Pictish Symbol Stones: religious imagery, heraldic arms or a language?

The Pictish Symbol Stones are a collection of finely carved stones dating from ca. 300 AD to 843 AD whose meaning has been lost in time. Despite attracting linguists and historians alike, as a collection they have been just too small to surrender their secrets. Now, however, **Rob Lee**, **Philip Jonathan** and **Pauline Ziman** have used a technique based on comparative entropy to unwrap some of the mystery of the stones.

Introduction

Enigmatic scripts from earlier societies are a source of fascination for linguists and historians alike. They can be found on artefacts left by many prehistoric societies. They range from inscribed Chinese Neolithic pottery, through clay tablets and seals of the Bronze Age Indus Valley to inscribed stones and metalwork of Late Iron-Age Scotland; typically they consist of short sequences of regularly-placed symbols, or sometimes a single symbol may stand alone. Their presence begs the question whether the symbols are a form of writing. One problem facing researchers is that the available total script is often small, from just a few hundred to a few thousand symbols. Furthermore, individual inscriptions are generally short, typically just one to three symbols in length. Establishing whether such scripts are language-like has until recently remained a challenge.

The Picts were an Iron Age society in Scotland that left a set of unknown symbols carved on rock and metal. They existed from Roman times (around 300 AD) to 843 AD when the Scot, Kenneth, son of Alpin, took the Pictish kingship. Picts and Scots had until then been separate societies. Picts inhabited central and eastern Scotland; Kenneth was king of a region based in western

Goose and fish on the Easterton of Roseisle Pictish stone, now in the Museum of Scotland, Edinburgh

(a)

(b)

Figure 1. (a) Grantown class I stone, inscribed image of a deer. (b) Aberlemno class II stone, bas-relief carved horsemen

Scotland and some of the western isles. After Kenneth, the Picts as a society disappear from history.

They had a language, Pictish, remnants of which are found in many Scottish place names. However, whilst the Picts are recorded in the writings of their contemporaries – the Romans, the Anglo-Saxons and the Irish – other than a copy of their King List, they left no written record of themselves. They did, however, leave a range of finely carved stones inscribed with glyphs of unknown meaning, known as "Pictish Symbol Stones". The Pictish Symbol Stones are found along the eastern side of Scotland from Edinburgh to Caithness, in the Western Isles and in Orkney and Shetland. Many are beautiful; all are enigmatic.

The Stones are categorised into two types, as shown in Figure 1:

- Class I stones, numbering between 180 and 195, consist of undressed stone with the symbols inscribed into the rock.
- Class II stones, numbering between 60 and 65 stones, contain the depiction of a cross, use dressed stone and relief carving for the symbols and may have other, often Christian, imagery.

Class I stones are taken to be the earlier tradition of the two types of Symbol Stone. Each stone contains between one and eight symbols, with stones containing two symbols being most common. The symbols are also found on pieces of fine metalwork, such as the silver plaques shown in Figure 2. Over the last century a wide variety of "meanings" for the symbols have been proposed, from pagan religious imagery to heraldic arms, but it is only recently that the question whether they might be a written language has been asked[1].

Figure 2. Silver plaques with red enamel inlaid symbols, from Norrie's Law bronze age burial ground, Fife

Scholars have proposed different symbol definitions for Pictish Stone inscriptions. In 1903 Allan and Anderson[2] made the earliest list of them; in 1997 Mack[3] made the most recent. In the work that follows we analyse their lists of symbols separately.

The problem

There are well-established statistical techniques for investigating unknown languages. They require large samples and are hampered by linguistic phenomena such as grammar, syntax and non-standardised spelling. For small script sets, such as the set of Pictish Symbols, a new methodology is required.

There are many different kinds of written communication. They include pictures of events (such as the Bayeux tapestry) or instructions (e.g. IKEA flat-pack assembly cartoons). Symbols may be purely allegorical, such as the Christian cross; they may be part of a system like heraldry, where complex rules of placement surround symbols which have no speech value in themselves but which do identify a person, position or a place. Such systems, in which the symbols are not associated with specific speech sounds, are known as *sematograms*. Pictorial road signs are also sematograms.

On the other hand, symbols may represent spoken sounds, either letters, as in modern alphabets, syllables, as in Egyptian hieroglyphs, or whole words, as in Chinese. The symbols can even subdivide letters, as in Morse code. Scripts whose symbols represent spoken sounds are known as lexigraphic, the symbols being *lexigrams*. We had first to determine whether the Pictish stones are a form of writing at all. If they are, the next problem was whether they are sematograms or lexigrams.

For the most part, placement of Pictish symbols upon rocks occurs with a strong implied direction, generally one above the next or more occasionally side by side. This directionality is symptomatic of speech-derived writing but is not unique to it. It implies that the symbols are not allegorical religious images; nor are they sematograms of the IKEA variety.

The problem that the Pictish script set poses could be addressed by considering two questions:

1. Are the symbols randomly ordered within the script set (admittedly unlikely, since they appear to have been carved for a purpose)?
2. If it is unlikely that they are randomly ordered, then what type of

communication do they convey? Are they:

(a) a sematographic system like heraldry – technically *directional semasiography* – where information is communicated without reference to verbal language forms but follows rules of placement? or,

(b) a lexigraphic script, where symbols embody the form of verbal language?[4]

We compared the Pictish script sets with a range of nearly 300 other scripts. These included sematograms (heraldic characters), logograms (Chinese), syllabaries (Linear B and Egyptian hieroglyphs), and alphabetic systems of different modern languages (English, Irish, Welsh, Norse, Turkish, Basque, Finnish, Korean) and ancient languages (Latin, Anglo-Saxon, Old Norse, Ancient Irish, Old Irish, Old Welsh). The languages were analysed at letter, syllable and word level; the texts we used cover prose, poetry, monumental inscriptions and genealogical lists (king lists, marriage and birth lists). All these script sets (apart from the Pictish) are known to convey information in either sematographic or lexigraphic form.

In our analysis, we refer to a single unit of script as a "unigram"; a unigram might be a single character in a set of characters, a word in a written language script, or a single symbol on inscribed stones. A "digram" is an ordered pair of adjacent unigrams.

Our basic data is sets (of different sizes) made up of sequences (of different lengths) of unigrams from a range of these different scripts.

Entropy: uncertainty in communication

A fundamental characteristic of any communication system is that there is a degree of uncertainty (also known as entropy or information) over the particular character or message that may be transmitted[5]. This uncertainty can be mathematically quantified – see box. It is known as the *Shannon entropy*.

For any given script set, the first-order entropy (E1) is a measure of the frequencies at which the different unigrams occur. It is greatest when all unigrams occur with equal frequency and thus with equal probability – implying that they might be random doodles or decoration rather than a written language. In written language, unigrams occur with unequal probabilities. For example, in English the letters "e" and "t" occur more frequently than the letters "x" and "z"; this lends some degree of predictability to the occurrence of a particular unigram (you will be right more often if you predict "e" than if you predict "z" as the next letter in an English script). This predictability means that in a written language, E1 will be less than in a random set of symbols.

But the fact that some symbols appear more often than others is not enough on its own to establish that we are looking at a language. On Western gravestones, for example, the cross occurs frequently – not of course because it is part of a written language but because it symbolises the Christianity that was central to the men who carved it. To establish writing we need an extra measure as well. Happily we have one.

In a written language, unigrams not only have different frequencies. There is also unigram-to-unigram dependence. For example, in English, "q" tends to be followed by "u". The digram "qu" would therefore occur more often that other digrams starting with "q". This second-order dependency is captured mathematically in the second-order entropy, E2. Where there is little second-order dependency, such as in a randomised set of unigrams, E2 will be larger and closer to E1; when there is much second-order dependency, as in English, E2 is lower[6]. This gives us a basis for investigating whether Pictish script sets are examples of writing.

Shannon entropy

Claude Shannon is considered by many to be the founder of modern information theory. His definitions of entropy follow: first-order Shannon entropy is given by

$$E1 = -\sum_{i=1}^{N} p_i \log_2 p_i$$

and second-order Shannon entropy by

$$E2 = -\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij} \log_2 p_{ij} - E_1$$

Here N is the total number of distinct unigrams in the sample, p_i is the empirical probability of the occurrence of the ith unigram, and p_{ij} is the empirical probability of the occurrence of the digram consisting of the ith unigram followed by the jth unigram. E2 is bounded above by E1 and equals it when unigrams occur randomly, i.e. $p_{ij} = p_i p_j$. E2 is bounded below by zero and equals it when knowledge of the first half of the digram completely determines the second, i.e. $p_{ij} = p_i$.

Digram dependence

Thus for the Pictish stones, we had to work out whether E2 was unusual compared to a system which had the same relative frequencies of unigrams (symbols) but with no unigram-to-unigram dependency.

We did this by comparing the value of E2 with the values of E2(R) we got from randomly mixing up the order of the symbols in the script set. The mixing of course would break any digram dependence. It might, however, introduce other, chance, "dependencies". Thus if "the cat sat on the mat" gets randomly reordered to "ta ta ta th th csnoeme", it appears that "t" must always be followed by "a" or "h". This particular reordering has a high second-order dependency, and a low value for E2(R). That, though, would be unusual; most random reorderings would give no such artificial patterns, and would give higher E2(R) values. For each of our 286 script sets – Chinese, Norse, Korean and the rest – we performed a thousand reorderings and calculated a thousand E2(R) values. For each script set the E2(R) values ranged, as anticipated, from mostly higher values to a few low ones.

We expected that the E2 from an original-order script with significant dependence would appear unusual compared to the values of E2(R) for that size and structure of script. This indeed turned out to be the case for a large proportion of the scripts sets (both sematographic and lexigraphic) that we analysed. The values of E2 for the two Pictish script sets are seen to be extreme with respect to the corresponding E2(R) distributions. We concluded that the Pictish script sets show dependence within digrams in the same way as the script sets which are examples of writing.

So what type of communication do the Pictish stones convey?

The analysis of digrams showed us that the Pictish symbols are not occurrences of random, independent unigrams, but are writing. The question is, what type of writing are they? In particular, are they sematograms, where the symbols convey information without connecting directly to the spoken word, or are they lexigrams – that is, speech-based writing? Unfortunately, with these small script sets, E2 cannot be used on its own to differentiate between the two; nor, in lexigraphic scripts, can it distinguish the different symbol types (words, syllables and letters).

The incomplete sample problem

When the script size is small, the empirical frequencies of unigram and particularly digram occurrence may be different from their true values. The script is "incomplete" in the sense that it is not extensive enough for all viable digrams to be represented. For example, taking the preceding sentence as a small sample of the English alphabet, the letters "j", "k", "q", "w", "y" and "z" do not occur at all; the letter "b" occurs only once and when it does it is followed by "l". The sample-based estimate for the probability of digram "bl" given first occurrence of "b" would be 1, and the estimate for any other digram starting with "b" is 0. Yet we know that these (conditional) probabilities of occurrence are not representative of the English language. As a consequence, the estimates of E2 we get from small samples are lower than estimates based on larger script sets.

For a given script set, the degree of incompleteness in the digram lexicon will be influenced by many phenomena of language, including:

1. the type of unigram used to code for the communication;
2. the size of the unigram lexicon that is used (e.g. texts with constrained, or limited, vocabularies, such as a list of kings, pull from a pool of available words that is limited to a fraction of a normal vocabulary);
3. the grammar of the unknown language (e.g. the system of inflection within the language);
4. the syntax of the unknown language (i.e. the word order);
5. the degree of standardised spelling (i.e. many inscriptions do not use standardised spelling).

As we said above, E2 alone does not provide a useful basis to distinguish between the different types of writing and symbol. However we have succeeded in identifying two summary statistics which do appear to be useful classifiers. These are:

1. a measure of entropy adjusted for diagram lexicon completeness, U_r;
2. a measure of unigram repetition adjusted for digram lexicon completeness, C_r.

Interestingly, U_r and C_r also have linguistic interpretations.

Entropy adjusted for digram completeness, U_r

An alphabet with N_u letters will have $N_u \times N_u = N_u^2$ possible combinations of two of them. Thus English has 276 = 26 × 26 two-letter pairs or digrams, from AA, AB, AC, … to ZX, ZY, ZZ. But if the sample of script that we have is incomplete, the number of digrams that our sample displays will be less than this. How much less will depend, among other things, on how complete or incomplete our sample is. In practice, N_d for actual lexicons will in any case be less than N_u^2 due to syntax and spelling rules. The ratio N_d/N_u provides a measure of digram lexicon completeness which we have found useful in practice. Figure 3 shows the relationship between E2 and N_d/N_u for over 400 script sets, each set containing between 30 and 50 000 unigrams. The sets span a wide variety of unigram types and scripts. Happily, the three types of lexigraphic symbols – words, syllables and letters – occupy different regions of Figure 3. This gives us a means to work out which of the three our Pictish symbols are – or would do, if only we could be sure that they were lexigrams and not sematograms. Sematograms, unfortunately, are not distinguishable from lexigrams in the chart. How, then, do we differentiate between directional sematograms, such as heraldry, and lexigrams?

A list of kings in an unknown language will contain only a fraction of the words that the language contains

Digram repetition factor, C_r

For that, we turn to the digram repetition factor, C_r. Whilst words, syllables and letters generally correspond to a fixed unit of language, sematograms and code characters do not. This makes them, by nature, intrinsically more repetitive than standard lexigraphic unigrams. For example, the Morse code uses two unigrams in combination with a space and thus the four digrams (dash-dash, dash-dot, dot-dash and dot-dot) repeat to build the 26 letters of the English alphabet. A measure of digram repetition may be a means to differentiate between sematograms and lexigrams – and this indeed

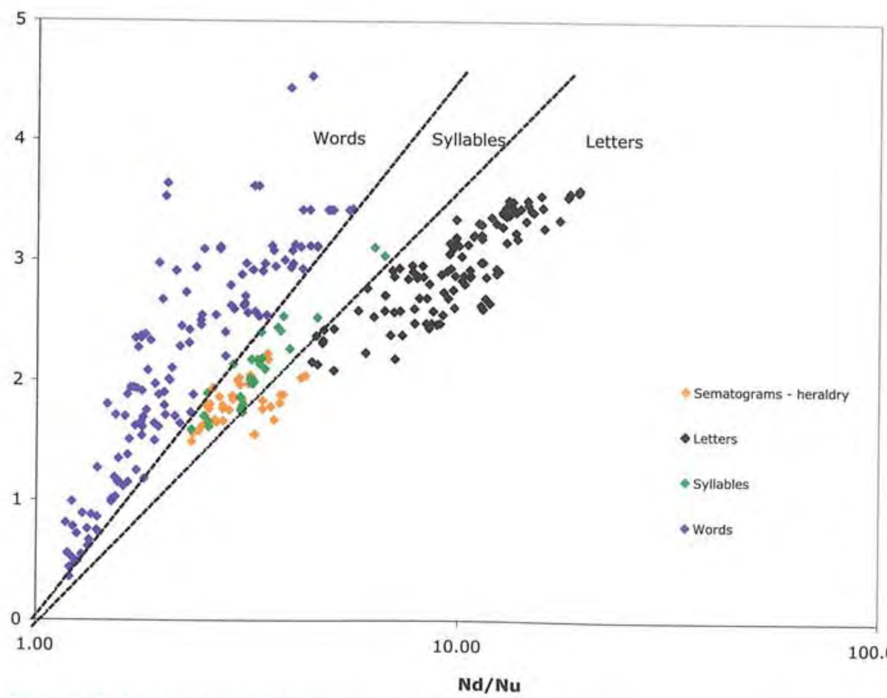

Figure 3. E2 plotted against N_d/N_u showing the separation of the different character types found in lexigraphic writing. The ratio $U_r = \mathrm{E2}/\log_2(N_d/N_u)$ is termed *entropy adjusted for digram completeness*

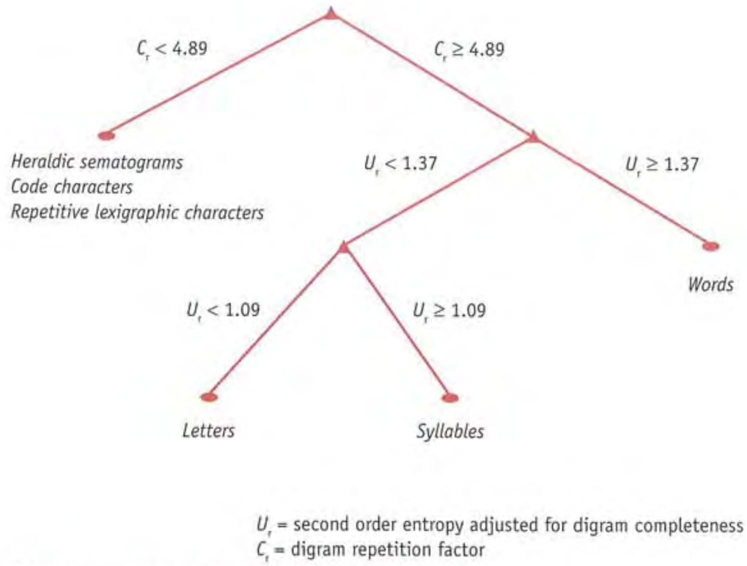

Figure 4. A decision tree to classify character type in scripts

turns out to be the case. We found that the ratio S_d/T_d, where T_d is the total number of digrams and S_d is the number of digrams that appear only once, helps to distinguish the two. For script sets with a high degree of digram repetition, S_d approaches zero. The degree of digram repetition is also related to digram completeness, since script sets with a relatively complete digram lexicon show greater digram repetition and a lower value of S_d. A linear combination of S_d/T_d and N_d/N_u defines a *digram repetition factor*, C_r.

Using a decision tree for script classification

The decision tree shown in Figure 4 uses the U_r and C_r values of a script set to classify it as one of the following character types: (a) words; (b) syllables; (c) letters; (d) other characters such as heraldic sematograms and lexigraphic code characters[7]. The predictive allocation success rate of the tree is greater than 99%.

So what of the Pictish Symbols? When the decision tree is applied to both classes of stones and to the Allen–Anderson and Mack lexicons, it suggests that the Pictish script sets are not heraldic or sematogram characters. They are lexigrams – true speech-based writings. So what type of lexigram are they? Here, strangely, the two lexicons give different answers. Table 1 shows the difference. According to Mack's lists of symbols, each symbol represents a syllable; according to the Allen and Anderson list, each symbol stands for a word. The difference in classification is probably due to the size of the symbol lexicon. Mack uses a narrow, restrictive list of symbols compared to Allen and Anderson and may thus be constraining the lexicon.

Syllables or words?

Since there are many complete stones inscribed only with a single symbol it seems unlikely, although not impossible, that the symbols are single syllables. But to clear up the issue, and so to define a system from which deciphering the stones can be initiated, a complete visual catalogue of the stones and the symbols will need to be created. The effect of widening the symbol set would also need to be investigated. However, demonstrating that the Pictish Symbols are writing, with symbols probably corresponding to words, opens a unique line of research for historians and linguists investigating the Picts.

We have shown that it is possible to use a comparative entropy technique to investigate the degree of communication in very small and "incomplete" written systems. It may be possible to extend this to other areas with similar characteristics. For example, animal language studies using Shannon entropies are often hampered by small sample data sets – the whistles of bottle-nosed dolphins are an example[8]. By building a similar set of data for spoken or verbal human communication it may be possible to make similar comparisons of the level of information communicated by animal languages.

The same techniques may work out the level of information in animal languages, such as the whistles of bottle-nosed dolphins

Table 1. The character classification of the Pictish symbols on the two types of stones by the different symbol lexicons

Stone Class	Symbol Type Set	U_r	C_r	Character classification
I	Mack	1.28	5.92	Syllable
II	Mack	1.36	6.11	Syllable
I	Allen and Anderson	1.39	5.64	Word
II	Allen and Anderson	1.45	6.16	Word

References

1. Forsyth, K. F. (1997) Some thoughts on Pictish Symbols as a formal writing system. In *The Worm, the Germ, and the Thorn* (ed. D. Henry), pp. 85–98. Balgavies: Pinkfoot.
2. Allen, J. R. and Anderson, J. (1993) *The Early Christian Monuments of Scotland*. Balgavies: Pinkfoot.
3. Mack, A. (2006) *Field Guide to The Pictish Symbol Stones*. Balgavies: Pinkfoot.
4. Powell, B. B. (2009) *Writing: Theory and History of the Technology of Civilization*, pp. 1–59. Chichester: Wiley-Blackwell.
5. Shannon, C. E. (1948) A mathematical theory of communication. *Bell Systems Technical Journal*, **27**, 379–423, 623–656.
6. Yaglom, A. M. and Yaglom, I. M. (1983) *Probability and Information* (transl. by V. K. Jain), pp. 44–100. Dordrecht: Reidel.
7. Lee, R., Jonathan, P. and Ziman, P. (2010) Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society* Series A. doi:10.1098/rspa.2010.0041.
8. McCowen, B., Hanser, S. F. and Doyle, L. R. (1999) Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphins whistle repertoires. *Animal Behaviour*, **57**, 409–419.

Rob Lee is a Professor in the School of Biosciences in the University of Exeter.