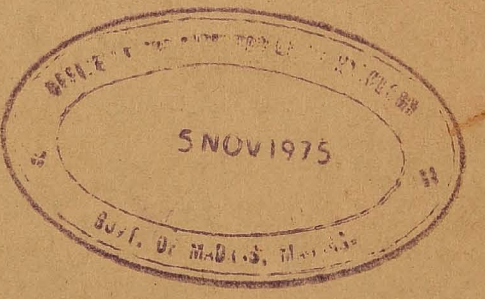
THE SCANDINAVIAN INSTITUTE OF ASIAN STUDIES

Joint Reprint Series Number Two

A METHOD TO CLASSIFY CHARACTERS OF UNKNOWN ANCIENT SCRIPTS

By

SEPPO KOSKENNIEMI, ASKO PARPOLA AND SIMO PARPOLA



COPENHAGEN 1970

The article enclosed is reprinted from *Linguistics*, 61 (September, 1970) by kind permission of Mouton & Co., Publishers, The Hague, The Netherlands.

A METHOD TO CLASSIFY CHARACTERS OF UNKNOWN ANCIENT SCRIPTS

SEPPO KOSKENNIEMI, ASKO PARPOLA AND SIMO PARPOLA

THE PURPOSE

The present paper has grown out of our efforts to decipher the Indus script used in NW India c. 2500-2000 B.C. Since no bilinguals so far have been unearthed, the only key to the script is the internal structure of the preserved texts, as it was in the case of the Cretan Linear B script that M. Ventris succeeded in decoding 15 years ago.

The method used by Ventris consisted in arranging the easily discerned syllabic signs into tables showing their components (C + V) on the ground of the textual behaviour of the signs alone and then giving phonetic values to the components on the basis of certain clue words (place names). E.g., Ventris got a series of signs containing the same consonant but a different vowel from words consisting of several signs and occurring with variants which differed only in the last sign and which apparently represented declined forms of the same word (cf. Latin *do-mi-nus, do-mi-ni, do-mi-no*).

In Linear B the total number of different signs is c. 250, of which c. 90 are used syllabically, the logograms being easily recognizable by means of word division: they occur alone and together with numerals in the texts. The syllabic signs consist, as Ventris presupposed, of open syllables only (type CV, V), and there were long sentences clearly cut into words many of which are of considerable length.

The thousand years older Indus script presents a much more difficult problem. One of the great obstacles to its decipherment is the shortness and one-sidedness of the preserved materials which consist of c. 2000 short inscriptions, mostly seals, the average length of which is 5 signs. Another difficulty is that the words are not separated from each other. This difficulty can, however, be removed to a certain degree, for, as we shall show in another paper, the separation is possible to a considerable extent; the normal word length varies from one to three signs. The shortness of the words, the number of different signs, approximately 300, and

the analogy between other contemporary writing systems shows that Indus script must have been of the logo-syllabic type. Hence there must have been closed syllables and logograms besides open syllables (types CV, V, VC, CVC, WORD), a fact that does not make decipherment easier.

We have approached the problem i.a., by starting from a selection of known ancient scripts. The idea has been to develop a method of analysis which would yield the same reasonable result when applied to any samples of known scripts and hence also when applied to an unknown script sample.

THE PROBLEM

Mathematically, the basic problem is to find a classification procedure which is based purely on statistical principles and can group the signs of written language in such a way that this classification would be constant. That is: the result must be independent of language. The only information that can be used for the procedure is the different frequencies and other statistics obtained from a sample of the written language that might be quite an unknown one. No grammatical or phonetical information is supposed to be available.

A good collection of data for classification purposes could comprise, for example, the frequencies of different signs, the position of signs within words and their occurrence in ligatured signs, and the pairwise appearance of signs.

Should a classification method be found which meets with these requirements it will very possibly have the following feature: applied to known languages it gives a grouping which may be very near to some grammatical or phonetical classifications.

However, the procedure need not lead to any *a priori* known classifications. The only feature required is that the result be constant.

THE METHOD

The input data for this classification procedure is composed only of the information obtained by collecting all pairwise frequences of the different signs. In other words, the procedure tries to group together signs which appear in similar surroundings and avoids grouping signs which, statistically, very seldom appear next to each other. We shall call the matrix of the pairwise frequences by F.

$$F = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & \\ \vdots & & & \\ f_{n1} & f_{n2} & & f_{nn} \end{pmatrix}$$

where f_{ij} shows how many times the signs called i and j appear next to each other (first i and then j). If two signs in this matrix have rows which are much alike we can say that these signs have similar right-hand-side-surroundings. Correspondingly, two similar columns indicate that the left-hand-side-surroundings of the signs are similar.

In order to eliminate the effect of the single frequencies of the signs themselves, we must normalize the binomial distributed random variables f_{ij} by their means and standard deviations. Assuming that the sample is large, we can use the formula

$$\bar{f}_{ij} = \frac{f_{ij} - s_{ij}}{\sqrt{s_{ij}}},$$

where

$$s_{ij} = \left(\sum_k^n f_{kj} \sum_l^n f_{il}\right) \Big/ \sum_k^n \sum_l^n f_{kl},$$

and where the new variables $\bar{f}_{ij}$ form the new double-frequency-matrix $\bar{F}$.

We now define the distance between two signs by the cross product of corresponding rows in the matrix $\bar{F}$ when the right-hand-side surrounding of the signs is concerned, and of columns in the matrix $\bar{F}$ when we are interested in the left-hand-side-surrounding of the signs. We normalize these cross products by the means and norms of the row (or column) — vectors and subtraction from one. In this way we get the distance matrix R which is the only data for the classification procedure.

We needed a new classification method because the known methods, such as factor and discriminant analysis, taxonomy, and so on, lack the correct criterion for our purposes. We are not searching for typical signs (that kind of classification is done in taxonomy) and we do not want signs with a strong negative correlation to appear in the same group (that happens when using factor analysis). We found out that a good criterion for our purposes is the total average within-groups distance (K) which is to be minimized:

$$K = \frac{1}{N} \sum_{k=1}^{m} \sum_{\substack{i \subset I_k \\ j \subset I_k}} r_{ij}$$

where

$$N = \sum_{k=1}^{m} \frac{n_k(n_k - 1)}{2}$$

and n_k = members (signs) in k'th class
m = number of classes
k = class-index
I_k = set of indexes of sings appearing in k'th class
N = total number of distances within all classes

The classification procedure is extremely simple and it is based upon the successive moving of signs from one class to another. Every move is to be selected from among all possible moves in such a way that the value of our criterion K after that step is minimal. This strategy will lead into a procedure which of course will converge since the value of K decreases all the time, but there is no proof of the optimality of the final classification. However, in practice a very good near-to-optimum solution is found.

Starting from two groups this procedure is then carried out by increasing the number of groups and considering the solution of every step as an initial classification for the next one. This method involves a great deal of computing and cannot be carried out without electronic computers. It is now programmed in FORTRAN language and a lot of material has been run at NEUCC[1] on IBM 7090.

THE RESULTS

The classification program has been applied to samples of five ancient scripts, each consisting of 8,000-10,000 signs. In this paper we are going to present the results of one sample *in extenso* in order to show the whole process. Of the other samples only the final results are given. It can be easily seen that in every case, the elements of the final groups stay together during the entire classification process. As a matter of fact the final results of basic groups are defined by means of three criteria:

(1) The frequencies of the elements of each basic group must not be small.

(2) The K-value (distance from the center of the group) must be small for every element which belongs to a basic group.

(3) Each basic group must hold together during the classification process.

[1] Northern Europe University Computing Center, Lyngby, Denmark.

In the listings of the results we present the percentage distribution of the signs, and the classifications into 2, 3 and so on groups. The symbols of the signs have been sorted out into ascending order according to the *K*-values. THIS MEANS THAT THE FIRST SIGNS ARE THE MOST SIGNIFICANT ONES.

1. *Elamite Cuneiform*

Elamite cuneiform of c. 1100 B.C. is a syllabic writing consisting of c. 130 syllabic signs (types CV, V, VC, CVC) and c. 30 logograms (used mostly as determinatives). The material analyzed[2] comprises the 47 stereotyped inscriptions (mostly dedications) from Tchogha-Zanbil published by M.-J. Stève in *Iranica Antiqua*, II (1962), 22-76, and two longer inscriptions of Šutruk-Nahhunte I and Šilhak-Inšušinak from F.W. König, *Die elamischen Königsinschriften* (=*Archiv für Orientforschung*, Beiheft 16)

TABLE 1

Frequency distribution of Elamite signs

Symbol	%	Symbol	%	Symbol	%	Symbol	%	Symbol	%
AK	4.78	HA	2.26	TU	1.36	ZA	0.82	AR	0.51
ME	4.18	IK	2.06	KA	1.26	LU	0.80	E	0.48
AN	3.98	ŠI	1.98	GI	1.23	UT	0.80	IM	0.46
A	3.93	IH	1.97	TI	1.16	DU	0.79	I	0.44
UN	3.64	RA	1.90	TA	1.13	AT	0.79	MU	0.44
ŠU	3.44	LI	1.82	SU	1.13	UL	0.75	BA	0.43
HU	3.38	IA	1.74	LA	1.08	NU	0.75	MI	0.41
NA	3.10	TE	1.72	UR	1.08	PU	0.70	EL	0.39
U_3	2.92	ŠA	1.52	IP	1.03	UŠ	0.69	ÁŠ	0.28
KU	2.92	UK	1.47	IR	1.02	H	0.67	UP	0.23
SI	2.59	PA	1.46	PI	0.95	RU	0.56	SA	0.21
KI	2.59	RI	1.43	IT	0.90	IŠ	0.54	PIR	0.21
NI	2.52	HI	1.38	EN	0.87	KA_4	0.52	DA	0.18
IN	2.38	MA	1.36	TAŠ	0.85	GA	0.51	AM	0.16

Total 70 different symbols
Total 6103 signs

[2] Unlike the other samples, this cuneiform has not been coded directly but from the transcription; this has caused an inexactitude with regard to the H-sign (thus in König's transcription): by Stève it has been transcribed with (the same) vowel (as in which the preceding syllable ends) + H.

(Graz, 1956), nos. 28 A and 45. All the 70 syllabic signs occurring in the sample have been included in the first two classifications according to the

TABLE 2

Classification of the Elamite syllables according to the text following the signs

Two groups

I	II
LA	AT
TA	UT
ŠA	ÁŠ
BU	MU
TAŠ	DU
KA	UR
IR	IK
ZA	IŠ
IH	AK
HA	AR
NU	LU
IT	EL
IM	AM
U_3	TU
IP	H
IA	PU
RI	RU
GI	PIR
MI	KU
PA	DA
KA_4	ŠU
RA	SU
LI	UK
A	GA
UL	E
TI	HU
MA	EN
HI	IN
UP	UN
NI	PI
SA	AN
TE	SI
ME	
I	
NA	
UŠ	
ŠI	
KI	

Three groups

I	II	III
LA	AT	TI
KA	ÁŠ	LI
TA	UT	RI
ŠA	MU	MI
KA_4	LU	IH
BA	IŠ	ME
ZA	UR	IR
PA	DU	NI
A	IK	ŠI
RA	AK	KI
IA	TU	HI
DA	AM	GI
TAŠ	PU	PI
HA	EL	TE
IP	H	UŠ
U_3	RU	IM
NU	KU	UP
NA	PIR	I
IT	ŠU	GA
AR	SU	UL
UK	HU	SI
MA	EN	AN
SA	IN	
E	UN	

Four groups

I	II	III	IV
LA	AT	LI	ŠU
KA	ÁŠ	TI	SU
TA	UT	RI	PU
ŠA	MU	MI	HU
KA_4	AK	IH	IN
ZA	IŠ	ME	SA
PA	IK	NI	E
BA	LU	IR	UL
A	UR	ŠI	GA
DA	DU	HI	I
TAŠ	AM	KI	AN
IA	TU	GI	MA
IP	EL	PI	EN
RA	H	UŠ	
HA	PIR	TE	
U_3	KU	IM	
NU	RU	UP	
IT	UN	SI	
NA			
AR			
UK			

Five groups

I	II	III	IV	V
LA	AT	LI	ŠU	AR
TA	AŠ	TI	SU	UK
KA	MU	RI	PU	AN
ŠA	AK	MI	KU	I
ZA	UT	NI	H	GA
BA	IŠ	ME	RU	EN
A	DU	ŠI	HU	E
PA	UR	HI	IN	UN
KA_4	AM	IH	SA	UP
TAŠ	TU	IR	UL	SI
IA	LU	KI	MA	
HA	IK	GI		
NU	PIR	UŠ		
RA	EL	PI		
IP		IM		
DA		TE		
U_3				
NA				
IT				

adjacent signs on each side (Tables 2 and 3). Then follows a classification which is based on these two analyses, but carried out anew with fewer variables, only the forty most frequent signs being included (Table 4).

TABLE 3

Classification of the Elamite syllables according to the text preceding the signs

Two groups

I	II
IR	KI
E	AK
IN	KA
IŠ	TAŠ
IP	GI
IM	NU
PI	PIR
IK	AR
EL	LI
IT	HI
DU	HA
TU	A
ŠI	AM
UK	ÁŠ
EN	TA
ŠA	H
MU	PU
SA	KA_4
UL	ZA
UT	SI
MA	RU
RI	RA
UP	AT
I	DA
HU	SU
TI	GA
IH	NI
UŠ	TE
KU	LU
LA	UR
MI	AN
BA	
UN	
IA	
U_3	
PA	
ME	
ŠU	
NA	

Three groups

I	II	III
IR	KI	UL
E	AK	ŠI
IŠ	GI	UK
IP	KA	UT
EN	TAŠ	SA
IT	PIR	TU
IK	A	PI
IN	LI	UP
IM	HA	UN
I	AR	UŠ
EL	NU	DU
IH	AM	RA
LA	HI	HU
TI	KA_4	RU
ŠA	PU	SI
TE	ÁŠ	KU
IA	AT	H
RI	LU	TA
NA	NI	ŠU
BA	DA	ZA
SU	GA	MA
ME	AN	UR
		MU
		MI
		U_3
		PA

Four groups

I	II	III	IV
IR	KI	UL	PA
E	AK	UT	RA
IŠ	TAŠ	ŠI	KU
IP	KA	UK	RI
EN	GI	SA	ŠA
IT	PIR	UŠ	ZA
IN	A	UP	SI
IK	LI	UN	TA
I	AR	PI	BA
IM	NU	TU	PU
EL	AM	HU	MI
IH	HA	H	ME
LA	HI	DU	DA
IA	KA_4	ŠU	GA
TI	AT	UR	SU
TE	ÁŠ	RU	U_3
NA	LU	MU	
	NI	MA	
	AN		

TABLE 4

Final classification of the Elamite syllables according to the text on both sides of the signs (only 40 most frequent signs)

Two groups

I	II
IP	ŠI
IR	NI
ŠA	HI
IT	HU
TA	SU
PI	GI
LA	ŠU
MA	UK
KA	UR
RI	AK
IK	TU
PA	HA
RA	AN
TI	LI
U_3	UN
IA	KU
NA	
KI	
TE	
A	
IH	
SI	
ME	
IN	

Three groups

I	II	III
IP	NI	IR
KA	ŠI	HU
TA	HI	IN
RA	LI	IK
PI	GI	ŠU
MA	HA	IH
KI	AK	SU
IA	UR	RI
LA	UK	NA
ŠA	ME	KU
IT	AN	TU
U_3		UN
TI		
A		
SI		
TE		
PA		

Four groups

I	II	III	IV
TA	NI	IR	RA
A	HI	IP	PA
ŠA	KI	IT	KU
LA	TI	IK	UK
NA	LI	HU	UR
IA	GI	IN	SI
MA	ŠI	U_3	TU
KA	PI	IH	UN
HA	ME	RI	AN
	TE	SU	
	AK	ŠU	

Five groups

I	II	III	IV	V
TA	NI	IR	RA	SU
ŠA	HI	IP	PA	HU
A	TI	IT	KU	ŠU
LA	ŠI	IK	SI	GI
NA	KI	U_3	UR	AK
MA	LI	RI	UK	HA
KA	ME	IH	UN	TU
IA	PI	IN		AN
	TE			

Six groups (these groups are no more relevant)

I	II	III	IV	V	VI
TA	KI	IP	PA	GI	UK
A	NI	IT	RA	SU	ŠI
ŠA	KA	IR	ME	ŠU	UR
NA	HI	IK	RI	HU	UN
LA	TI	U_3	PI	AK	TU
MA	TE	IN	SI	HA	AN
IA	LI	IH	KU		

2. *Cretan Linear B Script*

Cretan Linear B script is archaic Greek of c. 1500 B.C. The (logo-)syllabic writing consists of c. 250 signs of which c. 90 have syllabic values (types CV and V alone). The material analyzed comprises the texts nos. 1-45; 47-103,4; 134-136,2 in *Tabellae Mycenenses selectae* (=*Textus minores*, 28), ed. by C. J. Ruijgh (Leiden, 1962).

TABLE 5

Relative frequencies of the Linear B signs

Symbol	%	Symbol	%	Symbol	%
RO	4.44	DE	1.64	DI	0.56
TO	4.20	U	1.62	ZE	0.52
KO	3.88	PE	1.59	MI	0.45
E	3.86	MA	1.59	ZA	0.45
JO	3.66	RI	1.55	A_2	0.43
O	3.55	PO	1.30	QI	0.40
A	3.46	NE	1.28	NU	0.38
JA	3.41	SO	1.26	JE	0.36
RE	3.23	KU	1.26	QA	0.36
TE	3.12	WA	1.23	TU	0.36
TA	3.05	TI	1.21	SI_2	0.25
RA	2.96	DA	1.19	A_3	0.25
KE	2.96	I	1.05	ZO	0.20
PI	2.76	RU	0.99	SU	0.16
WE	2.74	WI	0.99	PTE	0.16
WO	2.69	DO	0.99	RA_3	0.16
PA	2.58	KI	0.92	RA_2	0.13
NO	1.97	MO	0.90	PU_2	0.13
ME	1.93	QO	0.85	DU	0.11
KA	1.86	PU	0.79	TA_2	0.09
SI	1.84	NI	0.74	RO_2	0.09
QE	1.77	SE	0.74	NWA	0.04
NA	1.75	SA	0.61		

TABLE 6

Classification of Cretan Linear B Script

According to the text following the signs

I	II	III	IV
MA	WI	SO	NA
A_2	SI	NO	QO
PU	TI	RO_2	RU
PO	NI	TA_2	TO
E	RI	DE	RA_3
MO	TE	NE	WO
PA	SE	SI_2	A_3
KE	KI	NU	WA
DU	DI	ZO	NWA
JE	JO	JA	ME
KA	I	QI	QE
PTE	PI	SA	A
PU_2	SU	RA_2	RA
KU	TA	TU	WE
ZA	RE	DO	ZE
U	PE	RO	O
MI	QA	KO	
DA			

According to the text preceding the signs[a]

I	II	III	IV
U	TE	TI	PE
WI	RO	JO	SO
PA	RA	ME	NO
SI	RE	O	TO
KA	RU	RI	NA
WE	NE	DE	KO
A	KI	JA	TA
DA	MA	KU	PO
WO	I	QE	
PI	DO	E	
WA	KE		

According to the text on both sides of the signs[a]

I	II	III	IV
WI	MA	RO	NO
SI	PA	KI	SO
I	PO	TE	KO
RI	KU	RA	PE
TI	KA	RU	NA
JO	KE	WA	NE
PI	E	RE	TO
O	U	DO	TA
ME	DA	WO	DE
A			JA
QE			
WE			

[a] 40 most frequent signs only were included.

3. *Neo-Assyrian Cuneiform*

Neo-Assyrian cuneiform of c. 700 B.C. is a logo-syllabic writing consisting of c. 500 signs of which c. 200 have a syllabic value (types CV, V, VC, CVC, VCV). Very many signs are used now syllabically, now as logograms; besides, one and the same sign can have different syllabic values. All this naturally hampers the classification procedure. The analyzed material comprises 43 letters (R. F. Harper, *Assyrian and Babylonian Letters* [Chicago, 1892-1914], nos. 1-2, 78, 85, 88, 90-1, 101, 126, 131, 134, 138, 140, 142, 144, 152, 154, 157, 167-8, 175, 178, 181, 386-7, 391, 406, 408, 410-1, 413-5, 419-421, 423, 531, 537-8, 541, 544, 561). The signs in table 7 (altogether 6948 out of 9089) have been included in the classification in table 8. The logographic values given are those occurring in these texts.

TABLE 7

Relative frequencies of Neo-Assyrian cuneiform

Syllabic value[a]	Logographic value	%
A	('water')	8.59
NI (LÍ, Ṣ/ZAL)	('oil')	6.07
NA		3.90
—	'king' (1)	3.74
AN	'god'	3.34
INA	'in'	3.31
I		2.99
ŠA		2.71
NU	('statue')	2.36
Ú, ŠAM	('plant')	2.33
ŠÚ	('totality')	2.32
(MEŠ, MÍŠ)	sign of pl.	2.16
—	'one'; sign of PN	2.16
TA	'with'	2.10
EN	'lord'	2.07
JA, JU, JÍ		2.04
U	10	1.96
—	'man'	1.96
LA		1.94
BE, BAD/T, TIL	many values	1.93
TÚ, UD/T, PAR (PIR, ḪIŠ, LAḪ)	'day', etc.	1.78
MU	'name'	1.77
BU, PU	('long')	1.77
LU, DIB/P	('sheep')	1.53
LI		1.51
KA	('mouth')	1.41
E		1.40
PA (ḪAT)	a god, etc.	1.32
BI		1.31
RI, D/TAL		1.25
KU	('to sit', etc.)	1.18
ŠI, LIM	'face'	1.09
KI, QÍ	'earth'	1.05
RU, ŠUB/P	'to cast'	1.05
ŠÁ	'to place'	1.02
BIT	'house'	0.96
SI	('horn')	0.94
SA		0.94
ḪI, ṬÍ	'good'	0.91
RA		0.86
MAN, NIŠ	'king (2); 20'	0.85
ŠU (QAT)	'hand'	0.85
TE	'to approach'	0.83
ME, ŠIB/P	(sign of pl.)	0.83
—	'heart'	0.82
DA, ṬA		0.81

[a] Rare values in parentheses

Syllabic value	Logographic value	%
(MUḪ)	'crown of the head'	0.81
—	'servant'	0.81
UR, TAŠ, LIK/Q		0.75
TI	('to live')	0.73
IŠ, (MIL)		0.66
BA	('to present')	0.66
DU	('to stand, to go')	0.65
SU	('to compensate')	0.63
IS/Ṣ/Z	'wood'	0.60
TU	'to enter'	0.59
KUR, MAT (ŠAT, LAT)	'country'	0.58
AD/T/Ṭ	'father'	0.56
ID/T/Ṭ	('arm')	0.53
ÁŠ		0.50
(GAL)	'big'	0.49
UN	'people'	0.45
[MA		0.00]

TABLE 8

Classification of the Neo-Assyrian cuneiform

According to the text following the signs

I	II	III	IV
ŠU	LA	NI	UN
KU	PA	MÍŠ	U
JU	TA	RI	BE
LU	SA	ÁŠ	ŠÁ
TU	BA	LI	AN
BU	*1*	SI	GAL
DU	IS	TI	MAT
NU	RA	IŠ	BI
TÚ	DA	ḪI	INA
ID	KA	EN	[MA]
E	I	ŠI	*HEART*
RU	MU	NA	ŠA
SU	MUḪ	ME	A
AD	Ú	TE	
ŠÚ	BIT	KI	
KING	*MAN*	*SERVANT*	
MAN			
UR			

According to the text preceding the signs

I	II	III	IV
SERVANT	ID	A	U
DA	IZ	NA	Ú
SI	E	*HEART*	SU
EN	IŠ	LI	UR
SA	ŠU	AD	UN
MAN	BIT	ÁŠ	RI
ŠÁ	ŠÚ	[MA]	MU
BA	LA	*KING*	TE
ŠA	I	ḪI	LU
TI	JU	BI	TÚ
INA	*I*	NI	BU
AN	ŠI		BE
TA	MAT		KI
TU	MUḪ		RU
GAL	MÍŠ		NU
RA	KU		
PA	ME		
MAN	KA		
DU			

According to the text on both sides of the signs

I	II	III	IV
ŠU	INA	SA	LI
E	UN	TI	AD
KU	AN	SI	ḪI
LU	EN	DA	KI
TÚ	ŠÁ	LA	[MA]
IS	U	TA	UR
ID	GAL	RI	*HEART*
ŠÚ	MAT	*I*	NI
DU	BE	*MAN*	BI
TU	RA	BIT	A
I	MÍŠ	PA	KA
MAN	*SERVANT*	BA	
BU	ME	MUḪ	
JU	ŠI	MU	
Ú	*KING*	NA	
SU	NU	ŠA	
IŠ		ÁŠ	
RU			
TE			

4. *Middle-Egyptian Hieroglyphs*

Middle-Egyptian hieroglyphs date c. 2000-1500 B.C. The logo-syllabo-alphabetic writing consists of c. 700 signs of which c. 100 have a phonetic value (types C, CC, CCC). The material has been taken from A. H. Gardiner's *Egyptian Grammar*, 2nd edition (London, 1950), *passim*. Only 26 alphabetic signs (60% of the sample) were included as variables in the classification. In examining the groups which, unlike the preceding samples, do not seem to follow any phonetic principle, it must be borne in mind that we do not know the vowels inherent in the consonant signs which were written alone.

TABLE 9

Relative frequencies

Symbol	%
N	10.13
T	9.97
R	5.48
F	3.79
M	3.79
W	3.63
İ	3.28
Ś	2.63
C	1.99
K	1.80
P	1.64
D	1.47
Ḥ	1.47
3	1.39
Ḫ	1.06
Y_2	1.01
Ḏ	0.98
S	0.86
B	0.78
Ṯ	0.75
Y_1	0.71
Q	0.50
Š	0.46
H	0.40
G	0.26
Ḥ	0.18

TABLE 10

Classification according to the text following the signs

I	II	III	IV	V
Q	Ṯ	C	W	İ
Ḥ	Ś	N	Ḫ	Y_2
F	P	R	B	Ḫ
G	S	Y_1	Š	Ḏ
M	T		3	
	H		D	
			K	

Classification according to the text preceding the signs

I	II	III	IV	V
F	W	Š	Ś	P
H	M	Ḥ	R	S
Ṯ	Q	3	C	İ
Ḏ	N	B	G	Y_2
K	Ḫ	Y_1	D	
T				
Ḥ				

Classification according to the text on both sides of the signs

I	II	III	IV	V
F	M	P	Š	Ś
H	W	Ṯ	3	C
K	Q	S	Ḫ	Ḫ
Ḥ	N	İ	B	D
T	R	Y_2	Y_1	
G			Ḏ	

5. *Sumerian Cuneiform*

Sumerian cuneiform of c. 2500 and 2100 B.C. is a logo-syllabic writing consisting of c. 700 signs of which c. 70 are constantly used as phonetic signs to indicate grammatical elements. We have coded a selection of Old Sumerian Royal Inscriptions (Edmond Sollberger, *Corpus des inscriptions 'royales' présargoniques de Lagaš* [Genève, 1956], pp. 37-38, 50-53) and the cylinder inscription of Gudea (F. Thureau-Dangin, *Les cylindres de Gudea* (=*Textes cunéiformes du Musée du Louvre*, VIII) [Paris, 1925]). The classification was first done with half of the last mentioned text only (Gudea A), then with the whole material. We present both results.

TABLE 11

Absolute and relative frequencies of the Sumerian cuneiform

Whole material

Symbol	Freq.	%
A	499	3.79
AN	490	3.72
NI	368	2.79
MU	358	2.72
NA	334	2.53
KA	316	2.40
DA	278	2.11
BI	246	1.87
BA	237	1.80
RA	219	1.66
E	208	1.58
MA	187	1.42
NE	186	1.41
GA	181	1.37
IM	164	1.24
ŠÈ	158	1.20
SU	132	1.00
KE_4	127	0.96
TA	126	0.96
LA	117	0.89
SI	116	0.88
MI	116	0.88
GÁ	110	0.83
ŠI	106	0.80
ME	102	0.77
ÀM	101	0.77
ZI	95	0.72
ŠU	89	0.68

Gudea-A only

Symbol	Freq.	%
A	228	4.10
MU	226	4.07
AN	172	3.10
NI	161	2.90
NA	153	2.75
KA	121	2.18
MA	105	1.89
RA	100	1.80
BA	98	1.76
IM	88	1.58
NE	84	1.51
E	77	1.39
GA	76	1.37
BI	73	1.31
ŠÈ	73	1.31
DA	70	1.26
MI	61	1.10
ŠI	61	1.10
GÁ	60	1.08
TA	57	1.03
KE_4	54	0.94
ŠU	51	0.92
ME	51	0.92
Ù	48	0.86
ZU	48	0.86
ZI	48	0.86
SI	44	0.79
LA	41	0.74

Whole material

Symbol	Freq.	%
NAM	88	0.67
Ù	80	0.61
GE	78	0.59
ZU	76	0.58
AB	66	0.50
LÁ	64	0.49
TU	63	0.48
RU	58	0.44
NU	54	0.41
KU	53	0.40
TE	49	0.37
RE	48	0.36
IB	44	0.33
ḪÉ	41	0.31
KAM	35	0.27
TI	33	0.25
LU	25	0.19
SA	22	0.17
LI	21	0.16
DAM	20	0.15
GU	17	0.13
ŠA	16	0.12
BU	13	0.10
AL	11	0.08
IR	10	0.08
UM	9	0.07
LUM	8	0.06
LAM	7	0.05
ZÉ	6	0.05
AR	5	0.04
Others	6240	47.74
TOTAL	13176	

Gudea-A only

Symbol	Freq.	%
ÀM	40	0.72
NAM	38	0.68
SU	37	0.67
AB	32	0.58
NU	30	0.54
GE	28	0.50
IB	25	0.45
RU	24	0.43
LÁ	23	0.41
TU	22	0.40
KAM	22	0.40
RE	21	0.38
KU	14	0.25
TE	14	0.25
TI	13	0.23
DAM	12	0.22
ḪÉ	12	0.22
LI	10	0.18
LU	10	0.18
ŠA	8	0.14
BU	7	0.13
AL	7	0.13
SA	6	0.11
UM	6	0.11
IR	5	0.09
GU	4	0.07
ZÉ	4	0.07
LUM	3	0.05
LAM	2	0.04
AR	2	0.04
Others	2549	45.87
TOTAL	5557	

TABLE 12

Classification of the Sumerian cuneiform

Whole material

According to the text following the signs

I	II	III	IV
ŠÈ	IM	ŠI	TI
DA	LUM	SA	SI
DAM	LAM	ME	SU
ÀM	UM	ŠA	BU
AL	NU	NI	GU
E	Ù	KU	LI
BI	TU	ZI	MA
ZU	NAM	MI	TE
A	ZÉ	AR	IR
ŠU	ḪÉ	IB	
AN	GÁ	RE	
TA	MU	KE_4	
RA		AB	
KAM			
KA			
LÁ			
NE			
GE			
LU			
RU			
GA			
NA			
BA			
LA			

Gudea-A only

According to the text following the signs

I	II	III	IV
DAM	IM	NA	TE
AM	LAM	AB	GA
AR	LUM	RA	LA
ŠÈ	NU	SU	MU
GE	ŠA	RU	ME
E	ZI	LU	IB
TA	SA	TI	ḪÉ
AN	TU	SI	IR
ZU	UM	MI	ZÉ
ŠU	NAM	BA	
A	ŠI	NI	
BI	Ù	MA	
DA		RE	
AL			
KAM			
NE			
KA			
LÁ			
KE_4			
GU			
LI			
BU			
KU			
GÁ			

According to the text preceding the signs

I	II	III	IV
ḪÉ	KAM	MI	AL
SU	DA	IR	LUM
AN	BI	TI	TE
ŠU	ÀM	IB	DAM
IM	ŠÈ	LÁ	GE
MU	LAM	MA	NI
GU	NA	LI	GA
Ù	ZU	BU	RU
NU	TA	UM	SI
NAM	KE_4	TU	KU
ZI	RA	LA	AR
BA	AB	GÁ	
LU	RE		
A			
ŠI			
E			
NE			
ŠA			
ME			
SA			
KA			
ZÉ			

According to the text preceding the signs

I	II	III	IV
SU	TA	IB	MI
ḪÉ	ZU	IR	MA
GU	RU	TU	GE
ZÉ	NA	BU	KE_4
ŠU	DAM	ŠÈ	SI
IM	LUM	ŠA	GA
NU	KAM	LA	UM
AN	DA	E	GÁ
MU	AB	RE	AL
ZI	LAM	LU	LÁ
Ù	RA	TI	AR
NE	BI	KU	
NAM	NI		
KA	SA		
A	ÀM		
BA			
ŠI			
ME			
LI			
TE			

According to the text on both sides of the signs

I	II	III	IV
ŠU	DA	IR	TI
AN	ŠE	LI	MI
ḪÉ	BI	LA	GU
A	KAM	BU	MA
MU	DAM	TE	NI
E	ÀM	IB	UM
NAM	TA	LA	TU
BA	ZU	ZÉ	GÁ
ZI	RA	SI	KU
NE	LAM	GA	AR
LU	AL	ḪA	RE
NU	NA		
ŠI	RU		
IM	LUM		
Ù	KE_4		
KA	AB		
GE			
SA			
SU			
ME			
ŠA			

According to the text on both sides of the signs

I	II	III	IV
ŠU	DA	NU	E
SA	ZU	KAM	IB
KA	GA	LAM	GÁ
ŠA	DAM	LUM	TI
ZI	GE	IM	AR
SU	LU	LI	MI
TA	AN	ÀM	NI
ḪÉ	RA	A	ŠÈ
NE	LA	TE	MA
RU	NA	BU	IR
AB	TU	NAM	ḪA
SI	AL	GU	UM
KU	ŠI	BI	BA
KE	ME	Ù	
RE	MU	ZÉ	
LA			

TABLE 13

The distribution of the signs according to their vowel components in the classifications into four groups[a]

Classification according to the text following the signs

Sample	Group	Type of the sign: V	CA	CU	CO	CI	CE	AC	UC	OC	IC	EC	CVC	Rest	Total
Elamite	I	A/U	13	1	—	0	0	1	1	—	2	0	1	0	21
	II	0	0	6	—	0	0	4	3	—	2	1	1	1	18
	III	0	0	0	—	11	2	0	2	—	3	0	0	0	18
	IV	E/I	3	4	—	0	0	1	1	—	1	1	0	0	13
Linear B	I	A/E/U	5	4	2	1	3	—	—	—	—	—	—	—	18
	II	I	2	1	1	8	4	—	—	—	—	—	—	—	17
	III	0	4	2	7	2	2	—	—	—	—	—	—	—	17
	IV	A/A/O	5	1	3	0	4	—	—	—	—	—	—	—	16
Neo-Assyrian	I	E	0	12	—	0	0	1	1	—	1	0	1	1	18
	II	I/U	8	1	—	0	0	0	0	—	1	0	2	2	16
	III	0	1	0	—	8	2	1	0	—	1	1	1	1	16
	IV	A/U	3	0	—	1	1	1	1	—	0	0	2	2	13
Sumerian	I	E/A	9	4	—	1	3	3	0	—	0	0	2	0	24
	II	U	1	3	—	0	2	0	1	—	1	0	3	0	12
	III	0	2	1	—	4	3	2	0	—	1	0	0	0	13
	IV	0	1	3	—	3	1	0	0	—	1	0	0	0	9

[a] N.B. The vowels inherent in the signs of Egyptian hieroglyphic writing being unknown, this sample has not been included.

Classification according to the text preceding the signs

Sample	Group	Type of the sign: V	CA	CU	CO	CI	CE	AC	UC	OC	IC	EC	CVC	Rest	Total
Elamite	I	E/I	3	0	—	1	1	0	0	—	8	2	0	0	17
	II	A	3	2	—	5	0	7	0	—	0	0	2	0	19
	III	0	2	6	—	2	0	0	7	—	0	0	0	1	18
	IV	U	8	3	—	3	1	0	0	—	0	0	0	0	16
Linear B	I	U/A	4	0	1	3	1	—	—	—	—	—	—	—	11
	II	I	2	1	2	1	4	—	—	—	—	—	—	—	11
	III	O/E	1	1	1	2	3	—	—	—	—	—	—	—	10
	IV	0	2	0	5	0	1	—	—	—	—	—	—	—	8
Neo-Assyrian	I	0	8	2	—	2	0	1	0	—	0	1	2	3	19
	II	E/I	3	3	—	1	1	0	0	—	3	0	4	1	18
	III	A	2	0	—	4	0	2	0	—	0	0	0	2	11
	IV	U/U	0	7	—	2	2	0	2	—	0	0	0	0	15
Sumerian	I	U/A/E	4	6	—	2	4	1	0	—	1	0	1	0	22
	II	0	4	1	—	1	3	2	0	—	0	0	2	0	13
	III	0	4	2	—	3	0	0	1	—	2	0	0	0	12
	IV	0	1	2	—	2	2	2	0	—	0	0	2	0	11

Classification according to the text on both sides of the signs

Sample	Group	Type of the sign													
		V	CA	CU	CO	CI	CE	AC	UC	OC	IC	EC	CVC	Rest	Total
Elamite	I	A	8	0	—	0	0	0	0	—	0	0	0	0	9
	II	0	0	0	—	8	2	1	0	—	0	0	0	0	11
	III	U	0	3	—	1	0	0	0	—	6	0	0	0	11
	IV	0	2	2	—	1	0	1	3	—	0	0	0	0	9
Linear B	I	I/O/A	0	0	1[a]	5	3	—	—	—	—	—	—	—	12
	II	E/U	4	1	1	0	1	—	—	—	—	—	—	—	9
	III	0	2	1	3	1	2	—	—	—	—	—	—	—	9
	IV	0	3	0	4	0	3	—	—	—	—	—	—	—	10
Neo-Assyrian	I	E/I/U	0	11	—	0	1	0	0	—	3	0	1	0	19
	II	U	2	1	—	1	2	1	1	—	0	1	3	3	16
	III	0	8	1	—	3	0	1	0	—	0	0	2	2	17
	IV	A	1	0	—	5	0	1	1	—	0	0	0	1	10
Sumerian (whole material)	I	A/E/U	4	5	—	2	4	1	0	—	1	0	1	0	21
	II	0	4	2	—	1	2	3	0	—	0	0	4	0	16
	III	0	4	1	—	2	2	0	0	—	2	0	0	0	11
	IV	0	2	3	—	3	1	1	1		0	0	0	0	11

[a] IO

CONCLUSIONS

The classification of the Elamite and Neo-Assyrian signs has given results that can be considered excellent. It appears that the vowels, especially I/E and U, play a predominant part, and that the groups classified according to the text following the signs generally consist of CV signs and those classified according to the text preceding the signs consist of VC signs. In some cases the classification method also reveals the actual pronunciation of a sign with several values: thus it seems fairly certain that, e.g., the Neo-Assyrian sign JA/JI/JU is to be read, in most cases, JU and not JA as has been done hitherto; likewise apparently TÚ and not UD.

Of the classification of the hieroglyphs we can draw only the conclusion that the consonants do not affect the grouping. The egyptologists are invited to consider the above given groups with regard to the unascertainable inherent vowels.

In the sample of Linear B, only the group containing syllables with the vowel I/E has clearly come out. It is possible that the consonants have affected the grouping here. The results of the Sumerian sample do not make much sense.

Hence we can conclude that the method is very effective when VC signs and CV signs are equally well represented in the script analyzed and when it is possible to eliminate most of the logograms. It is, however, not wholly useless in less favourable circumstances as is shown by the Linear B script which contains open syllables only, and it should be worth while to try it, e.g., also in the decipherment of Linear A.

We have learnt that the method is not enough for the successful decipherment of an unknown script. But the very fact that it is possible to attain useful results by mechanical means shows that we are on the right way. This line should now be continued and new methods based on other criteria be developed alongside the one now described until the same combined will yield the solution.

APPENDIX

On the classification of the phonemes of a language

The classification of the phonemes of a language using as a criterion their ability to combine with the other phonemes in the speech chain was suggested already by E. Sapir. Thinking that our program might prove useful in this field, too, we have tentatively applied it to samples of modern Finnish (from Mika Waltari's *Kuun maisema*),[3] English (Agatha Christie's *Murder in Mesopotamia*)[3] and French (Albert Camus' *La peste*),[3] and two samples of Latin (Tacitus' *Annales*, I, 1-3 and III, 1-3), comprising 3200 to 5500 letters each. It appeared that the material was insufficient for an effective statistical analysis (the two samples of Latin did not give identical results) and hence for a finer classification, and that the inexactitude of the conventional orthography gives some trouble (cf. the good results of Finnish representing an almost phonetic transcription with the results of the other languages). But that the vowels were forthwith separated from the consonants in all the samples gives a promise that the results can be better if these shortcomings are annulled.

[3] From the first pages.

Classification of the letters in a sample of Finnish

Relative frequency of Finnish letters

Symbol	%
A	12.50
I	11.47
T	9.86
N	9.59
E	7.28
S	7.28
L	6.02
U	5.56
K	5.43
Ä	4.61
O	4.55
M	3.08
V	2.69
R	2.06
J	2.06
H	1.93
P	1.66
Y	1.53
D	0.46
Ö	0.27
G	0.10

Classification according to the text following the letters

2 groups

I	II
P	E
K	I
G	A
V	Ä
M	Y
D	U
J	O
T	Ö
N	
S	
H	
L	
R	

3 groups

I	II	III
P	E	L
K	I	R
J	A	S
V	U	Ö
G	Ä	
M	O	
T	Y	
D		
N		
H		

4 groups

I	II	III	IV
P	A	R	Y
K	E	L	Ö
V	I	S	H
J	U		
G	Ä		
M	O		
T			
D			
N			

5 groups

I	II	III	IV	V
V	I	L	Y	D
J	A	O	Ö	G
M	E	R	H	N
P	U			S
K	Ä			
T				

6 groups

I	II	III	IV	V	VI
V	I	L	Y	D	R
J	A	O	Ö	G	
M	E		H	N	
P	U			S	
K	Ä				
T					

7 groups

I	II	III	IV	V	VI	VII
V	A	L	Y	D	S	T
J	I	O	Ö	G	R	H
M	E		Ä	N		
P	U					
K						

8 groups

I	II	III	IV	V	VI	VII	VIII
V	A	L	Y	D	S	T	Ä
J	I	O	Ö	G	R	H	
M	E			N			
P	U						
K							

Classification according to the text on both sides of the letters

2 groups

I	II
V P	E
M L	U
J	I
N	Ä
K	A
H	Y
R	O
T	Ö
S	G
D	

3 groups

I	II	III
V	E	P
J	I	L
N	A	Ö
K	U	G
M	Ä	
H	O	
T	Y	
R		
S		
D		

4 groups

I	II	III	IV
V	E	L	S
J	I	P	N
K	A	Ö	D
M	U		G
H	Ä		
R	O		
T	Y		

5 groups

I	II	III	IV	V
J	E	M	S	Y
V	A	P	N	Ö
K	I	L	T	G
H	U		D	
R	Ä			
	O			

6 groups

I	II	III	IV	V	VI
J	E	M	S	Y	D
V	A	P	N	Ö	G
K	I	L	T		
H	U				
R	Ä				
	O				

7 groups

I	II	III	IV	V	VI	VII
J	I	M	S	Y	D	U
V	E	P	N	Ö	G	O
K	A	L	T			
H	Ä					
R						

8 groups

I	II	III	IV	V	VI	VII	VIII
J	I	M	S	Y	G	U	L
V	E	P	N	Ö		O	R
K	A	D	T				
H	Ä						

Classification of the letters in Sample I of Latin

Relative frequency of the letters

Symbol	%
I	11.77
E	10.75
U	8.84
A	8.50
T	7.61
S	7.03
R	6.62
M	6.04
N	5.94
O	5.77
C	3.99
L	3.86
P	3.62
D	2.56
B	1.91
G	1.30
V	1.19
Q	1.16
F	0.65
X	0.61
H	0.31

Classification according to the text following the letters

2 groups

I	II
C L P	O
R F S	E
V B M	A
H X N	U
T G	I
D	Q

3 groups

I	II	III
H L	O	S
V F	E	X
R G	U	B
T P	A	N
C M	I	Q
D		

Classification according to the text preceding the letters

2 groups

I	II
N X Q	A P
S F L	I
B G	O
C R	E
D T	U
M V	H

3 groups

I	II	III
N S	I	B
C R	E	V
X T	A	P
G	O	Q
D	U	H
M		L
F		

Classification according to the text on both sides of the letters

2 groups

I	II
C G	O
D S	A
V L	E
T M	I
F X	U
R P	Q
B	
H	
N	

3 groups

I	II	III
D	O	B
C	E	S
H	A	N
T	I	X
F	U	Q
V		P
G		
R		
L		
M		

Classification of the letters in Sample II of Latin

Relative frequency of the letters

Symbol	%
I	11.60
E	10.52
A	9.53
U	8.09
T	7.91
R	7.55
S	6.88
N	6.34
M	5.93
O	5.35
C	4.00
L	3.42
P	3.01
D	2.38
B	1.71
Q	1.53
G	1.44
V	1.26
F	1.03
X	0.31
H	0.22

Classification according to the text following the letters

2 groups

I	II
R F	O
L B	E
V M	A
D S	U
T X	I
H N	G
C P	Q

3 groups

I	II	III
V	O	T
H	E	B
D	A	M
R	U	F
L	I	P
X		Q
S		G
N		
C		

Classification according to the text preceding the letters

2 groups

I	II
D X	E
N C	I
B R	A
G L	O
M Q	U
S T	F
H P	V

3 groups

I	II	III
D	E	P
N	I	Q
M	A	V
B	O	T
S	U	F
G		
H		
X		
C		
R		
L		

Classification according to the text on both sides of the letters

2 groups

I	II
D X	A
H S	E
L C	O
R T	I
B G	U
N F	Q
M P	
V	

3 groups

I	II	III
D S	E	T
H V	O	C
R B	A	G
N	I	Q
L	U	F
X		P
M		

Classification of the letters in a sample of French

Relative frequency of the letters

Symbol	%
E	17.22
S	9.70
N	7.76
A	7.64
T	6.53
L	6.50
I	6.41
U	6.25
R	5.98
O	5.48
D	4.00
M	3.33
C	3.23
P	2.68
V	1.26
Q	1.14
B	1.08
F	1.08
G	0.83
H	0.65
Y	0.55
J	0.37
X	0.31

Classification according to the text following the letters

2 groups

I	II
H T	E
D R	O
L S	A
B M	U
G C	I
V F	Q
J Y	N
P	X

3 groups

I	II	III
H P	O	R
D S	E	Y
B C	A	N
L M	I	X
J	U	F
V	Q	
G		
T		

Classification according to the text preceding the letters

2 groups

I	II
V D	O
G B	A
Y T	E
N S	H
X C	I
R L	Q
J M	U
F	
P	

3 groups

I	II	III
V R	O	Y
T S	E	B
X L	A	N
G	H	M
C	Q	P
D		I
J		U
		F

Classification according to the text on both sides of the letters

2 groups

I	II
V S	O
G M	M
D T	E
R X	Q
L P	I
B C	H
J F	U
Y N	

3 groups

I	II	III
D X	O	Y
G C	A	B
V M	E	I
T	Q	P
S	H	N
R		F
L		U
J		

Classification of the letters in a sample of English

Relative frequency of the letters

Symbol	%
E	12.05
T	9.35
A	8.36
O	7.49
S	7.40
I	7.37
N	6.94
R	6.33
H	5.72
D	3.86
L	3.77
U	3.13
C	2.47
P	2.38
M	2.32
Y	2.03
F	1.97
G	1.94
W	1.57
B	1.39
V	1.09
K	0.78
J	0.23
Q	0.09
Z	0.03

Classification according to the text on both sides of the letters

2 groups

I				II	
M	W	N	F	I	Y
C	S	P	G	A	H
B	T	J	L	E	
V	K	Q		O	
R	Z	D		U	

3 groups

I			II	III	
M	F	S	A	D	Y
B	J		I	G	H
R	Z		E	K	
C	Q		O	T	
V	N		U	L	
W	P				

POSTSCRIPT

This paper was completed in October 1968. The code of the Indus script was broken in January 1969. The three preliminary reports published so far (July 1970) have been summarized by Asko Parpola in a paper entitled, "The Indus Script Decipherment: The Situation at the End of 1969", in the *Journal of Tamil Studies* (Madras), II, 1, (April, 1970), where further references may be found.

PRINTED IN THE NETHERLANDS
MOUTON & CO., PRINTERS
THE HAGUE