Quantitative Methods for Classifying Writing Systems

Gerald Penn
University of Toronto
gpenn@cs.toronto.edu

Despite linguists' necessary reliance upon writing to present and preserve linguistic data, writing systems have remained a largely neglected corner of linguistics. While the linear and normative aspects of Gelb's [1963] teleology have been unceremoniously rejected by more recent work on the classification of writing systems, an acknowledgement that more than one dimension may be necessary to characterize the world's writing systems has not come easily. The ongoing polemic between Sampson [1985] and DeFrancis [1989], for example, while addressing some very important issues in the study of writing systems,[1] has been confined exclusively to a debate over which of several arboreal classifications of writing is more adequate.

Sproat's [2000] classification, to our knowledge, was the first multi-dimensional one. While acknowledging that other dimensions may exist, Sproat [2000] arranges writing systems along the two principal dimensions of *Type of Phonography* and *Amount of Logography*. This is the departure point for our present study.

The goals of our research programme can then be stated as follows: **(1)** Assuming Sproat's classification grid as a correct characterization of the structure of writing systems, are there quantitative methods that would allow us to posit a given writing system within this 2-dimensional space? This involves "measuring," at least in a relative sense with respect to other writing systems, its type of phonography and amount of logography. **(2)** Given such quantitative methods, can we distributionally classify the world's writing systems on this grid in such a way that corroborates or casts doubt upon the adequacy of Sproat's grid, or his placement therein of several specific writing systems [Sproat, 2000, p. 142]? **(3)** Given such quantitative methods and the putative correctness of Sproat's grid, can we distributionally classify the world's writing systems in a way that informs us of other underlying principles or latent variables upon which a writing system's position in several dimensions might possibly depend? If no such dependence exists, then these other variables would presumably imply additional dimensions for this grid. If such a dependence did exist, it could in principle reduce the dimensionality of the grid, and thus the classification's complexity.

The holy grail in this area, of course, would be a quantitative procedure that could classify entirely unknown writing systems to assist in attempts at archaeological decipherment, but more realistic applications do exist, particularly in the realm of managing on-line document collections in heterogeneous scripts or writing systems.

No previous work exactly addresses this topic. None of the numerous descriptive accounts that catalogue the world's writing systems, culminating in Daniels and Bright's [1996] outstanding reference on the subject, count as quantitative. The one computational approach that at least claims to consider archaeological decipherment Knight and Yamada [1999], curiously enough, assumes an alphabetic and purely phonographic mapping of graphemes at the outset, and applies an EM-style algorithm to what

[1] These include what, if anything, separates true writing systems from other more limited written forms of communication, and the psychological reality of our classifications in the minds of native readers.

is probably better described as an interesting variation on learning the "letter-to-sound" mappings that one normally finds in text analysis for text-to-speech synthesizers. The cryptographic work in the great wars of the early 20th century applied statistical reasoning to military communications, although this too is very different in character from deciphering a naturally developed writing system.

Amount of logography

Of the two dimensions defined in Sproat's grid, amount of logography is the more difficult — even to define, let alone measure. Roughly, logography is the capacity of a writing system to associate the symbols of a script directly with the meanings of specific words rather than indirectly through their pronunciations. No one to our knowledge has proposed any justification for whether logography should be viewed continuously or discretely. Sproat [2000] believes that it is continuous, but acknowledges that this belief is more impressionistic than factual. In addition, it appears, according to Sproat's [2000] discussion that amount or degree of logography, whatever it is, says something about the relative frequency with which graphemic tokens are used semantically, rather than about the properties of individual graphemes in isolation. English, for example, has a very low degree of logography, but it does have logographic graphemes and graphemes that can be used in a logographic aspect. These include numerals (with or without phonographic complements as in "3^{rd}," which distinguishes "3" as "three" from "3" as "third"), dollar signs, and arguably some common abbreviations as "etc." By contrast, type of phonography predicts a property that holds of every individual grapheme — with few exceptions (such as symbols for word-initial vowels in CV syllabaries), graphemes in the same writing system are marching to the same drum in their phonographic dimension.

Amount of logography is also difficult to measure because it is not entirely independent of phonographic type. As the size of the phonological units encoded by graphemes increases, a threshold is crossed at some point, after which the unit is about the size of a word or other meaning-bearing unit, such as a bound morpheme. When this happens, the distinction between phonographic and logographic uses of such graphemes becomes a far more intensional one than in alphabetic writing systems such as English, where the boundary is quite clear. Egyptian hieroglyphics are well known for their use of *rebus signs*, for example, in which highly pictographic graphemes are used not for the concepts denoted by the pictures, but for concepts with words pronounced like the word for the depicted concept. There are very few writing systems indeed where the size of the phonological unit is word-sized and yet the writing system is still mostly phonographic;[2] it could be argued that the distinction simply does not exist.

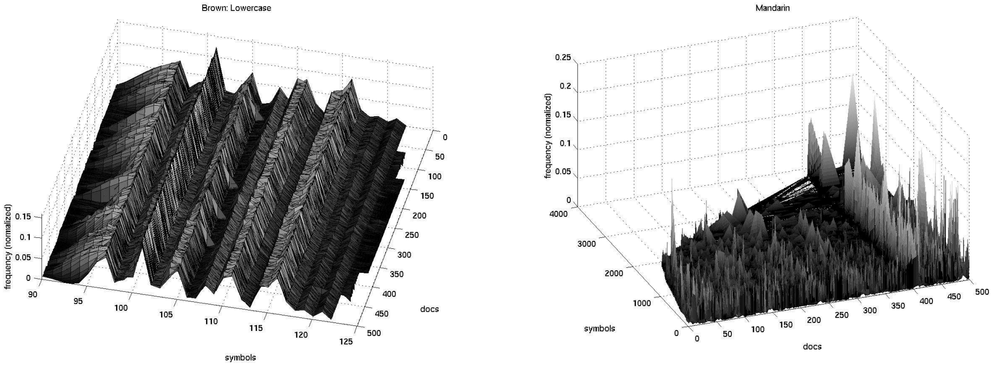
Nevertheless, one can distinguish *pervasive* semantical use from *pervasive* phonographic use. Compare the two figures in 1(a) and (b). These are length-normalized with-document grapheme counts from the English Brown corpus and a GB5-encoded online Mandarin Chinese newspaper, respectively. By counting within documents, graphemes that are pervasively used in their semantical respect will "clump" semantically just as words are known to do when counted across large document collections

[2] Modern Yi is one such example, although the history of Modern Yi is more akin to that of a planned language than a naturally evolved semiotic system.

(newspaper articles about sports use many concepts — and thus logograms — that pertain to sport, for example). This impression of clumpiness conveyed by these figures can be very easily quantified by using *sample correlation coefficients*. Given two random variables, X and Y, their correlation is given by their covariance, normalized by their sample standard deviations:

$$
\begin{aligned}
&corr(X,Y) = \frac{cov(X,Y)}{s(X) \cdot s(Y)} \\
&cov(X,Y) = \frac{1}{n-1} \Sigma_{0 \leq i,j \leq n} (x_i - \mu_i)(y_j - \mu_j) \\
&s(X) = \sqrt{\frac{1}{n-1} \Sigma_{0 \leq i \leq n} (x_i - \mu)^2}
\end{aligned}
$$

For our purposes, each grapheme type is treated as a variable, and each document represents an observation. Each cell of the matrix of correlation coefficients then tells us the strength of the correlation between two grapheme types.



(a)

(b)

Figure 1: Normalized character counts by document for (a) English and (b) Mandarin Chinese.

References

P. Daniels and W. Bright. *The World's Writing Systems*. Oxford University Press, 1996.

J. DeFrancis. *Visible Speech: The Diverse Oneness of Writing Systems*. University of Hawaii Press, 1989.

I. Gelb. *A Study of Writing*. University of Chicago Press, 2nd edition, 1963.

K. Knight and K. Yamada. A computational approach to deciphering unknown scripts. In *Proc. of ACL Workshop on Unsupervised Learning in NLP*, 1999.

G. Sampson. *Writing Systems*. Stanford University Press, 1985.

R. Sproat. *A Computational Theory of Writing Systems*. Cambridge University Press, 2000.