Computer Concordance of Proto-Indian Signs

I. Mahadevan and K. Visvanathan

INTRODUCTION

Almost any type of statistical-positional analysis of an undeciphered script presupposes the preparation of a concordance of sign-occurrences and the tabulation of positional frequencies of signs and sign-combinations. This paper describes an elementary computer technique devised at the Computer Centre, Fundamental Engineering Research Establishment, College of Engineering, Guindy, Madras, for the preparation of a concordance of the Proto-Indian signs from the coded texts. The technique is of general application in dealing with the characters of any unknown script, with the help of smaller computers. (The IBM 1620 Model II Electronic Digital Computer was used for the present programme.) A Soviet team has also used computational mathematical procedures (Kondratov, in Field and Baird 1965); similar procedures developed by Parpola (1971) and Parpola *et al* (1969) have also published.

DISTRIBUTION OF INSCRIBED OBJECTS

The data for analysis consist of inscriptions found on 2457 objects excavated from 11 Proto-Indian sites and 12 other objects reported earlier from Ur and other West Asian sites. The distribution of the inscribed objects according to sites is given in Table 1. The data include texts from 158 unpublished objects copied directly from the originals in the collections of the Archaeological Survey of India and the National Museum. These unpublished objects are from the later

excavations of Mohenjo-daro (113) and Harappa (41) as well as from Lothal (2) and Kalibangan (2).

TABLE 1. *Distribution of inscribed objects according to sites.*

Sites	*No. of inscribed objects*	*Percent*
MAJOR SITES		
Mohenjo-daro	1398	56.62
Harappa	891	36.09
Chanhudaro	67	2.71
Chanhudaro		
Lothal	54	2.19
Kalibangan	37	1.50
MINOR SITES		
Alamgirpur	3	0.40
Desalpar	2	
Jhukar	1	
Lohumjodaro	1	
Rojdi	2	
Ropar	1	
WEST ASIAN SITES	12	0.49
TOTAL	2469	100.00

TYPES OF INSCRIBED OBJECTS

The inscribed objects are mostly seals with the texts engraved on them or sealings (which are impressions in relief made from the seals or moulds). The other inscriptions are graffiti on miscellaneous objects. Table 2 indicates the distribution of the inscribed objects according to type.

LINES OF TEXT

The unit of coding is a line of text. The number of lines included in the analysis is 3013; 23 of these lines are wholly illegible but included in the coding as they are part of multi-linear texts, other portions of which are legible. It was not possible to consider a text as the unit, as we were not supposed to know at the time of coding, the real order of lines appearing on different sides of an inscribed object. The question whether different lines appearing on the same side as well as on different sides of an inscribed object have continuity of sequences or are to be considered as separate texts can be decided only on the basis of analysis of the results. The serial numbers given to the sides of an

inscribed object and the lines appearing thereon are therefore arbitrary and do not necessarily represent the real order (if any) of the sides or the lines. Single-line texts constitute the bulk of the material with 1967 lines (65.28 per cent). The maximum number of signs found in a single line of text is 15. However, most of the lines are much shorter, their average length being about 4 signs.

TABLE 2. *Distribution of inscribed objects according to type.*

Type	*No. of inscribed objects*	*Percent*
Stamp seals	1600	64.80
'Tiny' seals	257	10.41
Sealings	372	15.07
Graffiti on pottery	82	3.32
Copper tablets	113	4.58
Inscriptions on bronze weapons	8	0.32
Inscribed dice	20	0.81
Miscellaneous objects	7	0.28
Unknown type	10	0.41
	2469	100.00

DIRECTION OF WRITING OF LINES

One of the few well-established and generally accepted facts about the Proto-Indian script is that the general direction of the script is from right to left (and, consequently, from left to right in the engraving of the seals with reversed direction). Several investigators including Gadd and Smith (in Marshall, 1931), Langdon (in Marshall, 1931, 437), Hunter (1934, 37), Alekseev (in Field and Laird 1969), Lal (1966), and Parpola *et al.* (1969) have demonstrated this from a study of the external features of the writing and, more importantly, from the evidence of sequences. However, a distinction has to be made between the general direction of the script and the specific direction of writing of particular lines of text, which may be from right to left or left to right and also in *boustrophedon* style in cases of multi-linear inscriptions. For purposes of coding, a preliminary study of each line was made with the help of the criteria established by the earlier investigators and, in particular, using the sequences set up by Hunter. The coding of the direction was treated as provisional and one of the aims of the analysis was to make an independent determination of the direction of writing of each line in the light of sequences built up from the pair-wise frequencies.

SIGNS

A sign-list was prepared prior to the coding of the texts. The script contains *c.* 400 signs. (The provisional total in our sign-list is 409; but no emphasis is laid on this number as we cannot always distinguish, at the commencement of the analysis, between non-significant variants of the same signs and different signs.) We have enumerated separately not only the basic signs but also the modified signs (basic signs modified by some markings which themselves do not appear as separate signs) and the compound signs (two or more signs joined by ligature or as circumgraphs or otherwise) which seem to function as independent and integral signs in the inscriptions.

SIGN-OCCURRENCES

The total number of extant and legible sign-occurrences included in the analysis is 11,303 making the present sample the largest so far analysed. Comparative figures are: Hunter (1934): *c.* 3750; Soviet team (in Field & Laird): *c.* 6300; Finish team (Parpola *et al.*, 1969): 9147.

CODING OF DATA

The coding of each line of text consists of three parts:—

1. Line number (in six digits occupying columns 1 to 6 at the left)
2. Background data (in 14 digits occupying columns 7 to 20 in the middle)
3. Text (in maximum of 45 digits, i.e., 15 signs occupying columns 21 to 80 at the right with one blank space preceding each 3-digit numeric code for signs).

A specimen of two coded lines of text is given in Annexure I to indicate the general format of the coded input data. A brief description of the format is given below.

LINE NUMBER

Each line was given a serial number consisting of six digits. The first digit (from the left) signifies the site and the next three digits the serial number of the object (as far as possible, the same as the publication number). The last two digits represent the side number and the line number. It was found necessary to distinguish between single-line inscriptions (which alone could be used in the first instance to determine the terminal signs) and the 'first' line in multi-sided or multi-linear inscriptions. This was done in the coding by allotting the code number '0' for single sides and single lines and the number '1' for 'first' sides and 'first' lines in the respective columns. Thus 1.001.01 stands for the first line

of the only side of seal No. 1 from Mohenjo-daro (Marshall, 1931, III, pl. CIII, No. 1).

BACKGROUND DATA

One of the principal objectives of the study is to correlate the known archaeological data with the patterns of frequency and distribution of signs, sign-combinations and texts to facilitate the interpretation of the texts. The significant data coded for each line of text included (1) the site, (2) the location of the object within the site, (3) stratigraphy, (4) associated 'field symbols' found inscribed on the object (e.g., the 'unicorn'), (5) type of the object, (6) number of inscribed sides, (7) number of inscribed lines, (8) direction of writing of each line, and (9) length of each line. The analysis of the background data did not form part of the present programme (Mahadevan, in prep.).

CODING OF TEXTS

The signs were coded as a series of 3-digit numbers between 001 and 449. Each line of text was coded as a sequence of these 3-digit numbers and punched on a separate card (having 80 columns) along with the line number and the relevant background data. The transcription of the lines of text in code runs in every case from left to right irrespective of the original direction in the inscription (which however is coded in a separate column for each line as part of the background data). The code number '000' was used to indicate lost, mutilated or otherwise illegible portions of the inscriptions. This has been done so that the non-terminal extant signs at either ends of broken lines are not counted as terminal and the signs separated in the inscriptions by mutilated portions are kept apart. While punching the cards, each set of 3-digit numeric codes representing one sign was separated from the others by a blank space for easy readability.

PRELIMINARY SORTING OF INPUT DATA

The input data cards were initially run on the IBM 082 Sorter to arrange the cards in the increasing order of line-lengths. (The code number '000' was treated as a 'sign' for the purpose.) The deck of cards was now ready for being fed into the computer.

THE PROGRAMME

The programme used by us for preparing the sign-concordance is extremely simple. It amounts to nothing more than a simple set of instructions to read the cards, realign the data and duplicate. The programme consisted of the following instructions:—

(1) to read the input cards serially (only the line numbers and the texts were to be read for this programme)
(2) to transpose the 3-digit numbers representing the signs and re-arrange

them in such a way that each sign-occurrence (designated as the 'reference sign'), commencing from the left end of each line, was successively brought once to a fixed position ('frame') in the middle of the output card (columns 42-44).

(3) to duplicate each input card as many times as there are signs on the line (The code number '000' was also treated as a 'sign' for the present purpose).

The number of output cards equalled the total number of sign-occurrences (including the code number '000'). The computer took a little over three hours to execute the programme. One of the typical programmes used by us (for 12-sign lines) is reproduced *in extenso* as Annexure II. A sample of the output data as it appeared in our initial listing is shown in Annexure III. The characteristic 'lozenge' formations of the lines of text due to successive transposition of signs may be noted.

FURTHER PROCESSING OF DATA

The output cards were fed into the Sorter and first arranged in the ascending numerical order of the line numbers. Then the cards were sorted in the ascending numerical order of the reference sign numbers occurring within the 'frame'. Finally the output cards in respect of each sign were further sorted in the ascending numerical order of the immediately preceding sign to the left. The results were printed out by the IBM 407 off-line Printer (which was suitably wired to introduce a blank space in between each group of numbers for easy readability). The entire procedure was repeated with respect to the sign immediately succeeding the reference sign to the right. These two concordances of all the sign-occurrences with their immediately preceding and succeeding signs sorted, represent the results of the present programme. Two specimen listings are appended (Annexures IV & V) to give an idea of the arrangement of the lists.

SOME PRELIMINARY RESULTS

It is beyond the scope of the present paper to deal with the analysis of the results of the programme. However, the basic data regarding the frequencies and distribution of single signs and pair-wise combinations which can be computed from the sign-concordance are briefly summarized below.

FREQUENCIES OF SIGNS

The Table 3 summarizes the total frequencies of individual signs in the sign-list. The figures show clearly how relatively few of the signs take most of the functional load in the script. At one end we have two signs (which appear to be suffixes from their positional characteristics) according for as much as 15.71 per cent of the total sign-occurrences. At the other end we have as many as 131 signs (32.03 per cent of the total) occurring only once. More than three-

fourths of the occurrences are accounted for by 53 signs which appear to constitute the core of the script, so far as we know from the available texts.

TABLE 3. *Frequencies of signs*

Frequency range of sign-occurrences	*No. of signs*	*Total sign-occurrences*	*Per cent*
100 or more	1	1190	10.53
500 to 999	1	585	5.18
100 to 499	27	5107	45.18
50 to 99	24	1790	15.84
25 to 49	30	1042	9.22
10 to 24	57	859	7.60
2 to 9	138	599	5.30
Only once	131	131	1.15
TOTAL	409	11303	100.00

POSITIONAL FREQUENCIES OF SIGNS

Computation of the positional frequencies of signs can give us important information as to the probable nature of the language of the inscriptions and of the linguistic units represented by the signs and their combinations. Table 4 summarizes the positional frequencies of signs (with respect to lines of text as units):

TABLE 4. *Positional frequencies of signs*

Position (w.r.t. line)	*No. of signs*	*Total frequency*
Solus	73	142
Initial	277	2558
Medial	278	6051
Final	164	2552
TOTAL	409*	11303

* A Sign can occur in more than one position.

PAIR-WISE FREQUENCIES

Table 5 summarizes the frequency ranges of pair-wise combinations:

TABLE 5. *Frequencies of pair-wise combinations*

Frequency range	*No. of pair-wise combinations*	*Total frequency*	*Per cent*
100 times or more	4	611	7.41
50-99 times	11	764	9.26
25-45 times	37	1256	15.22
10-24 times	91	1284	15.56
2- 9 times	805	2692	32.63
Only once	1643	1643	19.92
TOTAL	2591	8250	100.00

From the statistical point of view, only those combinations which occur more than once are significant. 948 pair-wise combinations occurring more than once have been recorded. Not all of them, of course, are necessarily 'true' combinations. Analysis of pair-wise frequencies has led to the separation of 'real' pairs (corresponding to linguistic entities) from random pairs. It is possible to achieve 'word-division' by pursuing this line of enquiry, without having to make any *a priori* assumptions as to the nature of the language of the inscriptions.

LONGER SIGN-COMBINATIONS

The number and frequency of sign-combinations decrease rapidly with length. There does not seem to be any combination longer than 5 signs in length. However, combinations of combinations (including repetition of whole lines) occurring more than once have been recorded upto a maximum length of 11 signs within a line.

ACKNOWLEDGEMENTS

The work reported here briefly is part of the Project for the Study of the Indus Script, made possible by the award of a Jawaharlal Nehru Fellowship to I. Mahadevan. The authors are grateful to the authorities of the Computer Centre, Fundamental Engineering Research Establishment, College of Engineering, Guindy, Madras, for making available computer time and providing all other facilities for conducting the study. We are grateful to the Director General of Archaeology for providing relevant photographs. We also thank Miss R. Kamala and Miss T. Rajeswari for their able secretarial assistance including coding of data, punching the cards and tabulation of results.

ANNEXURE 1
A SPECIMEN OF CODED LINES OF TEXT (INPUT DATA)

1-6	7-8	9-11	12	13	14-16	17-18	19-20	21-80				
100102	23	-03	1	3	501	05	04	000	125	131	008	147
100101	23	-03	1	1	501	01	01	346				

EXPLANATION

Col. No	*Data*	*Code*	*Key*
1	Site	1	Mohenjodaro (Marshall)
2- 4	Object Number	001	Marshall (1931) seal No. 1
5	Side Number	0	Only side
6	Line Number	1, 2	Lines 1 and 2
7- 8	Location	23	Lower city — HR Area
9-11	Level	-03	3′ below surface
12	Type	1	Stamp seal with engraved signs
13	Direction of Writing	1	Right to left (as on impression)
14-16	Field Symbol	501	Animal(unicorn?)facing right (as on impression)
17-18	Number of Positions in the Line	05	5 positions (including the Code number '000');
		01	(one position in the second line)
19-20	Length of line	04	4 signs (extant) in line 1
		01	1 sign in line 2
21-80	Coded text	000	Broken portion of line 1
		Other 3-digit Numbers	Each number represents one sign in our Sign List. A blank space precedes each 3-digit Sign Number.

ANNEXURE II

COMPUTER ANALYSIS OF PROTO-INDIAN TEXTS
PROGRAMME TO FIND PAIR-WISE FREQUENCIES
FOR 12-SIGN LINES

```
      DIMENSION A(12)
      READ 200, N2
200   FORMAT (12)
999   READ     100, I1, I2, (A (I), I = 1, N2)
      PUNCH      1, I1, I2, (A (I), I = 1, N2)
      PUNCH      2, I1, I2, (A (I), I = 1, N2)
      PUNCH      3, I1, I2, (A (I), I = 1, N2)
      PUNCH      4, I1, I2, (A (I), I = 1, N2)
      PUNCH      5, I1, I2, (A (I), I = 1, N2)
      PUNCH      6, I1, I2, (A (I), I = 1, N2)
      PUNCH      7, I1, I2, (A (I), I = 1, N2)
      PUNCH      8, I1, I2, (A (I), I = 1, N2)
      PUNCH      9, I1, I2, (A (I), I = 1, N2)
      PUNCH     10, I1, I2, (A (I), I = 1, N2)
      PUNCH     11, I1, I2, (A (I), I = 1, N2)
      PUNCH     12, I1, I2, (A (I), I = 1, N2)
      GO TO     999
100   FORMAT    (2I3, 15X, 12 (A3, 1X) )
  1   FORMAT    (1X, 2I3, 1X, 33X, 12A3)
  2   FORMAT    (1X, 2I3, 1X, 30X, 12A3)
  3   FORMAT    (1X, 2I3, 1X, 27X, 12A3)
  4   FORMAT    (1X, 2I3, 1X, 24X, 12A3)
  5   FORMAT    (1X, 2I3, 1X, 21X, 12A3)
  6   FORMAT    (1X, 2I3, 1X, 18X, 12A3)
  7   FORMAT    (1X, 2I3, 1X, 15X, 12A3)
  8   FORMAT    (1X, 2I3, 1X, 12X, 12A3)
  9   FORMAT    (1X, 2I3, 1X,  9X, 12A3)
 10   FORMAT    (1X, 2I3, 1X,  6X, 12A3)
 11   FORMAT    (1X, 2I3, 1X,  3X, 12A3)
 12   FORMAT    (1X, 2I3, 1X, 12A3)
      END
```

ANNEXURE III

SPECIMEN OF COMPUTER OUTPUT

Each line registers one occurrence of reference sign in the central columns of the lozenge formation reference—seals 1, 5, 14 and 19 in Vats, II, 1940.

```
400100                                                  367305384162386229068147382193
400100                                                   36730538416238622906814738219 3
400100                                                    3673053841623862290681478219 3
400100                                                     367305384162386229068147382193
400100                                                      36730538416238622906814738219 3
400100                                                       3673053841628862290614738219 3
400100                                                        367305384162386229068147382193
400100                                                         36730538416238622906814738219 3
400100                                                          3673053841623862290681473821 93
400100                                                           367305384162386229068147382193
400500                                                  265384127125118069097082305147
400500                                                   26538412712511806909709230514 7
400500                                                    2653841271251180690970923051 47
400500                                                     265384127125118069097092305147
400500                                                      26538412712511806909709230514 7
400500                                                       2653841271251180690970923051 47
400500                                                        265384127125118069097092305137
400500                                                         26538412712511806909709230514 7
400500                                                          2653841271251180690970923051 47
400500                                                           265384127125118069097092305147
401400                                                  254254150072099190118406280147
401400                                                   25425415007209919011840628014 7
401400                                                    2542541500720991901184062801 47
401400                                                     254254150072099190118406280147
401400                                                      25425415007209919011840628014 7
401400                                                       2542541500720991901184062801 47
401400                                                        254254150072099190118408280147
401400                                                         25425415007209919011840628014 7
401400                                                          2542541500720991901184062301 47
401400                                                           254254150072099190118406280147
401900                                                  176385069383118097092305066147
401900                                                   17638506938311809709230506614 7
401900                                                    1763850693831180970923050661 47
401900                                                     176385069383118097092305066147
401900                                                      17638506938311809709230506614 7
401900                                                       1763850693831180970923050661 47
401900                                                        176385069383118097092305066147
401900                                                         17638506938311809709230506614 7
401900                                                          1763850693831180970923050661 47
401900                                                           176385069383118097092305066147
```

ANNEXURE IV

SPECIMEN LIST OF CONCORDANCE OF SIGNS WITH PRECEDING SIGNS SORTED.

REFERENCE SIGN 203 (No. 165 IN GADD-SMITH SIGN LIST)

LINE NUMBER	PRECEDING SIGNS	REF. SIGN
281510		203 069 000
203900		203 069 013
432310		203 069 013 000
476110		203 069 013 000
476210		203 069 013 000
476310		203 069 013 000
476410		203 069 013 000
476510		203 069 013 000
476610		203 069 013 000
434110		203 069 013 194
246100		203 069 013 383 057 025 254 254
272710		203 072 013 194
272910		203 072 013 194
970700		203 099 382 147 147
531100	099 129 131 147 352 068	203 069 013
140100	343 382 069 125	203 069 008 147
498600	343 382 069 125	203 069 008 147
410500	127 129 089 147	203 069 013
543910	000 386 019 196	203 000
462800		381 203 069 013
488100		381 203 069 013
217000	000	382 203 069 013
429700	229 068 147	382 203 069 013
401700	153 275 147	382 203 069 013
550200	383 275 147	382 203 069 013
202300	308	382 203 069 013
428200	265 384	382 203 069 013
143600	343 384	382 203 069 013
301000	386 254 254 188 384 129 162 386	203 215
105100		394 203 216 384 074

ANNEXURE V

SPECIMEN LIST OF CONCORDANCE OF SIGNS WITH SUCCEEDING SIGNS SORTED.

REFERENCE SIGN 203 (No. 165 IN GADD-SMITH SIGN-LIST)

LINE NUMBER	REF SIGN	SUCCEEDING SIGNS
543910	203	000
281510	203	069 000
140100	203	069 008 147
498600	203	069 008 147
203900	203	069 013
531100	203	069 013
410500	203	069 013
462800	203	069 013
488100	203	069 013
217000	203	069 013
429700	203	069 013
401700	203	069 013
550200	203	069 013
202300	203	069 013
428200	203	069 013
143600	203	069 013
432310	203	069 013 000
476110	203	069 013 000
476210	203	069 013 000
476310	203	069 013 000
476410	203	069 013 000
476510	203	069 013 000
476610	203	069 013 000
427610	203	069 013 194
434110	203	069 013 194
246100	203	069 013 383 057 025 254 254
272710	203	072 013 194
272910	203	072 013 194
970700	203	099 382 147 147
301000	203	215
105100	203	216 384 074

BIBLIOGRAPHY

FIELD, H. and LAIRD, E. M. (Eds.) 1969. *Soviet Studies on Harappan Script* (Florida).

KOSKENNIEMI, S., PARPOLA, A. and S. 1970. A method to classify characters of unknown ancient scripts, *Linguistics*, 61, 65-91.

LAL, B. B. 1966. The direction of writing of the Harappan Script, *Antiquity*, 40, 52-55

MARSHALL, J. (ed.) 1931. *Mohenjodaro and the Indus Civilization*, 3 vols. (London).

PARPOLA, A., KOSKENNIEMI, S., PARPOLA, S. and A. 1969. *Decipherment of the Proto-Dravidian inscriptions of the Indus Civilization* (Copenhagen).

PARPOLA, A. 1971. Computer techniques in the study of the Indus Script, *Kadmos*, 10(1), 10-15.

VATS, M. S. 1940. *Excavations at Harappa*, 2 vols. (Calcutta).