Computers and the Study of Lost Languages

Michael P. Oakes

University of Wolverhampton, England

Abstract

This paper presents a review of the computational and statistical work of previous authors in the structural analysis of three undeciphered languages, namely the Indus signs from India and Pakistan, the Phaistos disk from Crete, and the Rongorongo texts from Rapanui (Easter Island). In new work, we use Baroni and Evert's zipfR package, which implements a class of models called Large Number of Rare Events (LNRE) models. Starting with the frequency spectrum of the existing corpus of a language (the sample), the LNRE models are used to estimate the number of characters there might be in the language as a whole (the population). We estimate the total number of Indus signs (both discovered and undiscovered) to be about 1396, the total number of symbols in the language of the Phaistos disk to be about 56, and the number of characters in Rongorongo to be in the range 800 to 1000.

Key words : Indus signs, Phaistos disk, Rongorongo, Vocabulary size, LNRE models.

1. Introduction

According to Robinson (2009:43), "Computers have made little impact on archaeological decipherment". The best-known decipherment, that by Michael Ventris on Linear B (Chadwick, 1958), came before the widespread availability of computers. Decipherment requires "a synthesis of logic and intuition … wide linguistic, archaeological, historical and cultural knowledge that computers do not (and presumably cannot) possess". However, computers offer a number of advantages for the study of texts in unknown languages. They are fast and accurate, which greatly eases routine tasks like collating and counting. Computers can facilitate the tasks of indexing scripts, studying the relative frequencies of signs and sign groups, and grouping together texts which deal with similar subjects (Packard, 1974:7). Computer concordancers can display each sign of interest in the corpus of writings in a language in all the contexts in which it occurs. Computers have the advantage of impartiality, and thus help in what is called structural analysis – describing the patterns in a text without any preconceptions about the meaning or the possible linguistic nature of the features of that text. Harris and Melka (2011a) describe a mixed methods approach, where qualitative observations are made about seeming patterns of signs in the texts, and then quantitative methods statistically measure the phenomenon to discover whether those patterns are regular occurrences. The greatest bottleneck in the computer processing of unknown languages is that we need to be able to accurately transcribe the signs of that language into a definitive set of numeric codes, which is often very difficult to do. This has been true of the Indus signs and even more so for the Rongorongo texts discussed later in this paper, due to the presence of such features as allographs (variant forms of the same sign) and ligatures (the combination of two signs into one). Given this restriction, in this paper we will consider progress using computers so far on the analysis of two scripts which have still not been deciphered, Rongorongo from Easter Island and the Indus Valley tablets. In fact, Rongorongo might never be deciphered, as we have only a small corpus of surviving texts in the world, and no real hope of finding any more. There is controversy surrounding the Indus signs in that some people say that they might not even be writing. We also consider the Phaistos Disk found on Crete, written in a script which has only be found with certainty on this single object.

2. Rongorongo

Rongorongo originates from Easter Island, which has the local name of Rapanui. It is unlike any other known script, being made up of a large alphabet of characters (or "glyphs") vaguely resembling plants and animals. The first foreigner to describe this script was Father Joseph Eyraud in 1864. It is unlikely that Rongorongo will ever be deciphered, since there are only a small number of extant texts, consisting of about 14,000 glyphs in total. These are written on about 25 artefacts now scattered throughout the world, including rectangular wooden tablets, a staff and a breastplate. Various systems of transliteration of glyphs into numeric codes have been proposed, but none is perfect and universally-agreed upon. Other difficulties are that there are no illustrations to give clues as to the meaning of the surrounding texts, and there are complex combinations of glyphs which are joined together, called ligatures, making it difficult to decide whether a complex glyph should be read as one glyph or several. The chronology of the texts is unknown, and one theory is that was developed after the coming of the Spanish in imitation of European writing. It may represent the Old Rapanui language, but this is substantially different from the post-missionary Rapanui which has loanwords from Tahitian, French, Spanish and English (Harris and Melka, 2011a). The one part of the Rongorongo corpus which can be deciphered with confidence is a section on a tablet called "Mamari" in which the configuration of crescent moon shaped glyphs corresponds to the Rapanui lunar calendar.

To transcribe the Rongorongo texts, Barthel (1958) devised a three digit numeric code for each glyph or group of glyphs found in the corpus. Each basic glyph type has its own code, and variants are denoted by alphabetic suffixes. The first digit is a numeral from 0 to 7, where for example 0 and 1 denote geometric shapes and inanimate objects; 2 denotes figures with "ears", and 6 is used for figures with beaks. A similar idea is used for the next two digits, so for example 206, 306, 406, 506 and 606 all have different heads but a downward pointing wing or arm on the left and a raised four-fingered hand on the right. Barthel estimated that there was an "alphabet" of 120 basic glyphs, with 480 others being allographs or ligatures. The full set of Barthel's symbols may be viewed at http://kohaumotu.org/rongorongo_org/signs/1.html . Richard Sproat (2003) uses the "reduced" Barthel coding system where all the various diacritics are removed. Thus a string like 600a-600.711-20cfy.246-50.711-606-1t.6 is simplified to 600-600-711-20-246-50-711-606-1-6. This simplified encoding was also used for the experiment on the size of the Rongorongo character set described in this paper.

The first task in a statistical analysis of an unknown natural language text is to try and determine whether the language is phonemic, syllabic or logographic. The simplest way of doing this is just to consider the number of letters in the alphabet – about 60 would suggest syllabic, so estimates of about 120 characters for Rongorongo would suggest that it is mainly syllabic with some logographs. In the 1950s, Federova compared the frequency distribution of the glyphs with the frequencies of words, syllables, and phonemes, in the old Rapanui language as represented by the text "Apai", but found that there was no glyph in the corpus as frequent as the morphemes "te" or "e". This provides evidence against an exact correspondence between the syllables of Old Rapanui and Rongorongo glyphs. Pozdniakov

(1996) found that throughout the artefacts of the Rongorongo corpus, the frequencies of the basic signs are very stable, which provides evidence that it is “real” writing.

It was first noticed by three Russian schoolboys that some sections of the Rongorongo corpus are repeated, the longest of these sequences being repeated three times. This is clearly shown in Richard Sproat’s (2003) “dotplot” analysis, where the x and y axes correspond to the distance in glyphs from the start of the corpus, and a dark dot is placed on the graph at the point (x, y) whenever a glyph on the x axis matches one on the y axis. A number of things can be learned from the existence of these overlapping sequences. Firstly, one theory had been that the glyphs did not encode sound, but were merely placed in sequence as aide memoires to a storyteller like section headers. Guy (1990) felt that the mere existence of repeated segments within a text showed that their function was not merely mnemonic. Secondly, the starts and ends of repeated sequences might reveal initial and final delimiters in the texts. Also, frequently recurring patterns would likely be meaningful in themselves (Melka 2008). The multiple correspondences show that many other variant glyphs are in fact the same. For example, if one version of a repeated sequence has glyph 204 in the same position as glyph 206, which is similar in appearance, in another version, it would suggest that these two glyphs are the same. This observation would help us to rule out any putative translation in which two variants of the same glyph are assigned radically different meanings (see Pozdniakov and Pozdniakov, 2007:11).

A concordance can show all the occurrences of a glyph in the corpus in the contexts in which it appears. For example, glyph 040 is the crescent moon in the “(“ direction, and 041 is the crescent moon in the “)” direction. A concordance for glyph 041 shows that this glyph can occur as a double, but not as a triple; It can be followed by glyph 040; and it occurs several times in the overlapping sequences 390-041-378-041-670 and 378-041-670-008-078. The two glyphs 040 and 041, although most frequent in the “lunar calendar” part of tablet C, also occur frequently at the start of tablet E. In the first 162 characters of tablet E, there are 14 occurrences of 040 and 6 of 041, suggesting this opening might also have some sort of calendar function (Oakes, 2014). Harris (2010) clustered the individual texts in the Rongorongo corpus by “genre”. He started with a matrix with glyphs for rows and artefacts for columns, and recorded in each (row, column) entry the number of times that glyph was found in that artefact. This matrix was input to a Factor Analysis, where the output was a map showing both which texts were deemed close to each other, and which were less related, and the glyphs which predominated in each group of texts.

3. The Indus Signs

The Indus Valley civilisation at its peak from 2500 to 1900 BC covered a wide area of modern Pakistan and Northern India. The first of many artefacts bearing an unknown script was found in the 1850s near Harappan, about 150 miles south of Lahore, and in 1906 there was a major excavation of the area. Altogether about 4000 inscribed objects have been found, about 60% being seals, but also other media such as pottery and copper tablets. The seals are about the size of postage stamps, and many are beautifully engraved with pictures of animals (MacGregor, 2010). Some of the signs which look like writing are geometric shapes, while others resemble stick men, fish, and other stylised animals. The direction of writing is known, since spaces and cramped signs appear on the left. Since these are seals, the impressions are mirror images, so the writing goes from left to right. Although many putative decipherments have been made, none is generally accepted, and we do not even know for sure that the signs definitely are writing.

Decipherment of the Indus signs is difficult, since we still have a relatively small corpus of about 15,868 signs, and a large number (about 676) signs have been found so far (Wells, 2011). The individual inscriptions are very brief, being less than 5 characters long on average. About 40% of the inscriptions are duplicates, which reduces the set of unique texts to work with. The names of places and rulers at the time are not known from any other sources; we do not know what calendar system they used; and we know little about the culture of the Indus civilisation. No "Rosetta Stone"-like bilingual text is available, and there are no word boundaries.

Farmer, Sproat and Witzel (2004) believe that the Indus signs are not writing. Their simplest, and most important, argument for this is that the sequences all share the one striking feature of "extreme brevity". We would expect much longer signs to be found from a literate civilisation. To counter this argument, others have said that longer texts did exist, but were written on non-durable materials which have since rotted away. However, other civilisations where this was the case have left many other markers, such as long texts on pots and cave walls, and the remains of writing implements. No such findings have been made with certainty at the Indus sites. Another anomaly noted by Farmer et al. was that although some signs are repeated many times in the Indus corpus as a whole, signs are rarely repeated in individual inscriptions. Of the 20 most frequent symbols in Mahadevan's concordance, 10 have zero or almost zero repetition rates in individual inscriptions throughout the corpus, and the repetition rates are low for the other 10 as well. This is evidence that the symbols are not used for encoding sound. Farmer et al. compared the within-inscription repetition rates in the Indus texts with those found in the similarly long "cartouches" (oval shapes containing significant names) in Egyptian hieroglyphs. Sign repetition rates were much lower in the Indus texts than in the cartouches. In a sample of 67 cartouches with an average of 6.9 hieroglyphs each, there was a total of 48 signs which were repeated within a single inscription. For the Indus seals, in a sample of 67 inscriptions found at the site of Mohenjo-daro (average length 7.4 signs) there were only 8 repeating signs, and the sample of 67 inscriptions found at Harappan (average 7.36 signs) there were only 7 signs which repeated. Those relatively few inscriptions which do contain repeated signs, do so in ways that do not suggest sound encoding. For example, there are cases where the same sign is repeated up to 4 times in a row. Farmer et al. conclude that the Indus scripts are not writing as such, but more probably a collection of political or religious symbols.

In support of the Indus signs being writing, Rao et al. (2009) used conditional entropy which a way of quantifying the flexibility in the ordering of symbols – the amount of randomness in the choice of a token given the preceding token. If we have two symbols in a text, X and Y, and know symbol X already, conditional entropy H(Y|X) is the amount of additional information that needs to be supplied for us to know that the next symbol will be Y. The maximum value that conditional entropy can take occurs for a random sequence of symbols.

Relative conditional entropy is the absolute conditional entropy for a sequence divided by this theoretical maximum. Rao et al.'s results for a comparison of "linguistic" and "non-linguistic" systems according to their conditional entropy were as follows: Artificial random text with random order 1, DNA 0.98, Protein 0.96, Sanskrit 0.67, English words 0.65, Sumerian (logosyllabic) 0.59, Old Tamil (alpha-syllabic) 0.58, Indus signs 0.57, English characters 0.53, Fortran 0.41, Artifical text with a rigid order (each sign has a unique successor) 0.09. Here the comparison of English words and English characters shows that the unit of text has more influence than the language family. Old Tamil is alphasyllabic, and

Sumerian is logosyllabic, suggesting that the Indus text might follow one of these systems. Farmer et al. (2009) countered that "conditional entropy is not and never before has been claimed to be a statistical measure of whether or not a sign symbol is linguistic or non-linguistic", to which Rao et al. (2010) replied that the similarity in conditional entropy for the Indus script and various linguistic sequences does not prove that the Indus script is linguistic, but it constitutes empirical evidence for the linguistic hypothesis. The conditional entropy method does distinguish non-linguistic sequences of DNA, protein and Fortran computer code from the samples of known languages. Other evidence for the Indus signs being writing is that the inscriptions are linear, and have directionality, distinguishing them from non-linguistic symbols such as mediaeval heraldry and road signs. Unlike the non-linguistic Kudurru signs, which are restricted to boundary stones, the Indus texts have been found on diverse media. As we will see, evidence has been found of syntactic structure, as shown by many signs having positional preferences such as those which start and end the scripts in which they are found (Rao, 2010).

A very small number of signs predominate in the Indus scripts, suggesting a frequently occurring core vocabulary and hundreds of rare symbols, as might be found in a logosyllabic script like Japanese. This finding was quantified by Yadev et al. (2010) by observing the "Zipf's law gradient". They plotted the logarithm of sign frequency (on the y axis) against the logarithm of sign rank, and found that the slope of the plot for the Indus signs (-2.59) was very much greater than that for English (-1.15), showing a greater tendency for the frequency of the rth most frequent symbol to fall away at higher values of the rank, r.

Rao et al. (2009) used the information theoretic measure of perplexity to provide evidence that the Indus script is neither a collection of equiprobable symbols in random order, nor a random sequence of symbols with differing probabilities, but approximates to a higher order Markov Model. A Markov model is one where the probability of each sign in a sequence depends on the nature of the previous n signs. Perplexity depends on the length n of sequences of symbols which are taken into account. The lower the perplexity, the better the model approximates the language under study. The perplexity (PP) of the models generated using the probabilities found in the Indus corpus as a function of the number n of preceding signs taken into account were: n = 0, PP = 68.82; n = 1, PP = 26.69; n = 2, PP = 26.09; n = 3, PP = 25.26; n = 4, PP = 25.26. The large drop in perplexity between n = 0 and n = 1, with little change thereafter, shows that most of the information about sign probabilities can be captured by a Markov model where the probability of each sign depends on the nature of the previous sign. Thus the symbols are not independent of each other, and their ordering must matter. Markov models cannot shed light on the semantics of a text but they can tell us about the syntax as they look at sequences of symbols. Another way of looking at syntax is to consider possible affinities between pairs of characters using a statistical measure based on the contingency table such as chi-squared or log likelihood (LL). Yadev et al. (2010) used log-likelihood to determine the association strength of pairs of Indus signs, and found the strongest association (with LL = 792.4, values > 10.83 being statistically significant) between sign 267 on the left and sign 99 on the right. Text segmentation is unlikely between pairs of signs with high affinity, as they almost certainly belong to the same text unit as each other. They also showed that sign 267 has most affinity with the start position, and sign 341 has most affinity with the end. These results show characteristics of syntax: different signs taking on the roles of start and end symbols, directionality shown by different LL values for transposed sequences, and significant affinities between certain signs**.**

4. The Phaistos Disk

The Phaistos disk is a round clay tablet found in 1909 in the remains of the Minoan palace of Phaistos on Crete. The disk was not written, but stamped with dies, and so may be the earliest printed text known. The disk is written in a spiral, and reading begins at the centre. There are 242 characters (one damaged), divided into 30 (recto) and 32 (verso) groups by vertical lines. Many of the signs are clear depictions of people, birds, mammals and tools, and there are also symbols resembling a shield, a flower head and a boat. Other symbols do not clearly represent well-known items.

Dieter Rumpel (1994) made two uses of the type-token ratio curve, produced by plotting the number of types encountered so far on the y axis against the number of tokens from the start of the text on the x axis. In a comparison of the Phaistos disk with some modern languages, he found that the one which plotted closest to the diagonal line where y = x was Chinese which is ideographic. Just to the left of the plot for Chinese was Japanese, which is mixed syllabic and ideographic. The line for German lay closest to the x axis. The type-token plot for the Phaistos disk lay between German and Japanese, which would be consistent with a syllabic text. Rumpel also visually estimated the asymptote (the number of types which will never be exceeded, however long the text is) of the type token curve. This method gave 55 to 65 symbols for the entire sign inventory, consistent with a syllabary.

5. Large Number of Random Events (LNRE) Models

Word frequency distributions are characterised by Large Numbers of Rare Events (LNRE), an idea originally developed by Khmaladze (1987) and expanded on by Baayen (2001:51-57). Most random events, such as the outcomes of spinning a coin, produce vocabulary growth curves similar to that in Figure 1 (Baayen, 2001:52), which shows the results of a simulation for the spinning of a fair die N times, where N was between 1 and 100. For each value of N, the total number of different faces seen was determined a large number of times (1000), and the average was taken. The black dots show the growth in the average of faces seen E[V(N)] as the number of throws N increases. The curves below made up of white dots show (from left to right) the average number of faces which have been seen exactly once, E[V(1,N)], the average number of faces which have been seen twice, E[V(2,N)], and so on up to E[V(5,N)]. Two characteristics of these curves put the die simulation in contrast to LNRE events. Firstly, the black dots quickly approach (although never quite reach) their maximum value of 6, the total number of faces, which is called the asymptote. Secondly, the smaller curves with white dots reach maximum values then fall back towards 0 at higher values of N. An example of a LNRE event is given by Baroni and Evert (2014:7), that of Italian words with the prefix “ri”. Here N is the number of word tokens with the “ri” prefix seen from the start of the corpus, V(N) is the number of different types of words prefixed with “ri” seen so far, and V_1N is the number of words with the “ri” prefix we have seen exactly once so far. Unlike the corresponding curves for E[V(N)] and E[1(N)] in the die throwing simulation, the V(N) curve does not reach its asymptote even after reading in 1.4 million words of the corpus, and likewise V_1 keeps increasing throughout the experiment. Baroni and Evert’s experiment is shown in Figure 2. Baayen (2001:51) summarises the situation as follows: “The LNRE zone can be described as the the range of sample sizes for which it is clear from the spectral curves that we have only just begun to sample the types available in the population. Even large corpora with tens of thousands of words are located in the LNRE zone”. Most word frequency distributions seen for real natural languages, give a pattern more typical of the LNRE zone. The vocabulary growth curves do not reach their asymptote by the time the whole corpus has been sampled, and it is not clear whether the curves would eventually reach a horizontal asymptote (suggesting that the number of types in the population is finite) or a diagonal asymptote (sloping upwards, and implying that the number of types in the population is infinite).

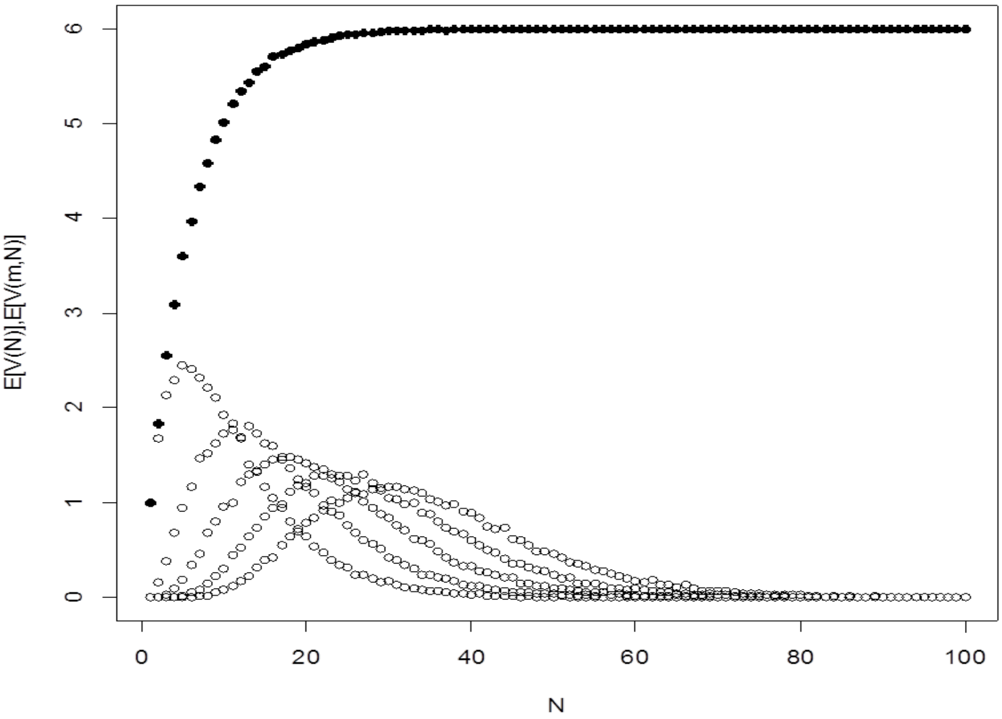

Figure 1. Mean vocabulary growth curves for 100 throws with a fair die.

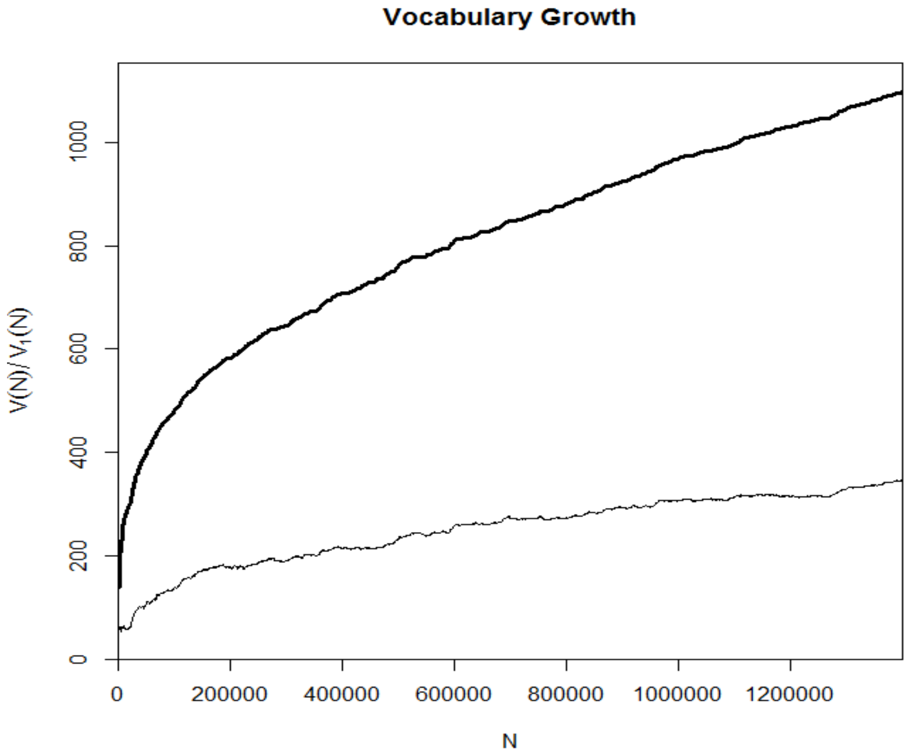

Figure 2. Vocabulary growth curves for words with the "ri" prefix in a corpus of newspaper Italian.

We can try to answer the question of whether the vocabulary is finite or infinite by extrapolating V(N) to larger sample sizes to find its value for the whole population (such as in the Italian language). For this we need parametric statistical models called LNRE models, three of which have been implemented in the zipfR toolkit by Baroni and Evert (see Evert and Baroni, 2007; Evert, 2004), which was used for the experiments described in this paper. These models are the Generalized Inverse Gauss Poisson (GIGP; Baayen, 2001: 135-160), Zipf-Mandelbrot (ZM; Evert, 2004) and the finite Zipf-Mandelbrot (FZM; Evert, 2004). In Section 6 of this paper, we extrapolate the number of characters found in finite corpora (In the cases of the lost language, these are the entire sets of extant texts) and extrapolate these values to estimate how many characters there might be in the entire language, including characters unseen in the extant texts.

6. Estimating the Entire Vocabulary for each Language

The data sets for the experiments designed to estimate the total number of vocabulary items in a language were as follows: The English character frequencies, based on a sample of 40,000 words, with any diacritics removed, were obtained from http://www.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html.

The frequency spectrum for the Indus texts (see Figure 3) was derived from the appendix of 342 Indus signs with their frequencies in David Wells "Epigraphic Approaches to Indus Writing" (2011). The entire text of the Phaistos disk is given by Andrew Robinson in "Lost Languages" (2009: 299-303), with the Evans numbered sign list. The entire Rongorongo corpus can be downloaded from the "Rongorongo" website at http://kohaumotu.org/rongorongo_org/corpus/1.html , and was simplified according to Sproat's recommendations (see section 2). The input data to each of the LNRE models is in the form of a frequency spectrum, such as the one for the Phaistos disk given in Figure 3. Vm is the number of characters in the corpus seen exactly m times, so in the Phaistos disk 10 symbols are seen exactly once, 9 are seen twice, and so on. The data needed for the frequency spectrum can be obtained from the raw corpus of symbol codes by using the R command `table(table(X))` where X is the raw corpus.

m	Vm
1	10
2	9
3	4
4	3
5	3
6	6
7	2
11	4
12	1
15	1
17	1
18	1
19	1

Figure 3. Frequency spectrum for the Phaistos Disk

Corpus	Corpus Statistics	GIGP	FZM	ZM

Italian "ri" (Baroni and Evert, 2014)	N = 1399896
	V = 1098
		P = 3411574
		p = 0.0016
		V(2N) = unavailable
			P = 78194057
			p = 0.045
			V(2N) = unavailable
				P = Infinite
				p = 0.064
				V(2N) = unavailable

English Characters	N = 182303
	V = 26
		P = 26.01
		p = 0
			Error	Error

Indus signs	N = 15868
	V = 676
		P = 1396.45
		p = 0.0088
		V(2N) = 843.23
			Error	P = Infinite
				p = 6.8 e -28
				V(2N) = 829.20

Phaistos disk	N = 242
	V = 46
		P = 55.78
		p = 0.36
		V(2N) = 50.61
			P = 51.09
			p = 0.044
			V(2N) = 50.92
				P = Infinite
				p = 2.2 e -08
				V(2N) = 62.94

Rongorongo	N = 14623
	V = 619
		P = 1003.13
		p = 0.00096
		V(2N) = 745.31
			P = 799.72
			p = 0.0023
			V(2N) = 723.92
				P = Infinite
				p = 2.3 e -26
				V(2N) = 804.03

Table 1. Sample and estimated population sizes for five corpora according to three LNRE models.

In order to compare LNRE models, the zipfR package (Baroni and Evert, 2014) implements a multivariate chi-squared test used to measure the goodness of fit of the LNRE model (once its parameters have been automatically optimised) to the supplied frequency spectrum, as described by Baayen (2001:118-122). The lower the chi-squared value, the higher the corresponding p value, and the better the fit. If p is less than 0.05, there is a significant difference between the frequency spectrum estimated by the model and the supplied frequency spectrum.

Data produced by the LNRE models for all the five corpora discussed in this paper are given in Table 1, where N = number of tokens in the existing corpus, V = number of types in the existing corpus, P = estimated number of types in the population at the asymptote, p = statistical significance of the multivariate chi-squared goodness-of-fit statistic, and V(2N) = estimated number of types in a text of twice the original corpus size. The ZM model assumes a population with an infinite number of types (Evert and Baroni, 2005). For their Italian "ri" experiments, this assumption is correct, as there are indeed a potentially infinite number of words with the "ri" prefix in Italian. Thus the goodness-of-fit p value is higher for ZM than the other LNRE models for this corpus. However, for the other four language corpora, goodness-of-fit was much better for the GIGP and FZM models which allow for a finite number of types in the population, and thus enable us to estimate population size. The GIGP model was the only one which was able to run with all five observed frequency spectra.

Using the GIGP model, the vocabulary growth curve for English characters (not shown here) rose very rapidly within the corpus sample to its asymptotic value of 26, and did not rise thereafter even with extrapolation, showing that there were no "undiscovered" characters outside the corpus. By extrapolating the vocabulary growth curve for the Indus signs to a very large value, as shown in Figure 4, a horizontal asymptote was eventually obtained, suggesting that the vocabulary of the entire language was finite, consisting of about 1396 signs. In Figure 5, the early part of extrapolation curve of Figure 4 is shown.

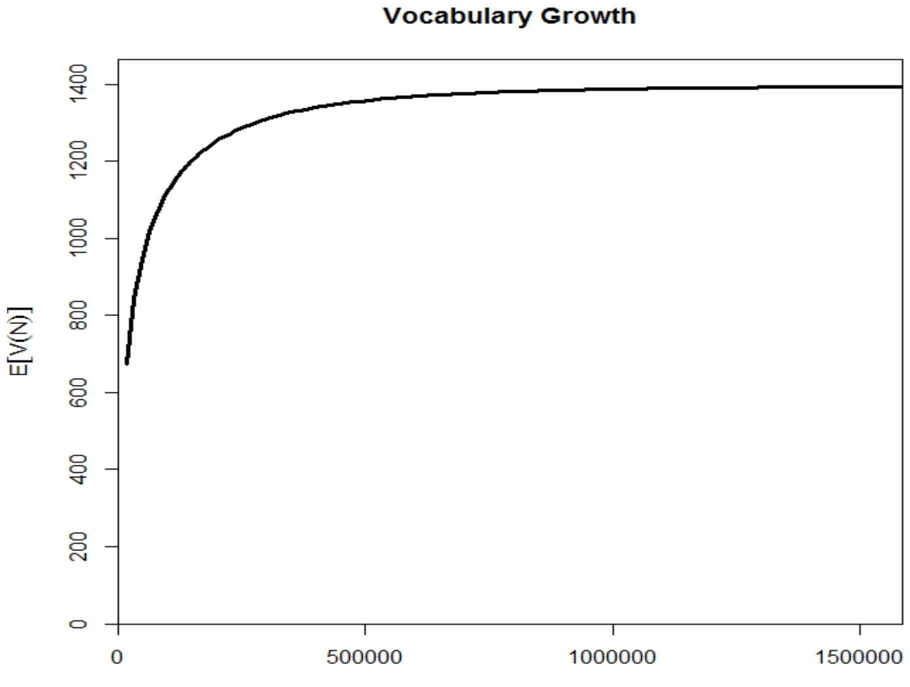

Figure 4. Extrapolated vocabulary growth curve for the Indus signs.

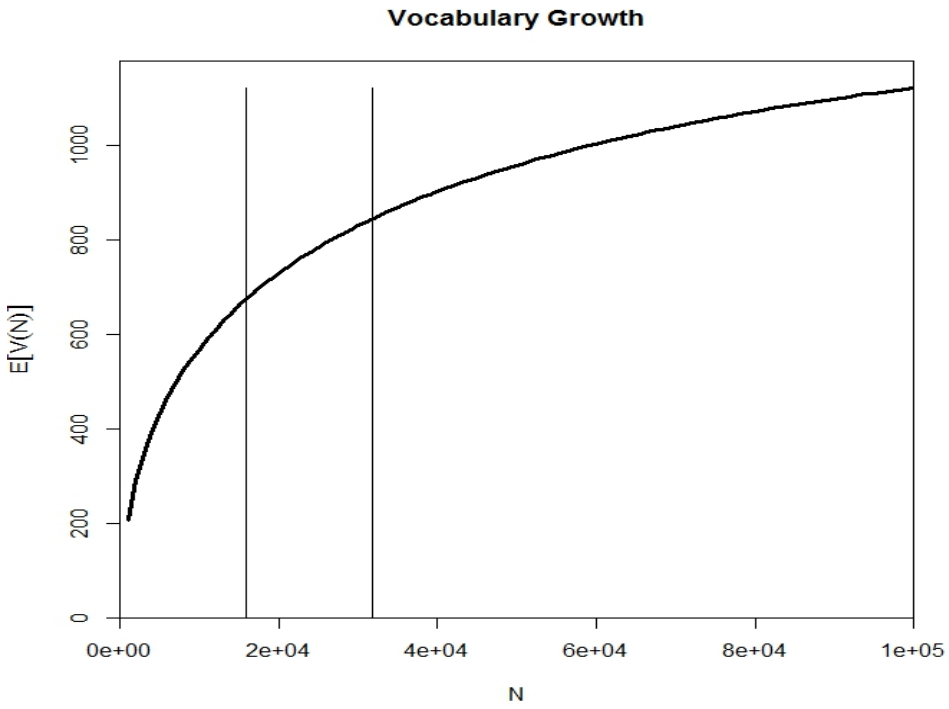

Figure 5. The leftmost portion of Figure 4 showing the region of greatest accuracy of the model.

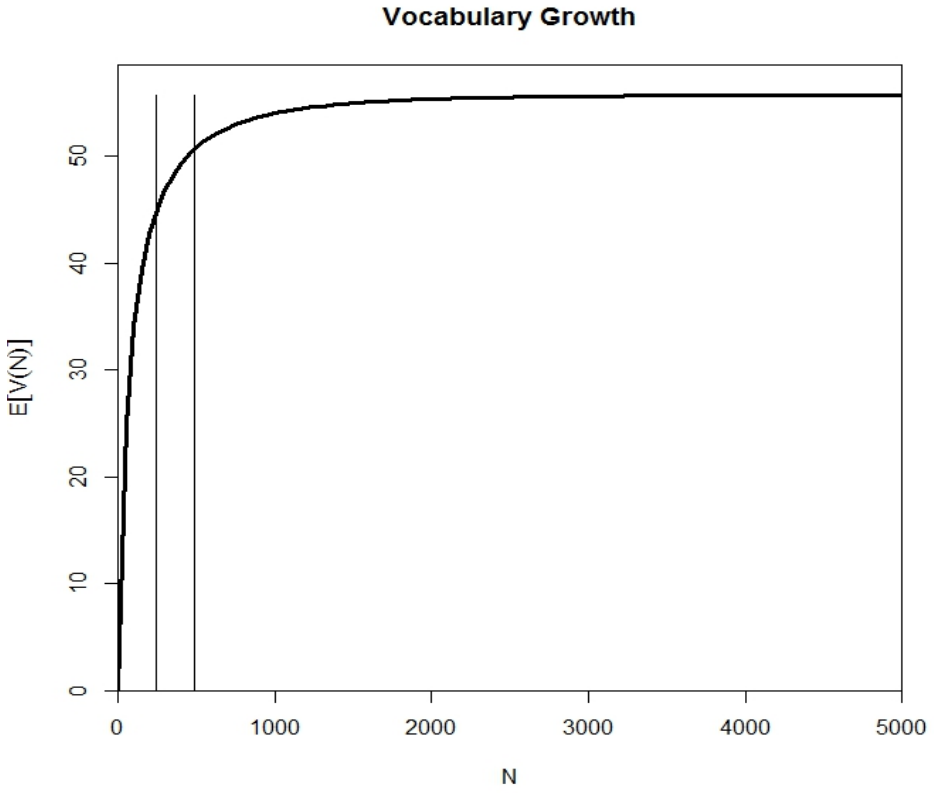

Figure 6. Extrapolated vocabulary growth curve for the Phaistos disk.

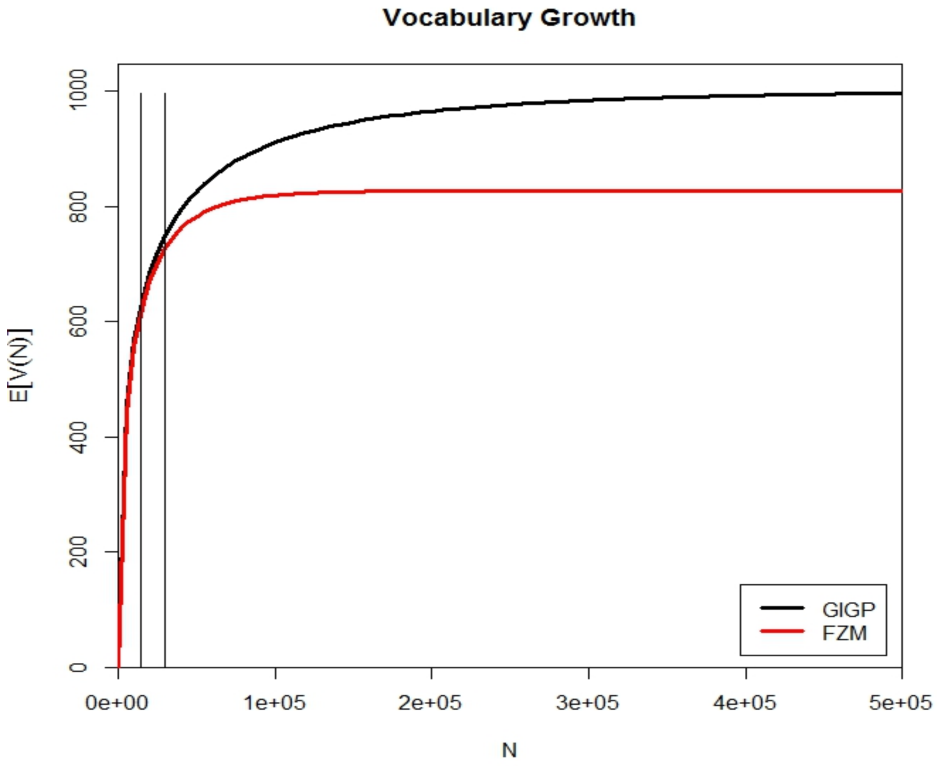

Figure 7. Extrapolated vocabulary growth curves for the Rongorongo corpus.

The vertical line in Figure 5 at N = 15868 corresponds with the corpus size, while the second vertical line at N=31736 corresponds with twice the corpus size. Evert and Baroni (2005) estimate that the vocabulary growth curves are accurate up to about twice the original corpus size, and so we can be reasonably sure that there are at least 843 symbols in the Indus language. Figure 6 shows the vocabulary growth curve for the Phaistos disk, which also reaches a horizontal asymptote, suggesting a finite number of symbols in the language of about 56. The estimated vocabulary at twice the corpus size is about 51. Figure 7 shows that the vocabulary growth curve estimated by the GIGP model for Rongorongo also reaches its horizontal asymptote of about 1003 characters in the language as a whole. The FZM model, which has better goodness-of-fit, reaches its horizontal asymptote at the much lower figure of about 800. Both these values are very rough estimates due to the difficulty of determining the number of different characters in the original corpus.

7. Conclusion

Although the computational approaches described in this paper have only scratched the surface of the decipherment task for the unknown scripts of Rongorongo, the Indus Valley script and the Phaistos disk, previous experiments have been able to use statistical measures to learn about the structural properties of these languages. The problem with Rongorongo and the Phaistos disk is that we have very little text in those scripts, and little prospect of finding more in the future. In contrast, we have many individual artefacts bearing Indus signs, though all of them are very short. New ones are being discovered all the time, so the corpus is continually increasing. There remains controversy as to whether the Indus signs are writing at all. Statistics alone cannot prove that something is written in a human language, but do provide empirical evidence about possible similarities between languages.

We have made use of a set of LNRE (Large Numbers of Rare Events) models to predict the sizes of the character sets in English, the Indus signs, the Phaistos symbols and the Rongorongo glyphs, in particular the GIGP (Generalised Inverse Gauss-Poisson model). Starting with the character frequency spectra for the known corpora (the sample), we were able to estimate the sizes of the character sets for the language as a whole (the population). We estimated the number of English characters in the alphabet to be 26.01. The number of Indus symbols in the entire language was estimated as 1396.4. This figure is much higher than the 676 characters given in Wells' (2011) appendix and Mahadevan's (reported in Robinson, 2009:281) estimate of about 425 allowing for allographs and ligatures. The higher estimate given here, which includes undiscovered signs, is consistent with a mixed syllabic and logographic language. The estimated number of Phaistos symbols here was 52.6, which accords well with the previous estimate made by Rumpel of 55 to 65 by estimating the asymptote of the type-token curve visually. This value is consistent with a syllabary.

References

Baayen R. H. (2001). *Word Frequency Distributions.* Kluwer Academic Publishers.

Baroni M. and Evert S. (2014). The zipfR package for lexical statistics: A tutorial introduction. 3 October 2014.

Evert, S. (2004). A simple LNRE model for random character sequences. In Proceedings of the 7th Journeés Internationales d'Analyse Statistique des Données Textuelles (JADT) (Louvain-la-Neuve, Belgium): 411-422.

Evert S. and Baroni M. (2005). Testing the extrapolation quality of word frequency models. *Proceedings of Corpus Linguistics* 2005.

Evert S. and Baroni M (2007). *zipfR*: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, Prague, Czech Republic: 29-32

Farmer S., Sproat R. and Witzel M. (2004). The collapse of the Indus script thesis: The myth of a literate Harappan civilization. *The Electronic Journal of Vedic Studies* 11(2): 19-57.

Farmer S., Sproat R. and Witzel, M. (2009). A refutation of the claimed refutation of the non-linguistic nature of Indus symbols: Invented data sets in the statistical paper of Rao et al (Science, 2009). http://www.safarmer.com/Refutation3.pdf.

Guy, J. (1990). On the lunar calendar of tablet Mamari. *Journal de la Societé des Océanistes* 91(2): 135-149.Harris M. (2010). Corpus linguistics as a method for the decipherment of Rongorongo. *M. Res. Dissertation*, Birkbeck University.

Khmaladze E.V. (1987). The statistical analysis of large number of rare events. *Technical Report* MS-R8804, Dept. of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

MacGregor N. *A History of the World in 100 Objects*. Harmondsworth: Penguin. Melka T. (2008). Structural observations regarding Rongorongo tablet "Keiti". *Cryptologia* 32(2): 155-179.

Pozdniakov K. (1996). Les bases du déchiffrement de l'écriture de l'île de Pâques. *Journal de la Societé des Océanistes* 103(2):289-303.

Pozdniakov K. and Pozdniakov I. (2007). Rapanui writing and the Rapanui language. Preliminary results of a statistical analysis. *Forum for Anthropology and Culture* 3: 3-36.

Oakes, M.P. (2014). Literary Detective Work on the Computer. John Benjamins.

Rao R., Yadav N., Vahia M. N., Joglekar, H., Adikhari, R. and Mahadevan I. (2009). Entropic evidence for linguistic structure in the Indus script. *Science* 324(5931): 1165. Online (with supplementary information) at http://homes.cs.washington.edu/~rao/ScienceIndus.pdf

Rao R., Yadav N., Vahia M. N., Joglekar, H., Adikhari, R. and Mahadevan I. (2010). Entropy, the Indus script and language: A reply to R. Sproat. *Computational Linguistics* 36(4), 795-805.

Robinson A. (2009). *Lost Languages*. Thames and Hudson.

Rao R., Yadav N., Vahia M. N., Joglekar, H., Adikhari, R. and Mahadevan I. (2009). Entropic evidence for linguistic structure in the Indus script. *Science* 324(5931): 1165. Online (with supplementary information) at http://homes.cs.washington.edu/~rao/ScienceIndus.pdf

Rumpel, D. (1994). Some quantitative evaluations of the diskos of Phaistos text. *Journal of Quantitative Linguistics* 1(2): 156-157.

Sproat, R. (2003). Approximate string matches in the Rongorongo corpus. http://clsu.ogi.edu/~sproatr/ror

Wells B. K. (2011). *Epigraphic Approaches to Indus Writing*. Oxbow Books.

Yadav N., Joglekar H., Rao, R. P. N., Vahia, M. N., Mahadevan, I and Adhikari, R (2010). Statistical analysis of the Indus script using n-grams. *PLOS One* 5(3).