A STATISTICAL APPROACH FOR PATTERN SEARCH IN INDUS WRITING[1]

Nisha Yadav[@], M N Vahia[@], Iravatham Mahadevan[#], H Joglekar[$]

[@] *Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005*
[#] *Indus Research Centre, Roja Muthiah Research Library, Chennai*
[$] *Oracle, Hyderabad*

Abstract

We search for potentially grammatical patterns in the Indus writing[2] based on the concordance of Mahadevan (1977). We make no assumptions about its structure or meaning. We *only* attempt to check if the Indus writing is meaningfully structured with specific rules to code useful information. In order to avoid possible errors in interpretation due to incompletely read or multiple copies of a piece of writing, and other possible sources of error, we create an Extended Basic Unique Data Set (EBUDS) based on the original electronic concordance of Mahadevan (1977). EBUDS consists of completely read single line texts on any side of the writing material. We exclude multi-lined text or partially read text. This gives us a set of 1548 lines of data consisting of 7000 signs. We

[1] Address for correspondence: y_nisha@tifr.res.in

[2] We use the word 'writing' in a generalised sense of the word in that it consists of information confined to a more durable medium than memory. While we do not imply linguistic writing, we do feel that any culture spread over a million square kilometres with such uniformity and a high degree of standardisation could not have indulged in random scribbling when they marked their writing material. We therefore feel that understanding their rules of writing is vital to understanding the culture. In this paper we do not comment on whether this is writing with sound or vowel writing or proto-writing of simple information. We leave that discussion to a later period. Similarly, when we use the words like phrases etc., we imply that a phrase is an information containing unit of one or more signs which do not appear without some form of prefixes or suffixes. However, we would like to draw attention to the compilation of Mahadevan (2002) on comparison of Indus writing with Aryan and Dravidian languages.

show that the ordering of the signs in the writing is much more significant than random association. The unit length of information is 2, 3 or 4 signs at a time. We then study the most frequently occurring two, three and four sign combinations in EBUDS. We find that in many cases, most common two-sign combinations are also parts of most common three-sign combinations, which in turn also appear in four-sign combinations. However, in texts with just 2, 3 or 4 signs we do not often find these frequent two, three or four sign combinations. We therefore conclude that while the information is given in units of two, three or four signs, these are more like phrases where an additional sign is required to complete the grammatical structure.

1. Introduction

Indus writing has been subjected to intense speculation and work by a variety of highly capable researchers (for recent reviews see Mahadevan, 2002, Subbarayappa, 1997, Parpola, 2005). However, there is no clear consensus on either the nature of writing or its possible content (for two extreme views see Farmer, Sproat and Witzel, 2004 and Fairservis, 1983), though some internally consistent studies do hold promise (e.g. Mahadevan, 2007). For a summary of various attempts see Possehl (1996). We therefore attempt to understand the writing *without any presupposition* to its content, language or nature. In the present study, we attempt to understand the *rules of writing*.

2. Data set

We use the concordance of Mahadevan (1977), henceforth referred to as M77, as the basic data set on which we use analytical, mathematical and computational tools to attempt to understand the rules of writing of the Indus script. We use the original electronic data set of the same.

Excluded data

M77 records 417 unique signs[3] in 3573 lines of 2906 texts. For the present study, we remove those texts that can have potentially ambiguous reading. We create an Extended Basic Unique Data Set (EBUDS). It removes all texts containing lost, damaged or illegible passages marked by the sign  and doubtfully read signs marked by asterisk. All texts from multi-lined sides are also removed. This is to avoid possible overflow in writing of information from one line of text to next. However, we assume that in multisided writing, the text on each side is independent of text on other side(s)[4]. It retains texts from those sides of multisided objects which have only one line of text. Texts appearing more than once are taken only once.

The unit of textual analysis for the study of distributional statistics is a line of text. There are two reasons why it is not possible to consider the whole text on a single side as a unit for this purpose. Firstly, there is no way of knowing beforehand whether different lines of an inscription appearing on the same object or even on the same side have continuity of sequence or to be regarded as separate texts. Secondly, it is not possible to ascertain beforehand the real order (if any) of the lines of text appearing on the same side (Mahadevan, 1977, p. 10).

Statistics

Based on this we read the texts and we get the following statistics.

Number of texts occurring only once: 1303
Number of texts occurring more than once: 245

[3] The serial number of the signs used in this paper is similar to that given by Mahadevan in his concordance (1977).

[4] An earlier Basic Unique Data Set (BUDS) did not consider multisided objects. However, this significantly affected the data set available for analysis. The analysis reported here was also done on BUDS which gave identical results with reduced statistical significance. These results are given in Yadav and Vahia (2007).

Therefore, Total number of independent texts: 1548

In EBUDS, 40 signs out of 417 present in the Sign List of Mahadevan do not make their appearance. Out of these removed 40 signs, one sign (sign number 374) appears 9 times, one sign (sign number: 237) appears 8 times, two signs (sign numbers: 282, 390) appear 3 times, three signs (sign numbers: 324, 376, 378) appear twice and thirty-three signs appear only once in M77.

Rules of reading the strings of script

We make no assumption about the content or meaning of Indus writing. M77 shows that the script is written right to left though this has no effect on our analysis. However, as a convention followed in the present paper, the texts depicted by pictures are to be read from right to left, whereas the texts represented by just strings of sign numbers are to be read from left to right. As our first emphasis is to attempt to *write in the script rather than read* we search for rules of writing without assigning meanings or interpretations. In the present work we only attempt to derive the grammar of writing in terms of search for patterns of *sequencing* of writing. We make no attempt to assign meanings to any of the signs.

We do not take into account the variation due to archaeological context of sites, stratigraphy and the type of objects on which the texts are inscribed.

3.1 Analysis 1: Preliminary analysis[5]

We studied the frequency distribution of the signs and sign pairs to check the consistency of our database with that of

[5] The analysis is done using computer programs written in PERL (Practical Extraction and Report Language) and shell script to extract required information from the digitized Indus data. The results were analyzed using Excel software.

Mahadevan (M77). Frequency distribution of the signs is given in Table 1.

The comparative statistics of sign occurrences is also given in Table 1. In the first 4 columns we reproduce the data from table 11.10 of M77. In EBUDS, the total sign occurrences are small and so they cannot be compared directly with M77. Hence, we study the frequency of n -signs where n is defined by column 2 of table 1. It should be noted that the signs that make up the frequent set are exactly the same in M77 and EBUDS for the signs that form the first 4 rows. In columns 5 to 7 we show that the most frequent sign in our data set accounts for 10.21% of our data set compared to 10.43 % in M77. Similarly, the second most frequent sign accounts for 5.39 % of our data set compared to 4.85 % in M77 and so on. Hence, comparing data from column 4 and 7, we see that that the frequency distribution of common signs in both the data sets are comparable and hence our reduced data set is a true representation of the data of M77, with a reduction in total sign occurrences, but not in the percentage of total sign occurrences. In table 2 we perform a similar test for sign pairs. The data of M77 is taken from table 11.12 of the concordance for sign pairs.

Table 1: Frequency distribution of Signs

Frequency range in M77	In M77			Present Work (EBUDS)		
	No. of signs	Total sign occurrences	Total sign occurrences (in percent)	No. of signs	Total sign occurrences	Total sign occurrences (in percent)
>1000	1	1395	10.43	1	715	10.21
999-500	1	649	4.85	1	377	5.39
499-100	31	6344	47.44	31	3230	46.14
99-50	34	2381	17.81	34	1243	17.76
49-10	86	1833	13.71	86	975	13.93
9-2	152	658	4.92	152	388	5.54
1	112	112	0.84	72	72	1.03
0	0	-	-	40	-	-
Total	417	13372	100.00	417	7000	100.00

Table 2: Frequency distribution of Sign pairs

Frequency Range in M77	In M77			Present Work (EBUDS)		
	No. of pair wise comb-nations	Total frequency	Total frequency (in percent)	No. of pair wise combinations	Total frequency	Total frequency (in percent)
> 100	6	949	9.69	6	472	8.66
99-50	13	898	9.17	13	438	8.03
49-25	38	1307	13.34	38	672	12.33
24-10	135	1953	19.93	135	1042	19.11
9-2	873	2998	30.60	873	1715	31.46
Only once	1693	1693	17.00	1113	1113	20.41
Total	2758	9798	100.00	2178	5452	100.00

Conclusion from Analysis 1:

The frequency distribution of the signs and sign pairs is consistent with M77. The result of our analysis shows that the manner of choosing the data set has not changed the pattern of occurrence of various signs in the data set and the results are consistent with the analysis of M77. We therefore proceed to the next step of analysis.

3.2 Analysis 2: Check against random order

In order to ascertain whether the pairing of signs in the Indus inscriptions or texts is significant or not, we adopt a statistical procedure. The 377 signs studied here appear in 1548 texts, giving a total sign set of 7000 signs. We take a random number generator and reshuffle them such that the total number of occurrences of each sign is identical to the genuine Indus data set (EBUDS). We then break this string of 7000 signs in such a way that they mimic the genuine Indus data set in terms of number of text lines with a specific number of signs. Hence, this reshuffling provides us with a randomised data set for Indus writing. In order to ensure that the relative frequency of occurrence of different signs is the same in randomised data set, we simply randomise the sequences rather than create an artificial set of texts. This method therefore gives an accurate comparative randomised data set to check the units of information in the genuine Indus data set without changing the

occurrence rate of different signs. We make 10 such random data sets and compare the frequency of most frequent combinations[6] of two- signs, three-signs, four-signs etc. in the randomised data set with the genuine Indus data set. If the Indus writing was simple pattern writing without information content, the sequence of occurrence of these signs would have a distribution comparable to random occurrence of different signs in the neighbourhood of another sign and the frequency of sign combinations of two-signs, three-signs etc. would be decided only by the statistics of absolute frequency of occurrence of a sign. The frequency with which a two or three sign combination is found in EBUDS data compared to random data indicates the level of significance of the writing. The observations are given in table 3 and figure 1.

Table 3: Result of Analysis 2

No. of signs in the sign combination	Frequency of most frequent sign combination											
	Random Data set											EBUDS
	1	2	3	4	5	6	7	8	9	10	Mean	
2	60	54	62	51	57	56	63	66	58	56	58.3	168
3	5	3	3	4	3	5	7	5	5	3	4.3	34
4	1	1	1	2	1	1	2	2	1	1	1.3	16
5	1	1	1	1	1	1	1	1	1	1	1	4
6	1	1	1	1	1	1	1	1	1	1	1	2

Table 3 above gives the result of Analysis 2. Here, the first column gives the number of signs in the sign combination. It is 2 for combination of two-signs, 3 for combination of three-signs etc. Columns 2 to 11 give the frequency of most frequently occurring sign combination of 2, 3, 4, 5 and 6 signs for the 10 random data sets. Thus, for the first random data set, the frequency of most frequently occurring two-sign combination is 60, the frequency of most frequently occurring three-sign combination is 5, the frequency of most frequently occurring

[6] We do not look at the information content of singlets as complete unit of information. This is because by its very nature EBUDS has unique data set. Hence if a singlet comes more than once, its subsequent appearances will not be seen by the data set. Also, the nature of randomisation applied here does not permit us to test for the possibility whether singlets contain complete information.

four-sign combination is 1, the frequency of most frequently occurring five-sign combination is 1 and the frequency of most frequently occurring six-sign combination is 1. Similarly, for second random data set, the frequency of most frequently occurring two-sign combination is 54, the frequency of most frequently occurring three-sign combination is 3, the frequency of most frequently occurring four-sign combination is 1, the frequency of most frequently occurring five-sign combination is 1 and the frequency of most frequently occurring six-sign combination is 1 and so on for other random data sets. Column 12 gives the mean of the frequencies of the most frequently occurring sign combinations of lengths 2, 3, 4, 5 and 6 of all the ten random data sets and the last column gives the frequencies of most frequently occurring sign combination of lengths 2, 3, 4, 5 and 6 for genuine Indus data set for comparison.

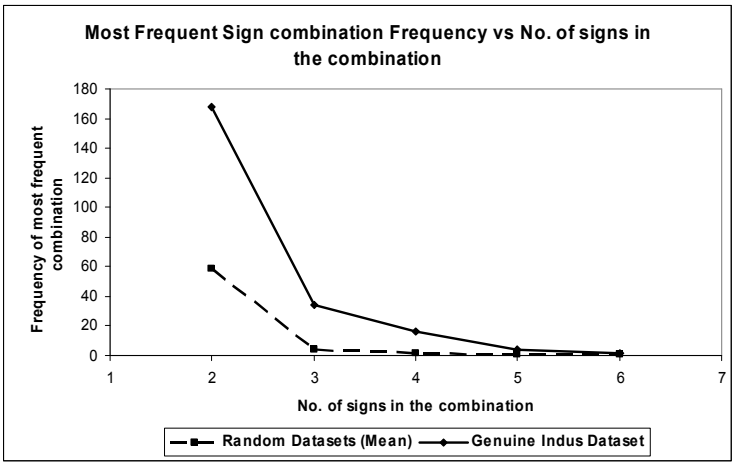

Fig. 1: A plot of the frequency of most frequently occurring sign combination of lengths 2, 3, 4, 5 and 6 versus the number of signs in the combination for genuine Indus database (bold line) and ten random databases (dashed line). By Genuine Indus data set we mean EBUDS.

Conclusion from Analysis 2:

The analysis shows that there is a significant difference in the frequency of the most frequent sign combinations of different sizes obtained from the random data sets in comparison

to that of EBUDS. The frequency of occurrence of the sign combinations of length two, three and four, in case of the ten random data sets, is considerably low in comparison to that obtained in case of EBUDS. Hence, it is justifiable to state that Indus texts followed certain rules and thereby meant something significant and meaningful.

3.3 Analysis 3: Positional analysis of most frequent two-sign, three-sign, four-sign combinations

In view of the importance of two-sign, three-sign and four-sign combinations from Analysis 2, we do a positional analysis of most frequent two-sign, three-sign, four-sign combinations present in EBUDS. We check whether these most frequent two-sign, three-sign, four-sign combinations tend to appear at fixed positions (i.e. left-end, middle or right-end) of the texts or they tend to appear as independent (or solo) texts. The positional analysis results for two-sign, three-sign and four-sign combinations are tabulated in tables 4, 5 and 6 respectively.

Table 4: Positional analysis of Two-sign Combinations

Two-sign Combination		Frequency	Solo (%)	Left (%)	Middle (%)	Right (%)
99	267	168	0.60	1.79	11.90	85.71
89	336	75	0.00	0.00	89.33	10.67
176	342	59	0.00	96.61	3.39	0.00
342	8	58	1.72	72.41	25.86	0.00
99	391	56	0.00	0.00	8.93	91.07
342	347	56	0.00	89.29	10.71	0.00
1	342	48	0.00	89.58	10.42	0.00
123	293	40	0.00	0.00	0.00	100.00
59	87	39	0.00	0.00	79.49	20.51
342	48	38	2.63	52.63	28.95	15.79
59	171	36	0.00	0.00	80.56	19.44
162	249	34	0.00	0.00	85.29	14.71
211	89	34	0.00	91.18	8.82	0.00
245	245	33	0.00	60.61	21.21	18.18
211	59	31	0.00	90.32	9.68	0.00
67	65	27	0.00	0.00	74.07	25.93
130	51	27	0.00	7.41	70.37	22.22
67	99	26	0.00	0.00	100.00	0.00
342	162	25	4.00	84.00	12.00	0.00
343	123	25	0.00	0.00	100.00	0.00

Table 4 gives the positional distribution of most frequent two-sign combinations (frequency of 25 or more). Several points immediately become clear.

1) All the two- sign combinations have a preferred location of occurrence (i.e. left-end, middle or right-end).
2) Their percentage of occurrence as independent texts (i.e. solo) is very small.
3) The most frequent two-sign combination (267, 99) whose total frequency of occurrence is 168 appears most frequently at the right-end of the texts, its percentage of occurrence at the right-end of texts being 85.71. Similarly, the two-sign combination (336, 89) whose total frequency of occurrence is 75, appears most frequently in the middle of the texts, its percentage of occurrence in the middle of texts being 89.33.

Table 5: Positional analysis of Three-sign Combinations

Three-sign Combination	Frequency	Solo (%)	Left (%)	Middle (%)	Right (%)
211 89 336	34	2.94	88.24	5.88	2.94
343 123 293	25	0.00	0.00	0.00	100.00
342 162 249	24	4.17	83.33	8.33	4.17
342 169 249	20	5.00	70.00	20.00	5.00
342 8 171	19	5.26	73.68	5.26	15.79
149 130 51	19	0.00	0.00	78.95	21.05
59 87 99	16	0.00	0.00	100.00	0.00
342 87 403	16	6.25	81.25	6.25	6.25
342 149 130	16	0.00	75.00	25.00	0.00
67 99 267	14	0.00	0.00	7.14	92.86
87 99 267	14	0.00	0.00	21.43	78.57
89 336 72	14	0.00	0.00	85.71	14.29
65 99 267	12	0.00	0.00	8.33	91.67
342 244 67	12	8.33	66.67	8.33	16.67
15 389 178	11	9.09	72.73	0.00	18.18
59 171 53	10	0.00	0.00	60.00	40.00
245 245 25	10	10.00	90.00	0.00	0.00

Table 5 gives the positional distribution of most frequent three-sign combinations (frequency 10 or more). The following points arise from this table:

1) The three-sign combinations tend to occur in specific positions in the text (i.e. left-end, middle or right-end).
2) Their percentage of occurrence as independent texts (i.e. solo) is very small.
3) The most frequent three-sign combination (336, 89, 211) whose total frequency of occurrence is 34 appears most frequently at the left-end of the texts, its percentage of occurrence at the left-end of texts being 88.24. Similarly, the three-sign combination (293, 123, 343) whose total frequency of occurrence is 25 always appears at the right-end of the texts.
4) Some of the important three-sign combinations arise as additive forms of frequent two-sign combinations listed in table 4. For example, the three-sign combination (336, 89, 211) arises from the two-sign combinations (336, 89) and (89, 211). However, this three-sign combination appears 34 times, as does the two-sign combination (89, 211), while the second two-sign combination (336, 89) appears a total of 75 times, out of which it appears in this three-sign combination only 34 times.
5) The second most frequent three-sign combination (293, 123, 343) consists of two-sign combinations (293, 123) and (123, 343). In this case also the frequency of occurrence of the three-sign combination (293, 123, 343) and the two-sign combination (123, 343) is identical with 25 occurrences.

Table 6: Positional analysis of Four-sign Combinations

Four-sign Combination				Frequency	Solo (%)	Left (%)	Middle (%)	Right (%)
342	149	130	51	16	6.25	68.75	6.25	18.75
59	87	99	267	9	0.00	0.00	33.33	66.67
89	336	59	171	6	0.00	0.00	83.33	16.67
15	389	178	98	5	0.00	100.00	0.00	0.00
342	53	230	175	5	20.00	80.00	0.00	0.00
342	169	249	65	5	20.00	20.00	20.00	40.00
211	89	336	72	5	0.00	80.00	0.00	20.00

Table 6 gives the positional distribution of most frequent four-sign combinations. The following points are evident.

1) The four-sign combinations tend to occur in specific positions in the text (i.e. left-end, middle or right-end).
2) Their percentage of occurrence as independent texts (i.e. solo) is very small.
3) The most frequent four-sign combination (51, 130, 149, 342) consists of very frequent three-sign combinations (51, 130, 149) and (130, 149, 342). This four-sign combination has the same frequency of occurrence as the three-sign combination (130, 149, 342).

Discussion on Tables 4, 5 and 6:

We discuss the combined importance of tables 4, 5 and 6. In table 4, the two-sign combinations (336, 89), (87, 59), (171, 59), (249, 162), (65, 67), (51, 130), (99, 67) and (123, 343) never appear solo and even within large texts they appear in the middle. This is in contrast to the two-sign combinations (267, 99), (391, 99) and (293, 123) which mostly appear at the right extreme of the texts. A third group of two-sign combinations (342, 176), (8, 342), (347, 342), (342, 1), (89, 211), (59, 211) and (162, 342) mostly appear at the left extreme of texts. The two-sign combinations (48, 342) and (245, 245) seem to be evenly distributed.

The two-sign combinations that appear at extremities clearly form whole or part of text-starting or text-ending sign combinations. We now look at the sign combinations that appear in the middle. The two-sign combination (336, 89) when combined with the sign 211 on the right comes solo 2.94 % of the times or at left end 88.24 % of the times (Table 5) indicating that if not complete text, it is a very decisive text ender sign combination.

The two-sign combination (87, 59) when combined with the sign 99 at the left, appears always in the middle (Table 5). However, when the three-sign combination (99, 87, 59) is combined with 267 at the left, it mostly appears at the right end of texts indicating that this four-sign combination is a text beginner sign combination. It is interesting that another four-

sign combination not listed in table 6 occurs 4 times and differs from this only by the rightmost sign where 267 is replaced by 391. This suggests that 267 has a meaning similar to 391.

The two-sign combination (171, 59) does not appear with high frequency in three-sign combination table (Table 5). Even when appearing in a four sign combination, this pair (171, 59) appears on the left of another significant two sign combination (336, 89) as (171, 59, 336, 89).

The two- sign combination (249, 162) when combined with 342 at the right comes solo 4.17% of the times and at the left extreme of the texts 83.33% of the times. It is interesting that the sign 169 shows similar behaviour to the sign 162.

Conclusions from Analysis 3:

1) The most frequent two-sign, three-sign and four-sign combinations appear at fixed positions.
2) The exact location varies from combination to combination.
3) However, frequently occurring two-sign, three-sign and four-sign combinations may be incomplete except of course when they occur as solo texts. It can be seen that two-sign, three-sign and four-sign combinations which are complete have typically one of the text-enders (mostly 342 or 211) at the end. This is confirmed by the solo occurrences of such texts.

3.4 Analysis 4: Comparison with other scripts

Naranan and Balasubrahmanyan (1992) have used the statistics of word frequencies in different languages. Quoting the work of Zipf (1935, 1949) they analyse the word frequency distribution and use the power law relation between W(k) and k, where W(k) is the number of different words or word types occurring k times in a discourse of words. They assume the relation $W(k) = Ck^{-\gamma}$. Naranan and Balasubrahmanyan have given the graph of W(k) vs. k for different languages and writing

styles and compared it with the Indus style. In figure 2 we have plotted the value of $-\gamma$ for different languages and authors and along with the value for the Indus writing (last point in figure 2). As can be seen from the figure, the slope for Indus writing is significantly different from that for other known writing styles.

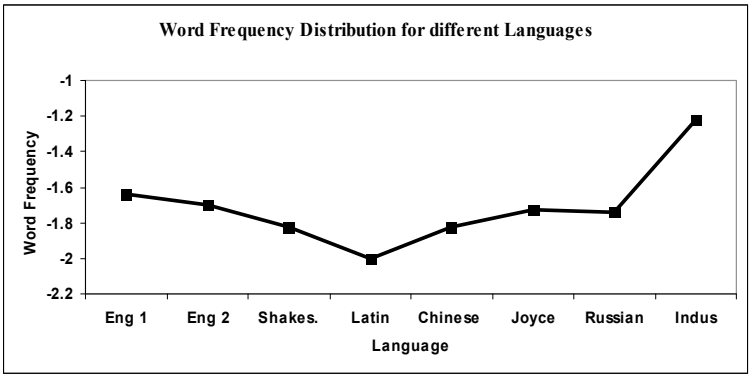

Fig. 2: Word frequency distribution of different languages (based on Naranan and Balasubrahmanyan, 1992)

Conclusions from Analysis 4:

1. Unlike a pictographic script such as Chinese writing where each sign is stand alone in meaning, the Indus script is not a simple incrementally built up script. Indus script is also not alphabetic. This is also clear from the fact that Chinese language has more than 3000 different signs compared with 417 signs in Indus script and 26 in English. However, some signs do seem to be pictorial nature.
2. Study of sign combinations may be more meaningful than a study of individual signs.

4.0 Summary

We therefore make the following conclusions about the Indus script.

1) The writing is highly ordered.

2) Typical length of information containing units is 2, 3 or maximum 4 signs. However, they are not always complete enough to exist as standalone piece of text. This also indicates a more complex grammar in the writing where information units need proper beginners or enders (see tables 4-6).

Acknowledgement

We wish to thank Dr. Subbarayappa for initiating us in this study with his stimulating interpretation of the script as numeric. We wish to thank Dr. Samudravijaya and Mr. Aniket Sule for their enthusiastic support for this work. We also wish to thank Ms. Prajakta Mahajani for digitising the data that was used to understand the basics the script. We wish to acknowledge the research grant of the Jamsetji Tata Trust that enabled us to do this work.

References

Fairservis, W.A., 1983, Scientific American March issue, pp. 44 - 52

Farmer, S., Sproat, R., Witzel, M., 2004, The Collapse of the Indus-Script Thesis: The Myth of a Literate Harappan Civilization, Electronic Journal of Vedic Studies (EJVS), Vol. 11, issue 2

Mahadevan, I., 1977, *THE INDUS SCRIPT Texts, Concordance and Tables*, Memoirs of the Archaeological Survey of India No. 77

Mahadevan, I., 2002, Aryan or Dravidian or Neither? A Study of Recent Attempts to Decipher the Indus Script (1995-2000), Electronic Journal of Vedic Studies (EJVS), Vol. 8, issue 1

Mahadevan, I., 2007, Agricultural terms in the Indus Script, to appear in Journal of Tamil Studies

Naranan, S. and Balasubrahmanyan, V.K., 1992, Current Science, Vol. 63, No. 6, pp. 297 - 305

Parpola, A., 2005, Study of the Indus Script, 50th ICSE Tokyo

Possehl, G.L., 1996, Indus Age: The writing System, Oxford & IBH Publishing Co. Pvt. Ltd.

Subbarayappa, B.V., 1997, INDUS SCRIPT: Its Nature and Structure, New Era Publications, Madras

Yadav, N. and Vahia, M.N., 2007, One more attempt to understand the Harappan writing, presented at the 1st International Conference of the Society of South Asian Archaeology (SOSAA), Mumbai, December 2006, to appear in the Journal of SOSAA