The comparative use of Shannon entropy to determine the level of communication expressed by prehistoric petroglyphs.

Rob Lee

School of Biosciences, Geoffrey Pope Building, University of Exeter, Stocker Road, Exeter, EX4 4QD, r.lee@exeter.ac.uk

Prehistoric ‘rock-art’ is found throughout the world, from the Palaeolithic cave art of France and Spain to the rock images of the Western Native Americans of North America. Common to archaeology and semiotics is the problem of whether these prehistoric petroglyphs are early examples of written languages. Unfortunately the petroglyph data sets are often small and this, coupled with the lack of a technique to systematically compare these undersampled datasets with known communication systems, has hampered the ability to determine if specific petroglyphs are forms of writing.

Calculation of the degree of uncertainty in being able to predict the next character in a communication system (2nd order Shannon entropy) gives a measure of the degree of information in the system and can be applied to any type of character within the communication. Plotting the 2nd order Shannon entropy against two undersampling measures for a range of communication systems from heraldry through Egyptian inscription hieroglyphs to modern language texts separates the different communication character types by their relative positions on the graphs. This paper reports on the development of this quantitative, comparative tool and its application to two rock inscribed petroglyph sets from Scotland; the Neolithic ‘Cup and Ring’ carvings and the Late Iron-Age Pictish Symbol stones, to try and determine whether they are forms of writing.

Prehistoric petroglyphs have a wide range of form, reflecting the different societies around the world that produced them. Petroglyphs vary from motifs painted on rock faces, such as the shamanistic paintings of the San of South Africa and Western Native Americans of North America (1), to rock inscribed carvings such as the Neolithic ‘Cup and Rings’ shapes of Atlantic Europe (2) and the Late Iron-Age symbols of Scotland (3, 4). A longstanding dilemma has been to determine whether any of the petroglyph sets might be an example of a written language (5). A number of problems have impeded progress in this area: the availability of reliable corpuses describing the specific petroglyphs, a lack of agreement on the definition of individual petroglyph types, small corpus sizes ranging from a couple of hundred to a few thousand petroglyphs, and lack of a technique to establish the level of communication of the petroglyphs in such undersampled systems (5). Thus, the interpretation of the petroglyphs has mainly remained the province of art historians and anthropologists. For known languages, comparative mathematical techniques such as phylogenetic methods have been used to aid in the reconstruction of ancient language histories (6, 7, 8) and the rates of linguistic evolution (9, 10). Likewise, comparative statistical techniques have been developed to authenticate artwork and date books and prints (11, 12). This paper describes a new, comparative technique that quantifies the level of communication in the petroglyphs, by determining the degree of petroglyph to petroglyph uncertainty, and thus establish whether they are a form of writing.

The Picts were an Iron Age society that existed in Scotland from ca. 300-843 AD when the Dalraidic Scot, Kenneth McAlpin 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 (4, 13). The Picts did, however, leave a range of finely carved stones inscribed with unknown glyphs, known as "Pictish Symbol Stones". The Pictish Symbol Stones are categorised into two types as shown in Figure 1: i) Class I stones, numbering between 180-195, consist of undressed stone with the symbols inscribed onto the rock, ii) Class II stones, numbering between 60-65 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. The stones contain between 1-8 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 (15). Whilst no modern visual catalogue of both the stones and the symbols exists, the Pictish Symbol Stones have recently been completely categorized by Mack (3), although he uses a smaller set of 43 symbol types compared to earlier workers (14, 15, 16). Over the last century a wide variety of 'meanings' for the symbols have been proposed, from pagan religious imagery to heraldic arms (3, 4, 14, 15), but it is only recently that the question as to whether they might be a written language has been asked (16, 17). However, in the absence of a suitable technique, the call for a comparative analysis to establish whether the symbols were a script and might represent names remains unanswered (16, 17).

A second type of petroglyph has also been investigated. The 'Cup and Ring' carvings of the Neolithic period are found throughout Britain (2). But some of the most complex and best preserved are to be found in Argyll, Scotland, with the RCAHMS offering an excellent visual record and categorisation of these carvings (18). The carvings consist of a shallow circular depression (the cup) that can exist on its own or is encircled (either completely or partially) by a number of rings. These petroglyphs are found scattered over rock faces in groups of 1-300. Current views hold that they had a wide rage of meanings including the secular marking of territories and the spiritual representation of alternative states of mind and passage graves (2).

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 (19). The simplest gauge of character to character information 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 (20):

$$F_2 = -\sum_{i,j} p(b_i,j) \log_2 p(b_i,j) + \sum_{i} p(b_i) \log_2 p(b_i) = -\sum_{i,j} p(b_i,j) \log_2 p(b_i,j) + F_1 \qquad (1)$$

in which: b_i is a block of uni-grams [single characters], j is an arbitrary character following b_i, $p(b_i,j)$ is the probability of the di-gram [pair of characters] b_i,j, F_1 is the uni-gram entropy. F_N is at a maximum when all the possible N-grams appear with the same frequency. Thus, as the ability to predict the next character increases, the di-gram entropy decreases. Since F_2 can be calculated for any type of communication system without any prior knowledge of the meaning of a system, F_2 is used as one of the measures with which to analyse petroglyph corpuses.

Absolute values of Shannon N-gram entropies have been used to compare complexity in different languages, but these have been calculated on large blocks of well-sampled texts (mean sample sizes of at least 10 per different N-gram) (20, 21). However the use of absolute di-gram entropy values for datasets such as the Pictish Symbols where the mean sample size is only 1.5-2.5 that of the number of different di-grams is not appropriate. Animal communication researchers have tried using an absolute value of an entropy gradient, based on F_0, F_1 and F_2, for comparison but this also suffered from undersampling effects, compounded by the fact that the different entropies, F_N, had different levels of undersampling (22). Thus a comparative technique is needed that takes into account the effect of undersampling while allowing discrimination between the different character types in communication systems. In order to enable meaningful comparisons of F_2 between datasets of different sizes and degrees of undersampling two measures of undersampling have been defined:

Undersampling ratio, $R_{p/s} = N_d/N_u$ (2)

Undersampling fraction, $D_f = 1-(S_d/T_d)$ (3)

Where: N_d = number of different di-grams, N_u = number of different uni-grams, T_d = total number of di-grams, S_d = number of di-grams that appear only once in the text. The maximum theoretical value of N_d is $(N_u)^2$ thus in a fully sampled system, $R_{p/s}$ => Nu. In a fully sampled system, the F_d => 0 and D_f => 1. At the limit of undersampling, $R_{p/s}$ =>1 and D_f = 0.

Written communications can be classified by the type of characters used to convey the information; i) 'standard' systems where the individual characters, generally, correspond to a fixed expression (such as letters or syllables which correspond to a sound and words which correspond to a sound and meaning/s), ii) 'sub-letter' systems where the characters have little or no fixed expression in themselves (such as the dot or dash character of the Morse code). Each of these character types operates at a different level of information. The size of the character lexicon will depend upon the degree of constraint imposed on the communication. Thus, constrained texts are ones in which the pool of available words has been limited to a fraction of a normal vocabulary, for instance the reduced vocabularies found in genealogical lists of names. A wide variety of data has been employed in order that a representative comparison can be made between the different communication system character types and their degree of constraint. English prose texts were produced as a base case (av. letters/word = 4.3). English prose texts where the average word length was limited to between 2.5-4.0 letters/word and modern English graveyard inscriptions were used as constrained vocabulary texts. A single text (United Nations Declaration of Human Rights, UNDHR) in a range of Indo-European and non-Indo-European languages (English, Latin, Irish, Welsh, Turkish, Finnish and Basque) was used as a linguistic comparison (23).

Inscriptions, written on stone from the pre-Norman British Isles were used as they bracket the Picts in both time and place and employ different alphabets (Roman/Latin memorials (24), Ancient Irish memorials written in ogam script (25, 26), Early Irish memorials (27), Anglo-Saxon memorials in Latin and in Old English (28), Early and later tradition Welsh and S. Scottish memorials, primarily in Latin (29, 30), Manx Norse memorials written in runes (31). Egyptian hieroglyphic monumental texts and Mycenaean lists were used as a comparison to hieroglyphic and syllabic languages (32, 33). A selection of genealogical family lists of kings and baronies in the British Isles were used as list texts containing a very constrained vocabulary (6, 34, 35, 36). Heraldic arms from 1086-1400 in Britain were used as one example of a 'sub-letter' character type since the majority of the charges (symbols) to be found in a coat of arms do not correspond to a fixed expression, but together they convey the information of a single name (37). Nine English texts of varying vocabulary constraint were also encoded using two types of sub-letter characters, the Morse code and a 3 sub-letter character code. A series of random "texts" were also generated.

Results

For a given text size and number of different uni-grams, random "texts" will, because of the nature of random data, tend to a more even distribution of uni-grams than written texts and thus have a higher F_1. Figure 2 shows this separation of the random data from the written texts and petroglyphs and thus confirming that the neither the Pictish Symbols or the Cup and Rings are random.

It is generally easier to predict the next letter than the next word because of; a) the spelling rules (e.g. q is usually followed by u in English), b) the constrained nature of the letter lexicon compared to word lexicon (26 letters in English, vs. word vocabulary of 100's for even the most constrained texts). This means that for a given undersampling ratio, $R_{p/s}$, we should expect F_2 for words to be larger than letters. Figure 3A shows this to be the case with words, syllables and letters all falling into separate bands. This separation of the 'standard' character types is independent of language with the 7 modern (i.e. UNDHR data) and the 7 ancient languages from the British Isles (i.e. Inscriptions) and elsewhere (i.e. Egyptian, Mycenaean data) following the same trends. Likewise the separation of the 'standard' character types is independent of the form of the basic alphabet with hieroglyphic and syllabic languages (i.e. Egyptian, Mycenaean data) obeying the same rules as letter based scripts. As would be expected, constraining the word vocabulary (e.g. king lists, genealogical lists and English constrained data) decreases F_2, as it is easier to predict the next word, moving the data closer to the syllabic band. A similar decrease in the 'letter F_2' is seen in these texts. In general, the inscriptions tend to have a higher 'word F_2' than normal English texts. This is partly due to language differences (see the spread of the data for the UNDHR text), but also because they tend to have a few words that dominate the text, e.g. "maqi" (meaning "son of"). This simultaneously lowers the word uni-gram entropy, F_1, and increases F_2 since it becomes harder to predict what word will follow "maqi".

Three types of 'sub-letter' heraldic characters are shown; a) the full symbol character set including colour, b) a constrained set using the full symbol character set but without colour, c) a second constrained set using a simplified system of base symbol characters without colour. Figure 3A also shows that, for a given $R_{p/s}$, when the heraldic lexicon of characters is constrained, F_2 also decreases and thus they follow a trend similar to the 'standard' characters. Unfortunately, the 'sub-letter' systems appear to overlap the lower regions of the syllable and letter bands. However, Figure 3B shows that the 'sub-letter' systems are separable from the ‘standard’ character systems since ‘sub-letter’ communications are highly repetitive in their character sequences leading to a higher undersampling fraction, D_f at a given $R_{p/s}$ when compared to ‘standard’ character systems. Although it is undoubtedly possible to drive standard character systems into the region of sub-letter systems shown in Figure 3B, by the addition of repetitious statements this will decrease F_1 and increase F_2 (as seen with the inscriptions) making it easy to differentiate between the systems. Thus using these two plots (or the combined 3D plot shown in supplemental Fig. 3) gives a comparative tool that differentiates between the different character types in communication systems.

Figures 3A and B show where the Pictish Symbols and Cup and Rings appear using the comparative plots. The Cup and Rings are clearly conveying information at a level similar to the ‘sub-letter’ systems. This may mean that, like the ‘sub-letter’ characters, their expression is not fixed. The corollary may be that their “meaning” is different in different places. On the other hand, the Pictish Symbols appear within the ‘standard’ written character bands. Figure 4 focuses on the written characters and uses a log-lin plot of $R_{p/s}$ vs F_2 to help clarify where the Pictish Symbol data appears. The Pictish Symbols, using Mack’s categorisation (red squares, the two data points are for Class I and Class II stones), fall in the syllable band, but close to the word band. However, Mack’s categorisation of the symbol types is much narrower than other workers (14, 15, 16). If Mack’s categorisation is incorrect then this will have the effect of constraining the symbol lexicon and lowering F_2. The larger symbol categorisation proposed by Allen and Anderson in Early Christian Monuments of Scotland (ECMS, red diamonds) implies that the Pictish Symbols are very constrained words, similar in constraint to the genealogical name lists, and appearing in or at the edge of the word/syllable banding. Thus it is likely that the symbols are actually words, but that Mack’s categorisation has lowered the symbol di-gram entropy such that the data falls in the syllable band.

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 a comparative technique to investigate the degree of communication in very undersampled written systems, it may be possible to extend this to other areas with similar problems. For example, animal language studies are often hampered by small datasets giving very undersampled data (22*)*. By building a similar set of comparative data for spoken or verbal human communication it should be possible to make similar comparisons of the level of information communicated by animal languages.

Materials and Methods

English texts. Fiction texts were written under varying degrees of word constraint (normal texts have ca. 4.3 letters/word, lightly constrained texts have between 3.6-4.0 letters/word and highly constrained texts have 2.5-3.0 letters/word). Graveyard texts from Kelsall CoE graveyard were used. The fiction texts were split into a wide range of smaller texts. A "start/

end” character was inserted at commas or full stops. All punctuation was removed and all spaces were removed (since many old inscriptions have little or no punctuation).

United Nations Declaration of Human Rights. The texts were split into sets of smaller texts by inserting a “start/end” character at commas or full stops. All punctuation was removed and all spaces were removed.

Inscriptions. 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 and McManus). Welsh inscriptions were split into an early tradition (Nash-Williams Class I) and later tradition (Nash-Williams Class II & III). Roman memorials were split into two groups, those found at Hadrian’s Wall and the rest. Inscribed stones, slabs, crosses and personal items from the Anglo-Saxon period were used.

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. The Egyptian hieroglyphs in these texts are primarily a mix of single and bilateral glyphs with some triliteral glyphs and these characters have been taken as equivalent to syllables.

Mycenaean lists. These were split into two groups; military lists and others.

King and genealogical lists. Contain only the names of child and parent(s).

Sub-letter heraldic. A normal distribution of arms from the Heraldic Arms of British Extinct peerages (1086-1400) was used. 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 ii) all different cross charges such as bourdonny and fleuretty are classified as 'cross' in the base symbol categorisation. Each arms was read "as observed" symbols from bottom to top.

Sub-letter coded systems. A range of English texts was transposed using morse code and a 3 character code for the letters.

Random. Randomly generated characters texts, ranging from sets of 2 to 100 different characters, were used.

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. The symbol data was taken only from complete stones

Cup and Ring. These were read in two ways: a) from carving to carving using the closest carving as the next character, b) from carving to carving letting the eye be drawn by possible patterns. Cup and ring carvings were defined by the number of complete or incomplete rings round the cup, whether there was a connecting line between the rings and whether there was a line joining disparate cup carvings. Other carvings were defined as per RCAHMS, i.e. whether they had radiating lines, a spiral, a double spiral or a lock shape.

Table 1 summarises the types of different characters used in the data analysis for the different text classes and the ranges of text size.

I thank Pauline Ziman, 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.

1. Lewis-Williams D (2002), *The Mind in The Cave*, (Thames and Hudson, London).
2. Bradley R (1997), *Rock Art and the Prehistory of Atlantic Europe*, (Routledge, London).
3. Mack A (1997, updated 2006), *Field Guide to The Pictish Symbol Stones*, (The Pinkfoot Press, Balgavies, Angus).
4. Wainwright FT, Feachem RW, Jackson KH, Pigott S, Stevenson RBK, (1980 reprint of 1955 original) *Problem of The Picts*, (Melven Press, Perth).
5. Bouissac PA (1997) in *Archaeology and Language I*, eds Blench R, Spriggs M, Eds. (Routledge, London) pp. 53-62.
6. Warnow T (1997), Mathematetical approaches to comparative linguistics. *Proc. Natl. Acad. Sci. USA 94: 6585-6590.*
7. Foster P, Toth A (2003), Toward a phylogenetic chronology of Ancient Gaulish, Celtic and Indo-European. *Proc. Natl. Acad. Sci. USA 100: 9079-9084.*
8. Dunn M, Terrill A, Reesink G, Foley RA, Levinson SC (2005), Structural phylogenetics and reconstruction of ancient language history. *Science 309: 2072-2075*.

9. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M (2008), Languages evolve on punctuational bursts. *Science 319: 588.*

10. Pagel M, Atkinson QD, Meade A (2007), Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature 449: 717-720.*

11. Lyu S, Rockmore D, Farid H (2004), A digital technique for art authentication. *Proc. Natl. Acad. Sci. USA 101: 17006-17010.*

12. Hedges SB (2006), A method for dating early books and prints using image analysis. *Proc. Royal Soc. A - Math. Phys. & Eng. Sci. 46: 3555-3573.*

13. Anderson MO (1973), *Kings and Kingship in Early Scotland*, (Scottish Academic Press, Edinburgh).

14. Diack FC (1944) in *The Inscriptions of Pictland*, eds Alexander WM, Macdonald J (Third Spalding Club, Aberdeen) pp 7-42.

15. Allen JR, Anderson J (1993 reprint of 1903 original), *The Early Christian Monuments of Scotland*, (The Pinkfoot Press, Balgavies, Angus).

16. Forsyth KF (1997) in *The Worm, The Germ, and The Thorn*, D. Henry, Ed., (The Pinkfoot Press, Balgavies, Angus) pp. 85-98.

17. Samson R (1992), The Reinterpretation of The Pictish Symbols. *J. Brit. Arch. Assoc.* 145: 29-65.

18. The Royal Commission on the Ancient and Historical Monuments of Scotland (1988), *Argyll Volume 6: Mid Argyll & Cowal, Prehistoric & Early Historic Monuments*, (RCAHMS, Glasgow).

19. Shannon CE (1993) in *Claude Shannon Collected Papers*, eds Sloane NJA, Wyner AD, (IEEE Press, Piscataway, NJ), pp. 5-83.

20. Shannon CE (1993) in *Claude Shannon Collected Papers*, eds Sloane NJA, Wyner AD, (IEEE Press, Piscataway, NJ), pp. 194-208.

21. Yaglom AM, Yaglom IM (1983), *Probability and Information*, transl. Jain VK, (D. Reidal Publ. Co., Dordrecht).

22. McCowan B, Hanser SF, Doyle LR (1999), Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour* 57: 409-419.

23. Office of the High Commissioner for Human Rights: The Universal Declaration of Human Rights, http://www.unhchr.ch/udhr/ (2008).

24. Collingwood RG, Wright RP (1965), *The Roman Inscriptions of Britain, Volume I, Inscriptions On Stone,* (OUP, Oxford).

25. Macalister RAS (1996 reprint of 1945 original), *Corpus Inscriptionum Insularum Celticarum, Volume I*, (Four Courts Press, Dublin).

26. McManus D (1991), *A Guide to Ogam*, Maynooth Monographs 4, (An Sagart, Maynooth)

27. Macalister RAS (1949), *Corpus Inscriptionum Insularum Celticarum, Volume II*, (Dublin Stationary Office, Dublin).

28. Okasha E (1971), *Hand-list of Anglo-Saxon Non-runic Inscriptions*, (CUP, Cambridge)

29. Nash-Williams VE (1950), *The Early Christian Monuments of Wales*, (University of Wales Press, Cardiff).

30. Thomas C (1991-2), The Early Christian Inscriptions of Southern Scotland. *Glasgow Archaeological Journal* 17: 1-10.

31. Page RI (1995), *Runes and Runic Inscriptions*, (Boydell, Woodbridge).

32. Zauzich K-T (2004), *Discovering Egyptian Hieroglyphs*, transl. Roth AM (Thames and Hudson, London, 2004).

33. Palmer LR (1998), *The Interpretation of Mycenaean Greek Texts*, (OUP, Oxford).

34. Montague-Smith PW (1992), *The Royal Line of Succession*, (Pitkin, Andover).

35. Byrne JF (1973), *Irish Kings and High-Kings*, (St Martins Press, New York).

36. Sanders IJ (1960), *English Baronies* (OUP, Oxford).

37. Burke B (1962), *A Genealogical History of the Dormant, Abeyant, Forfeited and Extinct Peerages of the British Empire*, (Burke's Peerage Ltd, London, Facsimile edition).

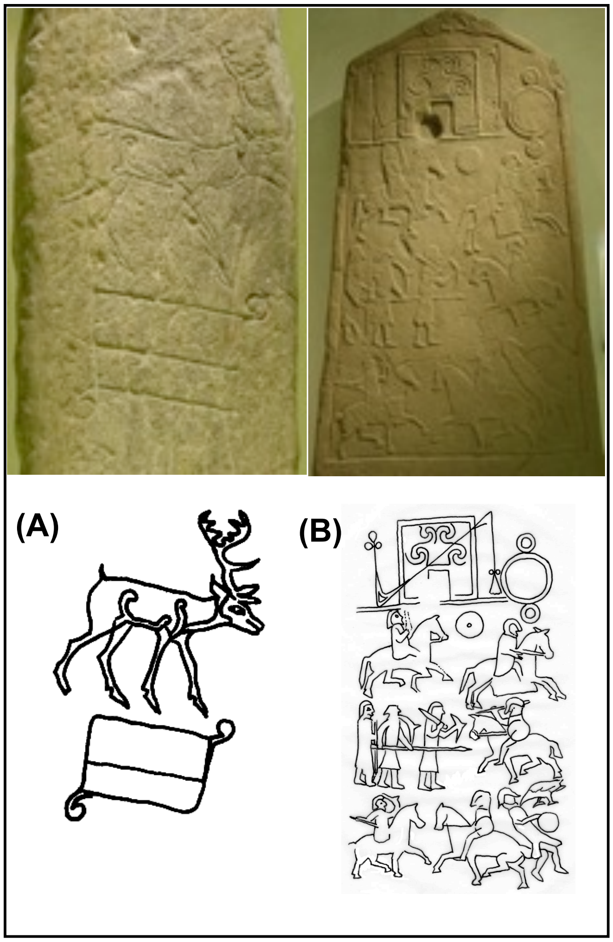


Fig. 1. Pictish Symbols Stones. (A), Class I stone, ‘Grantown’, with two symbols – stag and rectangle. (B), ‘Aberlemno 2’, a Class II stone with two symbols – divided rectangle with Z rod and triple disk, as well as other imagery (a battle, the cross is on the other face).

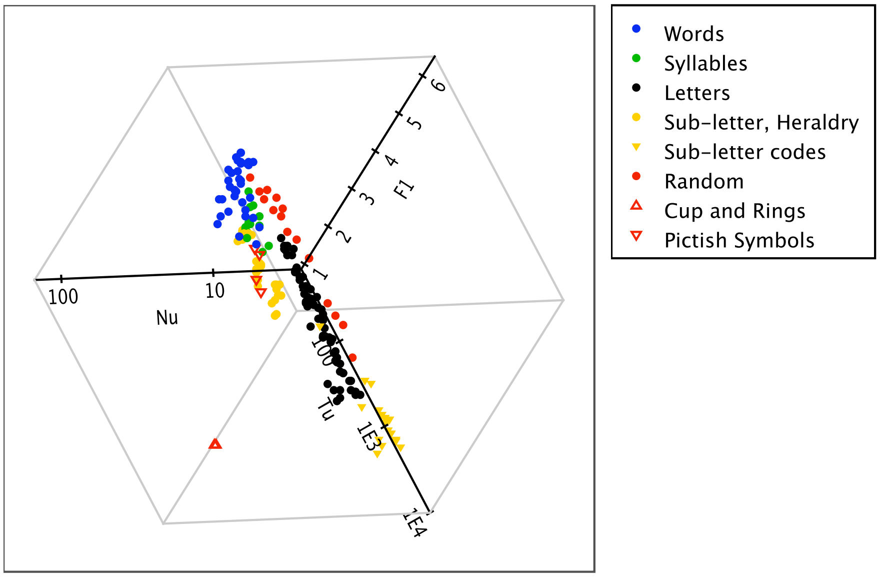


Fig. 2. Comparative analysis of the uni-gram character types of written communication systems using a 3D projection of F_1 (uni-gram entropy) vs T_u (uni-gram text size) vs N_u (Number of different uni-grams) showing that that random data falls in a plane which has the highest uni-gram entropy for a given N_u and T_u. Figure 2 shows a good separation between the random data and the Pictish Symbols and the Cup and Rings data confirming that the petroglyphs are not random in nature

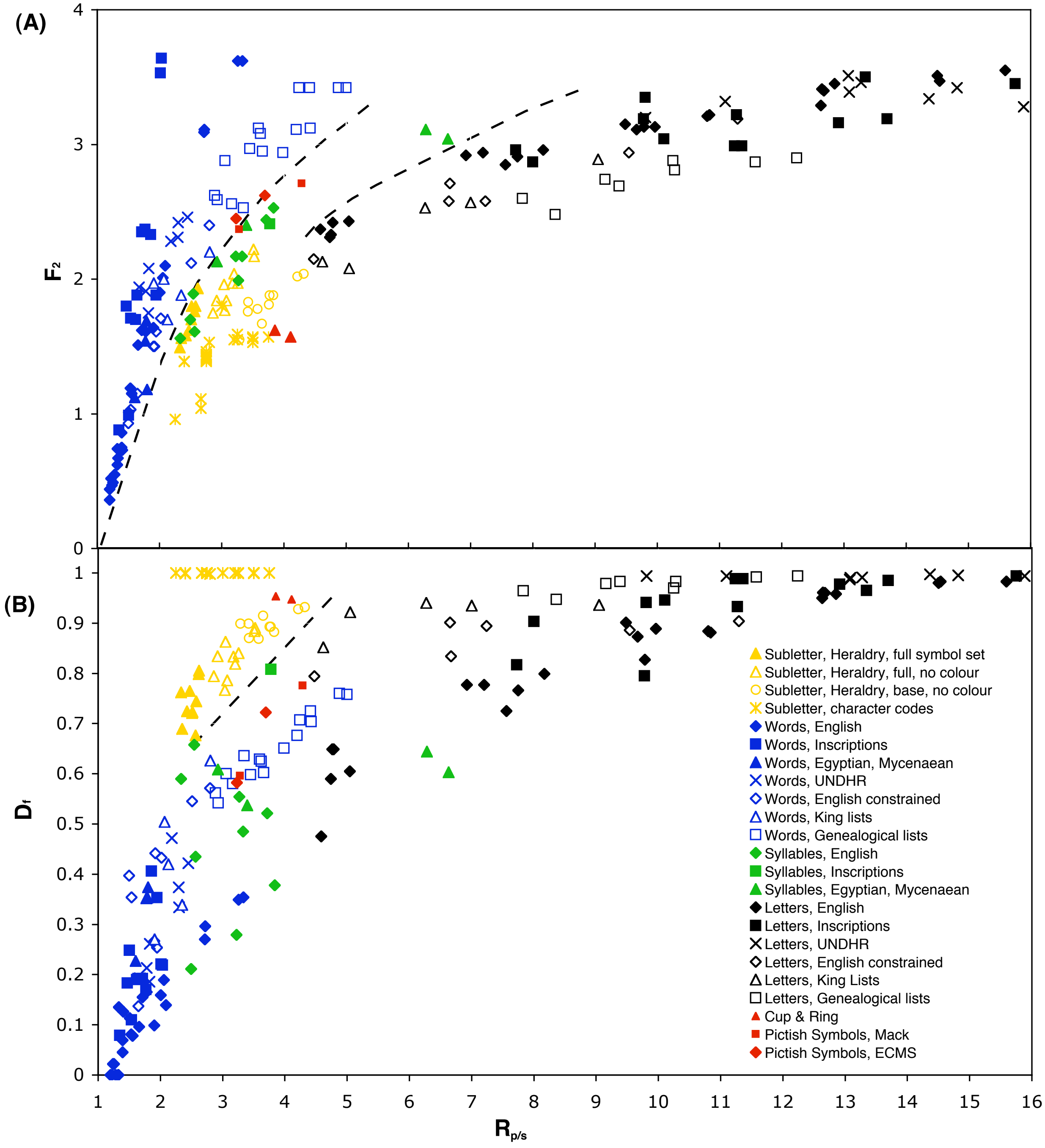


Fig. 3. Comparative analysis of the di-gram character types of written communication systems. Symbols: yellow – sub-letter characters, blue – word characters, green – syllables, black – letters, red – petroglyphs Triangles - Cup and Ring, squares and diamonds - Pictish symbols). (A), dependence of di-gram entropy (F_2) with the undersampling ratio ($R_{p/s}$) for the different character types showing separation (indicated by dashed lines) of the word, syllable and letter character types at similar levels of undersampling ratio. (B), undersampling fraction (D_f) vs. undersampling ratio ($R_{p/s}$) showing the separation (indicated by the dashed line) of the subletter character types from standard written characters (words, syllables, letters). Both (A) and (B) show that the Pictish Symbols (red squares {Mack} and diamonds {ECMS}) are a form of writing.

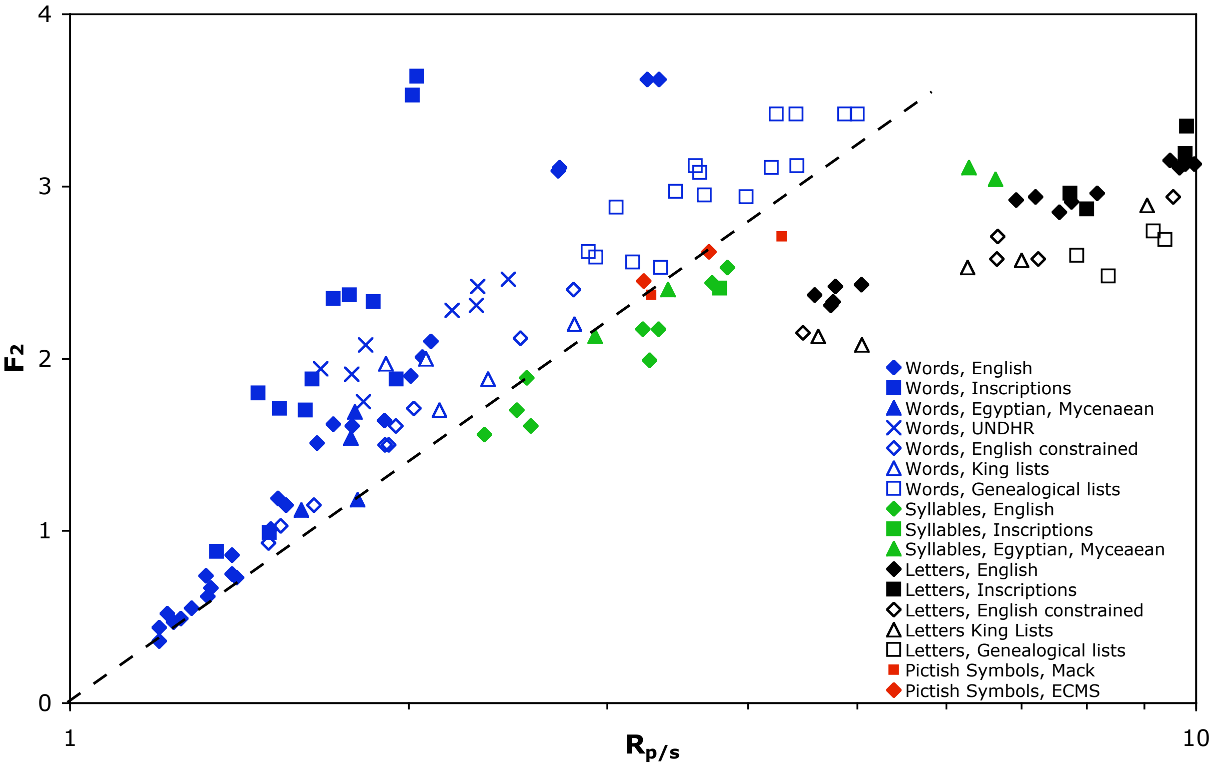
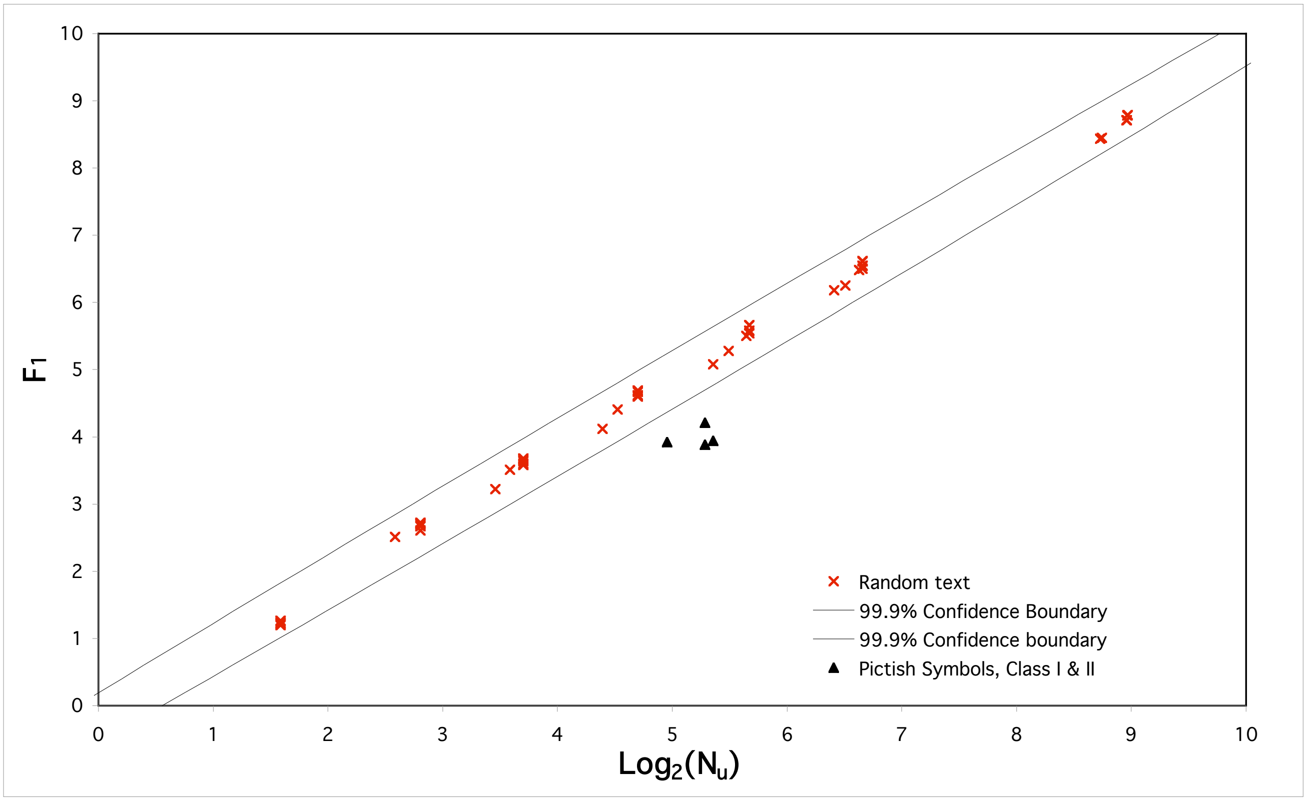


Fig. 4. Comparative analysis of the di-gram character types of standard written communication systems. Symbols: blue – word characters, green – syllables, black – letters, red – Pictish Symbols. Separation of the word and syllable character types indicated by the dashed line. Figure 4 shows that using the larger symbol set of Allen and Anderson (15) implies that the symbols may be very constrained words.

	LETTERS	SYLLABLES	WORDS	OTHER	Nu (words)	Tu (words)	Nu (other)	Tu (other)
ENGLISH	Y	Y	Y		30-2400	35-10000		
UNDHR	Y		Y		500-750	1275-2000		
INSCRIPTIONS	Y	Y	Y		50-780	75-2100		
EGYPTIAN		Y	Y		100-135	230-240		
MYCENAEAN		Y	Y		200-240	440-530		
KING LISTS	Y		Y		20-75	50-130		
GENEALOGICAL LISTS	Y		Y		45-150	200-1100		
HERALDRY				Y			25-165	350-1000
SUB-LETTER CODES				Y			2-3	700-3600
RANDOM				Y			2-100	25-10000
CUP & RING				Y			75	3600
PICTISH SYMBOLS				Y			31-43	140-420

Table 1 gives a summary of the types of different characters used in the data analysis for the different text types as well as the ranges of text size (words or other characters) and number of different N-gram word or other characters used in the comparative analysis.

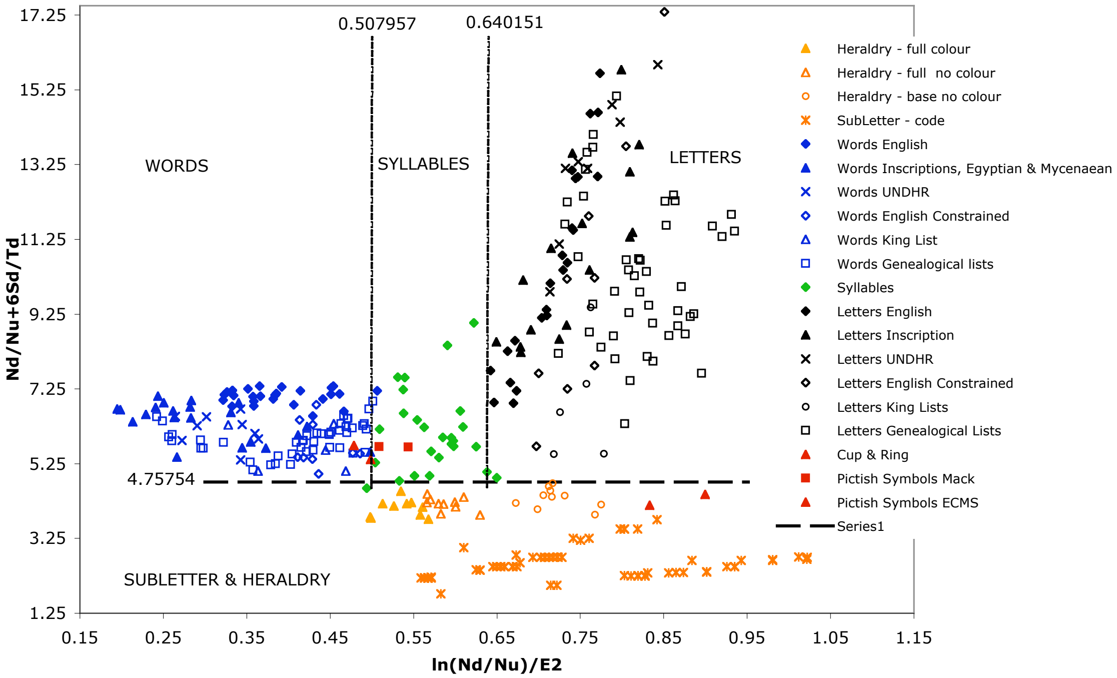


Plot of F_1 (uni-gram entropy) vs Log_2N_u (Number of different uni-grams) showing the 99.9% confidence limits for random data. The figure shows that there is a <0.1% chance that the Pictish Symbols are random in nature.

In a set of N_u characters, the first order entropy (F_1) is given by:

$$F_1 = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Where p_i is the probability of occurrence calculated from the data set. In a truly random set all uni-grams appear with the same frequency, so $p_i = {}^1/_{Nu}$, thus $F_1 = \log_2 N_u$. Small sets of random characters will deviate from this because $p_i \sim {}^1/_{Nu}$. Texts based on written communication have an uneven distribution of characters, generally resulting in a lower F_1 for any value of N_u when compared to random sets and thus allowing written systems to be separated from random systems.



Two parameter model that separates standard character types and subletter-heraldry characters. Boundary conditions for the character classes are at the 99.2% confidence level.