STATISTICAL METHODS IN LINGUISTICS

4

SPRÅKFÖRLAGET SKRIPTOR

ON THE TYPE-FOUNT OF THE PHAISTOS DISC.

Alan Mackay

> *"What song Syrens sang, or what name Achilles assumed when he hid himself among women, though puzzling questions, are not beyond all conjecture."*
> *— Sir Thomas Browne (1605—82), Urn Burial Ch.5.*

The physical appearance of the Phaistos Disc has been frequently described since L. Pernier's finding of it in 1908, (*Ausonia*, III, p. 261, 1908) and this is the only aspect of studies of the Disc in which there is a reasonable measure of unanimity. The account by A. Evans (*Scripta Minoa*, I pp. 273–93, 1909) probably has priority as the definitive description in English (see also: "The Palace of Minos", I, pp. 647–68, 1921).

The inscription, which exactly fills both sides of a circular disc of baked clay, consists of 241 characters and was made by pressing pieces of type in succession into the plastic clay. 45 different symbols (apart from 'word-dividers') occur and in addition a dash made with a stylus (to call it a 'virama' would be to prejudge its function) appears 15 or 16 times. In what follows we reckon there to be 256 characters of 46 different kinds. Each impressed character has a definite orientation with respect to the direction of writing. The individual symbols are pictographic and rather complex (consisting of some 30 strokes in many cases) and it is clear that this method of "typewriting, not printing" (J. Chadwick's analogy) was much faster than one of drawing each symbol separately. An experiment showed that a text of 250 characters could be impressed on clay, using a fount of 52 different symbols, in only 14 minutes, to make an object resembling the Phaistos Disc. This corresponds to 3.5 seconds per character.

We will here not consider the difficult questions of why the text so exactly fills the space available and in what direction it should be read, as these are not relevant for the present analysis. However, we accept the demonstration by E. Grumach (*Kadmos, Bd.* I, *Heft* I, 16–26, 1962) that corrections were made while the clay was still plastic and that thus overlapping of the characters associated with the corrections is not admissible as evidence for the direction of writing.

STATISTICAL METHODS IN LINGUISTICS NO. 4, 1965 SKRIPTOR FACK STOCKHOLM 40

There appears to be no evidence to support the suggestion that several pieces of type were used for the same symbol, but examination of the object itself in Iraklion Museum would settle this question unequivocally.

We here suggest that the type may not have been made by direct engraving on wood, bone, bronze or other material, but that a more advanced technique was employed and that the type was perhaps cast from moulds. The process suggested is the inscription of a set of symbols on clay with a stylus, followed by the drying and hardening of the clay by baking, and the casting of the pieces of type in bronze or clay from these matrices. We find that Evans (*loc. cit.*) was of a similar opinion and suggested steatite matrices but adduced no evidence.

The basis for our suggestion is the nature of the symbols evident on the Disc. They consist predominantly of lines, such as would be produced by a stylus drawing on clay, and do not have large raised or lowered areas. To cut directly a punch, which would consist of thin raised lines on a uniform background, while not beyond the technology of the times, is much more difficult than the cutting of larger raised and lowered areas. Designing in lines is characteristic of the scribe, but working in intaglio marks the punch or seal-maker. The tradition of early bronze-working in Crete makes the proposal plausible and the recovery of a fount of type conceivable. The pieces of type appear to have been made, by whatever technique, without shoulders. With the progress in archaeology achieved since Evans's discussion of the objects portrayed by the symbols it has perhaps become possible to localise the cultural milieu giving rise to the Disc from its iconography.[1)]

E. Grumach (*loc. cit.*) has quoted with approval the statement of A. E. Kober (*Amer. Jour. Arch.*, 52, 87, 1948) that: "Forty-five different signs are found; how many there were in the script to which it belongs cannot even be conjectured". We give grounds below for conjecturing that the type-fount in question contained about 10 symbols not represented on the Phaistos Disc. A statistical discussion follows which may have wider application than to this immediate problem.

We make the general assertion that, in a small sample of an alphabetic or syllabic writing system consisting of L_1 characters of M_1 different kinds, the probable number of symbols in the alphabet or syllabary is, subject to various restrictions discussed below, given approximately by $L_1^2/(L_1-M_1) - L_1$.

Given that the probability distribution of the symbols in the writing system as a whole (measured from a large sample) is known, the probable number of symbols accidentally unrepresented in a small sample can be calculated and confidence limits can be attached to this estimate. However, we must stress that an estimate of the size of the type-fount from a single sample would be gravely vitiated if the sample were very atypical. The statistical calculation thus gives no more than the minimum value of the probable error. If, for example, the typist's test sentence – 'The quick brown fox jumps over the lazy dog' – were thought to be a typical English sentence, then the method to be outlined would give 111 as an estimate of the probable number of symbols in the English alphabet.

It is shown that a number of known alphabets and syllabaries follow the

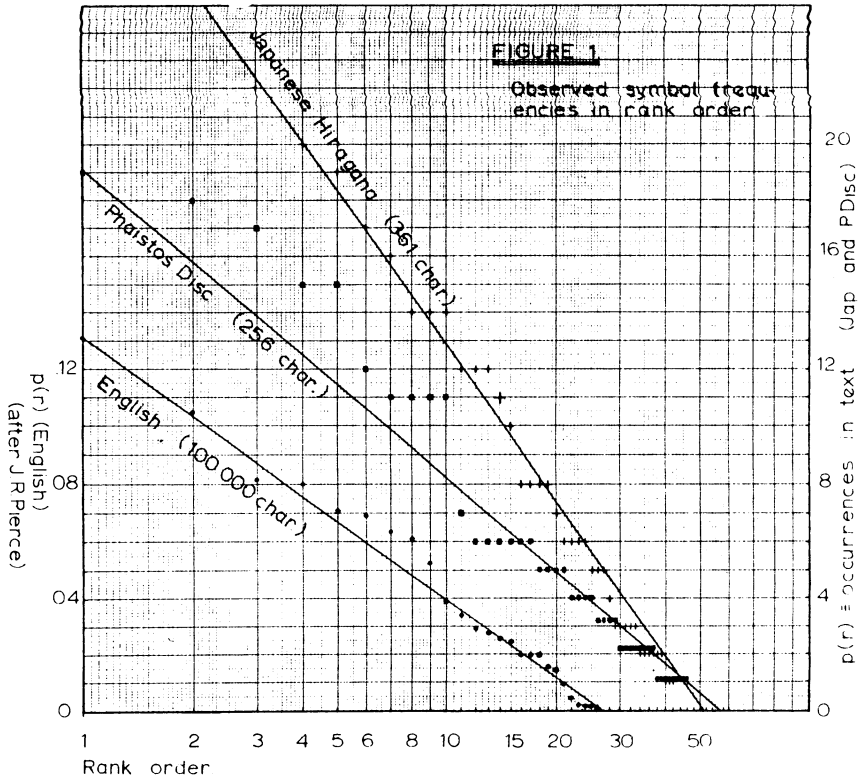
same general frequency distribution. It is assumed that the system of the Phaistos Disc follows the same kind of law and evidence for this assumption is adduced. With the further assumption that the text of the Disc is a typical specimen of the language, the probable number of missing letters and an estimate of the reliability of this number have been calculated.

Table I shows the number of times (t) which each different symbol of the Phaistos Disc inscription occurs. Evans's numbering (*loc.cit.*) of the symbols is used, but the number 46 is allocated to the dash. N_t is the number of different symbols which occur t times each. $N_2=8$ means that eight different symbols occur twice each. M_t is the number of different symbols which occur at least t times. M_t is the cumulative total $N_{19}+N_{18}+\ldots+N_t$, and L_t is the further cumulative total $M_{19}+M_{18}+\ldots+M_t$. Thus L_3 is the number of characters remaining if one of each kind of symbol observed is deleted and the same process is carried out again on the remainder. N_o is the number of symbols which exist in the writing system but do not happen to appear in the inscription under consideration because of its brevity. M_o is the total number of symbols in the writing system. M_1 is the number of different symbols actually observed in the sample and L_1 is the number of characters observed (the length of the text).

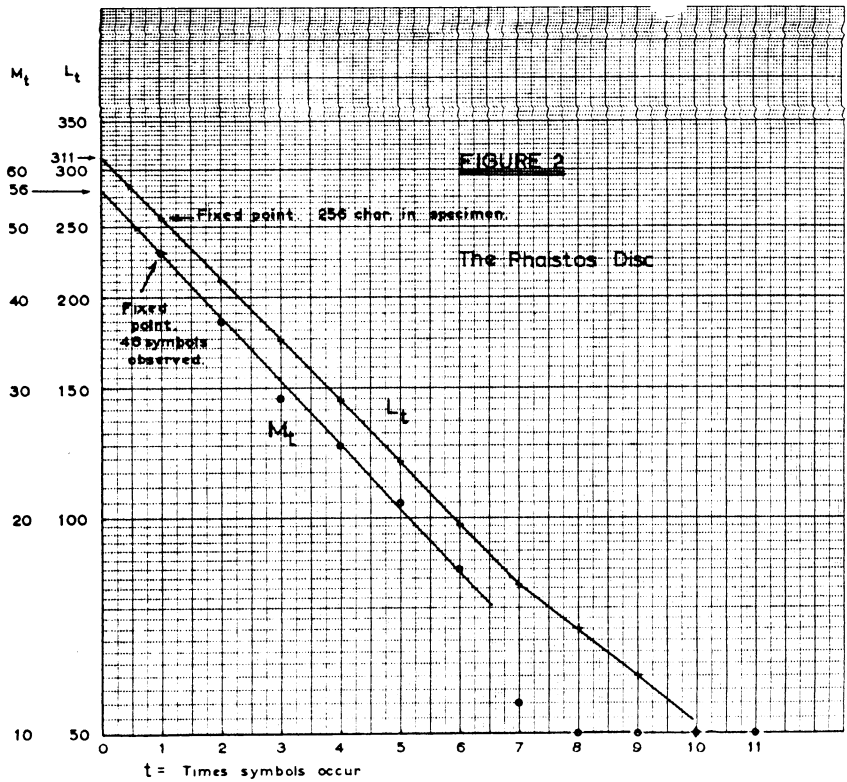
The observed symbols can also be arranged in rank order of frequency. Figure I shows the frequency of occurrence p_r plotted against the logarithm of the rank r. This has been done for the text of the Phaistos Disc and for a specimen of Japanese written in the hiragana syllabary. The latter specimen was spelled out in this way for the use of children and is deficient in alien words. The frequency/rank distribution observed in a large sample of English-language text has been plotted also. These graphs show that the frequency distributions of letters in small samples of the languages considered are of the same type. If we formulate the hypothesis that this type of frequency distribution, given by $p_r/p_1=1-(\log r)/(\log R)$, where R is one more than the number of symbols, is followed by small samples of all writing systems, then the graphs of figure I tend to confirm this hypothesis. Large samples of text are required to show deviations from this distribution. It is sufficient for our present purpose, however, to show that the Phaistos Disc inscription follows this law.[2)]

We require to find M_o, the number of different symbols in the writing system which have a probability of occurrence greater than zero.

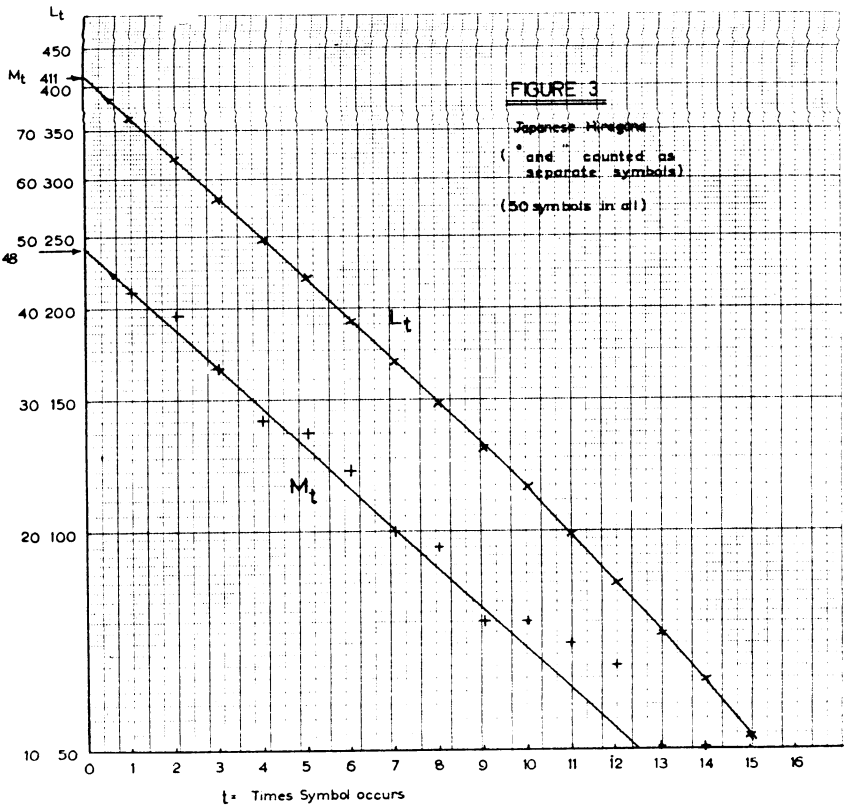
Although a plot of the observed frequencies of symbols gives a straight line, it is not suitable for accurate extrapolation to give the rank of the



symbol with lowest probability. After trials we find it best to use the functions M_t, and L_t, extrapolation of which to $t=0$ give M_o and L_o respectively. Plots of log M_t and log N_t against t give remarkably straight lines. From $p_r/p_1 = 1 - (\log r)/(\log R)$ it follows that $M_t = M_o \exp(-Kt)$ and $L_t = L_o \exp(-Kt)$, where K is a constant. These lines thus have the same slope. Log L_t plotted against t gives such a straight line that the intercept at $t=0$ can be calculated simply from $L_o/L_1 = L_1/L_2$ (which gives the rule mentioned above). It will be noted that this formula uses the best data available, namely, the total number of characters in the specimen and the total number of different symbols observed. If N_1 is small compared with other terms of N_t it may indicate that few symbols are unused; the graph of log M_t against t will then be slightly curved as it approaches $t=0$ and graphical extrapolation of M_t gives, in these circumstances, the best estimate of M_o, as experiments showed.

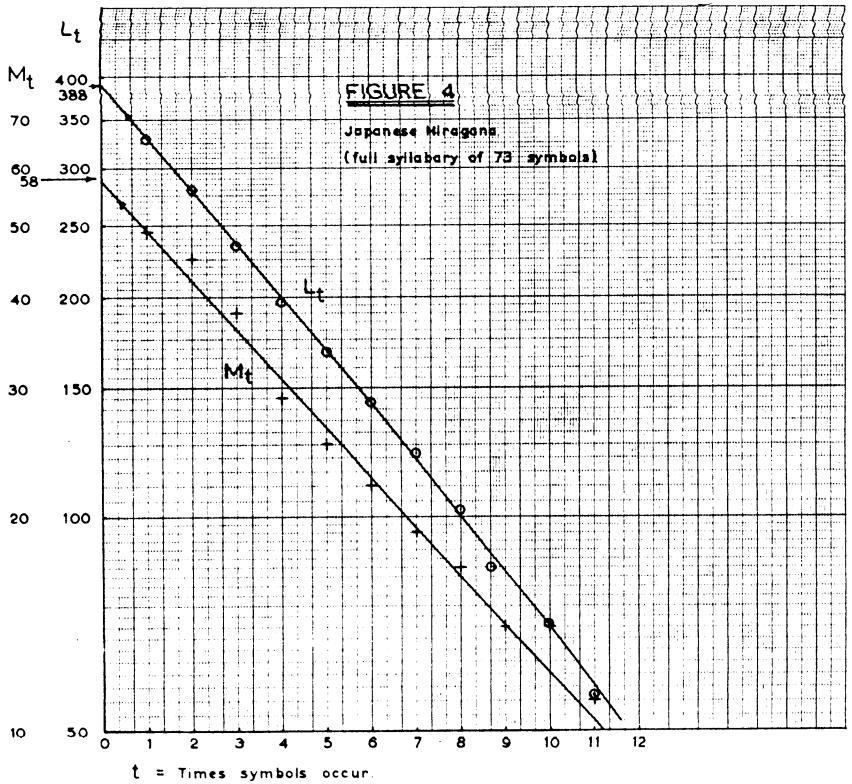


The procedure described has been carried out for the Phaistos Disc (figure II) and for Japanese Hiragana (figures III and IV), as well as for about ten other writing systems, the results of which are not illustrated but are listed in Table II. The lengths of the samples were chosen to have roughly the same average number of occurrences per symbol as those on the Disc. It will be seen that the three variants of our method – extrapolation of L_t graphically, extrapolation of M_t graphically and the use of the formula $L_o/L_1 = L_1/L_2$ – agree fairly well with each other. It will be seen, too, that the actual number of non-observed symbols in various samples corresponds reasonably well with the estimates thus obtained. Notes to Table II discuss some exceptions. The agreement should add weight to the estimate of 10 as the expected number of extra symbols in the writing system of the Phaistos Disc.



A minimum value for the probable error in our estimate of 10 can be obtained as follows:

We take a stright line representing p_r against log r for the Phaistos Disc. That is, we assume $p_r/p_1 = 1 - (\log r)/(\log R)$ exactly and put in the values $R = 57$ and $p_1 = 19.2/256$. The former comes from our estimate of the value of M_0 at 56 and the latter from the condition that the sum of the probabilities of all the possible symbols must be one. From these values of p_r the expected number of missing symbols E(O) in an inscription of 256 characters can be calculated from $E(O) = \sum_{r=1}^{r=56} (1 - p_r)^{256}$. This comes to 9.418 which is not far from our estimate of 10. The variance of this estimate, which we will expect to be similar to the variance of our estimate from the actual frequencies of the Disc, is calculated from



$$V(O) = \sum_{i=1}^{56} \sum_{j=1}^{56} (1 - p_i - p_j)^{256} - \sum_{i=1}^{56} (1 - 2p_i)^{256} - E(O)^2 + E(O).$$

Computation gives V(O) = 4.218. This means that the true value of E(O) is more likely than not to lie between 11.5 and 7.4.

We repeat, however, that this gives only an estimate of the inherent, "mathematical" error of the process, and cannot include any estimate of the errors due to the unrepresentative nature of the sample which, as we have shown above may, in special conditions, be overwhelming.

The treatment of the Japanese syllabary may illustrate some of the circumstances which may invalidate the above analysis. The basic syllabary consists of combinations of 10 consonants with 5 vowels. There is also a single character "–n" which may close a syllable. As three possible combinations are missing the total number of kana is 48. However, by adding

the nigori sign ゛, 20 of the syllables with unvoiced consonants can be converted to those with the approximately corresponding voiced consonants and by using the mark °, h can be converted to p. These modifications give an extra 25 signs making 73. Should the number of signs in the syllabary be counted as 48 plus " and °, or as 73? We suggest that the former course is the proper one and estimates of the total fount from small samples are wide of the mark in the latter circumstance (see figures III and IV). By using modifying marks whole blocks of symbols have been created and preserved irrespective of whether they are all individually required. It is only to be expected then that there will be, in such a case, an abnormal number of signs of very low frequency. Some of the symbols are, in fact, only required for representing foreign words. In Japanese the voiced and unvoiced consonants are not entirely distinct but are related by the phonology of the language. Thus 'kana' and 'gana' commute as readings of the same character, and the syllables 'ka' and 'ga' are not distinct entities. A more extreme example can be seen if capitals and lower case letters in a modern alphabet are taken as distinct symbols. It is clear that letters beginning the first words of sentences (plus proper names, etc.) do not belong to the same population as the small letters making up the bulk of the words. Among other considerations, the capitals are some fifty times rarer than the small letters. As an extreme example, four letters of the Russian alphabet never occur as capitals. The combined population cannot, therefore, be expected to follow a common probability distribution. Similarly it is clear that numerals should not be included with alphabetic signs, nor ideographic with phonetic, as they will follow different probability distributions and the proportions of each type of symbol will vary from text to text.

We suggest that consciously impressing a matrix pattern on the symbols representing the phonetics of a language may force on it more symbols than there is need for. The 2×26 matrix of English capital and lower case letters makes a total of 52 symbols, but half of these are merely an orthographic device and are not independent symbols.

The notation of the Phaistos Disc shows no indication of numerical symbols although some symbols might be ideographic rather than phonetic. However, circumstances (its uniqueness for one) make it likely that the notation is homogeneous and coherent. This contrasts with the example of the statistics of the linear B sample where there are far more symbols of low frequency in the syllabary than would be predicted by our technique.

This is perhaps explicable by the inhomogenuity of the corpus[3] used in compiling the syllabary table of 87 signs, of which 16 are not allocated phonetic equivalents. These 16 do not occur at all in the specimen used and only very rarely in the vocabularies of Ventris and Chadwick ('Documents in Mycenean Greek'). Our estimate of a type-fount of 60 for Linear B instead of 87 is thus not so wildly wrong.

The implications of our conclusion that the type-fount of the Phaistos Disc consisted of about 55 symbols (plus the dash) will not be discussed here, but it is clear that this is a very acceptable number of symbols for a syllabary of the consonant + vowel type to have. The above must be considered as only a preliminary essay in the field and to have raised many more questions than it has answered.

The author wishes to thank Dr H. D. Miller for statistical advice and other colleagues for compiling a computer programme and for patiently inducting him into the rudiments of a new field.

TABLE 1

The frequency of occurrence of symbols on the Phaistos Disc.
(Reference number of symbol in Evans's list)

	44																			
	43	41																		
	42	28																		
	30	21				40														
	17	20				45														
	15	16	38	39	36	33				35										
	11	14	34	37	31	26				29										
	5	9	32	10	22	24				23			46							
	4	3	19	6	8	13	25			1	18		27		12	7	2			
t=0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Number of symbols which each occur t times

N_t = ?	9	8	4	4	4	6	1	0	0	0	4	1	0	0	2	0	1	1	1	0

Number of symbols which each occur at least t times

M_t = ?	46	37	29	25	21	17	11	10	10	10	10	6	5	5	5	3	3	2	1	0

Cumulative total of M_t

L_t =	256	210	173	144	119	98	81	70	60	50	40	30	24	19	14	9	6	3	1	0

TABLE 2

Summary of results.

	Number of characters in the specimen — L_1	Number of different symbols observed — M_1	Mean number of occurrences per symbol.	Total fount estimated by extrapolation of M_1	Total fount estimated by extrapolation of L_1	Estimate from $L_1^2/L_2 - L_C$	$M_C - L_0 - L_1$	Best estimate of absent symbols.	Actual number of absent symbols.	Actual type-fount.
The Phaistos Disc.	256	46	5.56	56—57	55	312.1	56.1	10	?	?
Linear B	299	50	5.98	58—60	58	359.1	60.0	9	37	87
Japanese Hiragana (o, ", as symbols)	361	42	8.60	47	49	408.5	47.5	6	8	50
Japanese Hiragana (full syllabary)	329	49	6.71	56—58	59	386.6	57.6	9	24	73
Cypriot (Enkomi) (Doc. Myc. Gk. page 62)	391	57	6.86	60—63	70	457.7	66.7	11	?	?
Arabic (modern)	130	24	5.42	31	30	159.4	29.4	6	4	28
Armenian (modern)	150	32	4.69	41	38	190.7	40.7	8	6	38
Russian (modern)	200	27	7.41	31—32	35	231.2	31.2	5	5	32
Greek (modern)	115	20	5.75	24	23	139.2	24.2	4	4	24
Greek (classical)	115	20	5.75	23—24	23	139.2	24.2	4	4	24
Turkish (modern)	108	24	4.50	32	28	138.9	30.9	7	5	29
English (modern)	125	22	5.68	26	26	151.6	26.6	4	4	26
English ("quick brown fox...")	35	26	1.35	—	—	136.1	111.1	85	0	26

[1]) A. Evans ("The Palace of Minos", I, p. 657, 1921) in discussing the objects depicted on the Disc remarks, a propos sign No. 24 (in his numbering): "The most significant figure, however, among the signs on the Disc is the pagoda-like building shown..." We suggest that the sign in question shows, in fact, a closer resemblance to the palanquin (a clay model of which was found in the Palace of Minos and is illustrated on p. 220 of Evans's account) which may sometimes have been fitted with a canopy (as shown in the fresco illustrated on p. 772 of vol. II).

[2]) It should be noted that the letter frequencies studied here do not follow Zipf's Law as do word frequencies. We see a fundamental distinction in the substantially infinite number of possible words, while the number of different symbols used in an alphabet or syllabary is strictly finite.

[3]) The sample text was the sum of nine tablets from Pylos quoted in J. Chadwick, 'The Decipherment of Linear B', pp. 151–5, Penguin edition, 1961.