

## COMPUTATIONAL TECHNIQUES FOR INFERRING THE SYNTAX OF UN- DECIPHERED SCRIPTS

NISHA YADAV\*

*Department of Computer Science, University of Mumbai, Santacruz (E), Mumbai - 400098, Maharashtra, India*  
&  
*Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai - 400005, Maharashtra, India*  
*y\_nisha@tifr.res.in*

AMBUJA SALGAONKAR

*Department of Computer Science, University of Mumbai, Santacruz (E), Mumbai - 400098,*  
*Maharashtra, India*  
*ambujas@gmail.com*

MAYANK VAHIA

*Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai - 400005, Maharashtra, India*  
*vahia@tifr.res.in*

Understanding the syntax of an undeciphered writing is a significant challenge. This can provide important clues to the nature of writing and guide potential decipherments. Here we evaluate a set of computational tools that can help us address this problem. We show that significant aspects of the writing can be inferred through this approach without making any assumption about the underlying content. We demonstrate the validity of these techniques using the example of the undeciphered Indus script widely used in the Indus Valley or Harappan Civilization that flourished in the north-western part of the Indian subcontinent from 2600 to 1900 BC.

*Keywords:* Computational linguistics; data mining; machine learning; statistical analysis, undeciphered scripts.

### 1. Introduction to Indus Script

The undeciphered Indus script is a creation of one of the largest and richest ancient Bronze Age civilizations that flourished in the north-western part of the Indian subcontinent and is generally referred to as the Indus valley or the Harappan civilization. Though the roots of the Indus valley civilization go back to around 7000 BC, it peaked

\*Address for correspondence:

*Department of Computer Science, University of Mumbai, Santacruz (E), Mumbai - 400098, Maharashtra, India*  
& *Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai - 400005, Maharashtra, India*  
*E-mail: y\_nisha@tifr.res.in*

around 2600 BC and went into decline around 1900 BC [Wright, 2010; Agrawal, 2007; Possehl, 2002; Kenoyer, 1998].

The objects inscribed with the Indus script are generally a few square centimeters in size (Fig. 1). They are catalogued in the three volumes of the *Corpus of Indus Seals and Inscriptions* [Joshi & Parpola, 1987; Shah & Parpola, 1991; Parpola, *et al.*, 2010]. On these objects, the Indus people have expressed several aspects of their art, their myths, their perspective of nature, abstract geometrical and symmetrical patterns and at times, even their daily life. In terms of art, aesthetic sense and expressions of symmetric, geometric as well as abstract patterns, these objects are unsurpassed in their quality [Yadav & Vahia, 2011; Vahia & Yadav, 2010]. One of the most creative aspects of their work on these inscribed objects is the Indus script. Hence, an understanding of their script will provide unprecedented insights into the minds of the Indus people. The script therefore holds a vital clue to understanding the Indus culture.

Reasons that make the problem of Indus script challenging are the brevity of the Indus texts, paucity of the data, lack of definitive knowledge about the language(s) that the Indus people spoke, and absence of bilingual inscriptions. In spite of these hurdles scholars have continued to attempt to understand the contents of the script. Possehl [Possehl, 1996] provides an excellent critical review of some of the various attempts to understand and interpret the script (see also [Mahadevan, 2002; Parpola, 2005]). In spite of these efforts, the problem of Indus script lies unresolved with no universal consensus on any of the interpretations.



Fig. 1: Some examples of Indus seals with the Indus script (Copyright Harappa Archaeological Research Project/J.M. Kenoyer, Harappa.com, Courtesy Dept. of Archaeology and Museums, Govt. of Pakistan).

## 2. Motivation and Approach

While several interpretations of the contents of the Indus script have been put forward, none of them are satisfactory and hence we are no closer to decipherment of the script than we were a century ago when the script was first discovered. The empirical frameworks are unlikely to provide satisfactory decipherment without a proper understanding of the syntax to guide and validate such frameworks in an objective manner.

Until recently, no generalized tools were available for this purpose. We have used various computational techniques to identify aggregate characteristics of the Indus script without making any assumptions about its content. Our study, summarized here, aims to define several constraints that any proposed interpretation must satisfy.

We infer the characteristics of the syntax of the Indus script by employing computational techniques related to machine learning, data mining and information theory. Our study aims to define a syntactic framework of the writing rather than read or interpret it. We hope that this will provide an objective testing ground for any claims of decipherment of Indus script.

We used Mahadevan's concordance [Mahadevan, 1977] as the basic data set on which we applied various analytical, mathematical and computational tools to understand the syntax of the Indus script. It records 417 unique signs in 3573 lines of 2906 texts. From this, we removed ambiguous texts and create a filtered dataset EBUDS (Extended Basic Unique Dataset, for details see [Yadav, *et al.*, 2008a]). MATLAB and Linux were used for various analyses.

### **3. Summary of Results**

Any systematic writing would have specific ordering of signs. The frequency of signs and sign combinations as well as their pattern of appearance in the texts elucidate the syntax of the writing. It helps define the rules and flexibility available to anyone writing in the script. Our focus was to identify these patterns.

Several analysis performed by us on the dataset suggest that the writing is structured with limited number of signs for text beginners and enders and significant constraints on the pairing of signs. However the writing is not completely rigid and seems to give the writer a certain amount of flexibility in coding information. We discuss the individual analysis and their results below.

#### **3.1 Analysis of syntactic patterns**

The pattern of usage of signs in the dataset can be studied using its sign frequency distribution. We find that the frequency of usage of signs in the dataset follows Zipf-Mandelbrot distribution [Yadav, *et al.*, 2010]. This suggests that the texts are dominated by a few signs and the frequency of signs can be best described by a power law.

Cumulative frequency distribution of text enders and text beginners reveals significant asymmetry. While just 23 signs account for 80% of all text enders, 82 signs account for 80% of all text beginners (Fig. 2). This suggests that in the Indus writing only a small number of signs were allowed to end the texts while a relatively larger number of signs could begin the texts.

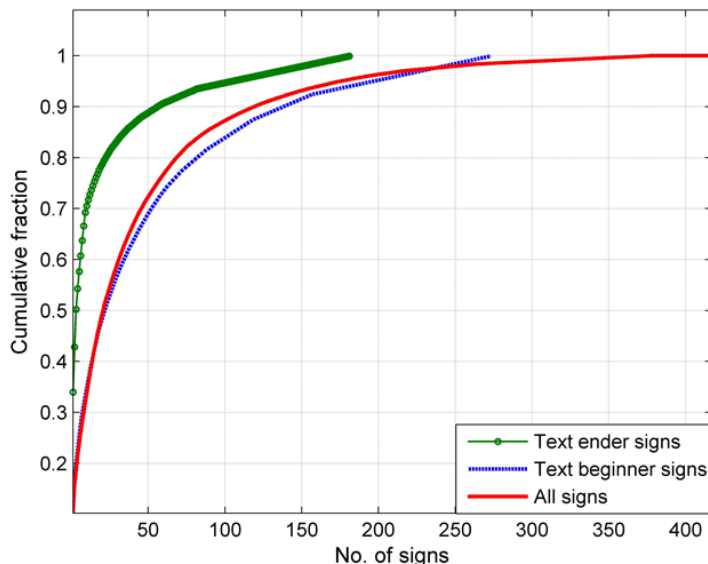


Fig. 2: Cumulative frequency plot for all signs, text-beginners and ends [Yadav, *et al.*, 2010].

To check if the sequencing of signs in Indus texts is significant, we compared the Indus script dataset with a randomized dataset [Yadav, *et al.*, 2008a]. Our study reveals that sign combinations of two, three and four signs appear with far higher frequency in the Indus script dataset than expected by chance. This suggests presence of correlations between signs in the Indus texts. It also indicates that the length of the information unit in Indus texts is two, three or four signs [Yadav, *et al.*, 2008a]. Further analysis of the distribution pattern of the sign combinations (pairs, triplets and quadruplets) in the Indus texts showed that they have preferred location in the Indus texts [Yadav, *et al.*, 2008a].

### 3.2 Segmentation of Indus texts

In order to investigate the possibility that the longer texts may consist of more than one unit of information we performed segmentation analysis on the Indus script based on the patterns identified in the earlier studies [Yadav, *et al.*, 2008b]. The length of text in Indus script varies from 1 to 14 signs per line. Moreover, the text beginners and ends are well defined. It is therefore significant to check whether the longer Indus texts consist of multiple units of information or highly complex (or detailed) but single unit of information. We find that about 88% of all texts of length five or more can be segmented into segments of length not exceeding four. Hence, Indus writing consists of multiple units of information written in a text. Our study firmly established that the longer strings of writing are a collection of several smaller units of information and not a long unit of

















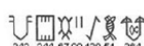
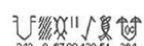
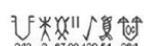


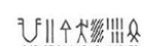






complex information. This provides additional clues to the possible purpose and contents of writing.

### 3.3 *n*-gram studies of the Indus script

The results above demonstrate that the Indus writing is highly ordered. We therefore subjected the dataset to formal techniques for analyzing sequences [Jurafsky & Martin, 2008; Manning & Schütze, 1999]. Probabilistic models such as *n*-gram models or Markov chains can be used to learn the sequential structure of the texts in an undeciphered script. In a general *n*-gram model, all correlations beyond the (*n*−1) preceding signs are discarded.

We developed a bigram model of the Indus script [Yadav, *et al.*, 2010]. In a bigram (or a first-order Markov model) the range of correlation is restricted to the nearest neighbor. We employ the bigram model of the Indus script for restoring signs in illegible Indus texts (Table 1), for generating sample Indus texts, and for comparing texts from Indus sites and West Asian sites.

Table 1: Restoration of doubtfully read signs in M77 [Yadav, *et al.*, 2010].

Text No.	Text	Incomplete Text	Most Probable Restoration	Probable Restored Sign
4312				
4016				
5237				
2653				
5073				
3360				
9071				

The model can restore signs with an accuracy of about 75% [Yadav, *et al.*, 2010]. We find that likelihood of many of the Indus texts found in the far off West Asian region under a model trained on texts from Indus sites is very low suggesting that the script may have been used for writing West Asian content [Rao, *et al.*, 2009b].

### 3.4 Comparison of flexibility in sign usage across different sign systems

We compared the flexibility of sign sequencing in the Indus script with sequencing in various linguistic and non-linguistic systems (viz. English, Sanskrit, Old Tamil, Sumerian, DNA, Protein, and Fortran) [Rao, *et al.*, 2009a]. We find that the conditional entropy (a measure of flexibility in the choice of a token given a preceding token) of the Indus script falls within the range of the various linguistic systems included in the study (Fig. 3).

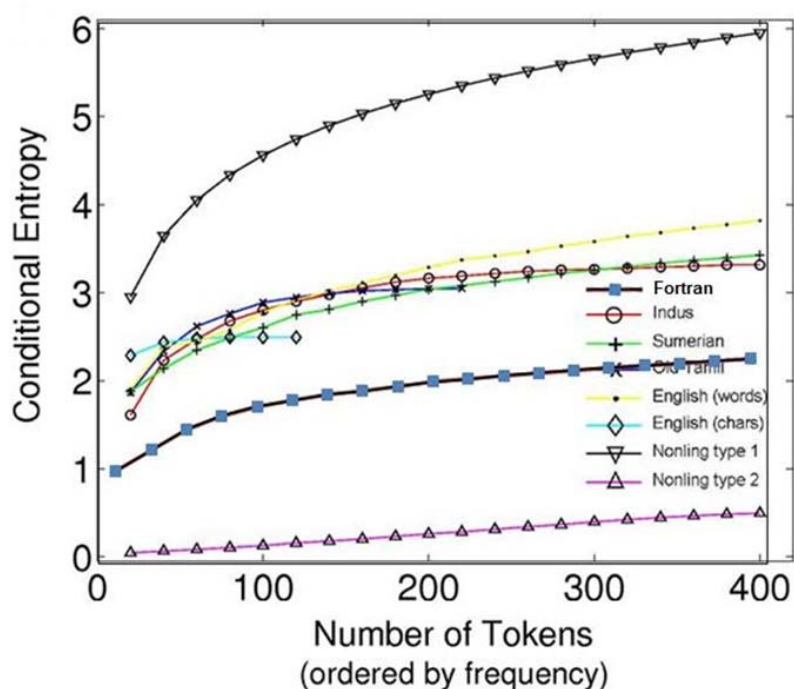


Fig. 3: Comparison of Indus data with linguistic and non-linguistic systems (From [Yadav, 2012], see also [Rao, *et al.*, 2009a]).

The result increases the evidence in favour of the linguistic hypothesis of the Indus script. It should however be noted that it does not prove it to be linguistic.

### 3.5 Study of design of Indus signs

The designs for Indus signs are distinct in terms of their complexity and the manner in which the signs share design characteristics. Several signs seem to have been designed by adding modifiers to signs or merger of several individual signs. We analyzed the design of individual signs of the Indus script in order to understand the general makeup and mechanics of the design of Indus signs [Yadav & Vahia, 2011]. We studied the designs of the 417 distinct signs in the sign list of Indus script [Mahadevan, 1977]. We visually

identified three types of design elements of Indus signs namely basic signs, provisional basic signs and modifiers. These elements combine in a variety of ways to generate the entire set of Indus signs. By comparing the environment of compound signs with all possible sequences of their constituent basic signs, we show that sign compounding (ligaturing) and sign modification seem to change the meaning or add value to basic signs rather than save writing space.

### **3.6 Site and medium sensitivity of Indus script**

The Indus writing has been found at several sites spread over an area of about a million square kilometers as well as on different types of objects. It is therefore possible that there were grammatical and stylistic differences between writing found at different sites. Similarly it is also possible that different medium of writing encoded different types of information. Using the technique of hierarchical clustering, we have investigated the variation in the usage of signs in the Indus script across sites and types of objects [Yadav, 2013]. Some of the major conclusions from this study are:

- (1) *Distribution of inscribed objects*: Study of the distribution of the inscribed objects with respect to their site of occurrence and type suggests that Mohenjodaro accounts for the highest percentage of seals and Harappa accounts for the highest percentage of sealings.
- (2) *Sensitivity of the Indus script to site and type of object*: There are no significant variations in the usage of signs at different sites or on different types of objects. However, subtle preferences in the usage of signs in the Indus writing on different type of objects and at different sites indicate the presence of some individualistic clues to their content.
- (3) *Clustering of sites and types of objects*: Using the technique of hierarchical clustering we compared various sites and types of objects based on different criteria such as their usage of signs or distribution of text lengths. Some of the significant conclusions from this analysis are:
  - (i) Mohenjodaro and Lothal share high level of similarity in their pattern of text length distributions and usage of signs.
  - (ii) Harappa is distinct in its sign usage from all other sites.
  - (iii) The pattern of text length distribution and usage of signs in West Asian sites is distinct from all other sites.
  - (iv) With respect to the usage of signs, sealings and miniature tablets are closest to each other.
  - (v) In usage of signs, seals share a high level of similarity with pottery graffiti.

## **4. Conclusions**

We conclude that the Indus texts have an underlying logic and syntax indicative of writing. We find that the ordering of signs is more rigid than random writing. There is significant asymmetry in the usage of text beginners and text enders. The structure of

writing is such that that even a first order Markov model of Indus script can accurately restore signs when they are intentionally removed from unambiguously read texts. We also find that the rules of writing were highly standardized over the entire civilization and medium of writing. Significant conclusions of our study are tabulated in Table 2.

Table 2: Major Conclusions.

Sl. No.	Test/Measure	Results	Conclusion
1	Zipf-Mandelbrot Law	Best fit for $a = 15.4$ , $b = 2.6$ , $c = 44.5$ (95% confidence interval)	A small number of signs account for bulk of the corpus while a large number of signs contribute to a long tail, a feature followed by many ordered systems.
2	Cumulative frequency distribution	69 signs: 80% of EBUDS, 23 signs: 80% of text enders, 82 signs: 80% of text beginners	Indicates asymmetry in the usage of text beginners and text enders. Suggests logic and structure in writing.
3	Comparison with randomized text	Sign sequences of size two, three and four appear far more frequently than that expected by chance.	There is significant order in sequencing signs in the Indus texts.
4	Extraction of frequent sign sequences	Frequent sign sequences extracted.	There are well-defined subunits within the broad framework of Indus writing.
5	Positional distribution of frequent sign sequences	Frequent sign sequences have a preferred location of occurrence within the Indus texts.	There are specific rules in ordering sequences of signs in the Indus script.
6	Segmentation of Indus texts	About 88% of texts of length five or more can be segmented into segments of two, three or four signs.	Indus writing seems to consist of multiple units of information written in a text.
7	Bigram probability	Conditional probability matrix is strikingly different from the matrix assuming no correlations.	Indicates presence of significant correlations between signs.
8	Conditional probabilities of text beginners and text enders	Restricted number of signs follow frequent text beginners whereas large number of signs precede frequent text enders.	Indicates presence of signs having similar syntactic functions.
9	Log-likelihood significance test	Significant sign pairs and triplets extracted.	The most significant sign pairs and triplets are not always the most frequent ones.



10	Entropy	Random: 8.70; EBUDS: 6.68	Indicates presence of correlations between signs.
11	Mutual information	Random: 0; EBUDS: 2.24	Indicates flexibility in sign usage even within the broad set of rules.
12	<i>Perplexity</i>	Monotonic reduction as $n$ increases from one to five.	Indicates presence of long range correlations in Indus texts.
13	Sign restoration	Restoration of missing and illegible signs.	Model can restore illegible signs by suggesting most likely replacements.
14	Evaluation of bigram model using <i>cross validation</i>	<i>Sensitivity</i> of the model = 75%	Model can predict illegible or missing signs in a text with 75% accuracy.
15	Comparison of likelihood of Indus texts from Indus Valley and West Asian sites	The median value of likelihoods for the Indus texts from West Asian sites is $6.40 \times 10^{-13}$ , which is 1,000,000 times less than the median value of $1.12 \times 10^{-7}$ for the Indus texts from Indus Valley sites.	Indicates difference in the pattern of sign sequencing for Indus texts coming from Indus Valley sites and West Asian sites. The Indus script may have been used for writing West Asian content.
16	Conditional and block entropy	Conditional and block entropy of Indus texts falls within the range of linguistic systems.	Flexibility in sign usage in Indus texts is closer to linguistic systems than to non-linguistic systems.
17	Sensitivity of Indus script to sites and types of objects	Variation in the usage of Indus signs is analyzed across sites and types of objects.	Indus texts at different sites and on distinct object types do have small individualistic clues to their content.
18	Classification of Indus signs based on their design elements	<u>Design elements:</u> Basic signs: 154, Provisional basic signs: 10, Modifiers: 21  <u>Types of Indus signs:</u> Basic signs: 154, Composite signs: 263 (Compound signs: 149 and Modified signs: 114)	Design elements of the Indus signs are identified and signs are classified into different categories based on their design complexity.
19	Analysis of compound signs	Pattern of occurrence of compound signs in the Indus texts is different from their constituent sign sequences, which occur rarely.	Compound signs are not shorthand but seem to have different meaning.
20	Evaluation of interpretative models for Indus script	The syntactic features of the Indus texts identified by our study can be used to evaluate proposed models of interpretation for Indus texts.	A proposed model for Indus script evaluated using these results was found to be internally inconsistent [Yadav, <i>et al.</i> , 2012].

## **5. Discussion and Future Work**

Recent advancements in computer science provide a large number of tools to address the problem of undeciphered scripts. Amongst the various algorithms developed to understand various types of data, it is necessary to identify the most appropriate methods to extract information about an undeciphered script. Currently there are no agreed criteria that can be applied to any written text to establish its nature. The current work is aimed at filling this lacuna. In this study, we have identified the computational methods that can be used to address this problem. We use the Indus script as a test case to show that the methods proposed here are comprehensive enough to provide deep insights into the syntax of an undeciphered script. Even though the computational techniques are insensitive to the semantics of the script, they can be effectively used to evaluate various semantic frameworks of decipherment. This, we believe is a significant enhancement in the field of machine interpretation of the human construct of writing.

In their present form, the results obtained by us can effectively rule out several possible interpretations and provide stringent limits on the direction in which the possible semantic decipherment will lie. We propose to extend the methodology to understand the subtleties of the Indus script in a content independent manner and evaluate proposed claims of decipherment based on different linguistic and non-linguistic systems. We plan to broaden our studies to analyze the variations in the Indus script due to various archaeological factors of the objects on which the texts are inscribed (see for example [Yadav, 2013]). Further attempts would be to design a logically consistent messaging system based upon our findings so far. This system need not be the one that the Indus people might have used. The system should serve a purpose for steganographic applications or as a language of minimal text in some specific domains. Similarly, the methodology developed in the field of semiotics can be explored to extract information on possible different types of contents that may exist in the Indus writing.

Another line of study that can potentially yield significant insights is to extend the comparison of various characteristics of the Indus script with other natural or artificial structured systems such as computer languages, natural languages etc. Similarly, techniques related to unsupervised learning of grammar can also be explored in future to test their applicability and potential utility for such purposes. It is hoped that the present work will provide further incentives to explore the relevance and power of computational methods in deciphering the texts when authentic background knowledge is inadequate.

## **6. Acknowledgement**

We are grateful to Harappa.com for their kind permission to use the images of the Indus seals in the paper and for their continuing support. We would like to thank Dr. Iravatham Mahadevan whose work and support provides the basis for the current work. We also wish to thank our other colleagues with whom we have worked and who have been co-authors of some of our work.

## References

- Agrawal, D. P. (2007). *The Indus Civilization: An Interdisciplinary perspective*. New Delhi: Aryan Books International.
- Joshi, J. P.; Parpola, A. (1987). *Corpus of Indus Seals and Inscriptions, 1. Collections in India*. Helsinki: Suomalainen Tiedeakatemia, also Memoirs of the Archaeological Survey of India No. 86.
- Jurafsky, D.; Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing*. 2nd Edition ed. s.l.:Pearson Prentice Hall.
- Kenoyer, J. M. (1998). *Ancient Cities of the Indus Valley Civilization*. Oxford: Oxford University Press.
- Mahadevan, I. (1977). *The Indus Script: Texts, Concordance and Tables*. New Delhi: The Director General, Archaeological Survey of India, also 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*, 8(1).
- Manning, C.; Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.
- Parpola, A. (2005). *Study of the Indus script*. Tokyo: The Tōhō Gakkai, s.n., pp. 28-66.
- Parpola, A.; Pande, B. M.; Koskikallio, P. eds., 2010. *New material, untraced objects, and collections outside India and Pakistan, Part 1: Mohenjodaro and Harappa*. Helsinki: Suomalainen Tiedeakatemia.
- Possehl, G. L. (1996). *Indus Age: The writing System*. New Delhi: Oxford & IBH Publishing Co. Pvt. Ltd..
- Possehl, G. L. (2002). *The Indus Civilization: A contemporary perspective*. New Delhi: Vistaar Publications.
- Rao, R. P. N.; Yadav N.; Vahia, M. N.; Joglekar, H.; Adhikari, R.; Mahadevan, I. (2009a). Entropic evidence for linguistic structure in the Indus script. *Science*, Volume 324, p. 1165.
- Rao, R. P. N.; Yadav N.; Vahia, M. N.; Joglekar, H.; Adhikari, R.; Mahadevan, I. (2009b). A Markov model of the Indus script. *Proceedings of the National Academy of Sciences*, 106(33), pp. 13685-13690.
- Rao, R. P. N.; Yadav N.; Vahia, M. N.; Joglekar, H.; Adhikari, R.; Mahadevan, I. (2010). Entropy, the Indus Script and Language: A Reply to R. Sproat. *Computational Linguistics*, Volume. 36, pp. 795-805.
- Shah, S. G. M.; Parpola, A. (1991). *Corpus of Indus Seals and Inscriptions, 2. Collections in Pakistan..* Helsinki: Suomalainen Tiedeakatemia, also Memoirs of the Archaeology and Museums.
- Vahia, M. N.; Yadav, N. (2010). Harappan Geometry and Symmetry: A study of geometrical patterns on Indus Objects. *Indian Journal of History of Science*, 45(3), pp. 343-368.

- Wright, R. P. (2010). *The Ancient Indus – Urbanism, economy and society*. New York: Cambridge University Press.
- Yadav, N.; Rao, R. P. N.; Vahia, M. N. (2012). Indus Script. *Current Science*, Vol. 103, pp. 1265-1266.
- Yadav, N. (2012). Statistical Studies of the Indus Script. *Man and Environment*, XXXVII (1), pp. 1-7.
- Yadav, N.; Joglekar, H., Rao, R. P. N.; Vahia, M. N., Adhikari, R.; Mahadevan, I. (2010). Statistical Analysis of the Indus Script using n-grams. *PLoS One*, 5(3).
- Yadav, N.; Vahia, M. N. (2011). Classification of patterns on Indus objects. *International Journal of Dravidian Linguistics*, 40(2), pp. 89-114.
- Yadav, N.; Vahia, M. N. (2011). Indus Script: A Study of its Sign Design. *Scripta*, Volume 3, pp. 133-172.
- Yadav, N.; Vahia, M. N.; Mahadevan, I.; Joglekar, H. (2008a). A statistical approach for pattern search in Indus writing. *International Journal of Dravidian Linguistics*, XXXVII (1), pp. 39-52.
- Yadav, N.; Vahia, M. N.; Mahadevan, I.; Joglekar, H. (2008b). Segmentation of Indus Texts. *International Journal of Dravidian Linguistics*, XXXVII (1), pp. 53-72.