Quantitative Methods for Classifying Writing Systems

Gerald Penn
University of Toronto
10 King's College Rd.
Toronto M5S 3G4, Canada
gpenn@cs.toronto.edu

Travis Choma
Cognitive Science Center Amsterdam
Sarphatistraat 104
1018 GV Amsterdam, Netherlands
travischoma@gmail.com

Abstract

We describe work in progress on using quantitative methods to classify writing systems according to Sproat's (2000) classification grid using unannotated data. We specifically propose two quantitative tests for determining the type of phonography in a writing system, and its degree of logography, respectively.

1 Background

If you understood all of the world's languages, you would still not be able to read many of the texts that you find on the world wide web, because they are written in non-Roman scripts that have been arbitrarily encoded for electronic transmission in the absence of an accepted standard. This very modern nuisance reflects a dilemma as ancient as writing itself: the association between a language as it is spoken and the language as it is written has a sort of internal logic to it that we can comprehend, but the conventions are different in every individual case — even among languages that use the same script, or between scripts used by the same language. This conventional association between language and script, called a *writing system*, is indeed reminiscent of the Saussurean conception of language itself, a conventional association of meaning and sound, upon which modern linguistic theory is based.

Despite linguists' necessary reliance upon writing to present and preserve linguistic data, however, writing systems were a largely neglected corner of linguistics until the 1960s, when Gelb (1963) presented the first classification of writing systems. Now known as the *Gelb teleology*, this classification viewed the variation we see among writing systems, particularly in the size of linguistic "chunks" represented by an individual character or unit of writing (for simplicity, referred to here as a *grapheme*), along a linear, evolutionary progression, beginning with the pictographic forerunners of writing, proceeding through "primitive" writing systems such as Chinese and Egyptian hieroglyphics, and culminating in alphabetic Greek and Latin.

While the linear and evolutionary aspects of Gelb's teleology have been rejected by more recent work on the classification of writing systems, the admission 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 (2000)'s classification was the first multidimensional 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*, both of which will be elaborated upon below. This is the departure point for our present study.

Our goal is to identify quantitative methods that

[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.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 117–120,
New York, June 2006.

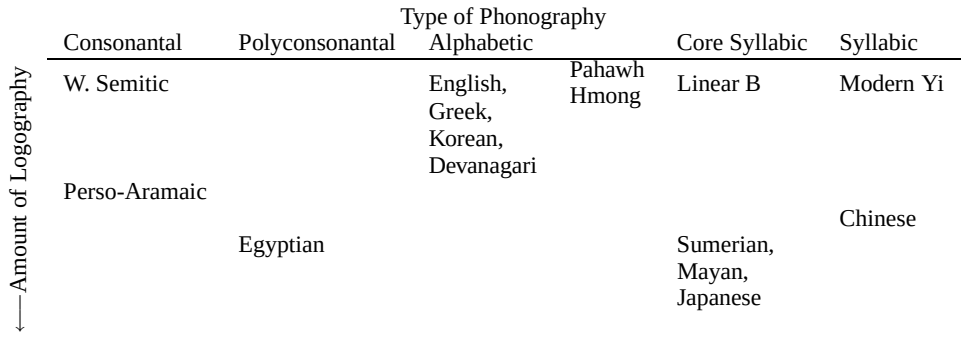

Figure 1: Sproat's writing system classification grid (Sproat, 2000, p. 142).

can assist in the classification of writing systems. On the one hand, these methods would serve to verify or refute proposals such as Sproat's (2000, p. 142) placement of several specific writing systems within his grid (Figure 1) and to properly place additional writing systems, but they could also be used, at least corroboratively, to argue for the existence of more appropriate or additional dimensions in such grids, through the demonstration of a pattern being consistently observed or violated by observed writing systems. The holy grail in this area would be a tool 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 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.

2 Type of Phonography

Type of phonography, as it is expressed in Sproat's grid, is not a continuous dimension but a discrete choice by graphemes among several different phonographic encodings. These characterize not only the size of the phonological "chunks" encoded by a single grapheme (progressing left-to-right in Figure 1 roughly from small to large), but also whether vowels are explicitly encoded (poly/consonantal vs. the rest), and, in the case of vocalic syllabaries, whether codas as well as onsets are encoded (core syllabic vs. syllabic). While we cannot yet discriminate between all of these phonographic aspects (arguably, they are different dimensions in that a writing system may select a value from each one independently), size itself can be reliably estimated from the number of graphemes in the underlying script, or from this number in combination with the tails of grapheme distributions in representative documents. Figure 2, for example, graphs the frequencies of the grapheme types witnessed among the first 500 grapheme tokens of one document sampled from an on-line newspaper website in each of 8 different writing systems plus an Egyptian hieroglyphic document from an on-line repository. From left to right, we see the alphabetic and consonantal (small chunks) scripts, followed by the polyconsonantal Egyptian hieroglyphics, followed by core syllabic Japanese, and then syllabic Chinese. Korean was classified near Japanese because its Unicode representation atomically encodes the multi-segment syllabic complexes that characterize most Hangul writing. A segmental encoding would appear closer to English.

3 Amount of Logography

Amount of logography is rather more difficult. 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.

Another reason that amount of logography is difficult to measure is that it is not entirely independent of the type of phonography. As the size of the phonological units encoded by graphemes increases, at some point a threshold is crossed wherein the unit is about the size of a word or another 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 (see Section 4).

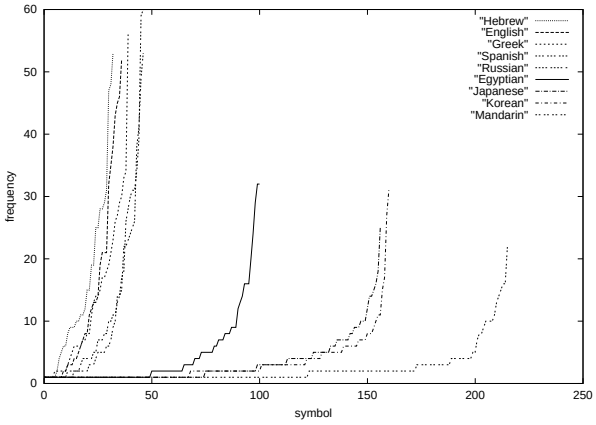

Figure 2: Grapheme distributions in 9 writing systems. The symbols are ordered by inverse frequency to separate the heads of the distributions better. The left-to-right order of the heads is as shown in the key.

Nevertheless, one can distinguish *pervasive* semantical use from *pervasive* phonographic use. We do not have access to electronically encoded Modern Yi text, so to demonstrate the principle, we will use English text re-encoded so that each "grapheme" in the new encoding represents three consecutive graphemes (breaking at word boundaries) in the underlying natural text. We call this *trigraph English*, and it has no (intensional) logography. The principle is that, if graphemes are pervasively used in their semantical respect, then they will "clump" semantically just like words do. To measure this clumping, we use *sample correlation coefficients*. Given two random variables, X and Y, their correlation is given by their covariance, normalized by their sample standard deviations:

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

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. For trigraph English, part of the correlation matrix is shown in Figure 3. Part of the correlation matrix for Mandarin

[2]Modern Yi (Figure 1) 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.

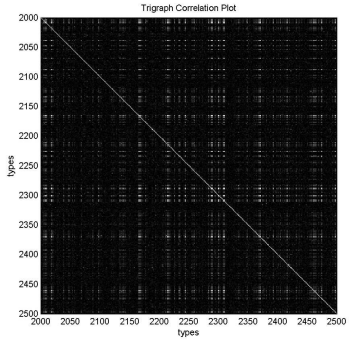

Figure 3: Part of the trigraph-English correlation matrix.

Chinese, which has a very high degree of logography, is shown in Figure 4. For both of the plots in

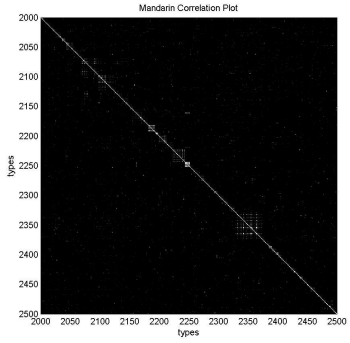

Figure 4: Part of the Mandarin Chinese correlation matrix.

our example, counts for 2500 grapheme types were obtained from 1.63 million tokens of text (for English, trigraphed Brown corpus text, for Chinese, GB5-encoded text from an on-line newspaper).

By adding the absolute values of the correlations over these matrices (normalized for number of graphemes), we obtain a measure of the extent of the correlation. Pervasive semantic clumping, which would be indicative of a high degree of logography, corresponds to a small extent of correlation — in other words the correlation is pinpointed at semantically related logograms, rather than smeared over semantically orthogonal phonograms. In our example, these sums were repeated for several 2500-type samples from among the approximately 35,000 types in the trigraph English data, and the approximately 4,500 types in the Mandarin data. The average sum for trigraph English was 302,750 whereas for Mandarin Chinese it was 98,700. Visually, this difference is apparent in that the trigraph English matrix is "brighter" than the Mandarin one. From this we should conclude that Mandarin Chinese has a higher degree of logography than trigraph English.

4 Conclusion

We have proposed methods for independently measuring the type of phonography and degree of logography from unannotated data as a means of classifying writing systems. There is more to understanding how a writing system works than these two dimensions. Crucially, the direction in which texts should be read, the so-called *macroscopic organization* of typical documents, is just as important as determining the functional characteristics of individual graphemes.

Our experiments with quantitative methods for classification, furthermore, have led us to a new understanding of the differences between Sproat's classification grid and earlier linear attempts. While we do not accept Gelb's teleological interpretation, we conjecture that there is a linear variation in how individual writing systems behave, even if they can be classified according to multiple dimensions. Modern Yi stands as a single, but questionable, counterexample to this observation, and for it to be visible in Sproat's grid (with writing systems arranged along only the diagonal), one would need an objective and verifiable means of discriminating between consonantal and vocalic scripts. This remains a topic for future consideration.

References

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

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

I. Gelb. 1963. *A Study of Writing*. Chicago, 2nd ed.

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

G. Sampson. 1985. *Writing Systems*. Stanford.

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