UNRAVELING THE STRUCTURE OF INDUS SCRIPT: A COMPUTATIONAL PERSPECTIVE

Dr. Nisha Yadav

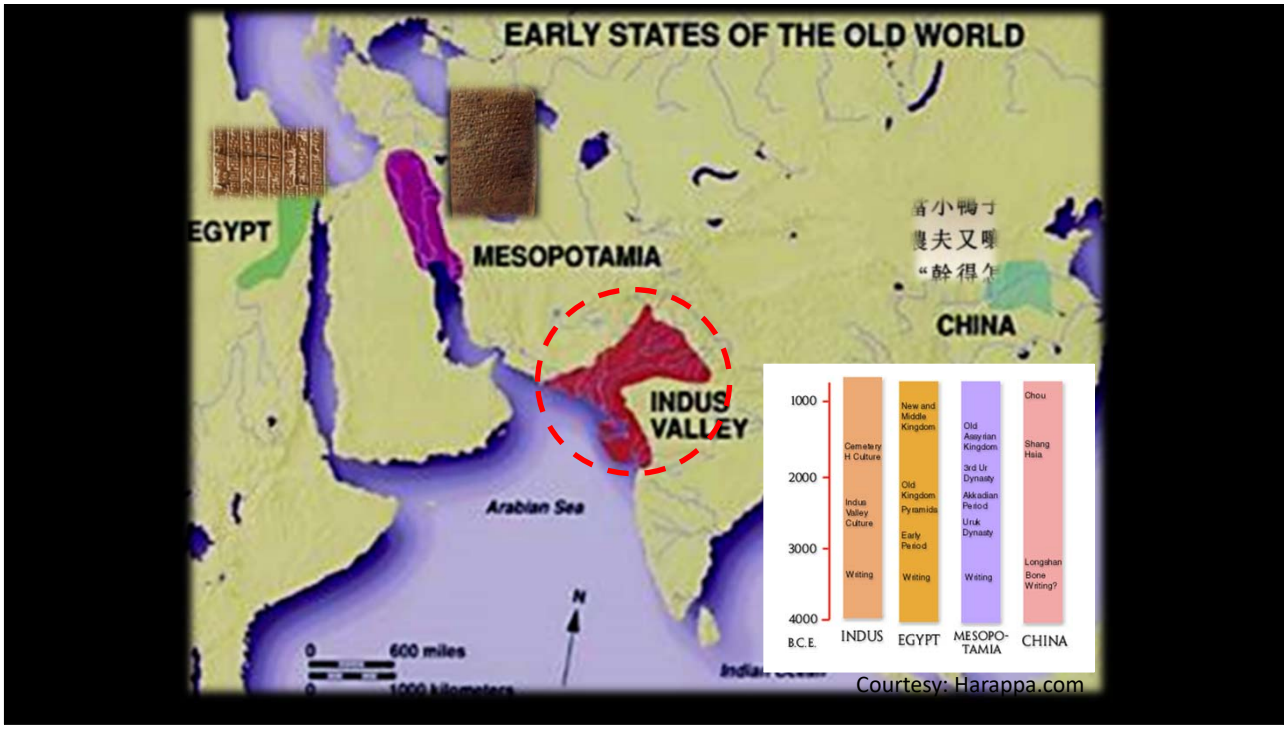
Tata Institute of Fundamental Research, Mumbai

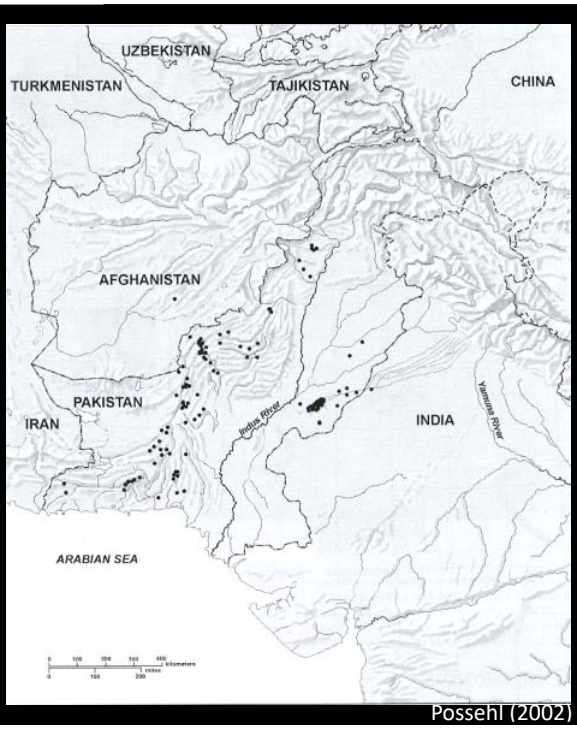
"The practice of writing and the development of a coherent system of signs, a script, is something which is seen only in complex societies....Writing, in other words, is a feature of civilizations."

-- Colin Renfrew

(Archaeology and Language, 1987)

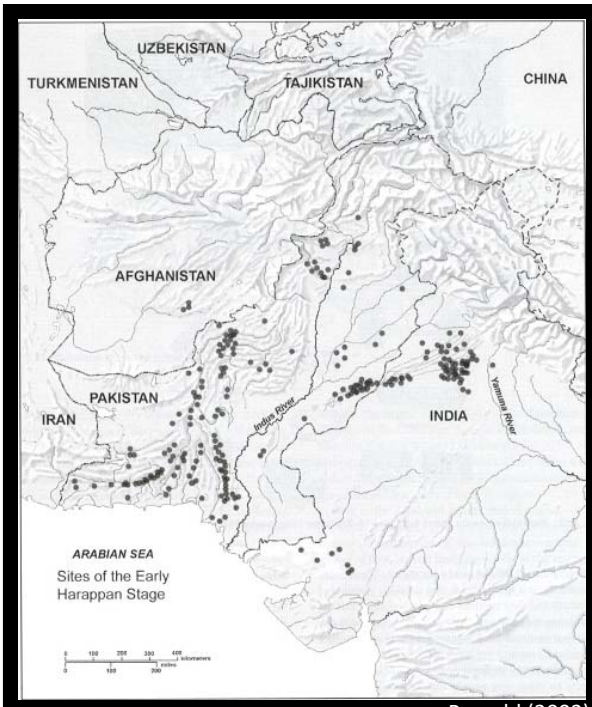

Courtesy: Harappa.com


EARLY STATES OF THE OLD WORLD
EGYPT
MESOPOTAMIA
CHINA
INDUS VALLEY
Arabian Sea
0 600 miles
N
1000
2000
3000
4000
B.C.E.
Cemetery H Culture
Indus Valley Culture
Writing
New and Middle Kingdom
Old Kingdom Pyramids
Early Period
Writing
Old Assyrian Kingdom
3rd Ur Dynasty
Akkadian Period
Uruk Dynasty
Writing
Chou
Shang Hsia
Longshan Bone Writing?
INDUS
EGYPT
MESOPO-TAMIA
CHINA
Courtesy: Harappa.com



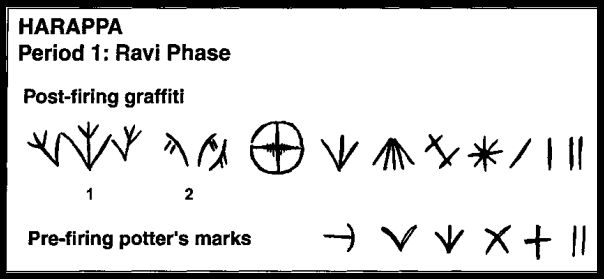
Possehl (2002)

7000 to 3200 BC

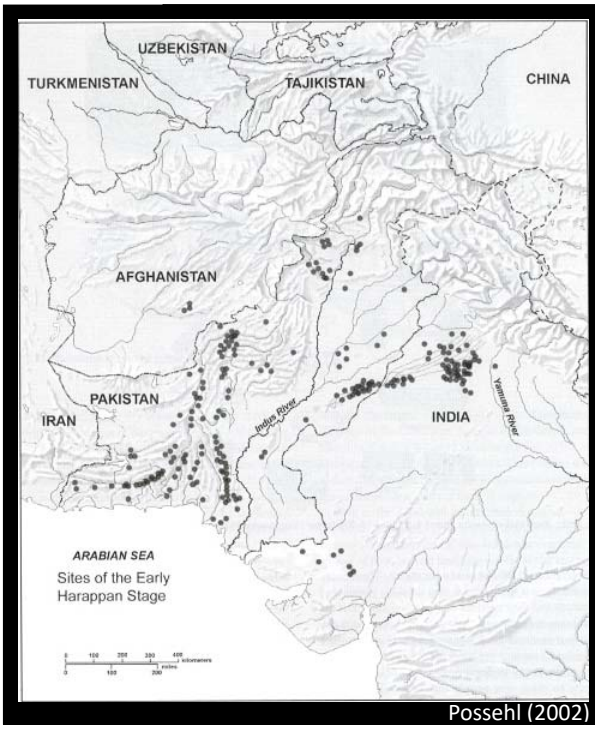


Possehl (2002)

Early Harappan
(Ravi phase)
3200 to 2800 BC

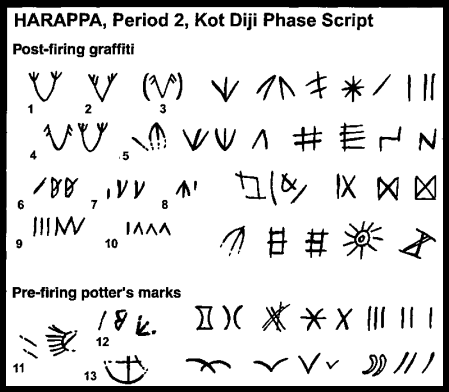


Kenoyer (2006)



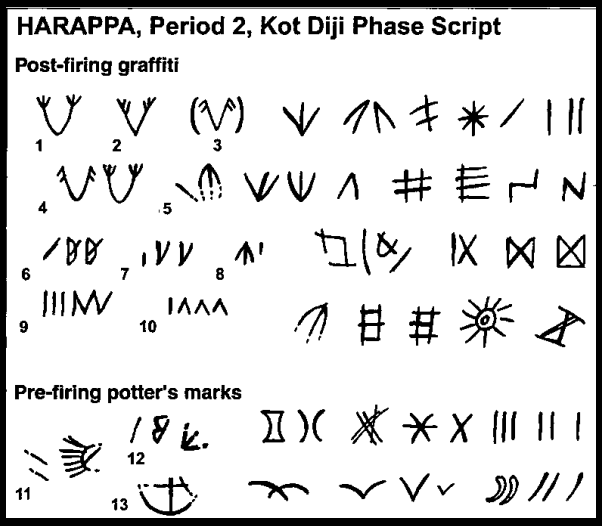
Possehl (2002)

Early Harappan (Kot Diji phase) 2800 to 2500 BC



Kenoyer (2006)

Kot Diji phase 2800 to 2500 BC



Kenoyer (2006)

Mature phase 2600-1900 BC Texts from Indus Seals

Mahadevan (1977)

INDUS SCRIPT ON VARIOUS OBJECTS

Seals (62%) & Sealing (19%)

~ 5 cm

Miniature Tablets (9%)

Copper Tablets (5%)

Pottery Graffiti (2%)

Ivory sticks (1%)

www.harappa.com, CISI Volumes

COMPONENTS OF INDUS SEALS

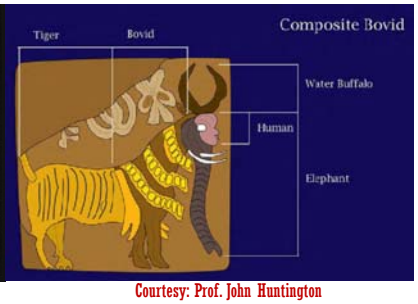
1. Script
2. Animal Motif
3. Manger
4. Narrative scenes
5. Human figures
6. Plant motifs
7. Mythical figures
8. Composite animals
9. Multi-headed animals
10. Geometric patterns
11. Abstract patterns
12. Crude patterns

Yadav & Vahia (2011)

SOME INTERESTING COMPONENTS

4.0 cm

3.0 cm



Courtesy: Prof. John Huntington

3.2 cm

CISI, www.harappa.com

SOME MORE PATTERNS...

3.5 cm

3.4 cm

Same object
opposite sides

3.3 cm X 1.5cm

2.7 cm

2.0 cm

Relatively to scale

Courtesy: CISI, Harappa.com

GEOMETRIC AND ABSTRACT PATTERNS

2 cm

3.2 cm

1.9 cm

2.0 cm

3.5 cm

CISI, Harappa.com

*"At their very best, it would be no exaggeration to describe them as **little masterpieces of controlled realism**, with a monumental strength in one sense out of all proportion to their size and in another entirely related to it."*

-- Wheeler, 1968

WHY DID ANCIENT PEOPLE WRITE?

- Writing allowed people to
  - store records (accounting)
  - keep track of time
  - facilitate communication over short and long distances (trade)
  - religious purposes
  - express superiority and power
- The ambiguity associated with the context and usage of Indus signs has made the problem of Indus writing system even more challenging.

MAJOR IMPEDIMENTS TO DECIPHERMENT

- Very short and brief texts –
  - average number of signs is 5
  - longest single line text: 14 signs
  - longest text: 26 signs running in 3 distinct lines

M-355 a

Longest single line text.

- Lack of information about their spoken language(s)
- Lack of bilingual or multilingual texts

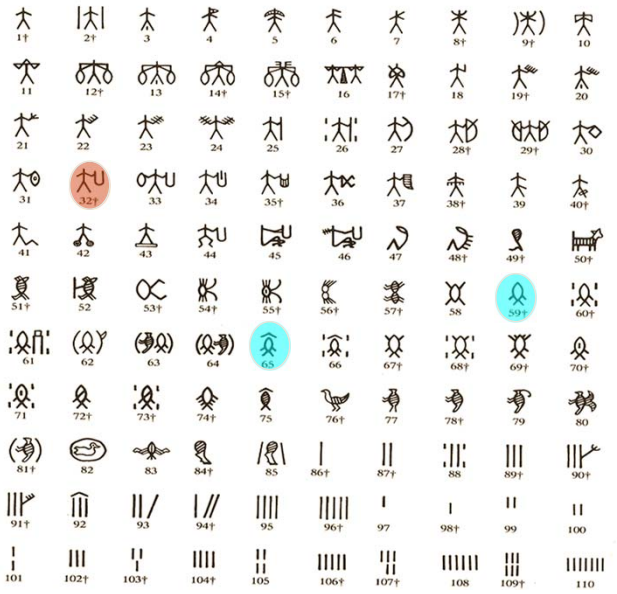
SEVERAL PAST ATTEMPTS

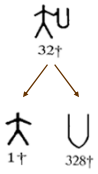
- Researchers from various disciplines have been attempting to read the Indus script with no clear answer.
- The sign system of Indus culture remains ambiguous, with contested claims of decipherment, but no consensus on any of them.

OUR APPROACH

- We make no assumptions about the nature, content or purpose of the Indus script.
- We use computational tools and techniques that can probe specific aspects of various types of data.
- The objective of our study is to identify the structure and nature of a collection of written material especially when the background knowledge is not enough.
- The present approach can be used to identify the syntactic framework for the Indus script which can be used to evaluate various claims of decipherment.

INDUS SCRIPT SIGNS (1 TO 110 OUT OF 417)





From Mahadevan (1977)

NUMBER OF SIGNS AND SCRIPT TYPE

S. No	Type of Scripts	Type of Signs	No. of Signs	Examples
1.	Logographic	Word-signs	Thousands	Chinese
2.	Logo-syllabic	Word-signs & Phonetic syllables	900-400	Sumerian, Egyptian
3.	Syllabic	(a) Closed & Open syllables (b) Open syllables	200-100 100-40	Elamite, Cuneiform Linear B, Old Persian
4.	Alphabetic	Single-sound signs	Below 40	Semitic, Greek, Latin

DIRECTION OF INDUS SCRIPT

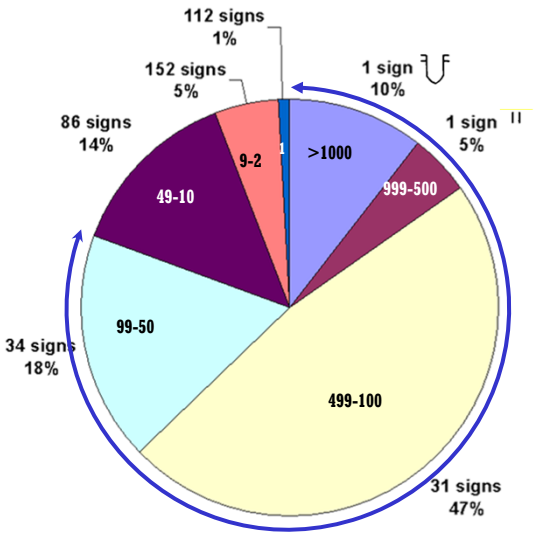
- Cramping of signs towards the left end of objects
- Overflow of signs at the left end of objects
- Gap towards the right end of objects

All this is indicative of right to left direction (83% of times), though there are a few exceptions.

Mahadevan (1977), Parpola (1992)

COMPUTATIONAL STUDIES OF INDUS SCRIPT

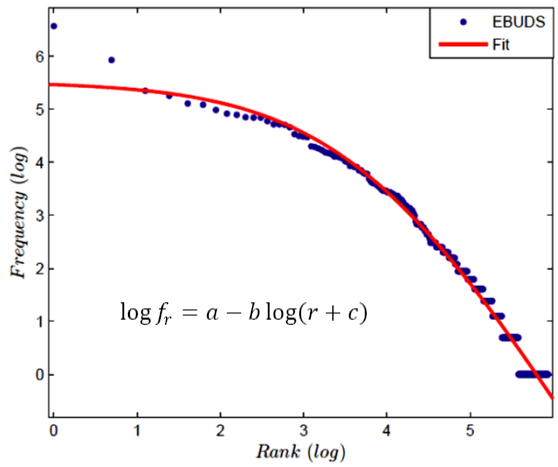
SIGN FREQUENCY DISTRIBUTION



Only 67 out of 417 signs account for over 80% of the writing.

Yadav et al. 2010 (PLoS One)

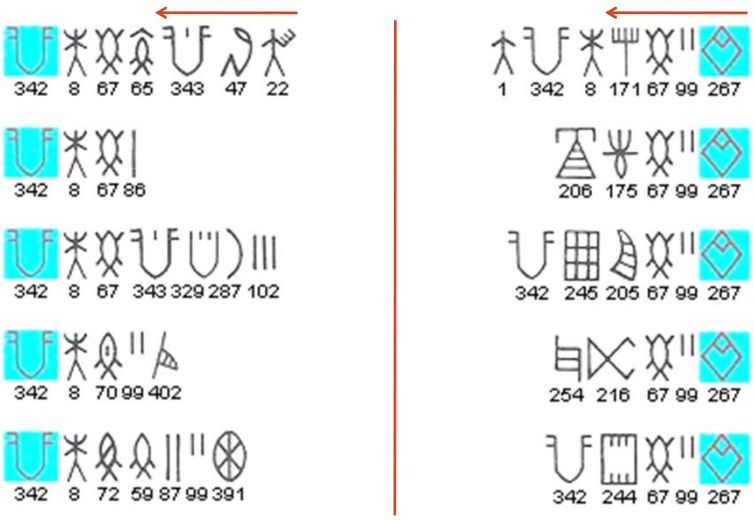
SIGN FREQUENCY DISTRIBUTION



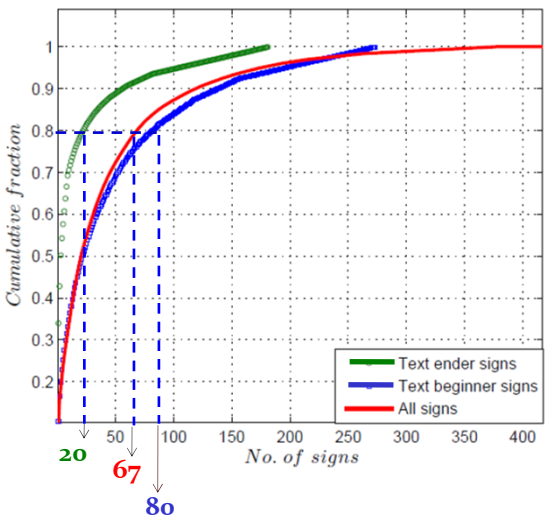
Sign frequency distribution follows Zipf-Mandelbrot law.

Yadav et al. 2010 (PLoS One)

TEXT BEGINNERS & TEXT ENDERS



BEGINNER ENDER ASYMMETRY



S5 S4 S3 S2 S1

80 signs: 80% of all text beginners

20 signs: 80% of all text enders

Yadav et al. 2010 (PLoS One)

Beginner-ender asymmetry provides clues about direction and presence of structure.

COMPARISON WITH RANDOMISED SEQUENCES

- The first question we wanted to answer was:

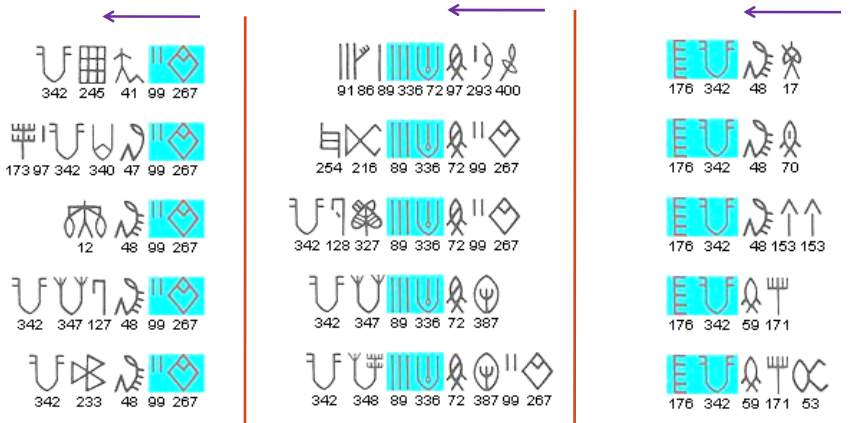
  "Are Indus texts just a collection of randomly ordered signs or is there any sequencing?"

- We shuffled the signs in the Indus texts and checked how often we get similar sequences by chance (in the randomized dataset) in comparison with the real Indus dataset.

- We found that the *sign sequences of 2, 3 or 4 signs appear far more frequently in the Indus dataset than what is expected by chance.*

Yadav et al. 2008a

POSITIONAL DISTRIBUTION OF SIGN PAIRS



The frequent sign sequences have preferred location in the Indus texts.

Yadav et al. 2008a

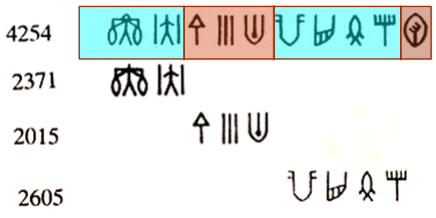
POSITIONAL ANALYSIS OF PAIRS

Two-sign Combination		Solo (%)	Left (%)	Middle (%)	Right (%)
		0.60	1.79	11.90	85.71
		0.00	0.00	89.33	10.67
		0.00	96.61	3.39	0.00
		1.72	72.41	25.86	0.00
		0.00	0.00	8.93	91.07
		0.00	89.29	10.71	0.00
		0.00	89.58	10.42	0.00
		0.00	0.00	0.00	100.00
		0.00	0.00	79.49	20.51
		2.63	52.63	28.95	15.79
		0.00	0.00	80.56	19.44

Yadav et al. 2008a

SEGMENTATION OF INDUS TEXTS

Comparing longer and shorter texts on different objects.



SEGMENTATION OF INDUS TEXTS

By using frequent sign sequences.

2476 

2168 

EXAMPLES OF SEGMENTATION

Object No.	Segments of Text						
1232:	P148	P86					
	 342 48	 99 267					
	1279	4441					
4254:	P53	T148	P116	PM9	389		
	 14 2	 211 89 336	 342 341	 59 171	 389		
	2371	2015	1226				
2537:	P41	PM14	67	PM9	389	344	PB1
	 342 149	 130 51	 67	 59 171	 389	 344	 123293
	8001		1093			4385	
2461:	T94	326	87	P131	178		
	 245 245 25	 326	 87	 15 389	 178		
	1437	2673	4560	2682			

Yadav et al. 2008b

COMPUTATIONAL MODEL OF INDUS SCRIPT

- Research in machine learning and data mining has led to new techniques for developing statistical models of sequences.
- These models are not sensitive to the semantic content of the sequences but, reveal the syntax, if any, that the sequences follow.
- These models can be used for pattern recognition and pattern completion.
- We created a statistical model of the Indus script.

N-GRAM & MARKOV MODELS

- Are probabilistic models which provide a very useful method of modelling different types of sequences.
- These models are not sensitive to the semantic content of the sequences but, reveal the syntax, if any, that the sequences follow.
- The order of the Markov model decides the length of correlation.
- The corpus of the Indus script was modelled with n-grams starting with n = 1 to n = 5, where n defines the length of correlation.
- The *perplexity* is reduced considerably when bigram correlations are taken into account.
- We study the bigram model of the Indus script.

HOW CAN WE MODEL ENGLISH ALPHABET IN BIGRAM?

S = {A, B, C, … , M, N, … , Z} ➔ 26 States

P =

sign	A	B	…	Z
A	$p_{1,1}$	$p_{1,2}$	…	$p_{1,26}$
B	$p_{2,1}$	$p_{2,2}$	…	$p_{2,26}$
…	…	…	…	…
Z	$p_{26,1}$	$p_{26,2}$	…	$p_{26,26}$

➔Transition matrix (26 × 26)

$\Pi = \{ p(X_0 = A), p(X_0 = B), \dots, p(X_0 = Z) \}$ ➔ Ini. state prob. vector

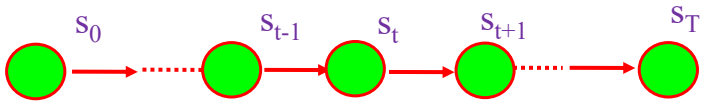
EXAMPLES FROM ENGLISH

- At the level of individual word, with each letter as an independent token, we find that
  - the letter 't' can be succeeded by lot of letters such as 'a', 'e', 'o', but not 'x' or 'z',
  - the letter 'q' usually succeeded by 'u' and so on.
- Similarly, if the tokens are words, then we find that the word 'the' can be succeeded by a large number of words, but not verbs such as 'eat'.

INDUS TEXTS ARE MAPPED TO SEQUENCE OF STATES

Indus Text

Sequence of states



BIGRAM MODEL OF THE INDUS SCRIPT

S = {  } → 417 States

P =

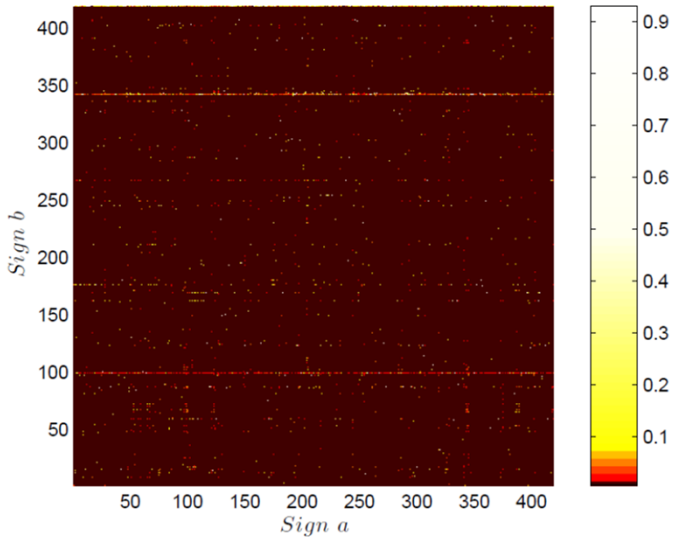
sign			...	
	$p_{1,1}$	$p_{1,2}$	⋯	$p_{1,417}$
	$p_{2,1}$	$p_{2,2}$	⋯	$p_{2,417}$
...	⋯	⋯	⋯	⋯
	$p_{417,1}$	$p_{417,2}$	⋯	$p_{417,417}$

→ Transition matrix (417 × 417)

Π = {p(X_0 =), p(X_0 =), … , p(X_0 =)}

→ Initial state prob. vector

BIGRAM MATRIX OF INDUS SCRIPT



APPLICATIONS OF THE MODEL

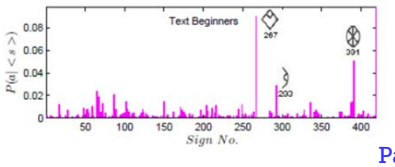
We use the probabilistic model of Indus texts to

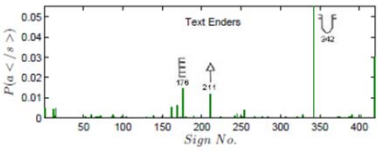
- Establish that there is an underlying order in the writing
- Fill-in damaged or illegible Indus texts
- To analyse variation in writing from different regions
- Generating Indus-like texts

ANALYSIS OF BIGRAMS

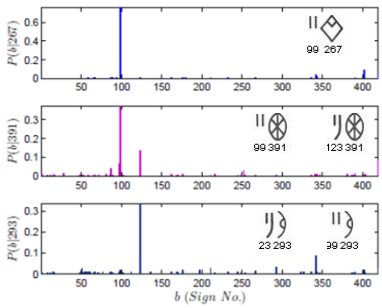
Beginners

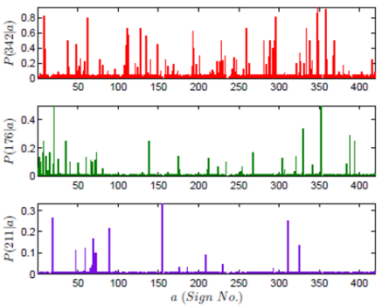
Enders





Pairs





→ Presence of signs having specific syntactic functions

RESTORATION OF TEXTS ON BROKEN SEALS

RESTORATION OF DAMAGED INDUS TEXTS

Text No.	Text	Incomplete Text
8302		
5317		
1193		
1407		
2179		
3396		
8101		

A bigram model of the Indus script can be used for restoration of damaged texts with about 75% accuracy.

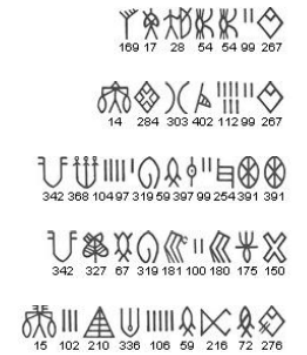
Yadav et al. 2010 (PLoS One)

MOST PROBABLE INDUS TEXTS: GENERATED BY MODEL

Text Length	Blank Text	Predicted Text	Text No.	Closest matching text from dataset
4	</s>  <s>		1232	
			2580	
5	</s>  <s>		2476	
6	</s>  <s>		1322	

Yadav et al. 2010 (PLoS One)

INDUS TEXTS GENERATED FROM MODEL



Yadav et al. (2010, PLoS One)

Courtesy: Harappa.com

INDUS SCRIPT BUT DIFFERENT GRAMMAR

Impression of a round stamp seal from West Asia

Sequence 

never occurs in Indus corpus of about 4000 texts from Harappan sites.

USE OF INDUS SCRIPT TO WRITE WEST ASIAN CONTENT?

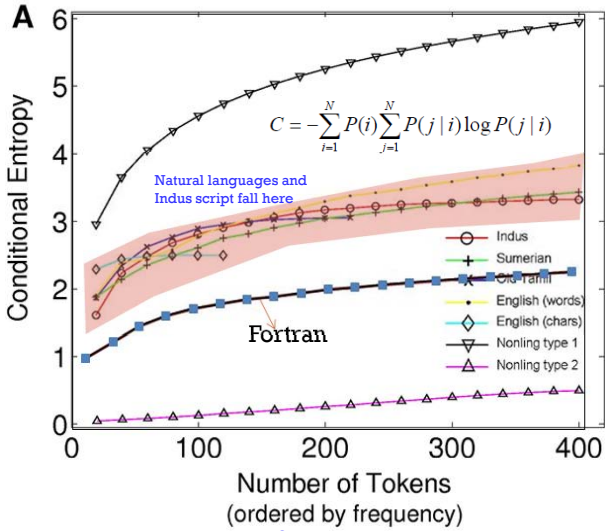
Indus text	 342 51 336 99 267	2.8×10^{-5}
Altered text	 336 51 99 267 342	~ 0
Indus seal	 342 53 345 297 150	1.4×10^{-6}
West Asian seal	 342 342 97 53 178	5.6×10^{-11}

West Asian seal text is approximately 100,000 times unlikely to be generated by the learned model of Indus texts.

COMPARISON WITH OTHER SIGN SYSTEMS

- In general, ordering of signs in a written text can be
  - random where the order of signs is not important
  - rigid where the ordering is precise and pre-determined
  - flexible where there exists *a degree of freedom* in choosing the order to allow expression of a variety of information.
- We compared the conditional entropy of Indus sign system with other linguistic and non-linguistic systems.
- Conditional entropy quantifies the *amount of flexibility* in choosing a sign given a preceding sign.

COMPARISON WITH OTHER SIGN SYSTEMS



Indus : Mahadevan (1977)

English : Brown Corpus

Sanskrit : Rig Veda

Old Tamil : Ettuthokai

Sumerian : Oxford Corpus

DNA : Human Genome

Protein : E. Coli

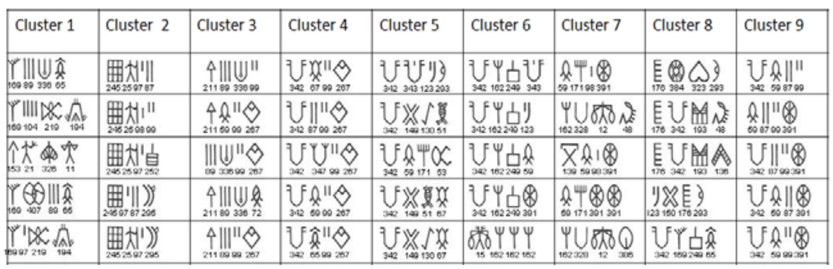
Fortran : FEM code

Conditional entropy of Indus script is close to linguistic systems.

Rao, Yadav et al. (2009, Science)

IDENTIFYING CLUSTERS IN INDUS TEXTS

- The associations between the signs in the Indus texts were explored using *clustering*, an unsupervised machine learning technique.
- Nine clusters of Indus texts were extracted such that the texts in each of these clusters were more similar amongst themselves than to texts belonging to other clusters.



Yadav et al. 2017

COMPARING CLUSTERS OF INDUS TEXTS

- The nine clusters were found to have their signature set of signs and sign sequences.
- The text clusters were not found to have any significant correlation to the sites of discovery or object types.
- The study suggested that the Indus writing had clusters of texts corresponding to distinct styles or contents.

A FRAMEWORK TO UNDERSTAND THE INDUS SCRIPT

Site of occurrence: Mohenjodaro, Harappa, Lothal, Kaliangan etc.

Context of occurrence: Stratigraphy, Location such as residence, workshop, street etc.

Indus Script

Structural studies:

1) Sign frequency
2) Sign sequencing
3) Text restoration
4) Sign design etc.

Object type: Seal, Sealing, Copper tablets, Clay tags, Pottery, Ornaments etc.

Relation with other components: Field symbol , Artistic expressions etc.

Yadav 2013 53

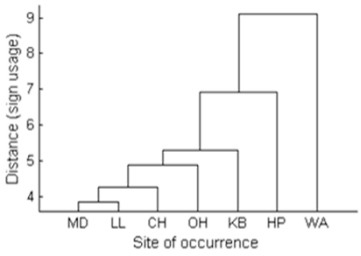
Source: http://pubweb.cc.u-tokai.ac.jp/indus/english/map.html

COMPARISON OF SIGN USAGE: SITES AND OBJECT TYPES

- One of the characteristics of writing is localized variations based on specific needs of region/context.
- We used hierarchical clustering to compare the usage of signs across various sites and types of objects.
- Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.
- Variation in the usage of signs on distinct types of objects and sites was studied.

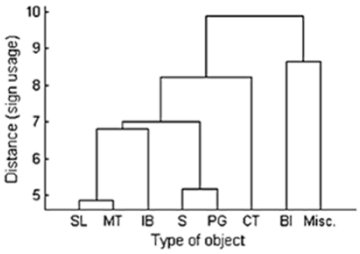
COMPARISON OF SIGN USAGE

Similarity of Sites



MD: Mohenjodaro
HP: Harappa
Ll: Lothal
CH: Chanhudaro
KB: Kalibangan
OH: Other sites
WA: West Asian sites

Similarity of Medium

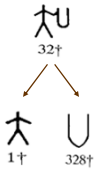


S: Seal
SL: Sealing
CT: Copper Tablets
MT: Miniature Tablets
PG: Pottery Graffiti
BI: Bronze Implements
IB: Ivory or bone rods

Yadav 2013

DESIGN OF INDUS SIGNS

INDUS SCRIPT SIGNS (1 TO 110 OUT OF 417)



Mahadevan (1977)

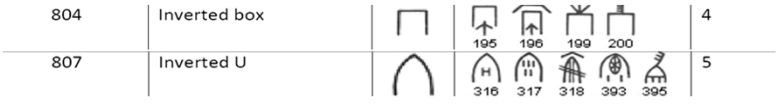
ANALYSING THE DESIGN OF INDUS SIGNS

- The designs of Indus signs have varying levels of complexity.
- We identify the design elements of the Indus signs and decompose the composite signs using them.
- We then studied the usage pattern of the compound signs to check if they were created to save writing space or they had some other function in the writing.

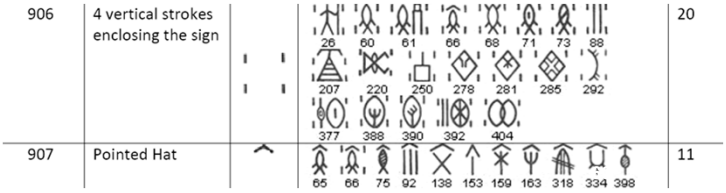
DESIGN ELEMENTS OF INDUS SIGNS

1. Basic signs (154)

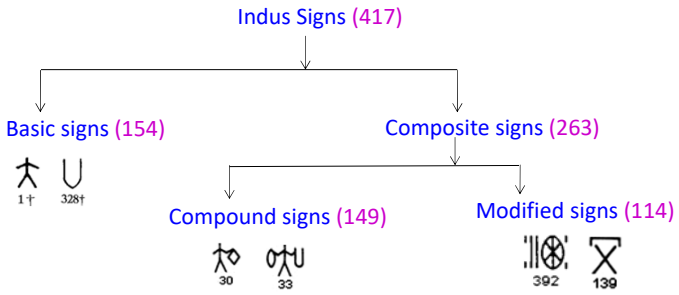
2. Provisional basic signs (10)



3. Modifiers (21)



CLASSIFICATION OF INDUS SIGNS



STUDY OF COMPOUND SIGNS

Compound sign	Freq.	Components			Sign combination found in M77	Freq.	Comments
21	2	1	162		1 162 162 1	2 1	No signs common on either side in both the cases.
22	3	1	176		176 1 1 176	6 1	No signs common on either side in both the cases.

- In *most* of the cases, the constituent elements in any combination and the composite signs do not share identical environment.
- The compound signs are not merely compacted version of the possible sign sequences created for brevity. They seem to have some other function in the Indus texts.

Yadav and Vahia (2011)

SUMMARY

- The script has a rich syntax with an underlying logic in its structure.
- There is a significant asymmetry in the usage of text beginners and text enders.
- Machine learning techniques enable us to identify syntactically valid writing and restore damaged Indus texts.
- The script may have been used for writing West Asian content.
- The flexibility of sign usage in Indus texts falls within the range of linguistic systems.
- The designs of Indus signs are intricate and sign compounding seem to add value to basic signs rather than save writing space.
- The texts can be optimally subdivided into nine distinct clusters using unsupervised machine learning techniques.
- The results provide significant constraints to any model of decipherment and can be used to evaluate the same.

THANK YOU