first verifiable writing, so if this were writing there would be no known contemporaneous systems. Haarmann claims that 'a script can be identified in terms of an operational technology even without being deciphered' (2008:13), yet why should anyone believe this? There is no question that quite a few people would be happy if this turned out to be a script, yet such wishful thinking, often driven by professional or nationalistic motivations, should not be considered seriously in scientific discourse.

Why then be so equally adamant that this was not a writing system? The extreme age, the Neolithic cultural context, and the simple fact that nonlinguistic symbol systems far outnumber linguistic systems across the world all conspire to make this the null hypothesis. That, and the fact that of the hundreds of inscriptions in this system, very few really 'look like' writing.

ities of the owner of the property. Our source for the deity symbol corpus was Seidl 1989. We picked texts from all stones where the depictions in Seidl 1989 were clear enough to read. Fortunately, Seidl 1989 includes a chart that lists, for each stone, the symbols that appear on that stone (though not in the order they appear); that chart proved useful for checking that the reading of the symbols was correct.

3.1.3. Pictish stones. **Classification: Unknown, possibly heraldic.**

The Picts were an Iron Age people (or possibly several peoples) of Scotland who, among other things, left a few hundred standing stones inscribed with symbols, with 'texts' ranging from one to a few symbols in length. The meaning of the symbols is unknown, but until recently, no one to our knowledge has argued that they are any form of writing. If nothing else, the Picts evidently had at least some literacy in the (segmental) Ogham script (Rhys 1892).

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, 1990, Royal Commission on the Ancient and Historical Monuments of Scotland 1994, 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.

3.1.4. Totem poles. **Classification: Heraldic, narrative.**

Totem poles are the product of a wide range of Native American cultures of the Pacific Northwest, dating from the nineteenth and twentieth 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, such as legends, genealogical information, or depictions of important events. Types of poles include memorial poles, house frontal poles, mortuary poles, 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 Newcombe & British Columbia Provincial Museum 1931, Garfield 1940, Barbeau 1950, Gunn 1965, 1966, 1967, Drew 1969, City of Duncan 1990, Stewart 1990, 1993, and 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 Nanasimget 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 (Wu et al. 2012) to represent the grouping.

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

3.1.5. Pennsylvania german barn stars. **Classification: Decorative.**

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

There is a common belief that 'hex signs' had a magical function, such as to ward off evil spirits, and indeed this belief was sometimes expressed by the German farmers who used the symbols on their barns (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 & 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 & Graves 2000).[2] One piece of evidence for this view is that the 'texts' are nearly always symmetric.

Our barn star corpus is based upon the slide collection of W. Farrell, who toured areas around Berks County in the 1940s 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, Pennsylvania. The most useful boxes, in terms of the 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 show only 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

[2] 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 nonlinguistic systems.

Second, much of the critique of Farmer et al. 2004 has depended on a fundamental confusion as to what is meant by the term 'nonlinguistic 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 article (Vidale 2007). Using the obviously decorative pottery designs of Shahr-e Sukhteh and elsewhere as his examples, he states:

> 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 Sokhta, '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 [sic] 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. (Vidale 2007:344)

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 nonlinguistic systems, though as we see in the main article, rates of local REDUPLICATIVE repetition relative to the total repetition rate DO seem to be characteristic of many nonlinguistic 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.

three six-pointed stars (St6). Farrell, however, does not distinguish some cases that Graves distinguishes: the main example of this is rosette, which Farrell classifies under six-pointed star; see Figure S1. In such cases I followed Graves's 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 S2. The 'four-slice pie' is so distinct, however, that it seems to deserve its own category.

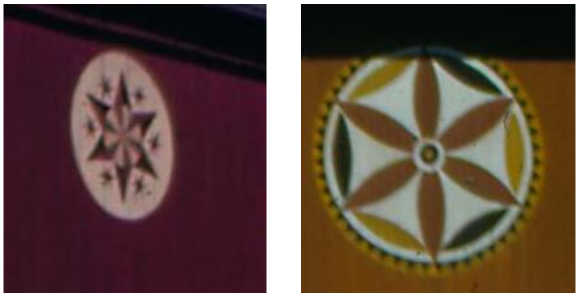


FIGURE S1. Six-pointed star (left) versus rosette.

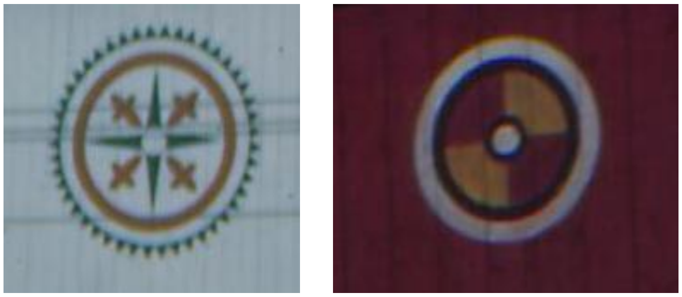


FIGURE S2. Four-pointed star (left) versus four-slice pie.

3.1.6. WEATHER ICONS. **Classification: Simple informative.**

Weather Underground (http://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, the icon for rain during the day is found at http://icons-pe.wxug.com/i/c/a/rain.gif. There are about twenty distinct icons.[3] 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 seventy-two days, giving us a corpus of 50,710 symbols.[4] Figure S3 gives an example of a weather 'text'.

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

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

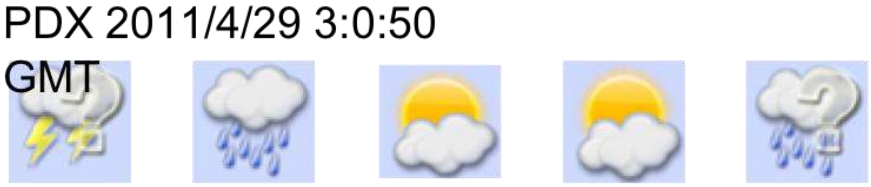


FIGURE S3. A sample weather icon sequence: forecast for Portland, Oregon, April 29, 2011.

3.1.7. ASIAN EMOTICONS. **Classification: Decorative.**

As part of a project on normalization of Twitter messages, my colleagues and I developed an analysis system to detect and parse Asian emoticons (KAOMOJI) (Bedrick et al. 2012). Unlike the familiar ninety-degree flipped ASCII 'smileys'— :-), ;-), :-(, 8-), and so forth—Asian emoticons (so-called because they were popularized by Japanese and other East Asian users) are oriented horizontally and make use of a much wider range of characters. Some examples can be seen in Figure S4. 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.

|^_^|	(๑•̑ت•̑๑)
[o_-]	(╏̆-╎̆ใ)
\(^v^)/	(*≧m≦*)

FIGURE S4. Some representative kaomoji emoticons.

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. Unlike in the symmetric 'texts' found with Pennsylvania barn stars, however, the mate characters found in Asian emoticons are different symbols, chosen because they are visually close mirror images. A statistical analysis of the symbol distributions would easily miss the fact that the texts are symmetric.

3.1.8. STATISTICS ON CORPUS SIZES. The number of texts, number of tokens, number of types, and the mean text length of the above-discussed corpora are given in Table S2.

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 S2. Number of texts, type and token counts, and mean text length for the nonlinguistic corpora.

3.1.9. CORPUS UNDER DEVELOPMENT: MATHEMATICAL FORMULAE. **Classification: Formal.**

Mathematical symbology is a clear case of a nonlinguistic system, and it also serves as a good example of the distinction between linguistic and nonlinguistic systems. Consider a formula such as S1.

(S1) $\int e^x \, dx = e^x + C$

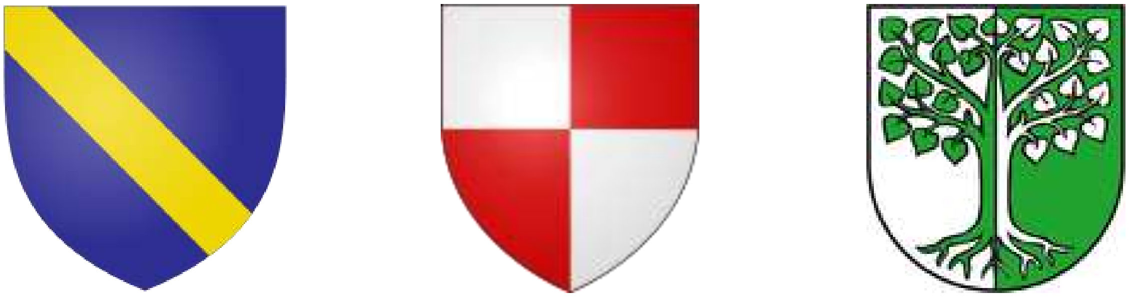
It is clearly possible to READ this expression using language: ‘the integral of e to the x, d x, equals e to the x plus C’. Thus it might seem that one could argue that mathematical symbols are linguistic. But this is missing a crucial point: while it is certainly possible to express mathematics using language, it is clear that the elements in a mathematical formula do not REPRESENT linguistic information. The symbol ∫ represents the mathematical concept of integration, not the English words *integrate*, or *integral*, or *integration*. Similarly, e^x represents the 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 LaTeX documents from the arXiv database (http://www.arxiv.org) were downloaded from http://www.cs.cornell.edu/projects/kddcup/. From these we extracted 414,492 equations delimited by \begin{equation} and \end{equation}.

3.1.10. CORPUS UNDER DEVELOPMENT: 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. It differs from English, however, 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 S5 for some examples of blazon and their corresponding arms.



Azure, a bend or.

Quarterly argent and gules.

Party per pale argent and vert, a tree eradicated counterchanged.

FIGURE S5. Examples of arms and their corresponding blazon (http://en. wikipedia.org/wiki/Blazon).

The data set 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 OCRed version at http://archive.org/details/generalarmoryofe00 burk, 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—that is, blazons of the form *quarterly first and fourth blazon*$_1$, *second and third blazon*$_2$, where *blazon*$_{1,2}$ are full blazons of the specified quarters.[5] 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 reasonably 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. To that end we developed our own specialized XML markup scheme for blazon, which will be described in a forthcoming data release.

3.1.11. ADDITIONAL CORPORA USED IN STATISTICS. In addition, we also included in our analysis a corpus of Indus bar seals from Harappa and Mohejo Daro, which we developed as part of the work reported in Farmer et al. 2004. The details on this latter corpus are given in S2.[6]

(S2)	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).

3.2. LINGUISTIC CORPORA FOR COMPARISON. For comparison with the nonlinguistic systems described above, I gathered a set of linguistic corpora from various existing sources. Just as I have tried to collect nonlinguistic corpora of a variety of types, so I have also tried to gather linguistic corpora that are varied along a couple of dimensions. The first dimension is age: we have both extremely ancient systems—Sumerian, Egyptian, Ancient Chinese, and Linear B (Mycenaean Greek)—and various modern systems. The second 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.

The following sections give brief descriptions of each of the corpora. Summary statistics can be found in Table S3.

[5] The upshot of this is that we have only 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 really is an artificial language. (We gratefully acknowledge Nate Bodenstab's help with training the BUBS parser on these data.)

[6] The reason for not using the same Indus data as that used in Rao et al. 2009a and Rao 2010 is that none of the Indus corpora that have been collected by various groups over the years have been made available to other researchers, which has in turn made it difficult to verify statistical results claimed for these corpora.

CORPUS	# TEXTS	# TOKENS	# TYPES	MEAN TEXT LENGTH
Amharic	111	17	219	159.9
Arabic	10,000	429	62	42.9
Chinese	10,000	136	2,738	13.6
Ancient Chinese	22,359	91	701	4.1
Egyptian	1,259	38	691	30.8
English	10,000	503	76	50.3
Hindi	1,000	61	77	61.2
Korean	10,000	223	984	22.4
Korean (jamo)	10,000	438	102	43.8
Linear B	439	5	220	12.4
Malayalam	937	46	80	49.6
Oriya	1,000	40	75	40.2
Sumerian	11,528	50	692	4.4
Tamil	1,000	38	61	38.5

TABLE S3. Linguistic corpora for comparison.

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

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

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

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

3.2.5. EGYPTIAN. Collection of texts transcribed using the JSesh hieroglyph editor (http://jsesh.qenherkhopeshef.org/) downloaded from http://webperso.iut.univ-paris8.fr/~rosmord/hieroglyphes. The text was tokenized at the glyph (individual symbol) level.[7]

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

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

[7] The following texts were used, all in http://webperso.iut.univ-paris8.fr/~rosmord/hieroglyphes/: CT160_S2P.hie, DoomedPrince.hie, HAtra.hie, HetS.hie, L2.hie, LC26.hie, Pachereniset.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.

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

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

3.2.10. LINEAR B. Transcription by the author of 300 tablets from Ventris & Chadwick 1956. Lines with uncertainties and/or omissions were omitted. The text was tokenized at the glyph level.

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

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

3.2.13. SUMERIAN. Tablets from Gudea ruler of Lagah (c. 2100 BC), 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. Lines with '#', '<', or '['were omitted, since these mark uncertainty/reconstructions.[8] 'Texts' consisted of individual lines in the CDLI transcription, meaning that the Sumerian 'texts' are rather short.[9]

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

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

We also included 1,000 'texts' from each of the two corpora—mathematical formulae and European heraldry, discussed more fully above—that are still under development. For mathematical formulae, we selected equations that consisted of single lines in the original LATEX. We processed the LATEX code so that commands (a string of alphanumeric characters preceded by a backslash) were kept as single tokens;

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

[9] One issue with Sumerian that will need to be resolved in future work is that the symbols are transcribed, following standard Sumerological practice, with romanized spellings of either the pronunciation of the symbol (if in lower case) or the morpheme denoted by the symbol (if in upper case), followed by a subscript. Thus *aya*$_2$ is a transcription of 𒀀, 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: 𒀀 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.

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 in our XML markup scheme (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 $\frac{r}{R}$ (§2.5 of the main article) of the heraldry corpus is quite high (0.71), whereas for the mathematics corpus it is low (0.027). For mathematics, it is relatively rare that the same symbol will follow itself, hence the low rate of local repetitions. By contrast, 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 S4.

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 S4. Number of texts, type and token counts, and mean text length for subsets of the mathematics and heraldry corpora.

All probability-based entropy statistics discussed below were computed using tools based on the OpenGrm NGram library (Roark et al. 2012), available from http://www.opengrm.org; these tools will be made available along with the corpora. We used Kneser-Ney smoothing, including the probability and thus entropy estimates for unseen cases. Start and end symbols to pad the text are implicitly computed by the software.

4. FIGURES AND TABLES REFERENCED IN THE MAIN ARTICLE. The following figures and tables are supplemental to the discussion in the main article and referenced there (§§2.4–3).

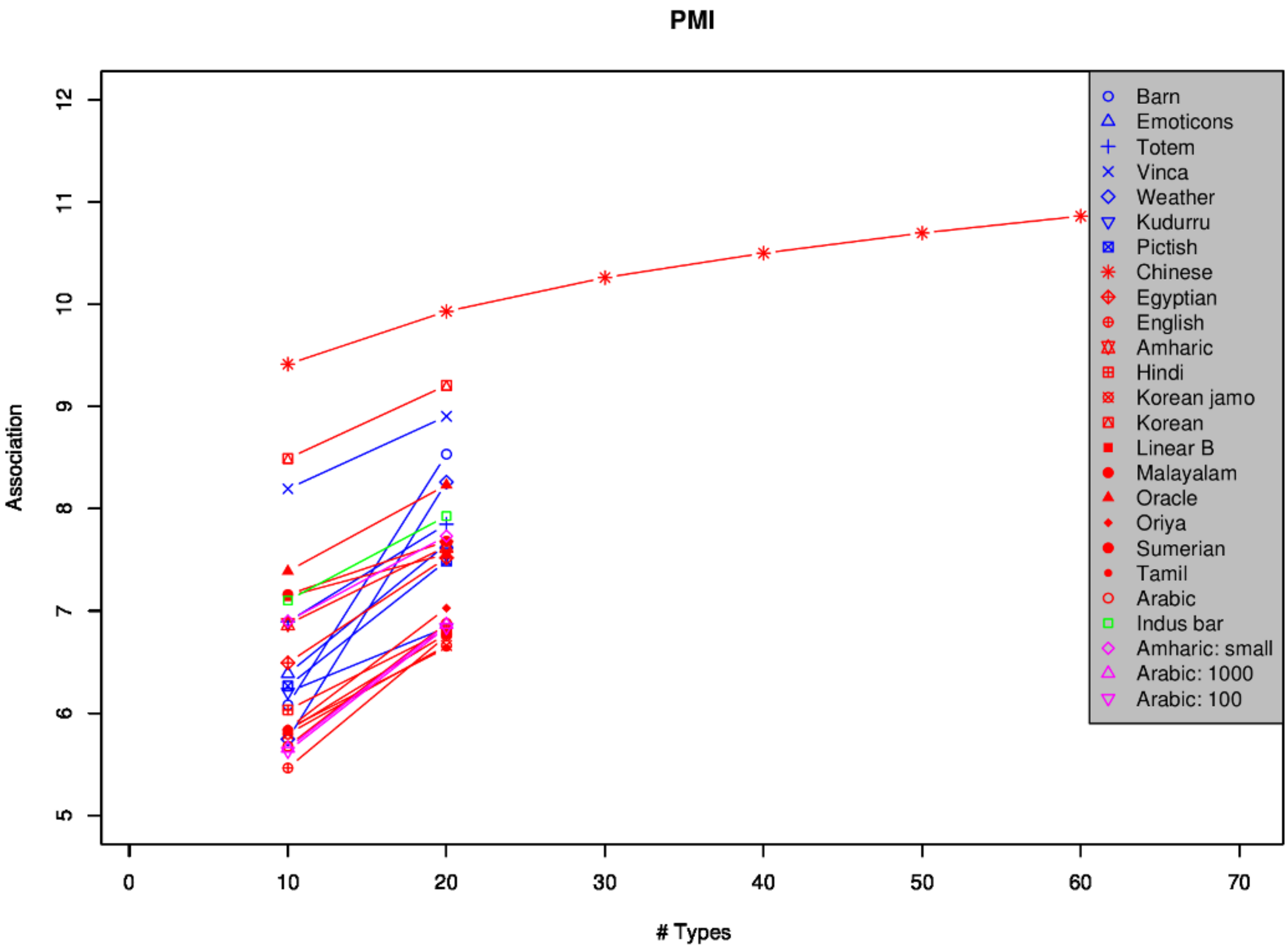


FIGURE S6. PMI association computations for our corpora, computed over subsets of each corpus starting with the ten most frequent symbols, the twenty most frequent, and so forth, up to 25% of the corpus. See the text for more detailed explanation.

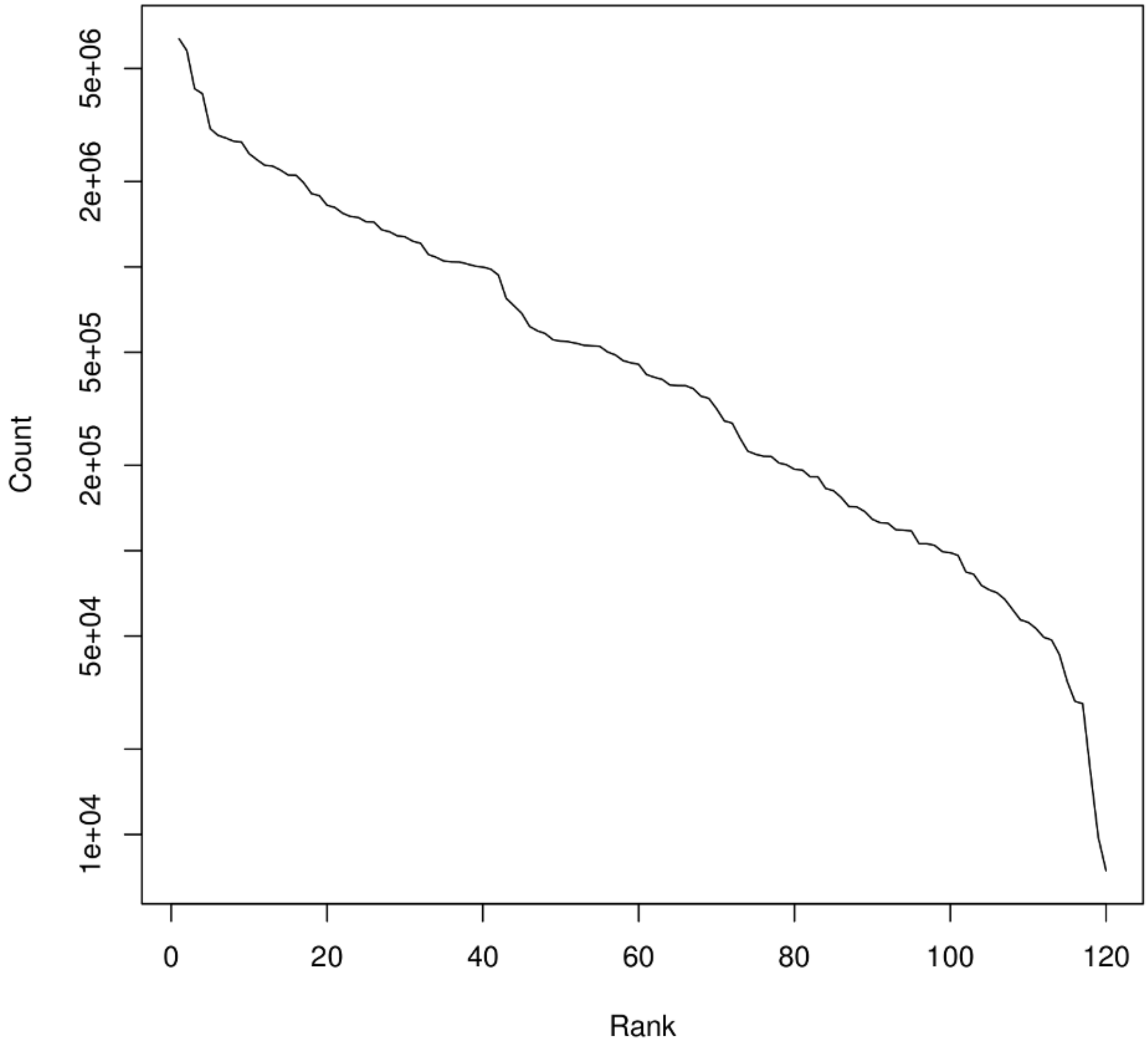


FIGURE S7. Unigram distribution of awarded Boy Scout merit badges, from http://meritbadge.org/wiki/index.php/Merit_Badges_Earned, showing the standard power-law inverse log-linear relationship between count and rank.

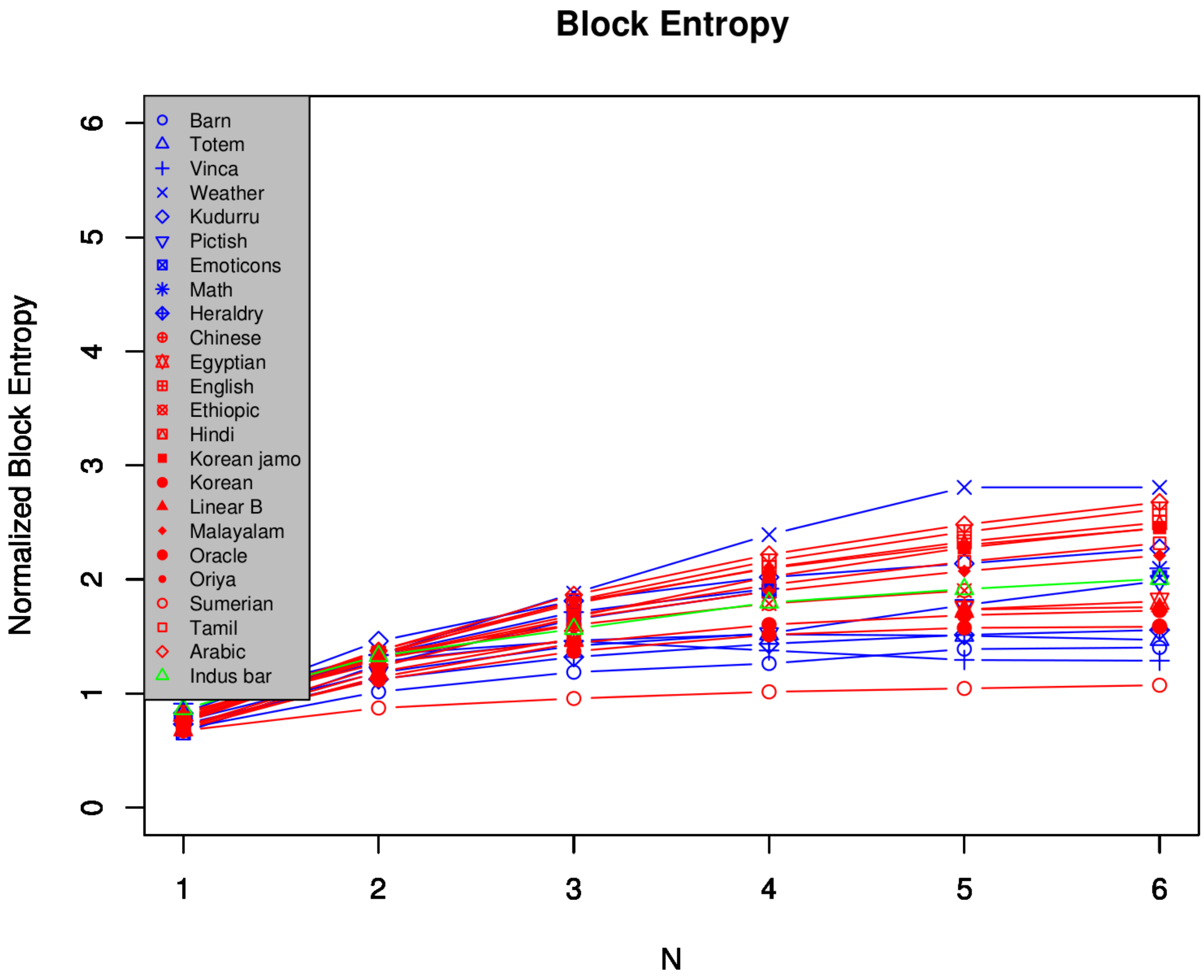


FIGURE S8. Block entropy values for our linguistic and nonlinguistic corpora, including mathematical formulae and heraldry, and the Indus bar seal corpus.

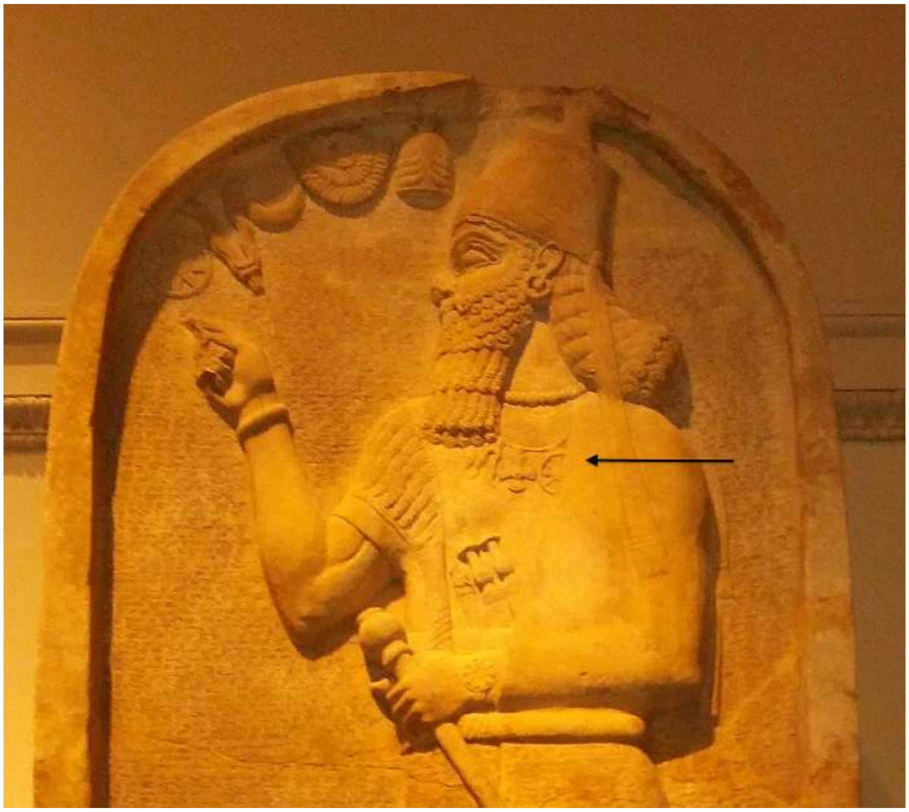


FIGURE S9. The linearly arranged symbols of the major deities of Aššurnaṣirpal II, indicated by the arrow. Note also the symbols on the necklace worn by the king. (Collection of the British Museum.)

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
Kudurrus	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 S5. Repetition rate $\frac{r}{R}$ for versions of the corpora with artificially shortened texts of length six or less, and no more than 500 texts.

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
Malayalam	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

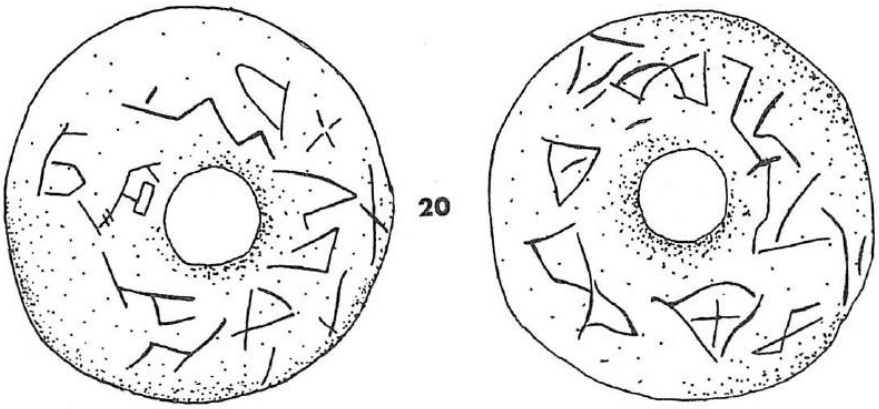
TABLE S6. Mean accuracy of results for each of the corpora when occurring in the training data.

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
Malayalam	0.80	21
Pictish	0.75	21
Weather icons	0.74	20
Egyptian	0.74	28
Mesopotamian	0.70	29
Sumerian	0.69	24
Asian emoticons	0.64	21

TABLE S7. Mean accuracy of results for each of the corpora when occurring in the test data.

5. SAMPLE TEXTS AND TRANSCRIPTIONS. In this section we give examples of texts, marked up with the XML markup system from Wu et al. 2012, from five of our nonlinguistic corpora: Vinča, Mesopotamian deity symbols (kudurrus), Pictish symbols, totem poles, and Pennsylvania German barn stars.

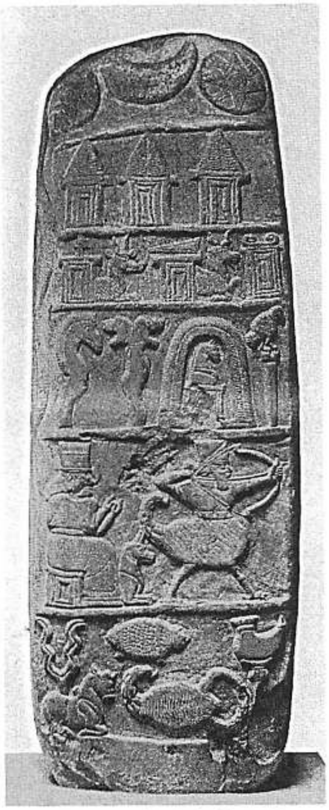
Vinča: Tordos Spindle Whorl #20 (Winn 1981:270)



```
<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>
```

Mesopotamian deity symbols: Kudurru stone #67, from Seidl 1989, plate 23a.

Used with permission.



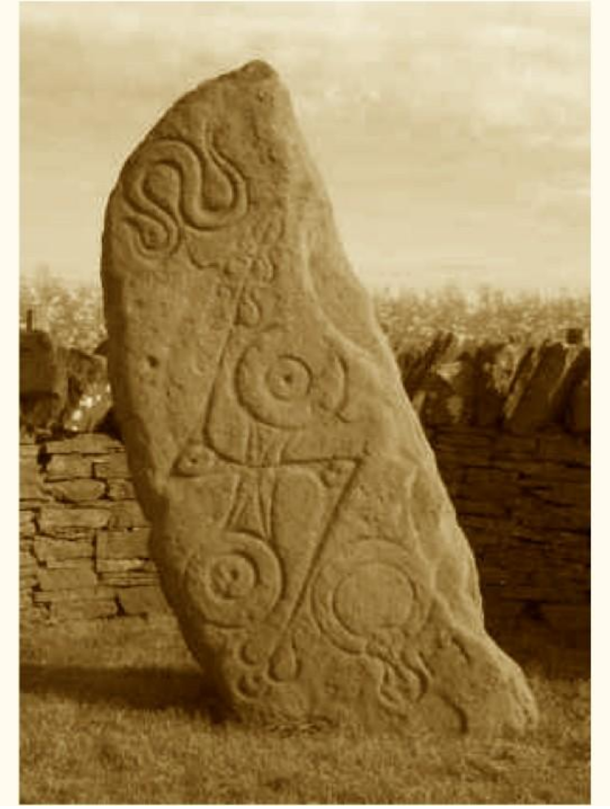
a. Nr. 67 Nabû-kudurrī-uṣur I.

```
<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>
             <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>
```

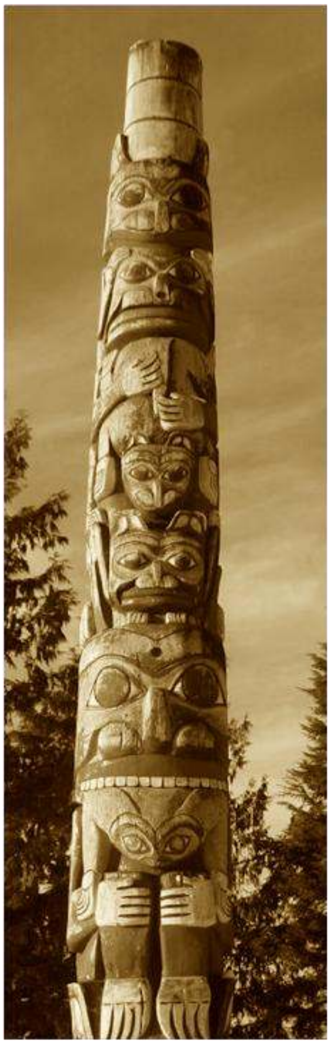
Pictish symbols: The Aberlemno 1 stone

(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>
```

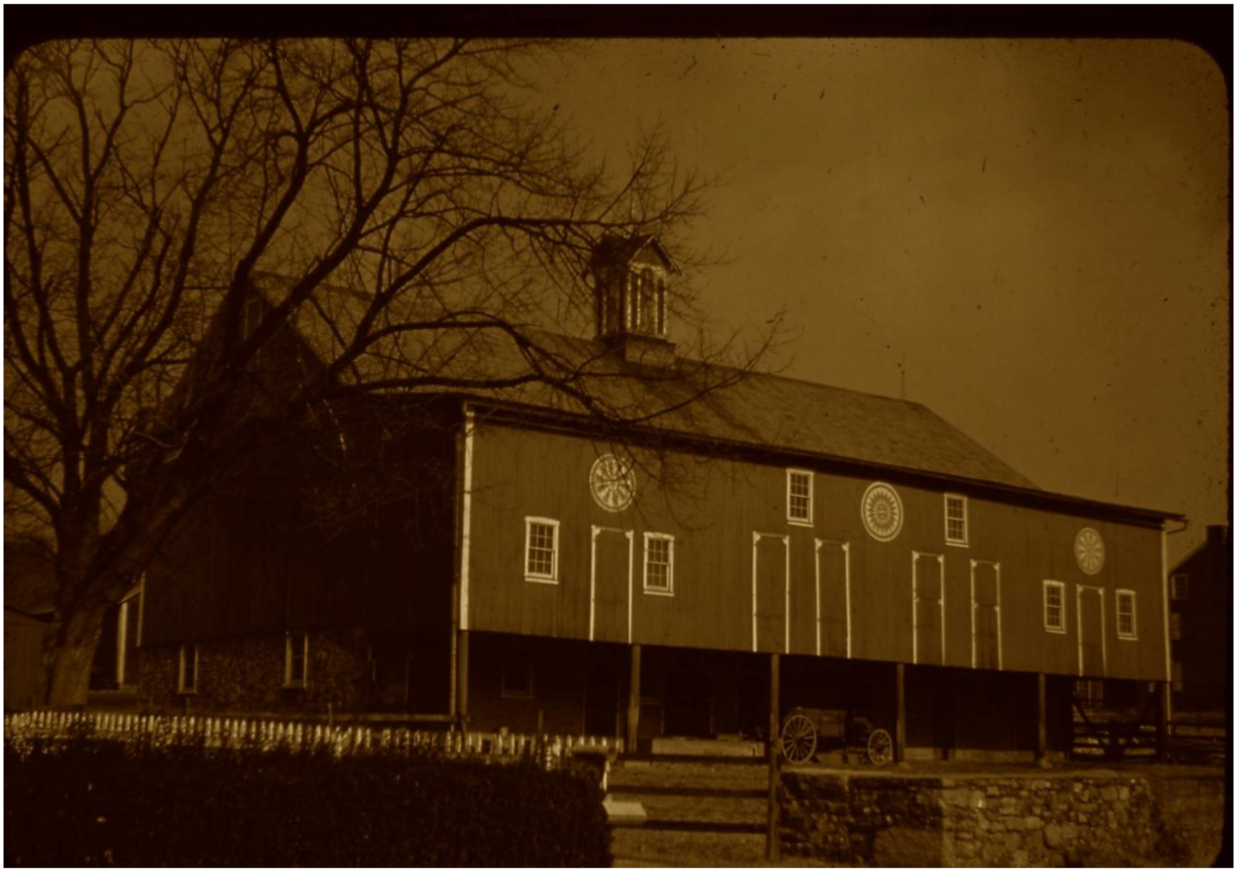
Totem poles: Grizzly Bear Pole of Yan (Drew 1969:16)



```
<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>
```

Pennsylvania German barn stars: Farrell Collection AR(2)F

Used with permission of the Historical Society of Berks County.



```
<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>
```

REFERENCES

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

BASSO, KEITH, and NED ANDERSON. 1973. A Western Apache writing system: The symbols of Silas John. *Science* 180.4090.1013–22.

BEDRICK, STEVEN; RUSSELL BECKLEY; BRIAN ROARK; and RICHARD SPROAT. 2012. Robust kaomoji detection in Twitter. Paper presented at the Workshop on Language and Social Media, Montreal, Canada, June 2012.

BODENSTAB, NATHAN; AARON DUNLOP; KEITH HALL; and BRIAN ROARK. 2011. Adaptive beam-width prediction for efficient CYK parsing. *HLT '11: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 440–49.

BURKE, BERNARD. 1884. *The general armory of England, Scotland, Ireland, and Wales; Comprising a registry of armorial bearings from the earliest to the present time*. London: Harrison & Sons.

CITY OF DUNCAN. 1990. *Duncan: City of Totems*. Duncan, BC.

DANIELS, PETER, and WILLIAM BRIGHT (eds.) 1996. *The world's writing systems*. New York: Oxford University Press.

DEFRANCIS, JOHN. 1989. *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

DREW, F. W. M. 1969. *Totem poles of Prince Rupert*. Prince Rupert, BC: F. W. M. Drew.

FARMER, STEVE; RICHARD SPROAT; and MICHAEL WITZEL. 2004. The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies* 11.2.19–57. Online: http://www.ejvs.laurasianacademy.com/ejvs1102/ejvs1102article.pdf.

FARNELL, BRENDA. 1996. Movement notation systems. In Daniels & Bright, 855–79.

FELDMAN, RICHARD. 2003. *Home before the raven caws: The mystery of the totem pole*. Cincinnati: Emmis Books.

GARFIELD, VIOLA. 1940. *The Seattle totem pole*. Seattle: University of Washington Press.

GELB, IGNACE JAY. 1952. *A study of writing: The foundations of grammatology*. Chicago: University of Chicago Press.

GRAVES, THOMAS. 1984. *The Pennsylvania German hex sign: A study in folk process*. Philadelphia: University of Pennsylvania dissertation.

GUNN, SISVAN WILLIAM. 1965. *The totem poles in Stanley Park, Vancouver, B.C*. 2nd edn. Vancouver: Macdonald.

GUNN, SISVAN WILLIAM. 1966. *Kwakiutl House and totem poles at Alert Bay*. Vancouver: Whiterocks.

GUNN, SISVAN WILLIAM. 1967. *Haida totems in wood and argillite*. Vancouver: Whiterocks.

HAARMANN, HARALD. 2008. The Danube script and other ancient writing systems. *Journal of Archaeomythology* 4.1.12–46.

HAARMANN, HARALD, and JOAN MARLER. 2008. An introduction to the study of the Danube script. *Journal of Archaeomythology* 4.1.1–11.

HARRIS, ROY. 1995. *Signs of writing*. London: Routledge.

JACKSON, ANTHONY. 1984. *The symbol stones of Scotland: A social anthropological resolution of the problem of the Picts*. Elgin: Orkney.

JACKSON, ANTHONY. 1990. *The Pictish trail*. Elgin: Orkney.

KAISER, DAVID. 2005. Physics and Feynman's diagrams. *American Scientist* 93.156–65.

LE NOVÈRE, NICOLAS, ET AL. 2009. The systems biology graphical notation. *Nature Biotechnology* 27.735–41. Online: http://www.nature.com/nbt/journal/v27/n8/full/nbt.1558.html.

LEE, ROB; PHILIP JONATHAN; and PAULINE ZIMAN. 2010. Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences* 466.2121.2545–60. Online: http://rspa.royalsocietypublishing.org/content/466/2121/2545.

LI, LINCAN. 2001. *Naxizu xiangxing biaoyin wenzi zidian*. Yunnan: Yunnan Minzu Chubanshe.

MACK, ALASTAIR. 1997. *Field guide to the Pictish symbol stones*. Balgavies: The Pinkfoot Press. [Updated 2006.]

MAHR, AUGUST. 1945. Origin and significance of Pennsylvania Dutch barn symbols. *Ohio History: The Scholarly Journal of the Ohio Historical Society* 54.1.1–32.

MALIN, EDWARD. 1986. *Totem poles of the Pacific Northwest coast*. Portland: Timber.

MALLERY, GARRICK. 1883. Pictographs of the North American Indians: A preliminary paper. *Fourth annual report of the Bureau of Ethnology*, 13–256. Washington, DC: Smithsonian Institution. Available online on Google Books.

MCCAWLEY, JAMES. 1996. Music notation. In Daniels & Bright, 847–54.

NEWCOMBE, WARREN ALFRED, and BRITISH COLUMBIA PROVINCIAL MUSEUM. 1931. *British Columbia totem poles*. Victoria: Charles F. Banfield printer to the King's most excellent majesty.

PARPOLA, ASKO. 1994. *Deciphering the Indus script*. New York: Cambridge University Press.

RAO, RAJESH. 2010. Probabilistic analysis of an ancient undeciphered script. *IEEE Computer* 43.3.76–80.

RAO, RAJESH; NISHA YADAV; MAYANK VAHIA; HRISHIKESH JOGLEKAR; R. ADHIKARI; and IRAVATHAM MAHADEVAN. 2009a. Entropic evidence for linguistic structure in the Indus script. *Science* 324.5931.1165.

RAO, RAJESH; NISHA YADAV; MAYANK VAHIA; HRISHIKESH JOGLEKAR; R. ADHIKARI; and IRAVATHAM MAHADEVAN. 2009b. A Markov model of the Indus script. *Proceedings of the National Academy of Sciences* 106.33.13685–90.

RHYS, JOHN. 1892. The inscriptions and language of the northern Picts. *Proceedings of the Society of Antiquaries of Scotland* 26.263–351.

ROARK, BRIAN; RICHARD SPROAT; CYRIL ALLAUZEN; MICHAEL RILEY; JEFFREY SORENSEN; and TERRY TAI. 2012. The OpenGrm open-source finite-state grammar software libraries. *Proceedings of the 50th annual meeting of the Association for Computational Linguistics, System Demonstrations*, 61–66.

ROGERS, HENRY. 2005. *Writing systems: A linguistic approach*. Malden, MA: Blackwell.

ROYAL COMMISSION ON THE ANCIENT AND HISTORICAL MONUMENTS OF SCOTLAND. 1994. *Pictish symbol stones: A handlist*. Edinburgh: Royal Commission on the Ancient and Historical Monuments of Scotland.

SAMPSON, GEOFFREY. 1985. *Writing systems*. Stanford, CA: Stanford University Press.

SEIDL, URSULA. 1989. *Die babylonischen Kudurru-Reliefs: Symbole mesopotamischer Gottheiten*. Freiburg: Universitätsverlag Freiburg.

SPROAT, RICHARD. 2000. *A computational theory of writing systems*. Cambridge: Cambridge University Press.

SPROAT, RICHARD. 2014. Non-linguistic symbol systems: A taxonomy and analysis of their statistical properties. New York, MS.

STEWART, HILARY. 1990. *Totem poles*. Seattle: University of Washington Press.

STEWART, HILARY. 1993. *Looking at totem poles*. Seattle: University of Washington Press.

SUTHERLAND, ELIZABETH. 1997. *The Pictish guide*. Edinburgh: Birlinn.

VENTRIS, MICHAEL, and JOHN CHADWICK. 1956. *Documents in Mycenaean Greek*. Cambridge: Cambridge University Press.

VIDALE, MASSIMO. 2007. The collapse melts down: A reply to Farmer, Sproat and Witzel. *East and West* 57. 333–66.

WADDINGTON, CAROLINE. 1974. Horse brands of the Mongolians: A system of signs in a nomadic culture. *American Ethnologis*t 1.3.471–88.

WINN, SHAN M. M. 1981. *Pre-writing in Southeastern Europe: The sign system of the Vinča culture, ca. 4000 B.C.* Calgary: Western.

WINN, SHAN M. M. 2008. The Danube (Old European) script ritual use of signs in the Balkan-Danube region c. 5200–3500 BC. *Journal of Archaeomythology* 4.1.126–41.

WU, KATHERINE; JENNIFER SOLMAN; RUTH LINEHAN; and RICHARD SPROAT. 2012. Corpora of non-linguistic symbol systems. Paper presented at the annual meeting of the Linguistic Society of America, Portland. Extended abstract online: http://elanguage.net/journals/lsameeting/article/ view/2845/pdf.

YODER, DON, and THOMAS GRAVES. 2000. *Hex signs: Pennsylvania Dutch barn symbols and their meaning*. Mechanicsburg, PA: Stackpole.

[rws@google.com]