Commentary and Discussion

A Response to Richard Sproat on Random Systems, Writing, and Entropy

Rob Lee*
School of Biosciences

Philip Jonathan**
University of Lancaster

Pauline Ziman†
PHS Consulting, Ltd.

In his article "Ancient symbols and computational linguistics" (Sproat 2010), Professor Sproat raised two concerns over a method that we have proposed for analyzing small data sets of symbols using entropy (Lee, Jonathan, and Ziman 2010): first, that the method is unable to detect random but non-equiprobable systems; and second, that it misclassifies *kudurru* texts. We address these concerns in the following response.

1. Random Systems

Random systems can contain unigrams drawn from an equiprobable or from a non-equiprobable distribution. For small data sets, random but equiprobable systems are likely to have a non-equiprobable actual frequency of unigram occurrence due to the sample size. A method for determining whether a data set is unlikely to be random but equiprobable was given in Lee, Jonathan, and Ziman (2010).

For a given script set, first order entropy (E_1) summarizes the frequencies at which unigrams occur. E_1 is maximized when all unigrams occur with equal probability. In written language, unigrams occur with unequal probabilities—for example, the letters *e* and *t* occur more frequently in English than the letters *x* and *z*, thereby lending some degree of predictability to the occurrence of a particular unigram, and reducing the value of E_1. Random script sets drawn from a non-equiprobable distribution could have the same actual frequencies of unigram occurrence as a written language script set. However, whereas there is unigram-to-unigram dependence in a language, there is no such dependency in a random system. For example, *q* tends to be followed by *u* in English. The digram *qu* would therefore occur more often than other digrams starting with *q*. This second-order dependency is captured in the second-order entropy, E_2. Thus it is one of the fundamental outcomes of Shannon's theory that the dependency in

* School of Biosciences, Geoffrey Pope Building, University of Exeter, Stocker Road, Exeter EX4 4QD, UK. E-mail: R.Lee@exeter.ac.uk.
** Department of Mathematics and Statistics, University of Lancaster, Lancaster LA1 4YF, UK. E-mail: p.jonathan@lancaster.ac.uk.
† PHS Consulting Limited, Pryors Hayes Farm, Willington Road, Oscroft, Tarvin, Chester CH3 8NL, UK.

language script sets reduces E_2 compared to random script sets with the same actual frequencies of unigram occurrences but no unigram-to-unigram dependency (Shannon 1948; Yaglom and Yaglom 1983). This provides a basis for investigating whether a script set is unlikely to be an example of a random system drawn from a non-equiprobable distribution of unigram occurrence.

2. Digram Dependence

The significance of dependence with digrams for a script set can be quantified by comparing its value of E_2 with the distribution of values $E_2(R)$ obtained from randomized permutations of the unigrams in the script set. A script set with significant dependence would yield a value of E_2 which was extreme in the distribution of $E_2(R)$. This is indeed the case for a large proportion of the scripts sets analyzed here (see Figure 1). To construct Figure 1, E_2 was calculated for both the original script set and 1,000 different randomizations, the latter giving rise to an empirical distribution of $E_2(R)$ for randomly generated script sets of the same size and structure as the original. For each script

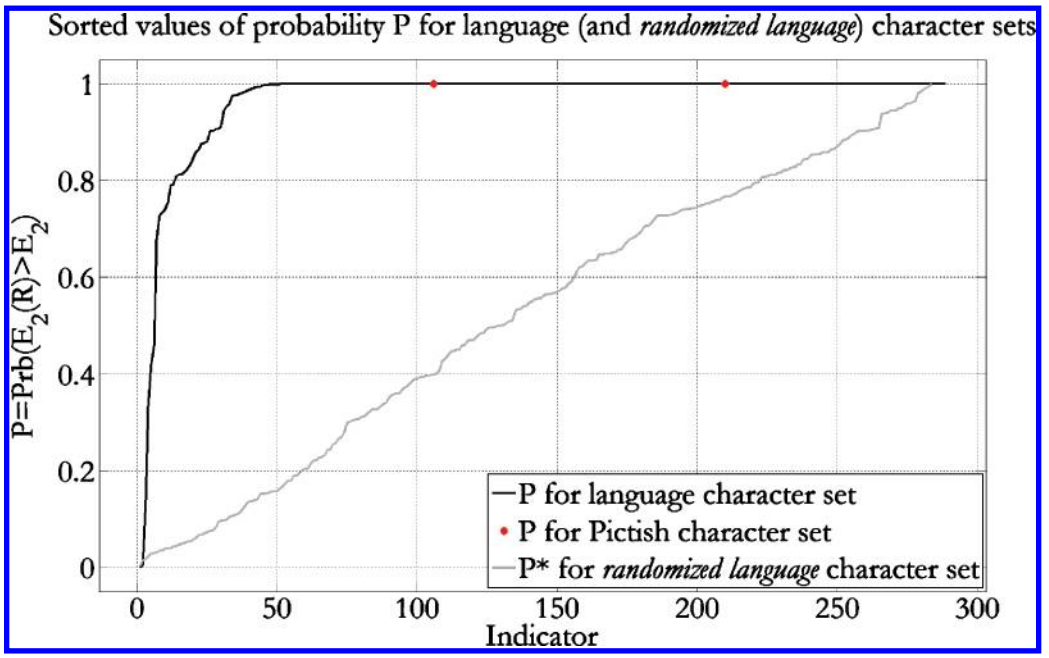

Figure 1
Characterizing dependence within digrams. For a given script set, the value of second-order entropy, E_2, is calculated and compared with the corresponding value, $E_2(R)$, for a **randomized script** R consisting of a randomized permutation of the unigrams comprising the original script set. The probability $P = Prob(E_2(R) > E_2)$ is estimated empirically using 1,000 randomized permutations. The sorted values of probability P are shown in black for the 286 script sets (of small unigram sample size) examined. For approximately 80% of the script sets, the value of probability P is unity, that is, E_2 is the smallest of the corresponding values $E_2(R)$ observed (for script sets of larger unigram sample size, this percentage would be expected to increase towards unity). For both Pictish script sets, the estimated value of probability P is unity (note that their positioning within the subset of script sets where P = unity is arbitary since all these sets have the same value of P). For comparison, we also calculate the probability P* corresponding to scripts sets which are themselves randomized permutations of the original script set. The sorted values of P* (in gray) are seen to be approximately uniformly distributed as expected, whereas the values of P are not.

set we estimate the probability that E_2 for the original script set is less than that for the corresponding randomized script sets. In most cases these estimated probabilities are unity. Figure 1 shows the probabilities as an ordered sequence of script sets. The lower line on Figure 1 depicts the corresponding ordered probabilities for script sets which are themselves randomized permutations of language script sets. As expected, the probability associated with randomized script sets approximately follows a uniform distribution. For a genuinely random but non-equiprobable script processed in the same manner, it is highly unlikely that this script set would yield a value of E_2 which was extreme in the distribution of $E_2(R)$. Figure 1 also shows the probabilities of the Pictish symbol script sets. The values of E_2 for the two Pictish script sets are seen to be extreme with respect to the corresponding $E_2(R)$ distributions. We conclude that the Pictish script sets show dependence within digrams in the same way as the other script sets analyzed which are known examples of dependent digram communication (given here as "language character set").

3. Kudurru

With regard to the question raised by the *kudurrus*, the issue appears to be a difference in viewpoint over terminology as to the definition of what constitutes "writing." Professor Sproat uses a stricter definition of writing than some other researchers, such as Powell (2009, page 13), who defines writing more broadly: "writing is a system of markings with a conventional reference that communicates information." For example,

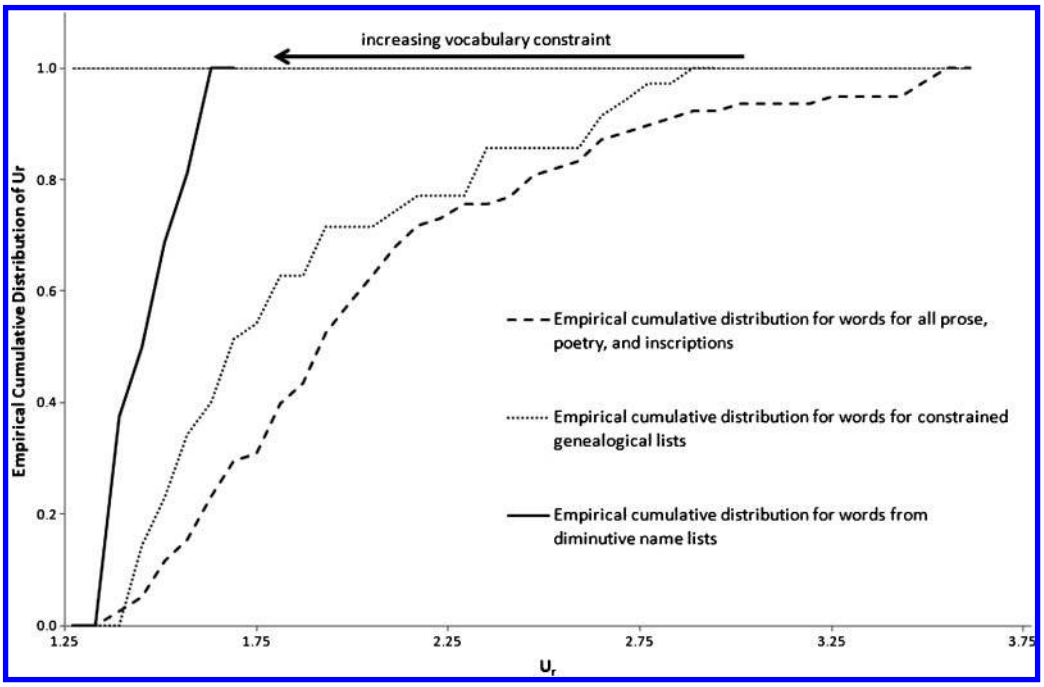

Figure 2
The effect on the empirical cumulative distributions of U_r of increasing the character constraint for words. As the vocabulary becomes constrained, the distribution of U_r becomes narrower and the mean value decreases.

genealogical name lists, which for individual inscriptions or persons may be very short (two or three names), would not be considered a full linguistic system and hence not meet Professor Sproat's criteria for writing. However, we have included these types of communication in the model along with less constrained linguistic systems, many of which would be classed as writing by Professor Sproat. Although the model does not differentiate between these different levels of linguistic systems, their effects upon E_2 can be observed using the structural variable U_r (Lee, Jonathan, and Ziman 2010). Figure 2 illustrates the effect on U_r of increasing constraint on the vocabulary and syntax in moving from prose and poetry to genealogical name lists (including king lists) to very constrained name lists utilizing only "diminutive" name stems (for the "diminutive" name stems data set, the names contained in the genealogical lists have been reduced to their "familiar" form, such as *Al* for *Albert*, *Alan*, and *Alfred*). This constraint further constrains the vocabulary by removing a multitude of names and replacing them with a much smaller and less diverse set. As stated in the paper (Lee, Jonathan, and Ziman 2010), one of the corpora of Pictish symbol types gives values of the structure variables (U_r and C_r) defined in the original paper that are consistent with digram communication encoding at the constrained vocabulary level such as name lists. The values that Professor Sproat calculates for the *kudurrus* data set places them in a similar level of communication.

References

Lee, Rob, Philip Jonathan, and Pauline Ziman. 2010. Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A*, 466:2545–2560.

Powell, Barry B. 2009. *Writing: Theory and History of the Technology of Civilization*. Wiley-Blackwell, London.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October. Reprinted in N. J. A. Sloane and A. D. Wyner (eds). [1993]. *Claude E. Shannon: Collected Papers*. IEEE Press, Piscataway, NJ.

Sproat, Richard. 2010. Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics*, 36(3):585–594.

Yaglom, Akiva M. and Isaak M. Yaglom. 1983. *Probability and Information* [Translated by V. K. Jain]. D. Reidal Publishing Co., Dordrecht.