PROCEEDINGS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Pictish symbols revealed as a written language through application of Shannon entropy

Rob Lee, Philip Jonathan and Pauline Ziman

Proc. R. Soc. A published online 31 March 2010

PROCEEDINGS OF THE ROYAL SOCIETY A

FirstCite® e-publishing

Proc. R. Soc. A
doi:10.1098/rspa.2010.0041
Published online

Pictish symbols revealed as a written language through application of Shannon entropy

By Rob Lee[1,*], Philip Jonathan[2] and Pauline Ziman[3]

[1]*School of Biosciences, Geoffrey Pope Building, University of Exeter, Stocker Road, Exeter EX4 4QD, UK*
[2]*Department of Mathematics and Statistics, University of Lancaster, Lancaster LA1 4YF, UK*
[3]*PHS Consulting Limited, Pryors Hayes Farm, Willington Road, Oscroft, Tarvin, Chester CH3 8NL, UK*

Many prehistoric societies have left a wealth of inscribed symbols for which the meanings are lost. For example, the Picts, a Scottish, Iron Age culture, left a few hundred stones expertly carved with highly stylized petroglyph symbols. Although the symbol scripts are assumed to convey information, owing to the short (one to three symbols), small (less than 1000 symbols) and often fragmented nature of many symbol sets, it has been impossible to conclude whether they represent forms of written language. This paper reports on a two-parameter decision-tree technique that distinguishes between the different character sets of human communication systems when sample sizes are small, thus enabling the type of communication expressed by these small symbol corpuses to be determined. Using the technique on the Pictish symbols established that it is unlikely that they are random or sematographic (heraldic) characters, but that they exhibit the characteristics of written languages.

Keywords: language; petroglyph; Pictish; prehistoric script; Shannon entropy; symbol

1. Introduction

Among the durable artefacts left by prehistoric societies, there are many instances of enigmatic scripts. These scripts typically consist of very short sequences of regularly placed symbols (or single symbols) and range from the inscribed pottery of the Chinese Neolithic pottery (Li *et al.* 2002), through the inscribed clay tablets and seals of the Indus Valley culture (Rao *et al.* 2009) to the inscribed stones of Late Iron Age Scotland (Wainwright *et al.* 1955; Mack 1997). A longstanding conundrum has been to determine whether any of the symbol sets might be an example of a written language. A number of problems have impeded progress in this area: the non-availability of reliable corpuses describing the specific symbols, a lack of agreement on the definition of individual symbol types, small corpus sizes ranging from a couple of hundred to a few thousand symbols, the often short nature of individual inscriptions (one to three symbols in length) and the lack of

*Author for correspondence (r.lee@exeter.ac.uk).

Received 27 January 2010
Accepted 26 February 2010

a technique to establish the level of communication of the symbols when sample sizes are small (Bouissac 1997). For known languages, statistical techniques such as phylogenetic methods have been used to aid in the reconstruction of ancient language histories (Warnow 1997; Foster & Toth 2003; Dunn *et al.* 2005) and the rates of linguistic evolution (Pagel *et al.* 2007; Atkinson *et al.* 2008). Recently, conditional entropies have been used to investigate the Indus script, but the conclusions were not definitive owing to the use of small, smoothed datasets, the comparative nature of the technique and its inability to differentiate between different lexigraphic systems (Rao *et al.* 2009). This paper describes a technique that incorporates linguistic functions in order to quantify the level of communication in these small, 'incomplete' symbol datasets and thus differentiate between the different possible character types of writing (the term incomplete is used here to describe text samples that have insufficient data to properly characterize the character lexicons).

The Picts were an Iron Age society that existed in Scotland from *ca* AD 300–843 when the Dalriadic Scot, Kenneth, son of Alpin, took the Pictish Kingship. The Picts are recorded in the writings of their contemporaries—the Romans, the Anglo-Saxons and the Irish but, other than a copy of their King list, they left no written record of themselves (Wainwright *et al.* 1955; Anderson 1973). The Picts 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 categorized into two types as shown in figure 1: (i) Class I stones, numbering between 180 and 195, consist of undressed stones with the symbols inscribed onto the rock and (ii) Class II stones, numbering between 60 and 65 stones, contain the depiction of a cross, use dressed stones 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 Stones. The stones contain between one and eight symbols, with the commonest syntax being one or two symbols. Over a century ago, Allen and Anderson visually catalogued the then known Pictish Symbol Stones and categorized their symbols (Allen & Anderson 1903). While no visual categorization catalogue of the possible different symbol types exists, the Pictish Symbol Stones have recently been completely categorized by Mack (1997), although he uses a smaller set of 43 symbol types than do earlier workers (Allen & Anderson 1903; Diack 1944; Forsyth 1997). Over the last century, a wide variety of 'meanings' for the symbols have been proposed, from pagan religious imagery to heraldic arms (Allen & Anderson 1903; Diack 1944; Wainwright *et al.* 1955; Mack 1997), but it is only recently that the question as to whether they might be a written language has been asked (Samson 1992; Forsyth 1997). However, in the absence of a suitable technique, the call for an analysis to establish whether the symbols were a script and that the stones might be memorial in character remains unanswered (Samson 1992; Forsyth 1997).

2. Theory

The problem that the Pictish symbols pose can be broken into a couple of questions. (i) Are they random in nature (admittedly unlikely since they appear to have been carved for a purpose)? (ii) If it is unlikely that they are random, then what type of communication do they convey: (a) semasiography, where

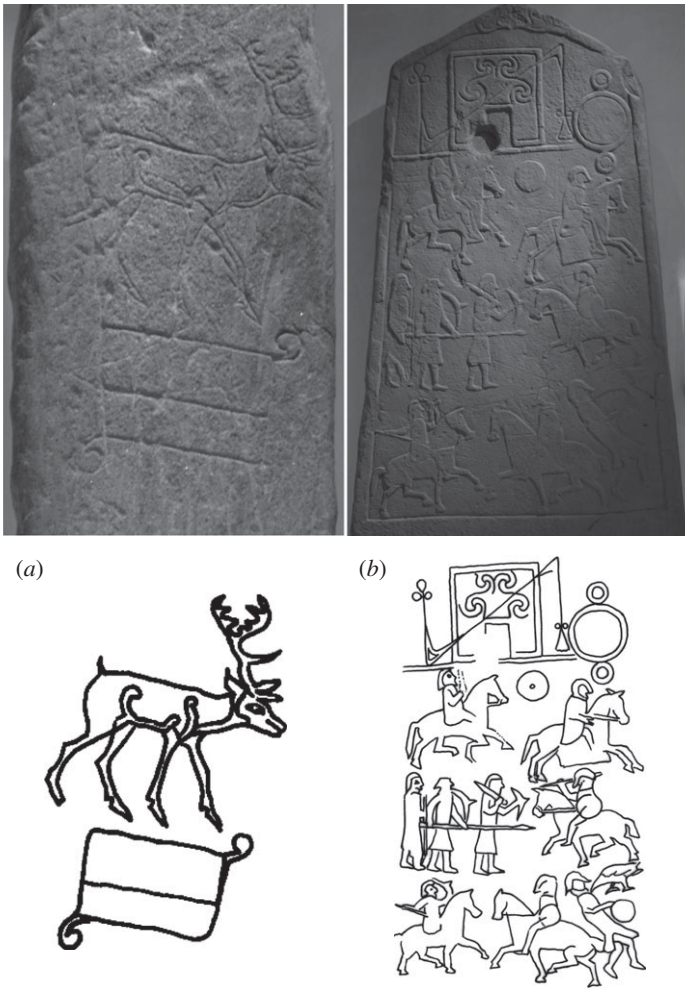

Figure 1. Pictish Symbol Stones. (*a*) Class I stone, 'Grantown', with two symbols—stag and rectangle. (*b*) 'Aberlemno 2', a Class II stone with two symbols—divided rectangle with a Z rod and triple disc, as well as other imagery (a battle, the cross is on the other face).

information is communicated without reference to verbal language forms (such as heraldic characters that have no lexigraphic value in themselves but identify a person, position and place) or (b) lexigraphic scripts, where the characters embody the form of verbal language (e.g. logograms representing words and syllables (non-phonetically), syllabograms representing syllables (phonetically), alphabetic signs representing letters (parts of syllables) and code characters (e.g. Morse code) representing parts of letters (Powell 2009)?

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 (Shannon 1993*a*). A measure of the average uncertainty of character occurrence is the uni-gram (single character) entropy, F_1 (Shannon 1993*b*). In a set of N_u different characters, the first-order

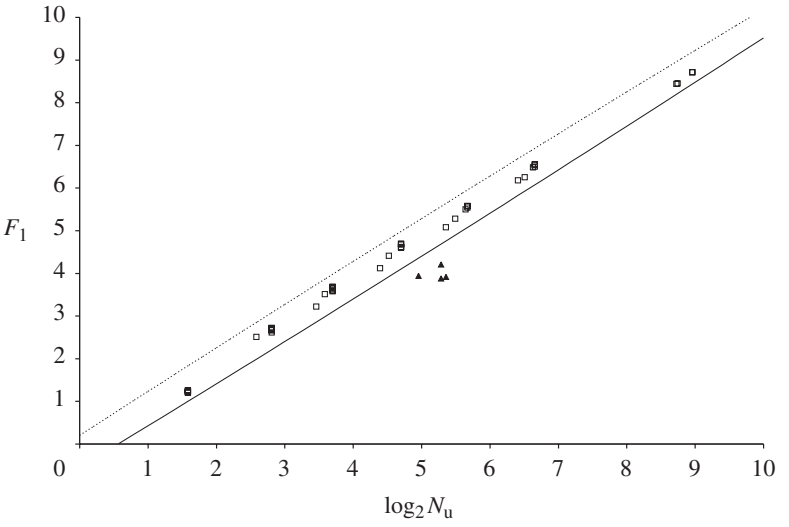

Figure 2. Plot of F_1 (uni-gram entropy) versus $\log_2 N_u$ (number of different uni-grams) showing the 99.9% confidence ellipse for prediction of the random data. This figure tests whether the stones correspond to similar-sized samples from a finite alphabet of equal relative frequency of unigram occurrence. It is extremely unlikely that the observed values for the Pictish Stones would occur by chance were they indeed a random dataset. Open squares, random data; filled triangles, Pictish Symbol Stones; dotted line, upper 99.9% confidence ellipse for prediction; solid line, lower 99.9% confidence ellipse for prediction.

entropy (F_1) is given by

$$F_1 = -\sum_{i=1}^{N_u} p_i \log_2 p_i, \tag{2.1}$$

where p_i is the relative frequency of occurrence of a character calculated from the dataset. In a large dataset of random characters (i.e. sampled with equal probability from a finite lexicon), all uni-grams appear with the same frequency, so $p_i = 1/N_u$, thus $F_1 = \log_2 N_u$. However, small sample sets of random characters will deviate from this since the incompleteness of the sample available will lead to unequal relative frequencies being observed. Thus, in small sample sets of random characters, $p_i \sim 1/N_u$ when estimated from the sample. Figure 2 confirms that $F_1 \sim \log_2 N_u$ for 40 sets of random data of small sample size ranging from 15 to 1000 characters. Systems for which F_1 is different from $\log_2 N_u$ (with respect to the confidence ellipse for prediction) can be identified as non-random and characteristic of writing.

The simplest gauge of character-to-character information in written communications is the di-gram entropy, F_2, the measurement of the average uncertainty of the next character when the preceding character is known. Shannon defined F_2 as (Shannon 1993*b*)

$$F_2 = -\sum_{i,j} p(b_i, j) \log_2 p(b_i, j) + \sum_i p(b_i) \log_2 p(b_i) = -\sum_{i,j} p(b_i, j) \log_2 (b_i, j) + F_1, \tag{2.2}$$

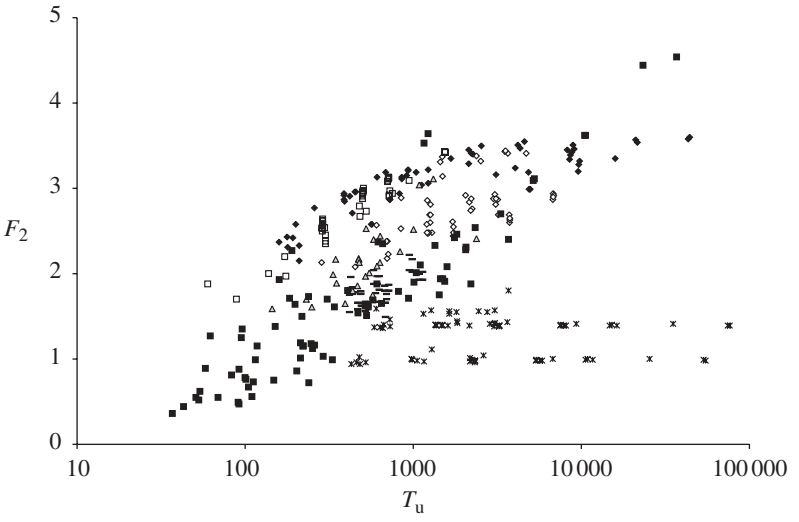

Figure 3. Plot of F_2 (di-gram entropy) versus T_u (text size based on the total number of uni-gram characters) for a wide range of texts and character types. The di-gram entropy is similar for different types of characters in datasets with small sample size owing to the incomplete nature of the di-gram lexicons. 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.

in which b_i is a uni-gram (single character), j is an arbitrary character following $b_i, p(b_i, j)$ is the relative frequency of the di-gram (pair of characters) b_i, j and F_1 is the uni-gram entropy where the summation is from 1 to N_u for a set of N_u uni-gram characters. F_2 is at a maximum when all the possible di-grams appear with the same frequency (Yaglom & Yaglom 1983). Thus, as the ability to predict the next character increases, the di-gram entropy decreases.

Figure 3 shows the di-gram entropy for over 400 datasets of scripts containing small samples of characters. Each dataset contains between 30 and 10 000 word equivalents for a wide variety of character types and scripts. The scripts analysed cover sematograms (Heraldic characters), logograms (Chinese), syllabaries (Linear B and Egyptian hieroglyphs), alphabetic systems (analysed at letter, syllable and word level) 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 texts cover prose, poetry, monumental inscriptions and genealogical lists (King lists, marriage and birth lists). Full details are given in §5. Unfortunately, figure 3 shows that, for systems containing only small samples of characters, the Shannon di-gram entropy as a function of text size (as given by the total number of uni-grams, T_u) cannot be used to differentiate between the different character types or even between the types of writing (semasiography or lexigraphic). The reason for this failure to differentiate between character type is that, at these small sample sizes, there are insufficient data to properly characterize the character lexicon, which affects the observed N-gram distributions and hence entropy. In this paper, the term

incomplete is used to describe text samples that are insufficiently representative to characterize the underlying character lexicons. For a text of a given size, the degree to which the N-gram lexicon is incomplete will be strongly affected by a number of linguistic phenomena, including

— type of character used to code for the communication,
— size of the character lexicon used (e.g. texts with constrained (or limited) vocabularies pull from a pool of available words that is limited to a fraction of a normal vocabulary),
— grammar of the unknown language (e.g. the system of inflection within the language),
— syntax of the unknown language (i.e. the word order), and
— degree of standardized spelling (i.e. many inscriptions do not use standardized spelling).

At very large datasets (100+ million words), these phenomena reduce the prediction ability of N-gram-based statistical language models used in such areas as speech and optical character recognition, document classification and machine translation (Rosenfeld 2000). As a consequence, linguistic-derived functions are used to make up some of the predictive deficiency (Rosenfeld 2000). Unfortunately, these functions are not appropriate for unknown systems with small sample size datasets and very incomplete character lexicons. In order to be able to compare the di-gram entropies of such datasets, a measure of the degree of 'incompleteness' or 'completeness' of the di-gram lexicons is needed. This paper proposes that a measure of the completeness of the di-gram lexicon (or its lack of incompleteness) can be derived from the number of different uni-grams and di-grams in the dataset.

For a text with a given number of different uni-grams, N_u, the number of different di-grams, N_d, will depend upon the incompleteness of the di-gram lexicon (which in turn is dependent upon the linguistic phenomena outlined above). Depending upon the degree of di-gram lexicon completeness, N_d will range between N_u (very incomplete) and $(N_u)^2$ (complete, but note that this is the theoretical maximum: actual lexicons will, in practice, always be less than $(N_u)^2$ since the rules of syntax and spelling will only allow a value less than $(N_u)^2$). Thus, a measure of the completeness in the di-gram lexicon for small samples can be obtained by calculating (N_d/N_u), where N_d is the number of different di-grams and N_u is the number of different uni-grams.

Figure 4 shows that the di-gram entropy is dependent upon this measure of the degree of completeness in the di-gram lexicon and shows differentiation between three types of lexigraphic character types (words, syllables and letters). Thus, this paper proposes normalizing F_2 by $\log_2(N_d/N_u)$ in order to define, for any text, a second-order function (U_r) adjusted for the di-gram lexicon completeness in a small sample size,

$$U_r = \frac{F_2}{[\log_2(N_d/N_u)]}. \tag{2.3}$$

As figure 4 shows, systems written in sematograms (heraldry) and lexigraphic code characters can have similar di-gram entropies to those of standard lexigraphic characters. However, while words, syllables and letters generally

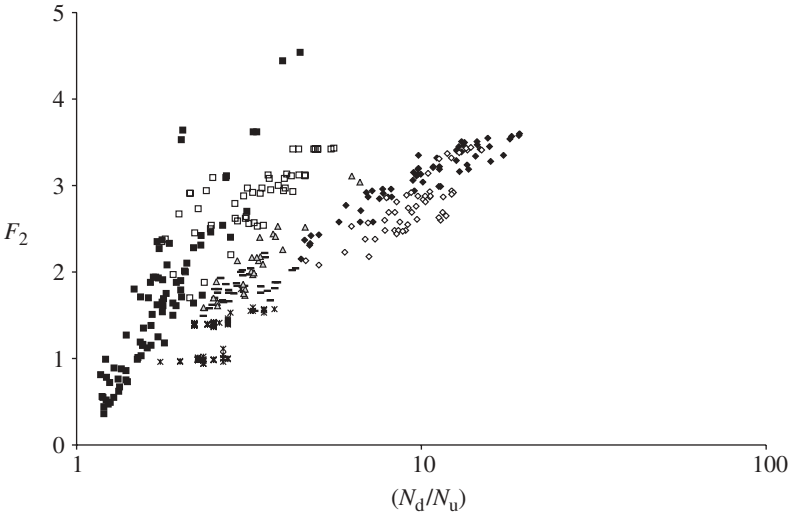

Figure 4. Plot of F_2 (di-gram entropy) versus N_d/N_u (degree of di-gram lexicon completeness) using a log-linear scale. The di-gram entropy for different types of characters is dependent upon the level of completeness of the di-gram lexicon. Dashes, sematograms—heraldry; filled diamonds, letters—prose, 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.

correspond to a fixed unit of language, sematograms and code characters have no fixed lexigraphic value in themselves, but are combined to produce a lexigraphic character. Thus, heraldic and code characters are, by their nature, intrinsically more repetitive than standard lexigraphic characters. For example, the morse code uses two characters in combination (with a space) and thus the four di-grams (dash–dash, dash–dot, dot–dash and dot–dot) are used repeatedly in order to build the 26 letters of the alphabet. By definition, repetitive di-grams are ones that appear more than once, thus a measure based on the degree of di-gram repetition in a text can be given by (S_d/T_d), where T_d is the total number of di-grams and S_d is the number of di-grams that appear only once in the text. For texts with a high degree of di-gram repetition, S_d approaches 0. The degree of di-gram repetition will have some dependency upon the degree of completeness in the di-gram lexicon—since texts with a relatively 'complete' di-gram lexicon will be more likely to have greater repetition of the di-grams and a lower value of S_d. Figure 5 shows that the degree of di-gram repetition (S_d/T_d) is dependent upon the degree of completeness in the di-gram lexicon (N_d/N_u) for texts using standard lexigraphic characters. Figure 5 also shows that heraldic and code characters do not follow the same dependency as standard lexigraphic characters. Thus, this paper proposes a di-gram repetition factor, C_r, defined as a linear combination of the two quantities

$$C_r = \frac{N_d}{N_u} + a\left(\frac{S_d}{T_d}\right), \tag{2.4}$$

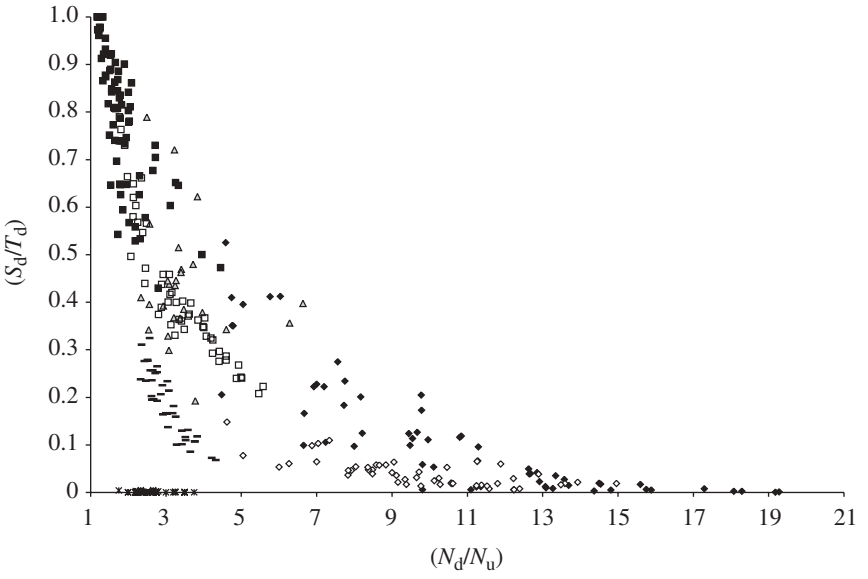

Figure 5. Plot of S_d/T_d (degree of di-gram repetition) versus N_d/N_u (degree of di-gram lexicon completeness). The degree of di-gram repetition is also dependent upon the level of completeness of the di-gram lexicon and that this dependency is different for standard lexigraphic characters compared with heraldic sematogram characters. 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.

where a is a constant estimated using cross-validation techniques in order to maximize the performance of a decision tree. Thus, the structure variables, U_r and C_r, are combinations of underlying linguistic variables that elucidate the key characteristics and structure of the data. Both U_r and C_r can be calculated for any type of communication system without any prior knowledge of the meaning of a system and have been used in a two-parameter decision tree to classify the following character types: (i) words, (ii) syllables, (iii) letters, and (iv) other characters such as heraldic sematograms and lexigraphic code characters.

3. Results

Figure 2 shows the 99.9 per cent confidence ellipse for prediction around 40 sets of random data. The datasets plotted in figure 2 were generated as follows: characters were sampled from a uniform distribution (i.e. with equal relative frequencies) into small units of text similar to the small units of glyphs seen on the stones. The key properties (total number of unigrams, number of different unigrams and the subsequent fraction of unigrams appearing only once) bracketed the corresponding properties observed in the stones. Figure 2 therefore tests whether the stones correspond to similar-sized samples from a finite alphabet of equal relative frequency of unigram occurrence. Texts based on written

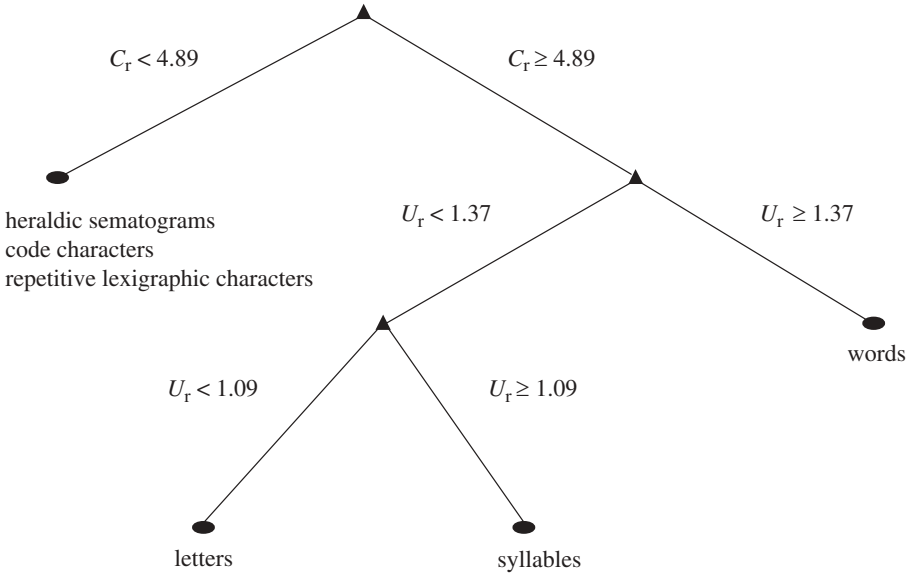

Figure 6. Two-parameter decision tree that separates repetitive text from non-repetitive text. This figure classifies the character types found in non-repetitive text into the three main lexigraphic character units (words, syllables, letters). Repetitive text consists of two main categories of characters: non-lexigraphic heraldic characters and lexigraphic code characters, as well as non-concordant letter, syllable and word character texts that are repetitive.

communication have an uneven distribution of characters that generally results in a lower F_1 for any value of N_u when compared with random sets. Figure 2 shows that the observed uni-gram entropy values for the Pictish symbols fall outside the 99.9 per cent confidence ellipse for prediction surrounding the random uni-gram dataset. Hence, it is extremely unlikely that the observed values for the Pictish stones would occur by chance were they indeed a random dataset.

The structure variables U_r and C_r have been used in a decision-tree analysis to differentiate the majority of the character types found in written communication. Figure 6 shows a decision tree, cross validated using the Gini diversity index, with a successful allocation rate of 99.1 per cent. (The performance of the classifier remains constant provided two decimal places of the partition values of the variables are retained, thus the classifier estimates are optimal to two decimal places. An estimate of the value of the parameter a in equation (2.4) was obtained by cross validation. Full details of the validation are given in §5.) Texts with $C_r < 4.89$, where $C_r = N_d/N_u + 7(S_d/T_d)$, are repetitive in nature and are consistent with Heraldic character systems (sematograms) and code-character systems, both of which are characterized by highly repetitive character sequences. Unfortunately, some repetitive lexigraphic texts also fall in this group and so if a text has a $C_r < 4.89$, we cannot determine what character type is present using this tree. If, however, the texts have a $C_r \geq 4.89$, then we classify the character types as lexigraphic and, depending upon the value of U_r, determine whether the characters represent words, syllables or letters. It is generally easier to predict the next letter than the next word because of: (i) the spelling rules (e.g. q is usually

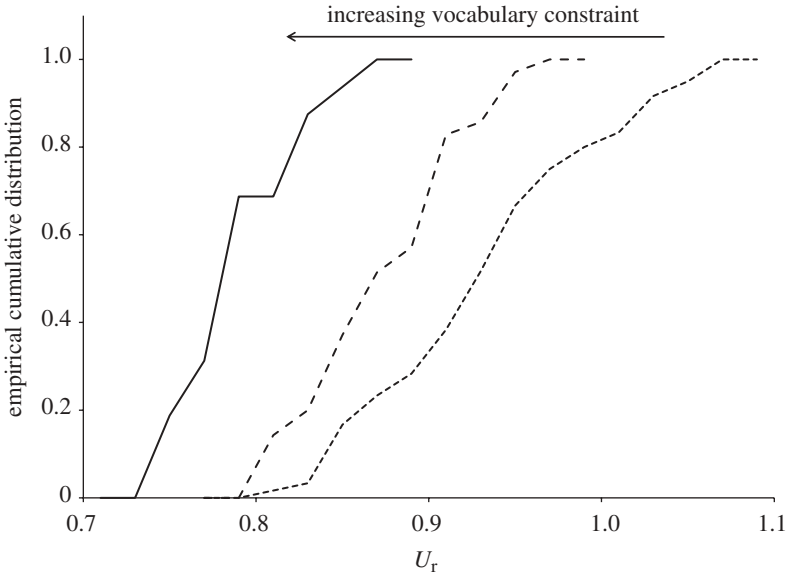

Figure 7. The effect on the empirical cumulative distributions of U_r ($F_2/\log_2(N_d/N_u)$) of increasing the character vocabulary constraint for letters. As the vocabulary becomes constrained, the distribution of U_r becomes narrower and the mean value decreases. Short-dashed line, empirical cumulative distribution for letter characters for all prose, poetry and inscriptions; long-dashed line, empirical cumulative distribution for letter characters for constrained genealogical lists; solid line, empirical cumulative distribution for letter characters from very constrained lists.

followed by u in English) and (ii) the constrained nature of the letter lexicon compared with word lexicon (26 letters in English versus word vocabulary of hundreds for even the most constrained texts). This means that for a given value of (N_d/N_u), we should expect F_2 for words to be larger than letters and thus U_r to be larger, and figures 3 and 6 show this to be the case. The separation of the lexigraphic character types is independent of language or sign type (i.e. alphabet, syllabogram and logogram scripts).

As a character vocabulary is constrained, it becomes easier to predict the next character, decreasing F_2 and U_r. The effect of constraining the character vocabulary upon the distribution of U_r is shown in figure 7. Within normal texts, there is a wide variety of vocabulary constraints. Constraining the character vocabulary (e.g. King lists and genealogical lists that are constrained to a vocabulary of names or genealogical lists using an even smaller vocabulary of familiar, diminutive names) gives a narrower distribution and a decreasing mean value of U_r.

The tree classifier developed suggests that the Pictish symbols are lexigraphic in nature because they have values of C_r in the interval [5.6, 6.2] (table 1). In particular, we infer that the Pictish symbols are not drawn from a distribution of heraldic characters. Table 1 shows that Mack's symbol categorization gives values of U_r that fall in the syllable side of the syllable/word boundary. However, Mack's categorization of the symbol types is much narrower than that of other workers (Allen & Anderson 1903; Diack 1944; Forsyth 1997). If Mack's categorization

Table 1. Values of C_r and U_r calculated for the Class I and Class II Pictish symbol stones using the symbol types given by Mack (1997) and by Allen & Anderson (1903).

stone class	symbol type set	C_r	U_r	character classification
I	Mack	5.92	1.28	syllable
II	Mack	6.11	1.36	syllable
I	Allen & Anderson	5.64	1.39	word
II	Allen & Anderson	6.16	1.45	word

is incorrect, then this will have the effect of artificially constraining the symbol lexicon, lowering F_2 and U_r. The larger symbol categorization proposed by Allen and Anderson in Early Christian Monuments of Scotland implies that the Pictish symbols are very constrained words, similar in constraint to the genealogical name lists. Thus, it is likely that the symbols are actually words, but that Mack's categorization has lowered the symbol di-gram entropy such that the data fall in the syllable band.

4. Discussion

Since there are many complete stones inscribed only with a single symbol, it seems unlikely (although not impossible) that the symbols are single syllables. In order to answer the question of whether the symbols are words or syllables, and thus define a system from which a decipherment can be initiated, a complete visual catalogue of the stones and the symbols will need to be created and the effect of widening the symbol set investigated. However, demonstrating that the Pictish symbols are writing, with the symbols probably corresponding to words, opens a unique line of further research for historians and linguists investigating the Picts and how they viewed themselves.

Having shown that it is possible to use an entropic 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 problems. For example, animal language studies using Shannon entropies are often hampered by small sample datasets (McCowan *et al.* 1999). By building a similar set of data for spoken or verbal human communication, it should be possible to make similar comparisons of the level of information communicated by animal languages.

5. Material and methods

(*a*) *Entropy calculations*

For all texts, a 'start/end' character was inserted at commas or full stops, otherwise all punctuation was removed and all spaces ignored (since many old inscriptions have little or no punctuation). F_0, F_1 and F_2 were calculated at the character levels appropriate for the text, e.g. alphabetic texts were mainly analysed at the letter and word level, syllabogram texts at the syllable and word level.

(*b*) *English texts*

Prose fiction texts were written under varying degrees of word constraint (normal texts have *ca* 4.3 letters/word, lightly constrained texts have between 3.6 and 4.0 letters/word and highly constrained texts have 2.5–3.0 letters/word). Graveyard texts from Kelsall Church of England graveyard were used. Text size varied from 35 to 10 000 words and were analysed at the letter, syllable and word level.

(*c*) *Chinese texts*

Prose and poetry texts from Yu Xuan Ji, Hong Lou Meng and Shijing texts were analysed (Lung 2009). Text size from 50 to 3000 word characters was analysed at the word level.

(*d*) *Universal Declaration of Human Rights text*

Languages analysed were English, Irish, Welsh, Norse, Turkish, Basque, Finnish and Korean at the word and letter level (UDHR 2008).

(*e*) *Ancient inscriptions from the British Isles*

Languages analysed were Latin, Anglo-Saxon, Old Norse, Ancient Irish, Old Irish and Old Welsh. Only whole words from translatable inscriptions were included. Each inscription was bracketed with a start/end character or a 'missing' character for incomplete inscriptions. All punctuation (if present) was removed. Ligated letters were separated into their constituent letters. Alphabet-specific characters were retained. Each corpus of specific inscription types was run as a single set. Irish inscriptions were split into an early tradition (ogam) and later tradition (uncial) with two different authors being used for the early tradition (Macalister 1945, 1949; McManus 1991). Welsh inscriptions were split into an early tradition (Class I) and later tradition (Class II and III) (Nash-Williams 1950). Roman memorials were split into two groups, those found at Hadrian's Wall and the rest (Collingwood & Wright 1965). Inscribed stones, slabs, crosses and personal items from the Anglo-Saxon period were used (Okasha 1971). Isle of Man Norse runic inscriptions (Page 1995) and Southern Scottish inscriptions (Thomas 1991–1992) were used. The text sizes ranged from 50 to 2200 words and were analysed at the letter, syllable and word level.

(*f*) *Egyptian monumental texts*

These were transcribed in two ways: using the standard modern spelling (which removes superfluous hieroglyphs and applies a standard spelling) and an 'as observed' reading of the hieroglyphs (Zauzich 2004). The Egyptian hieroglyphs in these texts are primarily syllabic in nature, being predominantly a mix of single and bilateral glyphs. Text size was 250 words analysed at the word and syllable level.

(*g*) *Mycenaean lists* (*Linear B*)

These were split into two groups: military lists and others (Palmer 1998). Text size was 450–600 words analysed at the word and syllable level.

(*h*) *King lists*

These contain only the names of child and parent(s) for the Pictish, Anglo-Saxon, Scottish, English, Cashel and Munster lineages (Anderson 1973; Byrne 1973; Montague-Smith 1992). Text size was 60–175 words analysed at the word and letter level.

(*i*) *Genealogical lists*

English baronial genealogies containing: (i) Christian names of the child and parent(s), (ii) Christian names of bride and groom, and (iii) surnames of bride and groom were used (Sanders 1960). A second set of lists was created using familiar, diminutive names instead (e.g. 'Al for Alan, Alfred and Albert'). Text size was 250–1500 words analysed at the word and letter level.

(*j*) *Sematogram heraldic*

A normal distribution of arms from the Heraldic Arms of British Extinct peerages (1086–1400) was used (Burke 1962). The charges (symbols) on the shield were used as characters for analysis. The colour of the charge was also used for analysis. A simplified set of characters was also generated using only the base symbols, e.g. (i) all the different lion charges such as rampant or passant are classified as a 'lion' character and (ii) all different cross charges such as bourdonny and fleuretty are classified as 'cross' in the base-symbol categorization. Each arms was read as observed symbols from bottom to top. Text size was 400–1200 symbols.

(*k*) *Subletter coded systems*

A range of English texts was transposed using morse code and a three-character code for the letters. Text size was 400–75 000 characters.

(*l*) *Random*

Randomly generated characters texts, ranging from sets of two to 100 different characters, were used with texts sizes of 15–1000 characters. The texts bracketed the values observed in the stones for the total number of uni-grams (T_{u}), the number of different uni-grams (N_{u}) and the fraction of uni-grams appearing only once.

(*m*) *Pictish symbols*

These were split into Class I and Class II symbols. The symbols were read as observed from top to bottom, left to right, using Mack's symbol set and the symbol set given in Early Christian Monuments of Scotland (Allen & Anderson 1903; Mack 1997). The symbol data were taken only from complete stones, which form the majority of the stones.

(n) Statistical analysis

The 99.9 per cent confidence ellipse for prediction was calculated from the random character data assuming a bi-variate normal distribution for F_1 and $\log_2 N_u$. (Histograms and normal probability plots of the marginal distributions show no obvious departure from normality.) The confidence ellipse is centred on the mean of the random data (Mardia *et al.* 1979).

Classification trees, constructed using the classification-tree methodology of Breiman *et al.* (1984), are non-parametric models to describe the variation in a response variable (the categorical character-class variable here) as a function of a number of explanatory variables (the continuous-structure variables U_r and C_r here) for a sample of data in a two-step approach. Firstly, the sample is partitioned, by means of successive binary partitions, such that the subsets eventually formed are as homogeneous as possible with respect to the response variable, quantified using one of a number of criteria (the Gini diversity index here). To avoid over-fitting, subsets of the partition can then be recombined if the resulting loss of homogeneity is not large (assessed using a 10-fold cross-validation strategy) in the second stage, known as pruning.

Cross validation is a well-known data resampling method to estimate model predictive performance, and possibly thereby the optimal values of one or more tuning parameters (Stone 1974; Picard & Cook 1984); for example, the value of parameter a in structure variable C_r, or the optimal extent of pruning. In cross validation, the sample set is partitioned into two or more subsets. One subset is typically withheld, while the remaining subsets are used to construct the model, in this case a classification tree. The withheld subset is then treated as an independent test set with which to estimate model performance, possibly as a function of one or more tuning parameters. The withheld subset is then reinstated, another subset withheld and the procedure repeated until each subset has been withheld exactly once. Overall model performance is then calculated by summing performance over all withheld subsets and the corresponding optimal value of tuning parameter selected. There are many possible refinements to the cross-validation procedure. In general, it is necessary to explore the sensitivity of cross-validation-based inferences as a function of the parameters of the cross-validation strategy. We found that, for the current application, inferences were generally insensitive to these choices—for instance, any value between 6 and 9 of the parameter a in structure variable C_r gives a cross-validation performance of greater than 99 per cent. The performance of the classifier remains constant provided two decimal places of the partition values of the variables are retained, and thus the classifier estimates are optimal to two decimal places.

Glossary

T_u: total number of characters (uni-grams) in a text. T_u is the text size for that character type, thus a text of 200 words may have a letter text size of 900 letters and a syllable text size of 520 syllables.

N_u: the number of different characters (uni-grams) in a text. Thus, a 200 word text might have 25 different letters, 100 different syllables and 130 different words.

T_d: total number of character pairs (di-grams) in a text. T_d is the character pair text size for that character type, thus a text of 200 words may have 199 word pairs and have a letter-pair text size of 899 letter pairs and a syllable-pair text size of 519 syllable pairs.

N_d: the number of different character pairs (di-grams) in a text. Thus, a 200 word text might have 270 different letter pairs, 390 different syllable pairs and 190 different word pairs.

S_d: the number of different character pairs that appear only once in a text.

We thank Nigel Tait, Clive McDonald, Richard Price and John Love for critical discussions and reading of the manuscript; Nigel Tait for technical help with the coding of the macros; and the referees for their help in improving the paper.

References

Allen, J. R. & Anderson, J. 1903 *The early Christian monuments of Scotland*. Balgavies, Angus: The Pinkfoot Press. (Reprinted by The Pinkfoot Press 1993.)

Anderson, M. O. 1973 *Kings and kingship in early Scotland*, pp. 119–204. Edinburgh, UK: Scottish Academic Press.

Atkinson, Q. D., Meade, A., Venditti, C., Greenhill, S. J. & Pagel, M. 2008 Languages evolve on punctuational bursts. *Science* **319**, 588. (doi:10.1126/science.1149683)

Bouissac, P. A. 1997 In *Archaeology and language I* (eds R. Blench & M. Spriggs), pp. 53–62. London, UK: Routledge.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 *Classification and regression trees,* ch. 2–8. London, UK: Chapman & Hall.

Burke, B. 1962 *A genealogical history of the dormant, abeyant, forfeited and extinct peerages of the British Empire*. London, UK: Burke's Peerage Ltd.

Byrne, J. F. 1973 *Irish kings and high-kings*, pp. 275–301. New York, NY: St Martins Press.

Collingwood, R. G. & Wright, R. P. 1965 *The Roman inscriptions of Britain*, *volume I: inscriptions on stone.* Oxford, UK: Oxford University Press.

Diack, F. C. 1944 In *The inscriptions of Pictland* (eds W. M. Alexander & J. Macdonald), pp. 7–42. Aberdeen, UK: Third Spalding Club.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. 2005 Structural phylogenetics and reconstruction of ancient language history. *Science* **309**, 2072–2075. (doi:10.1126/science.1114615)

Forsyth, K. F. 1997 In *The worm, the germ, and the thorn* (ed. D. Henry), pp. 85–98. Balgavies, Angus: The Pinkfoot Press.

Foster, P. & Toth, A. 2003 Toward a phylogenetic chronology of Ancient Gaulish, Celtic and Indo-European. *Proc. Natl Acad. Sci. USA* **100**, 9079–9084. (doi:10.1073/pnas.1331158100)

Li, X., Harbottle, G., Zhang, J. & Eang, C. 2002 The earliest writing? Sign use in the seventh millennium BC at Jihua, Henan Province China. *Antiquity* **77**, 31–45.

Lung, M. 2009 Chinese text initiative, University of Virginia library. See http://etext.virginia.edu/chinese/.

Macalister, R. A. S. 1945 *Corpus inscriptionum insularum celticarum*, vol. I. Dublin, Ireland: Dublin Stationery Office. (Reprinted by Four Courts Press 1996.)

Macalister, R. A. S. 1949 *Corpus inscriptionum insularum celticarum*, vol. II. Dublin, Ireland: Dublin Stationery Office.

Mack, A. 1997 *Field guide to the Pictish symbol stones*. Balgavies, Angus: The Pinkfoot Press (updated 2006).

Mardia, K. V., Kent, J. T. & Bibby, J. M. 1979 *Multivariate analysis*, 1st edn, ch. 2, pp. 38–40. London, UK: Academic Press.

McCowan, B., Hanser, S. F. & Doyle, L. R. 1999 Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Anim. Behav.* **57**, 409–419.

McManus, D. 1991 *A guide to Ogam*. Maynooth Monographs 4. Maynooth, Ireland: An Sagart.

Montague-Smith, P. W. 1992 *The royal line of succession*. Andover, UK: Pitkin.

Nash-Williams, V. E. 1950 *The early Christian monuments of Wales*. Cardiff, UK: University of Wales Press.

Okasha, E. 1971 *Hand-list of Anglo-Saxon non-runic inscriptions*. Cambridge, UK: Cambridge University Press.

Page, R. I. 1995 *Runes and runic inscriptions*, pp. 207–244. Woodbridge, VA: Boydell.

Pagel, M., Atkinson, Q. D. & Meade, A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)

Palmer, L. R. 1998 *The interpretation of Mycenaean Greek texts*. Oxford, UK: Oxford University Press.

Picard, R. & Cook, D. 1984 Cross-validation of regression models. *J. Am. Stat. Assoc.* **79**, 575–583. (doi:10.2307/2288403)

Powell, B. B. 2009 *Writing: theory and history of the technology of civilization,* pp. 1–59. Chichester, UK: Wiley-Blackwell.

Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R. & Mahadevan, I. 2009 Entropic evidence for linguistic structure in the Indus script. *Science* **324**, 1165. (doi:10.1126/science.1170391)

Rosenfeld, R. 2000 Incorporating linguistic structure into statistical language models. *Phil. Trans. R. Soc. Lond. A* **358**, 1311–1324. (doi:10.1098/rsta.2000.0588)

Samson, R. 1992 The reinterpretation of the Pictish symbols. *J. Brit. Arch. Assoc.* **145**, 29–65.

Sanders, I. J. 1960 *English baronies*. Oxford, UK: Oxford University Press.

Shannon, C. E. 1993*a* A mathematical theory of communication. In *Claude Shannon collected papers* (eds N. J. A. Sloane & A. D. Wyner), pp. 5–83. Piscataway, NJ: IEEE Press.

Shannon, C. E. 1993*b* Prediction and entropy of printed English. In *Claude Shannon collected papers* (eds N. J. A. Asloane & A. D. Wyner), pp. 194–208. Piscataway, NJ: IEEE Press.

Stone, M. 1974 Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B* **36**, 11–147.

Thomas, C. 1991–1992 The early Christian inscriptions of southern Scotland. *Glasgow Archaeol. J.* **17**, 1–10.

UDHR 2008 The Universal Declaration of Human Rights. Library of translations, Office of the High Commissioner for Human Rights. See http://www.unhchr.org/.

Wainwright, F. T., Feachem, R. W., Jackson, K. H., Pigott, S. & Stevenson, R. B. K. 1955 *Problem of the Picts*. Perth, WA: Melven Press. (Reprinted by Melven Press 1980.)

Warnow, T. 1997 Mathematical approaches to comparative linguistics. *Proc. Natl Acad. Sci. USA* **94**, 6585–6590. (doi:10.1073/pnas.94.13.6585)

Yaglom, A. M. & Yaglom, I. M. 1983 *Probability and information*, pp. 44–100 [transl. V. K. Jain]. Dordrecht, The Netherlands: D. Reidal Publishing Co.

Zauzich, K.-T. 2004 *Discovering Egyptian hieroglyphs* [transl. A. M. Roth]. London, UK: Thames and Hudson.