Markov chain model for the Indus script

Ronojoy Adhikari
The Institute of Mathematical Sciences
Chennai

Outline

- Statistical models for language.
- The Indus civilisation and its script.
- Difficulties in decipherment.
- A Markov chain model for the Indus script.
- Statistical regularities in structure.
- Evidence for linguistic structure in the Indus script.
- Applications

Collaborators

References

- "Entropic evidence for linguistic structure in the Indus script", Rajesh P. N. Rao, Nisha Yadav, Hrishikesh Joglekar, Mayank Vahia, R. Adhikari, Iravatham Mahadevan, Science, 24 April, 2009.

- "Markov chains for the Indus script", Rajesh P. N. Rao, Nisha Yadav, Hrishikesh Joglekar, Mayank Vahia, R. Adhikari, Iravatham Mahadevan, PNAS, 30 Aug, 2009.

- "Statistical analysis of the Indus script using n-grams", Nisha Yadav, Hrishikesh Joglekar, Rajesh P. N. Rao, Mayank Vahia, R. Adhikari, Iravatham Mahadevan, Plos One under review (arxiv.org/0901.3017)

- Featured in Physics Today, New Scientist, Scientific American, BBC Science in Action, Nature India and in other news media.

- http://indusresearch.wikidot.com/script

Disclaimer

We have not deciphered the script!

Statistical properties of language : al Kindi

source : wikipedia

"One way to solve an encrypted message, if we know its language, is to find a different plaintext of the same language long enough to fill one sheet or so, and then we count the occurrences of each letter. We call the most frequently occurring letter the ‘first', the next most occurring letter the ‘second', the following most occurring the ‘third', and so on, until we account for all the different letters in the plaintext sample".

"Then we look at the cipher text we want to solve and we also classify its symbols. We find the most occurring symbol and change it to the form of the ‘first' letter of the plaintext sample, the next most common symbol is changed to the form of the ‘second' letter, and so on, until we account for all symbols of the cryptogram we want to solve" - "A Manuscript on Deciphering Cryptographic Messages" (~800 CE)

al Kindi noted that language has statistical regularities in terms of letters.

He also introduced the Indian numerals and methods calculation to the Arab world.

Statistical properties of language : Zipf

Ranked frequency of words

Rank

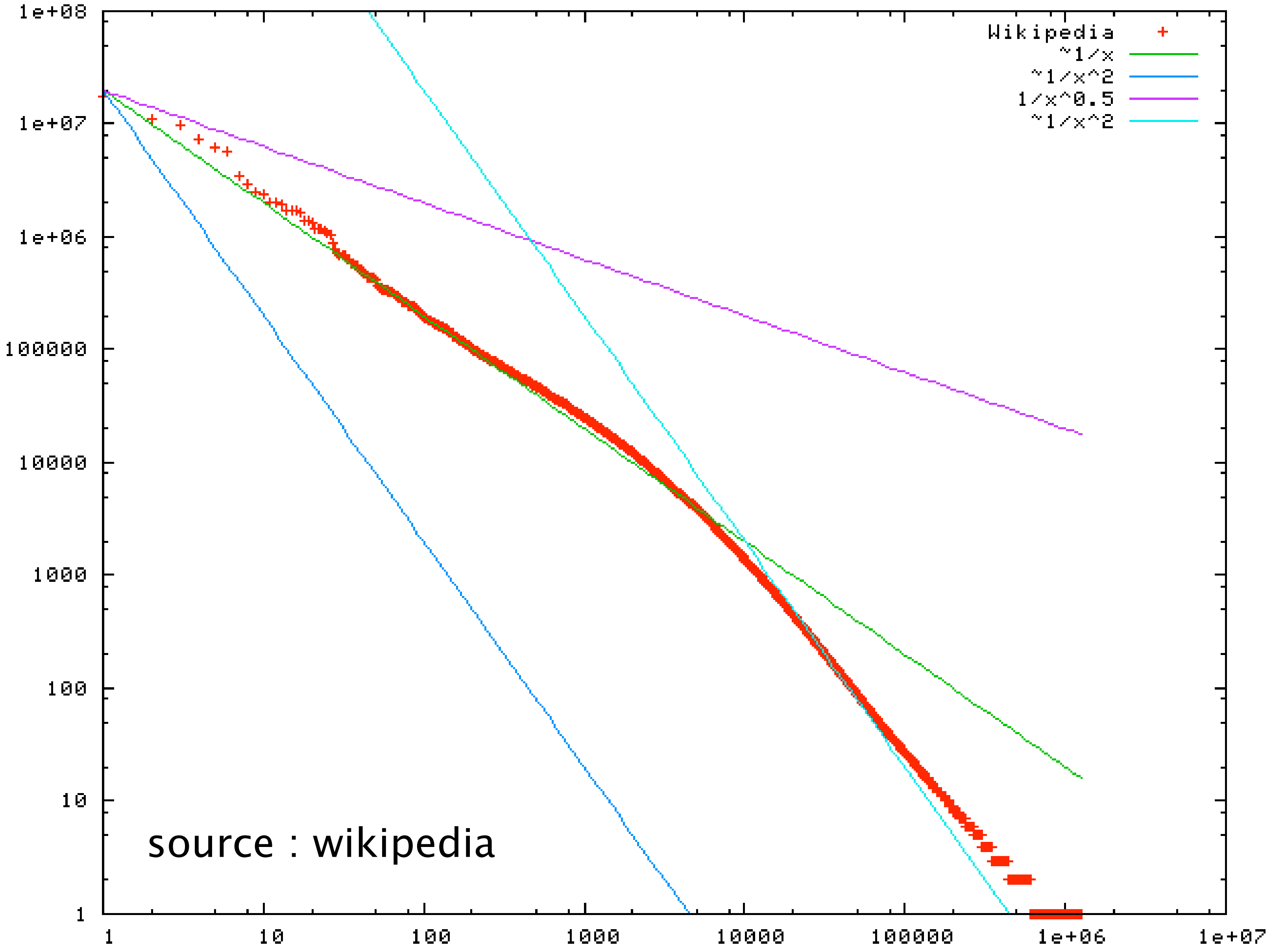
$$f_r \sim \frac{1}{r}$$

For the Brown Corpus

r = 1 : “the”
r = 2 : “and”
r = 3 : “of”
……

$$\log f_r = a - b \log(r + c)$$

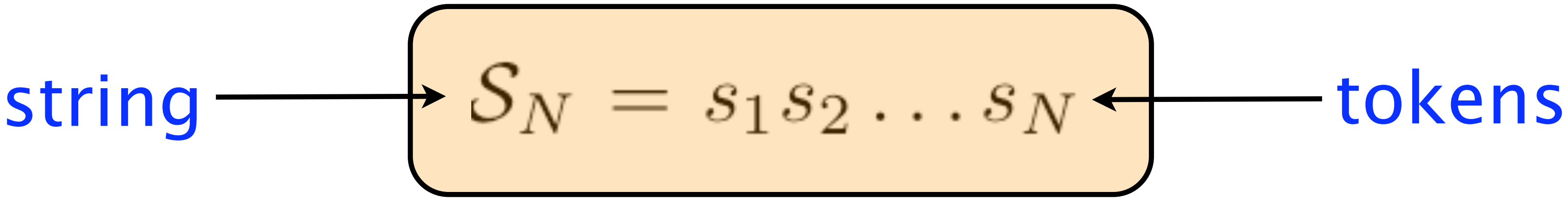
Zipf-Mandelbrot law



source : wikipedia

For the “Wikipedia Corpus”

Markov chains and n-grams



string → $S_N = s_1 s_2 \ldots s_N$ ← tokens

Andrei Markov was a founder of the theory of stochastic processes.

letter sequences

markov = m|a|r|k|o|v

word sequences

to be or not to be = to|be|or|not|to|be

tone sequences

doe a deer = DO|RE|MI|DO|MI|DO|MI|

many other examples can be given.

Unigrams, bigrams, ... n-grams.

unigrams	$P(s)$
bigrams	$P(s_1 s_2)$
trigrams	$P(s_1 s_2 s_3)$
n-grams	$P(s_1 s_2 s_3 \ldots s_N)$

$$P(s_1 s_2) = P(s_2 | s_1) P(s_1)$$

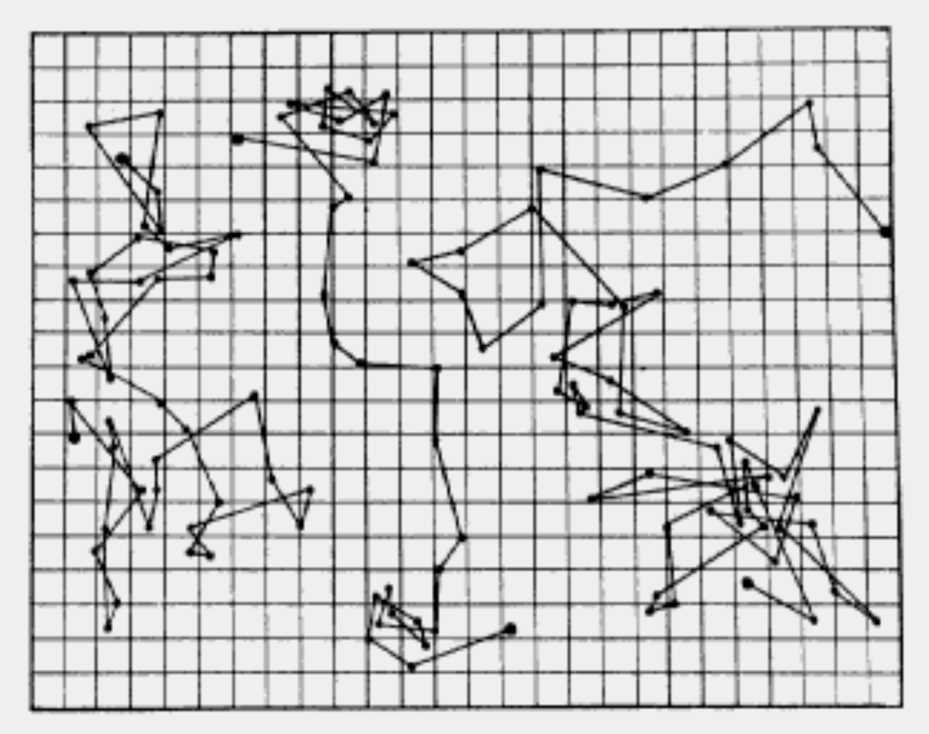
conditional probabilities

$$P(s_N | s_{N-1} \ldots s_1) = P(s_N | s_{N-1})$$

$$\begin{aligned} P(s_1 s_2 \ldots s_N) &= P(s_N | s_{N-1}) \\ &\times P(s_{N-1} | s_{N-2}) \\ &\vdots \\ &\times P(s_2 | s_1) \\ &\times P(s_1) \end{aligned}$$

A first-order Markov chain approximation to a sequence of tokens, in terms of bigram conditional probabilities.

Markov processes in physics



source : wikipedia

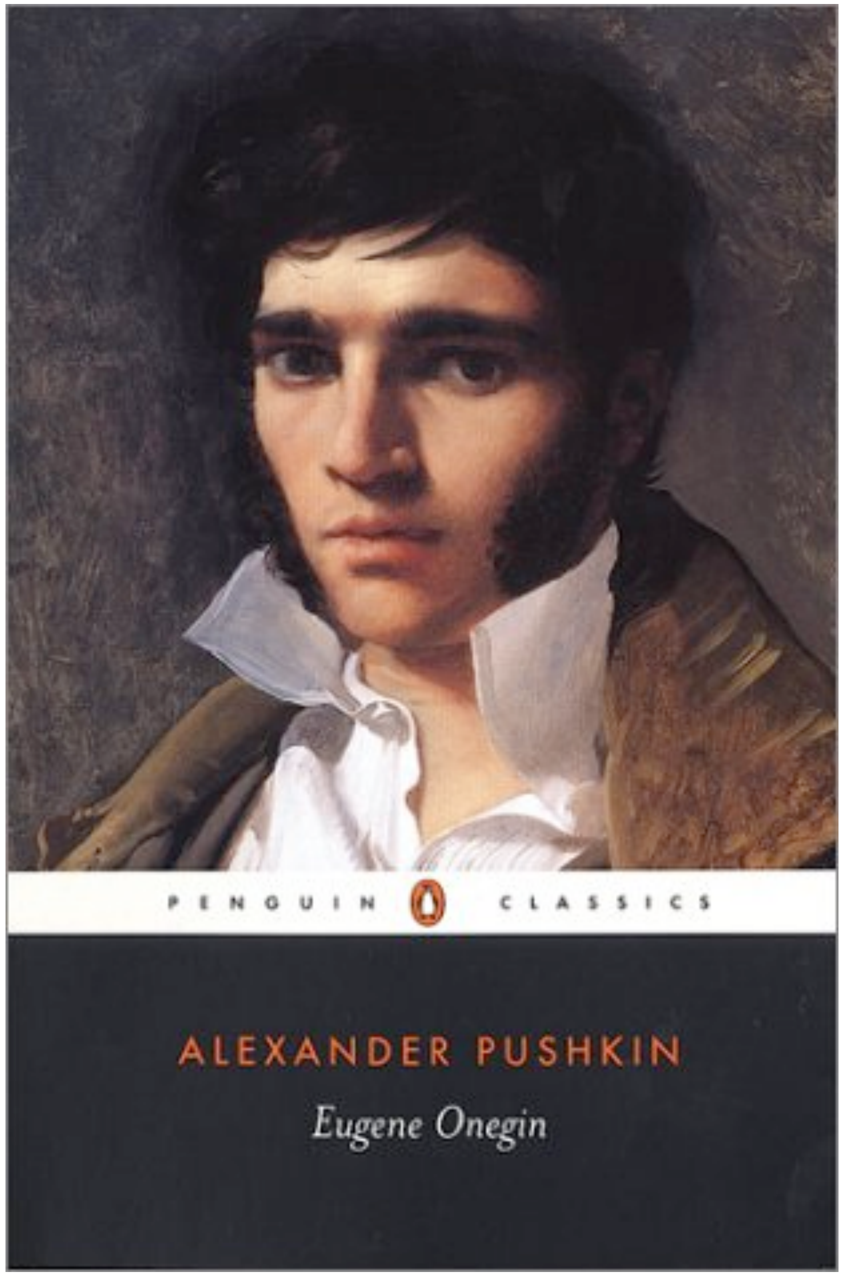
Brownian motion : Einstein (1905)

$$P(x_1, x_2, \dots, x_N) = P(x_N | x_{N-1}) \dots P(x_2 | x_1) P(x_1)$$

$$P(x' | x) = \frac{1}{\sqrt{2\pi D\tau}} \exp\left[\frac{-(x' - x)^2}{2D\tau}\right]$$

We have no “microscopic” model for language. The conditional probabilities are, therefore, empirical.

Markov chains and language : Evegeny Onegin



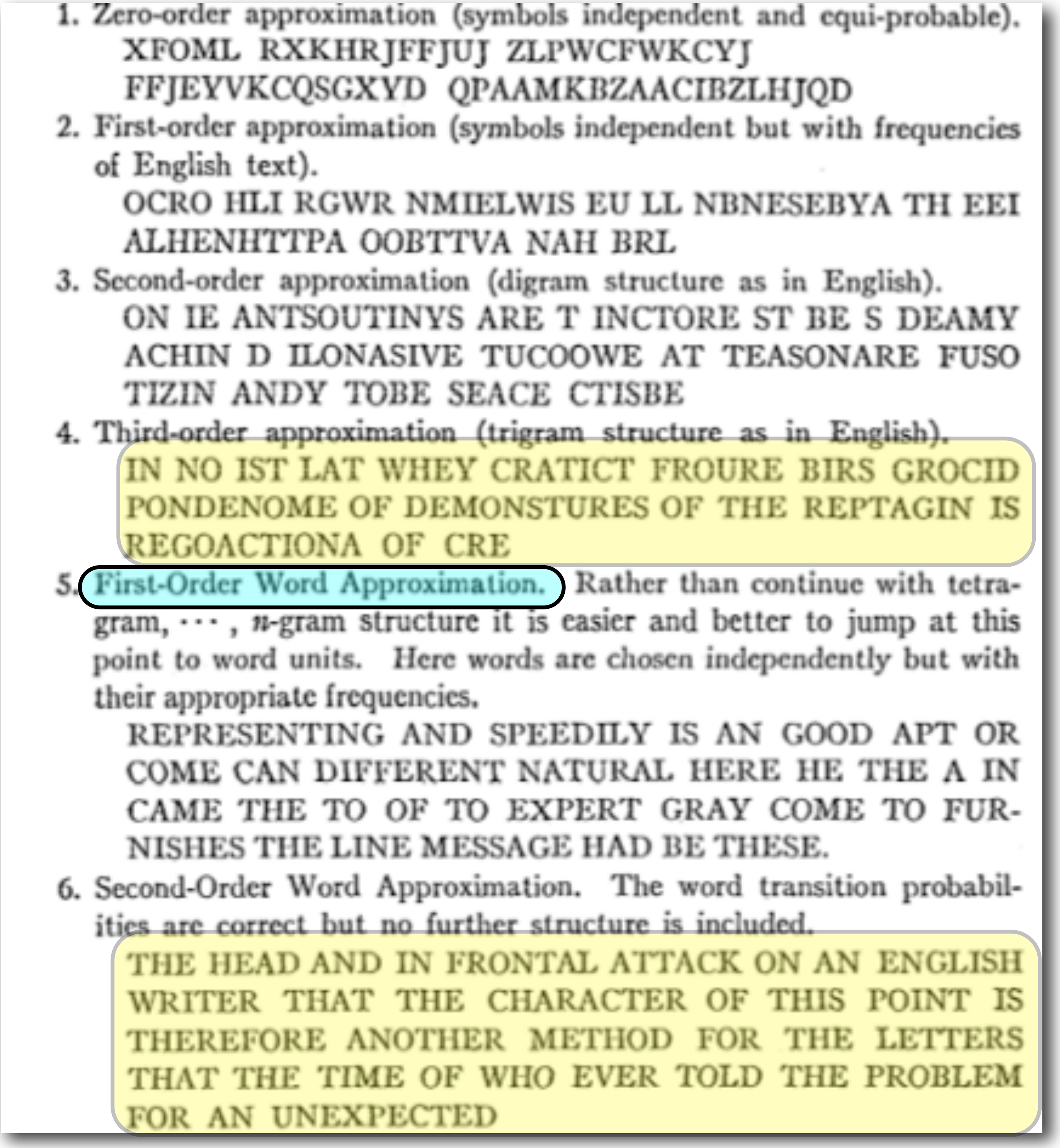
What is the probability of co-occurences of vowels and consonants ?

$$P(v|v)P(v|c)$$

$$P(c|v)P(c|c)$$

First known use in language modelling (1911)

Markov chains, n-grams and the Shannon entropy



1. Zero-order approximation (symbols independent and equi-probable).
   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
   FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD
2. First-order approximation (symbols independent but with frequencies of English text).
   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
   ALHENHTTPA OOBTTVA NAH BRL
3. Second-order approximation (digram structure as in English).
   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
   ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
   TIZIN ANDY TOBE SEACE CTISBE
4. Third-order approximation (trigram structure as in English).
   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID
   PONDENOME OF DEMONSTURES OF THE REPTAGIN IS
   REGOACTIONA OF CRE
5. First-Order Word Approximation. Rather than continue with tetragram, · · · , n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.
   REPRESENTING AND SPEEDILY IS AN GOOD APT OR
   COME CAN DIFFERENT NATURAL HERE HE THE A IN
   CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.
6. Second-Order Word Approximation. The word transition probabilities are correct but no further structure is included.
   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
   WRITER THAT THE CHARACTER OF THIS POINT IS
   THEREFORE ANOTHER METHOD FOR THE LETTERS
   THAT THE TIME OF WHO EVER TOLD THE PROBLEM
   FOR AN UNEXPECTED

Claude Shannon introduced the idea of entropy as a measure of missing information in his seminal 1948 paper on communication theory.

$$H = -\sum_a p(a) \ln p(a)$$

Fun with Markov chains

http://www.eblong.com/zarf/markov/

Dissociated Press algorithm.

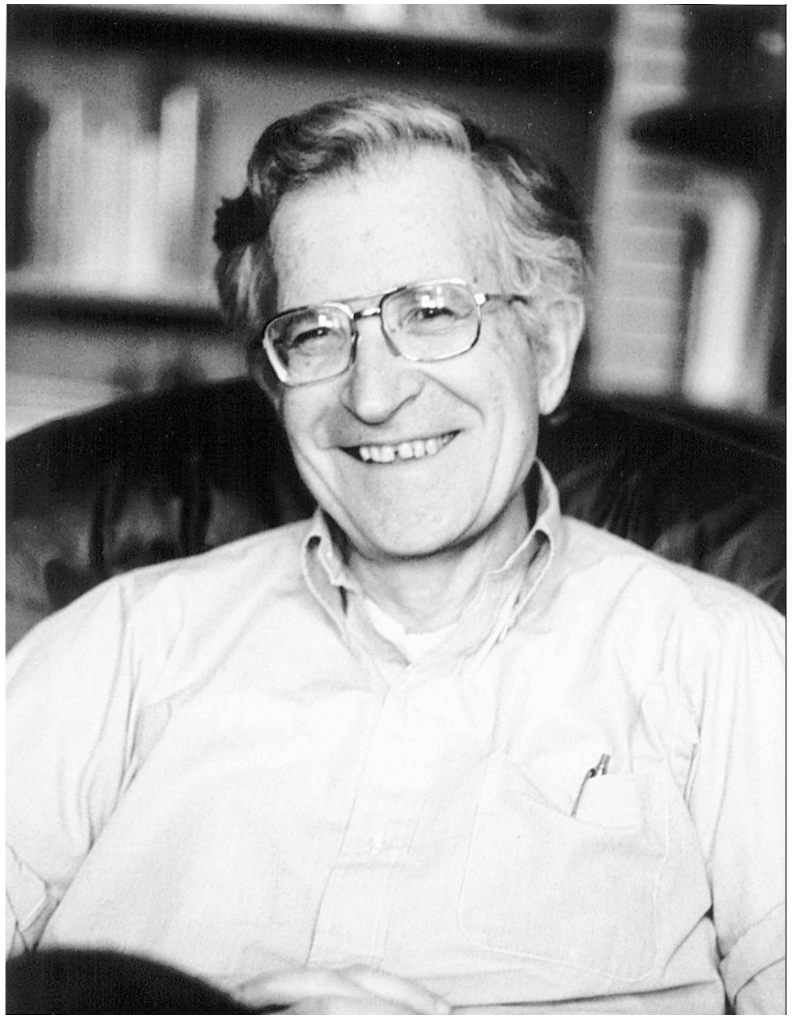
Sampling from a Markov Chain

$$
\begin{aligned}
P(s_1 s_2 \ldots s_N) &= P(s_N|s_{N-1}) \\
&\times P(s_{N-1}|s_{N-2}) \\
&\vdots \\
&\times P(s_2|s_1) \\
&\times P(s_1)
\end{aligned}
$$

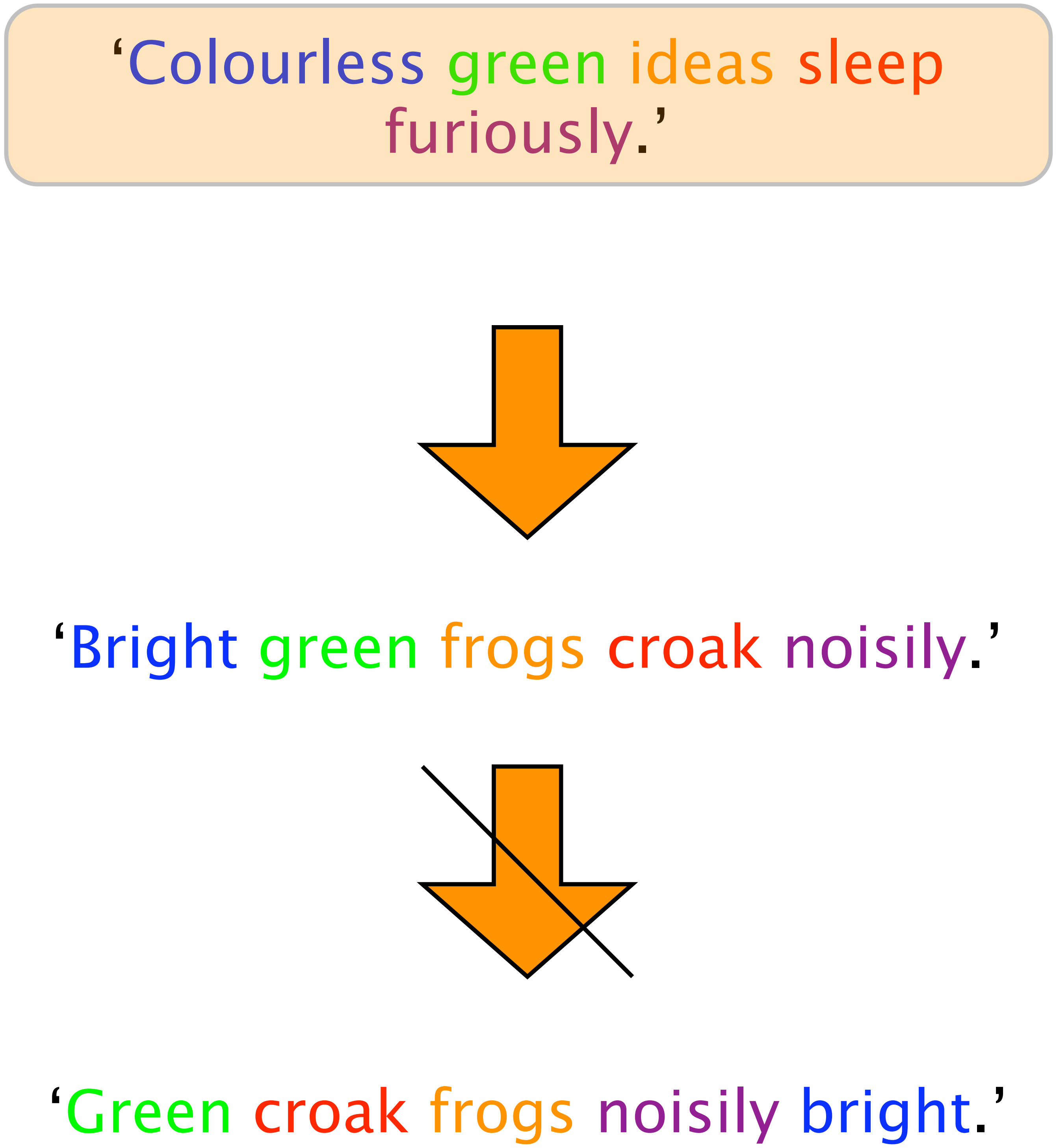
> Now And the sun rose up his father, and they said, O LORD judge betwixt us concerning that is his brother Abel. And the likeness of the field which his wife; and the God set by her.
>
> And Mahalaleel eight days of thine only bring them unto us? one that Ishmael his wife, and hath also heard thee: bury thy dead; none other but the land whereon thou fearest God, the choice of the two hundred and eat, and with him, Because the flocks by force thy face, and he ungirded his father, and behold the LORD. And she said unto his dead, and sent messengers before him forth jewels of Padanaram, for I pray thee, drink also: and Esau said, Behold, in the LORD hath given no man is life, and we said, I establish my father speak unto Zoar. Then again bare Abram and which returned that is my service which he took a wife took one that are these things, and daughters: And I give all thy brother, and Methusael begat sons and I pray thee, if now done in the same is the ground. And God went out, and the sons of Ellasar; four hundred pieces of Abram's brother's name Asher. And I pray thee. And Jared were sons of them unto my son of the LORD said unto him in the name Seth: For Sarah saw the LORD scatter again into the younger. And Enoch walked with thee a keeper of millions, and twelve princes shall thirty years, and came to pass, when he commanded Noah. http://www.toingtoing.com/?p=79

Markov Chain models can only capture syntax. They are “dumb” as far as semantics goes.

Syntax versus semantics



Noam Chomsky led the modern revolution in theoretical linguistics.



‘Colourless green ideas sleep furiously.’

‘Bright green frogs croak noisily.’

‘Green croak frogs noisily bright.’

“Nonsense” poetry.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal sword in hand:
Long time the manxome foe he sought—
So rested he by the Tumtum tree,
And stood awhile in thought.

And as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

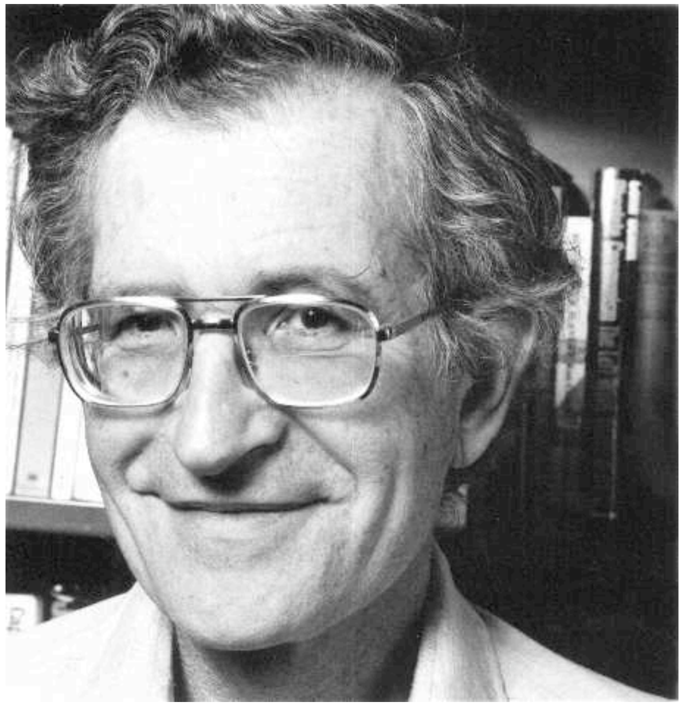
One, two! One, two! and through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went galumphing back.

"And hast thou slain the Jabberwock?
Come to my arms, my beamish boy!
O frabjous day! Callooh! Callay!"
He chortled in his joy.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

“slithy” – adjective
“gyre” – verb
.....

Markov chains for language : two views



“But it must be recognised that the notion ‘probability of a sentence’ is an entirely useless one, under any known interpretation of the term”. - Chomsky

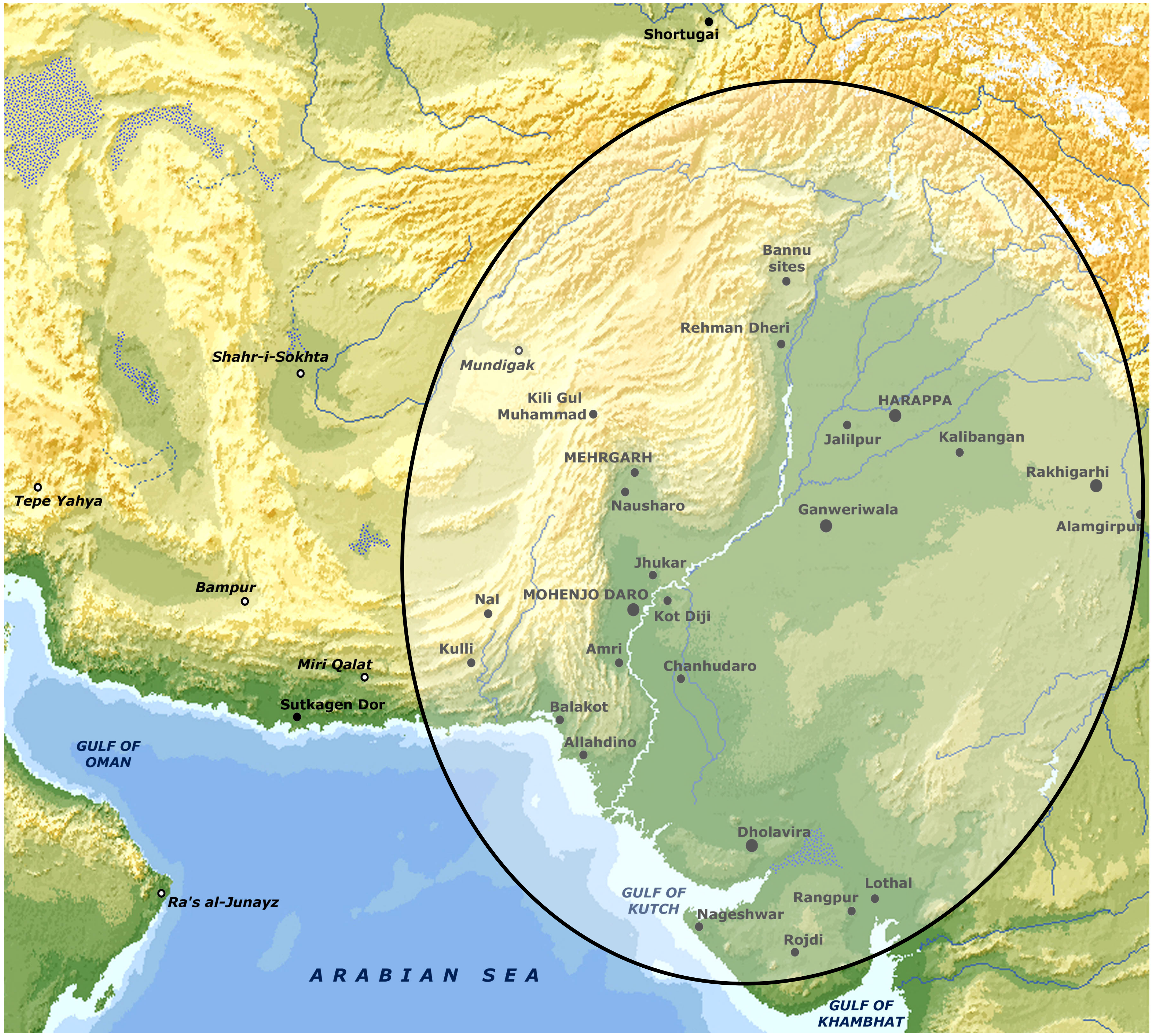
“Anytime a linguist leaves the group the recognition rate goes up”.- Jelenik

We analysed the Indus script corpus using Markov chains.

This is the first application of Markov chains to an undeciphered script.

Is it possible to infer if a sign system is linguistic without having deciphered it ?

The Indus valley civilisation



Largest river valley culture of the Bronze Age. Larger than Tigris-Euphrates and Nile civilisations put together.

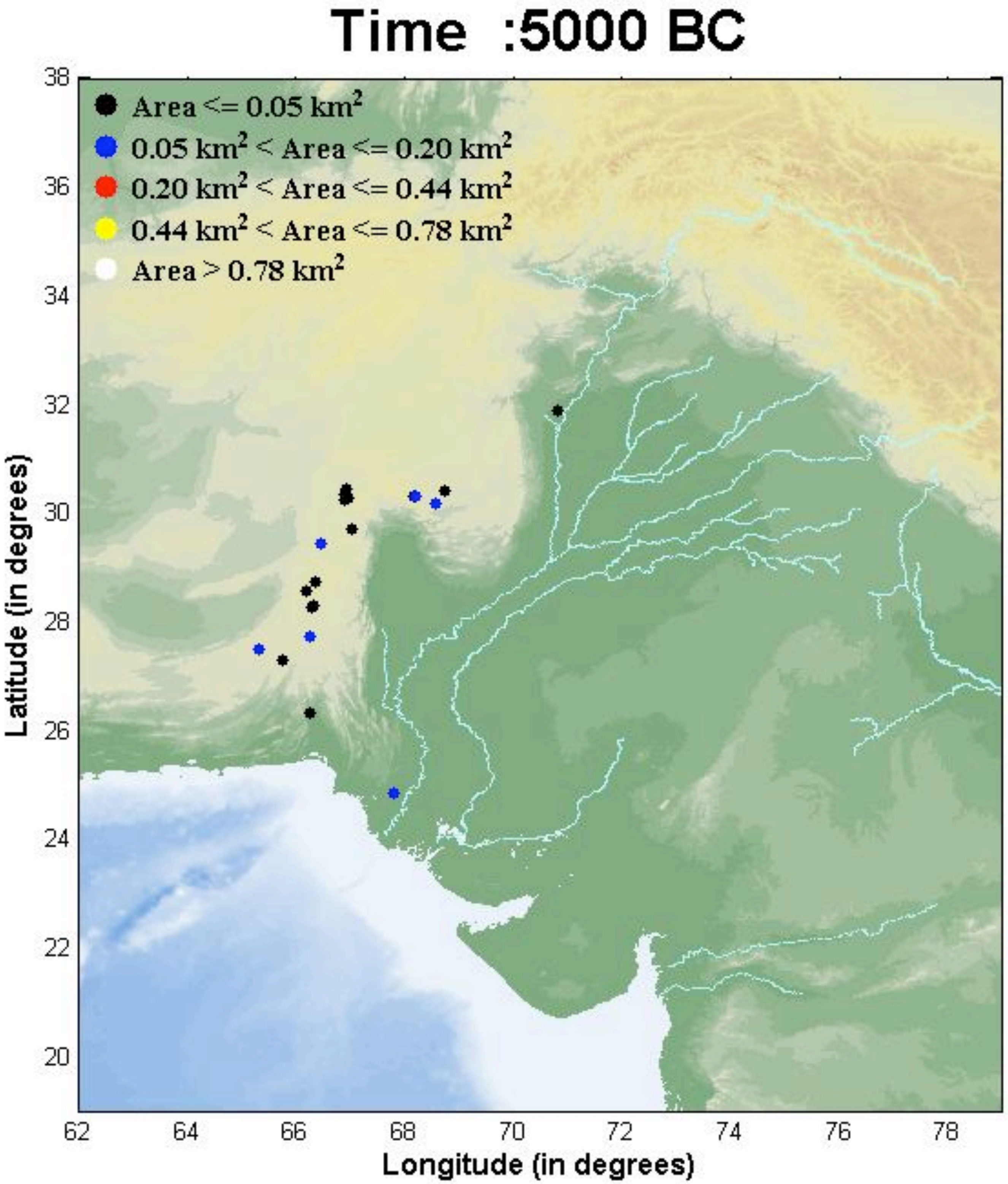
Spread over 1 million square kilometers.

Antecedents in 7000 BCE at Mehrgarh.

700 year peak between
2600 BCE and 1900 BCE.

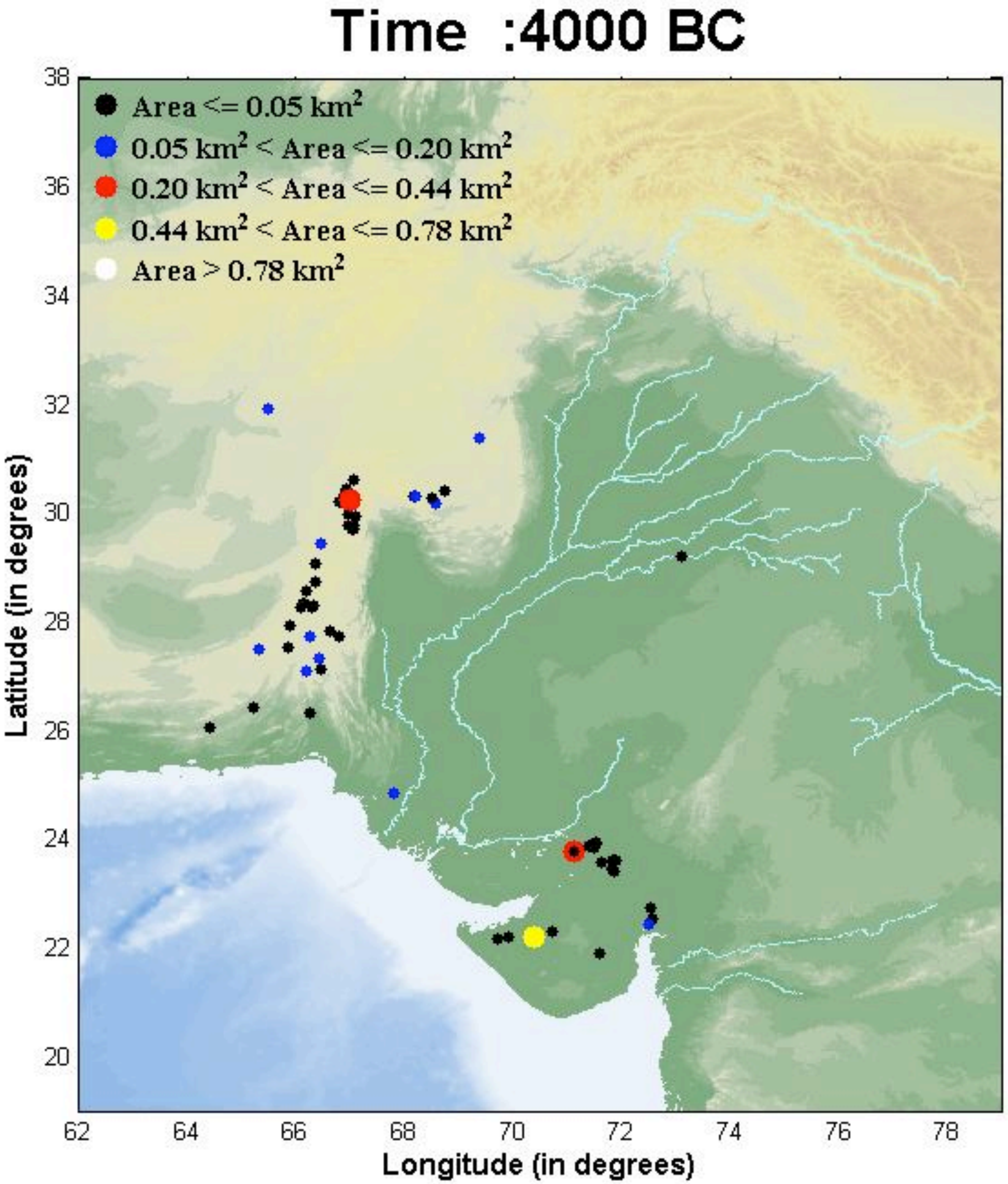
Remains discovered in 1922.

The Indus civilisation : spatio-temporal growth

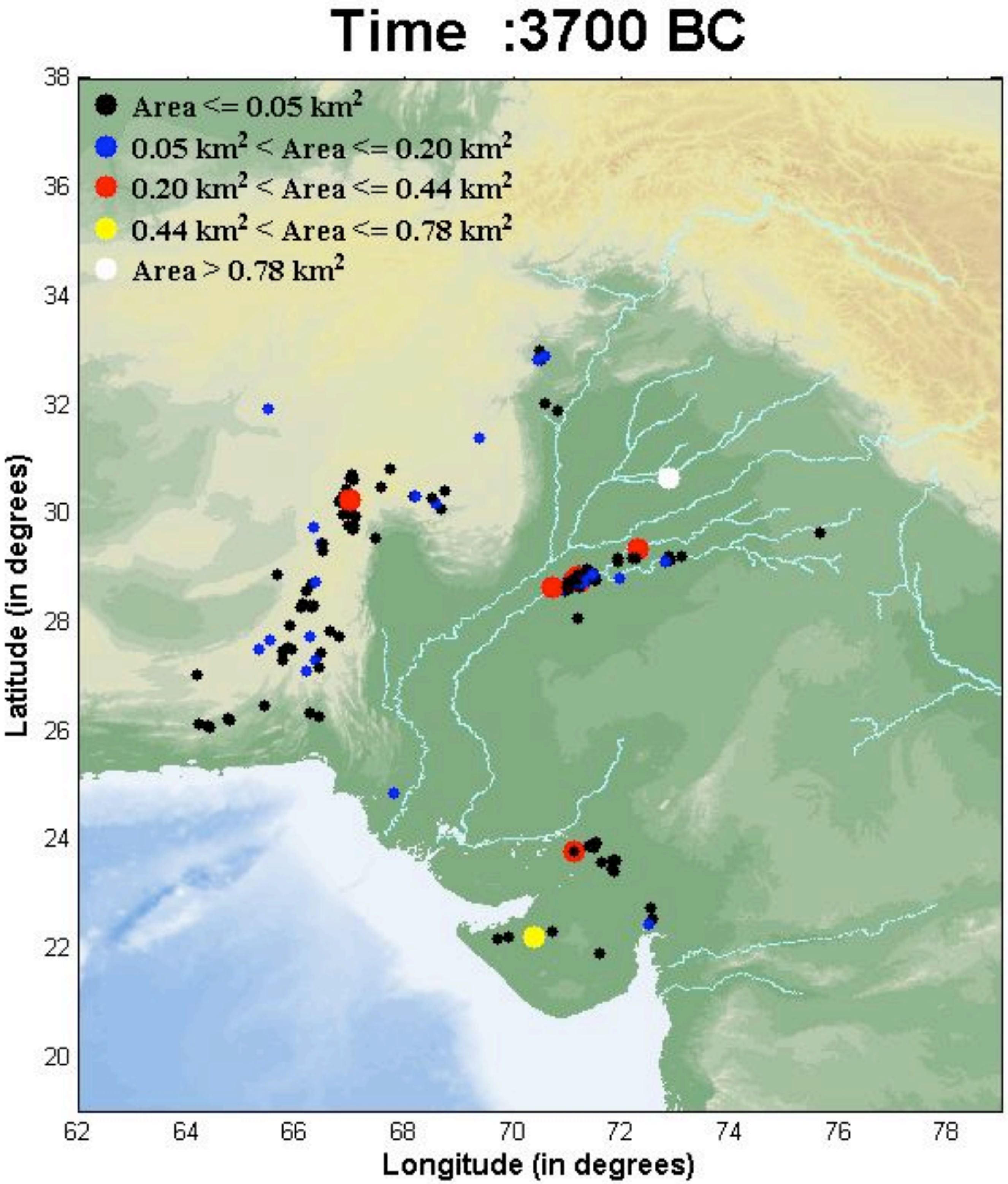


Acknowledgemen
t :
Kavita Gangal.

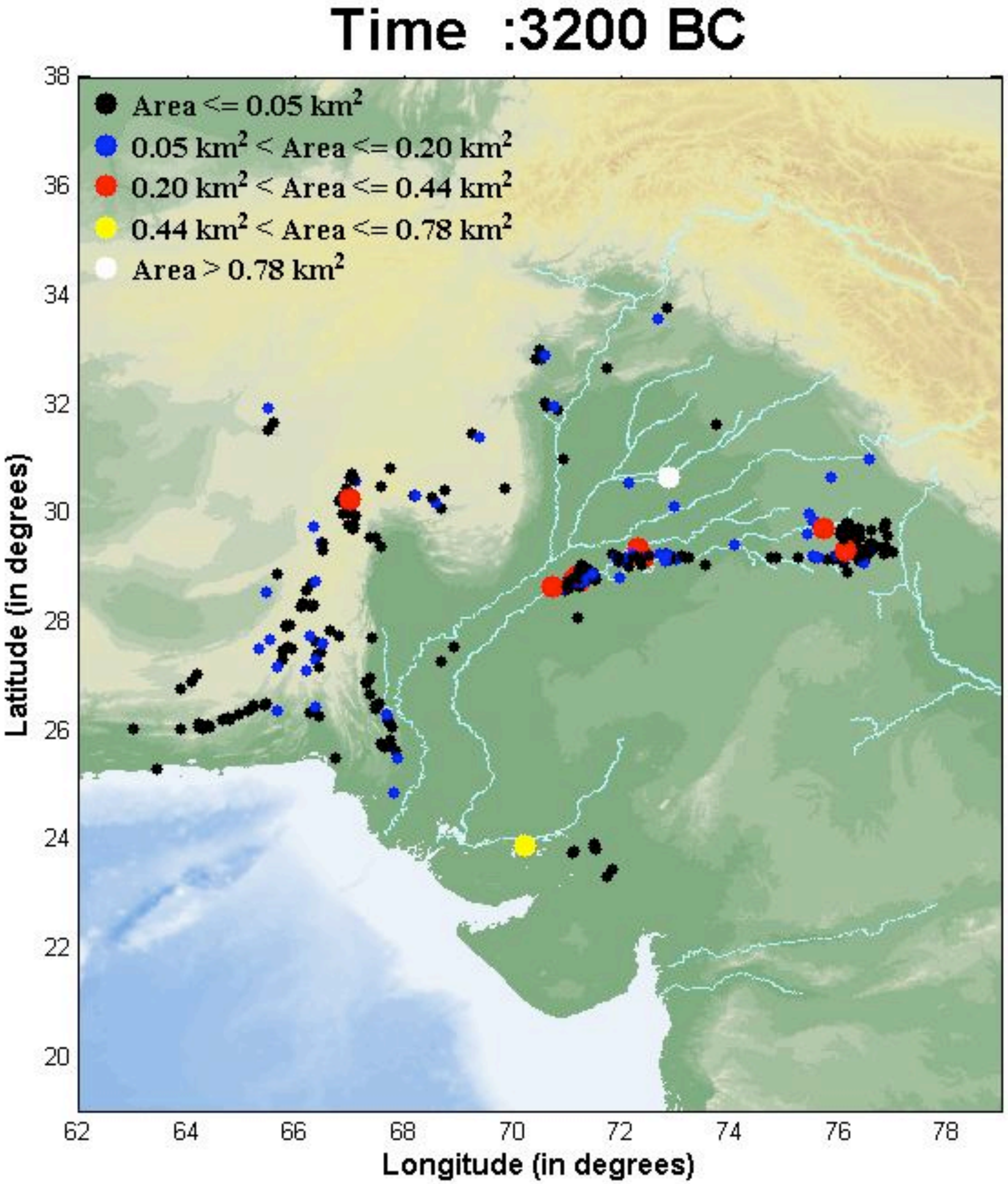
The Indus civilisation : spatio-temporal growth



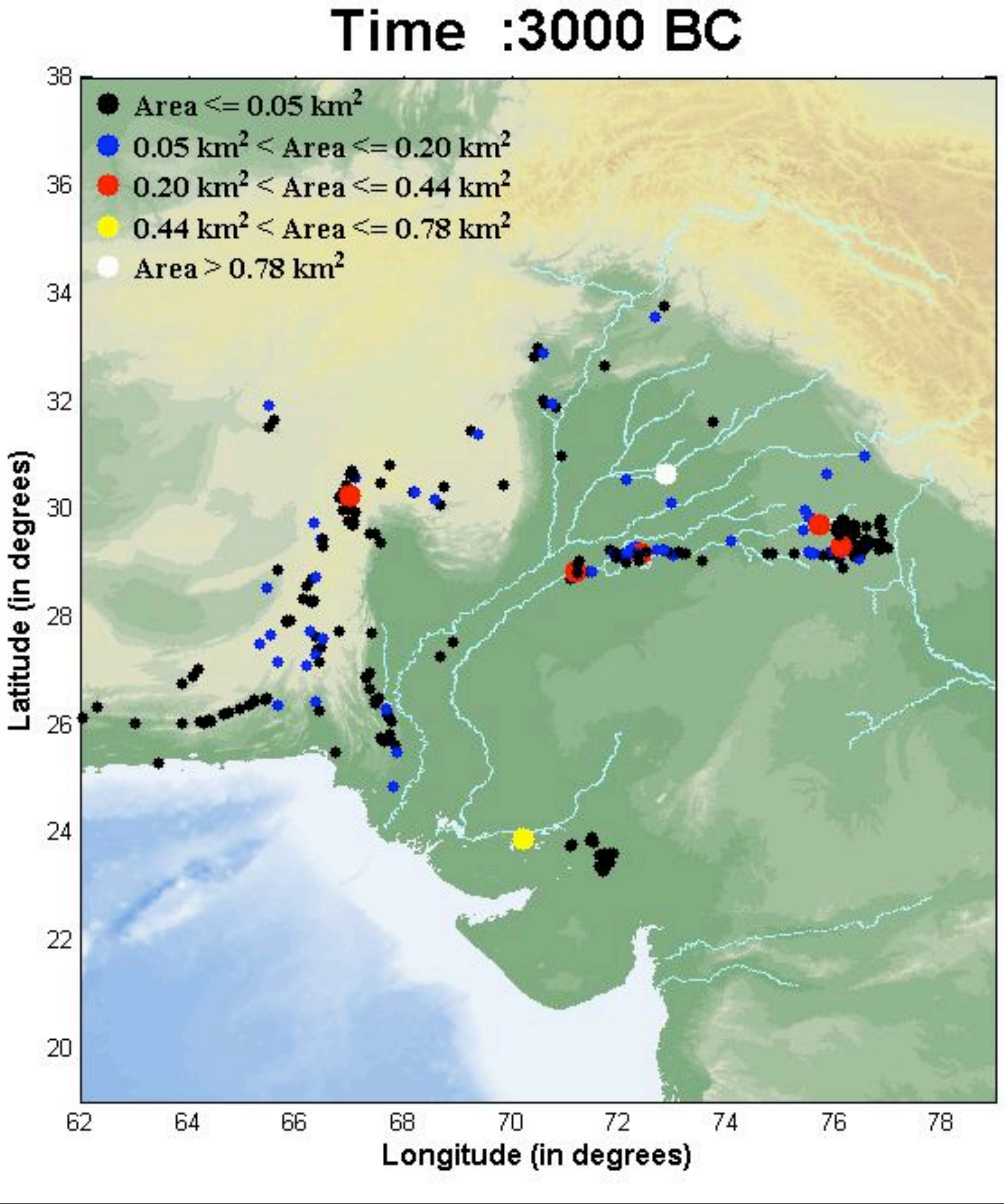
The Indus civilisation : spatio-temporal growth



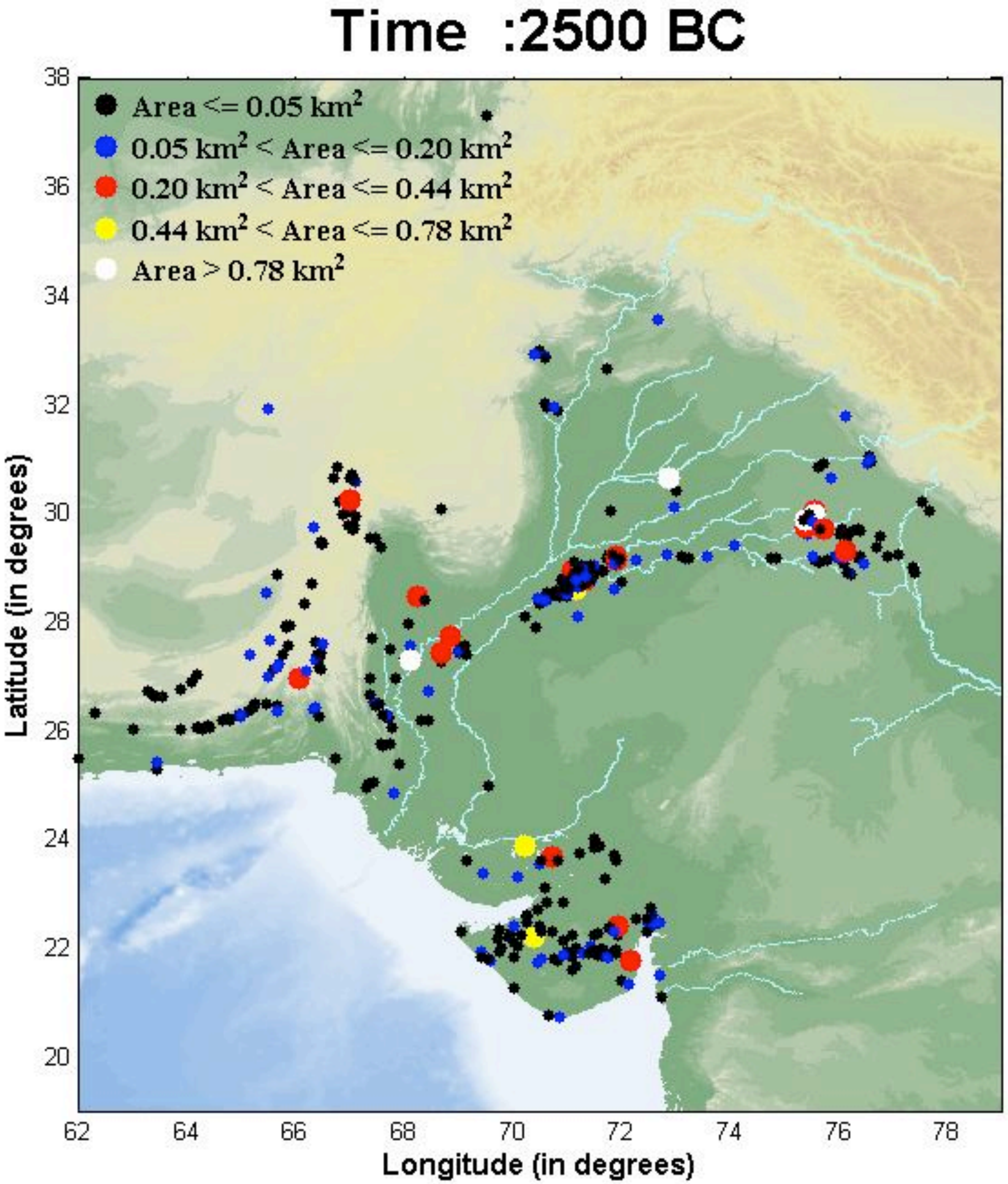
The Indus civilisation : spatio-temporal growth



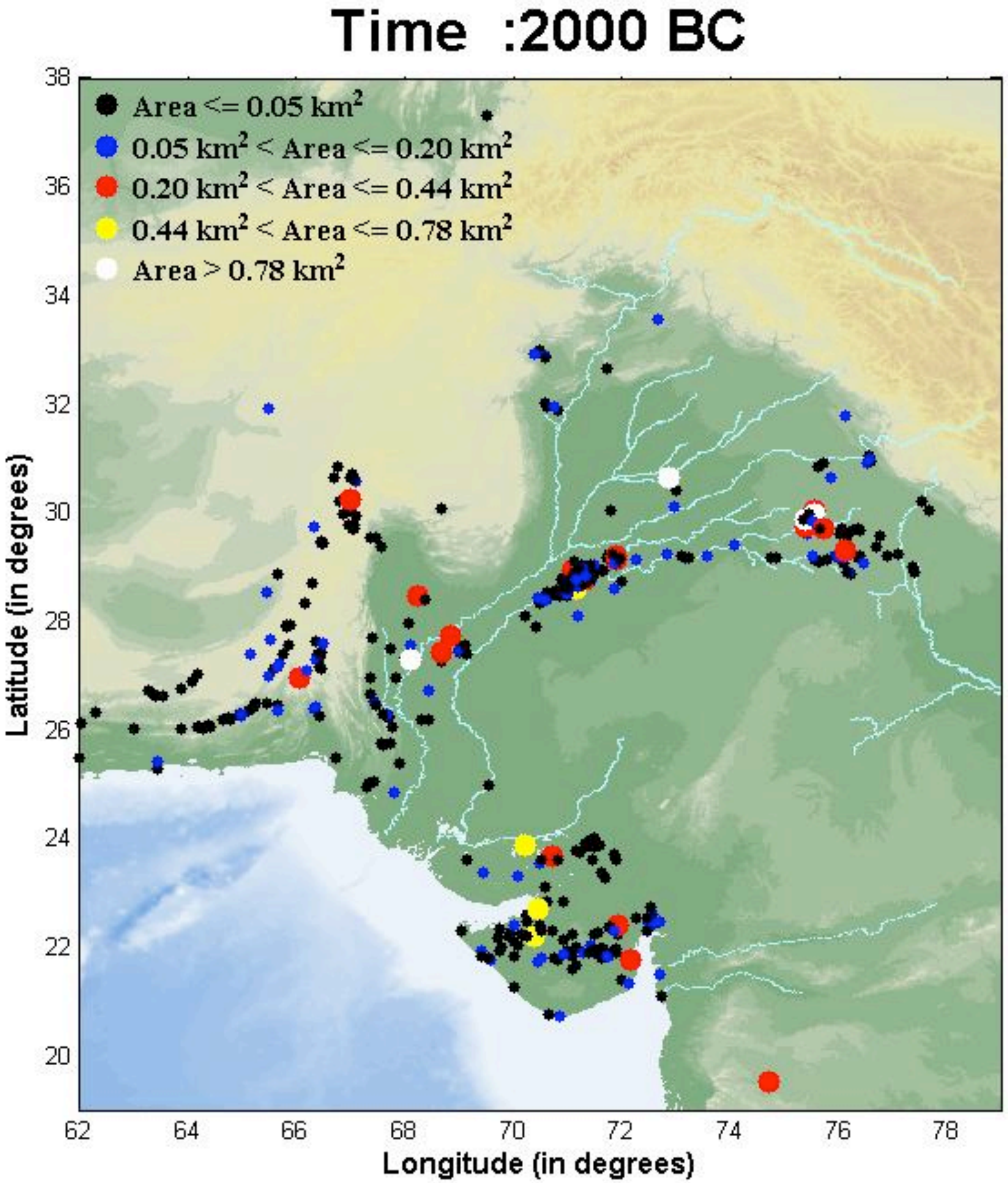
The Indus civilisation : spatio-temporal growth



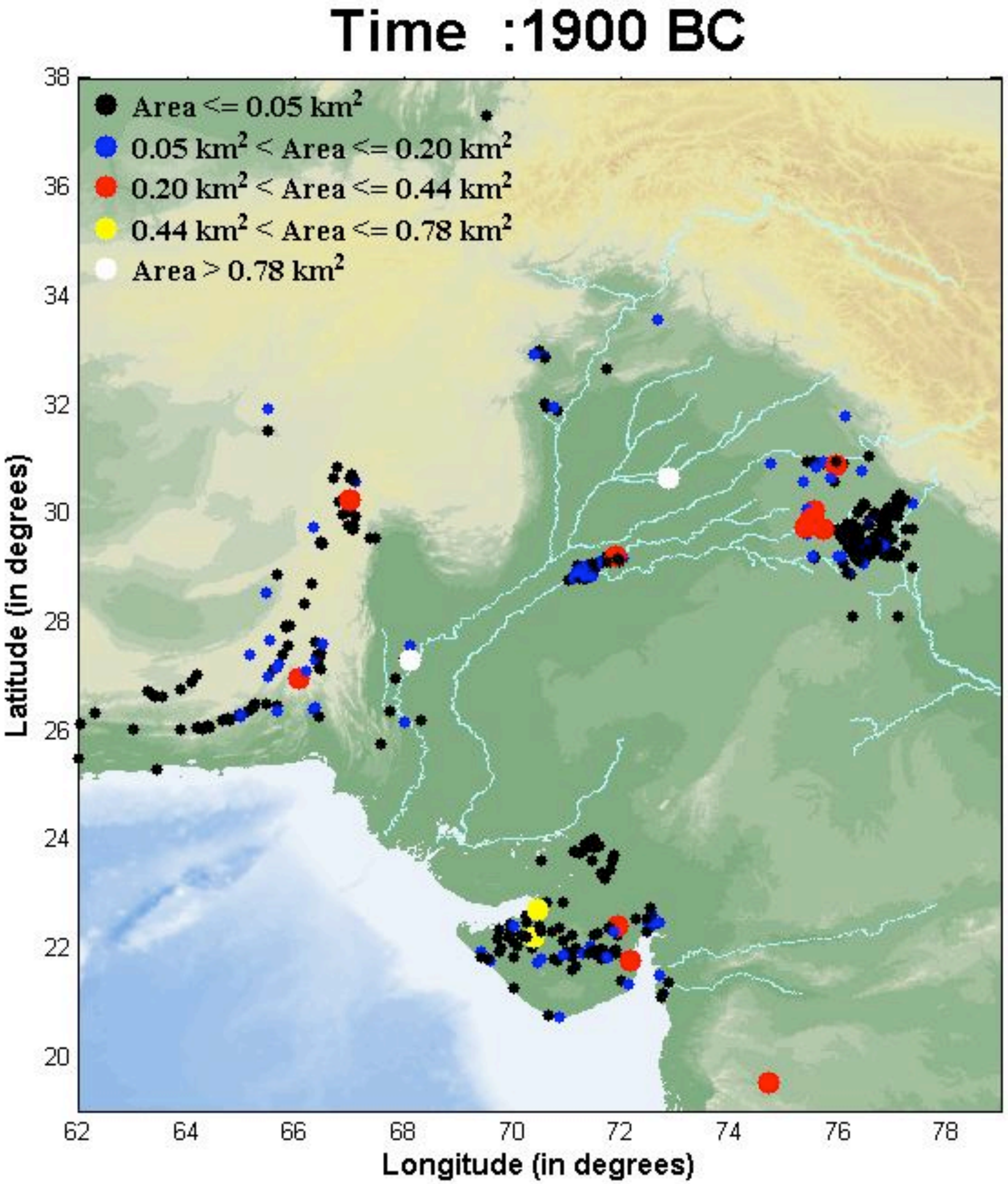
The Indus civilisation : spatio-temporal growth



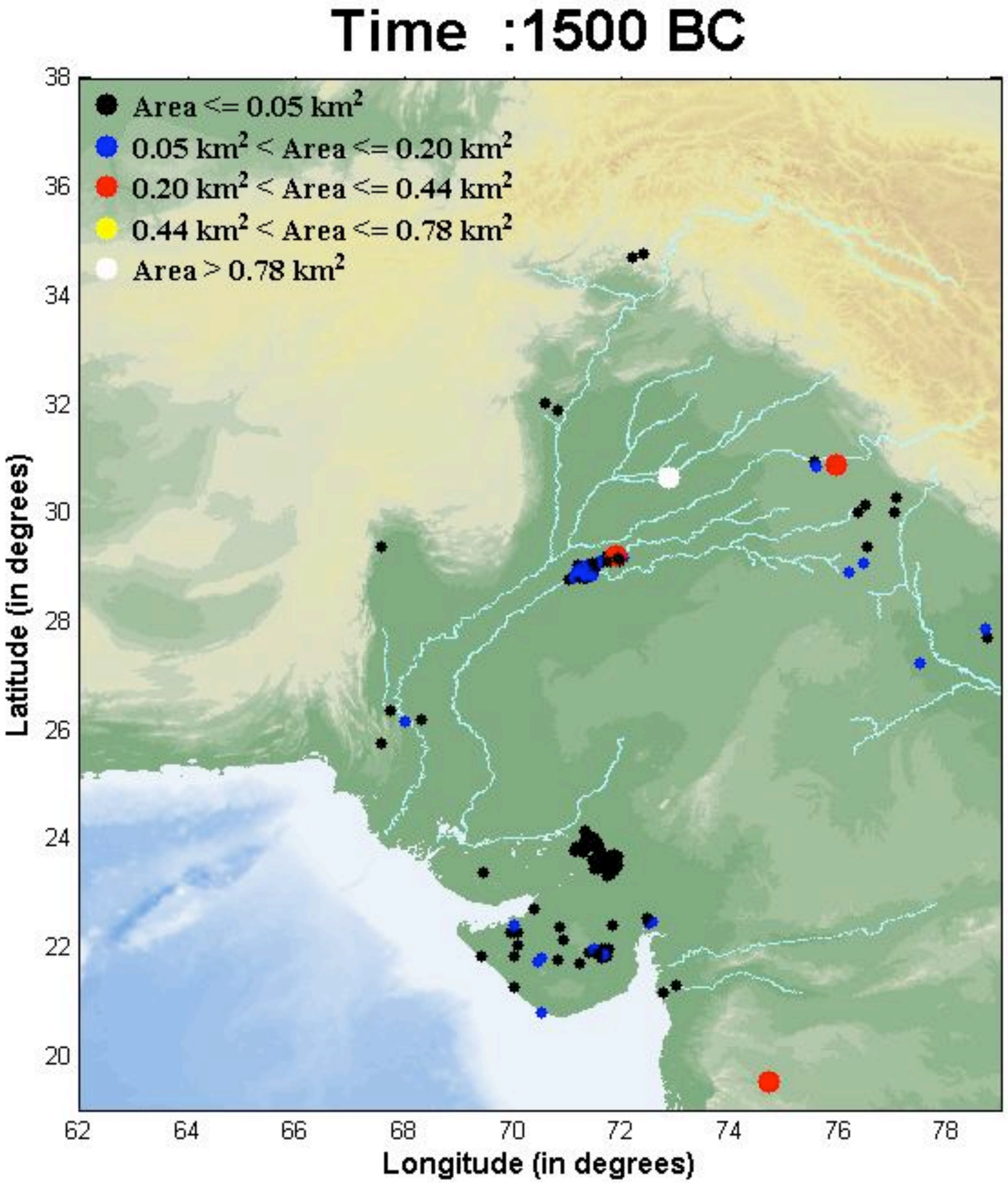
The Indus civilisation : spatio-temporal growth



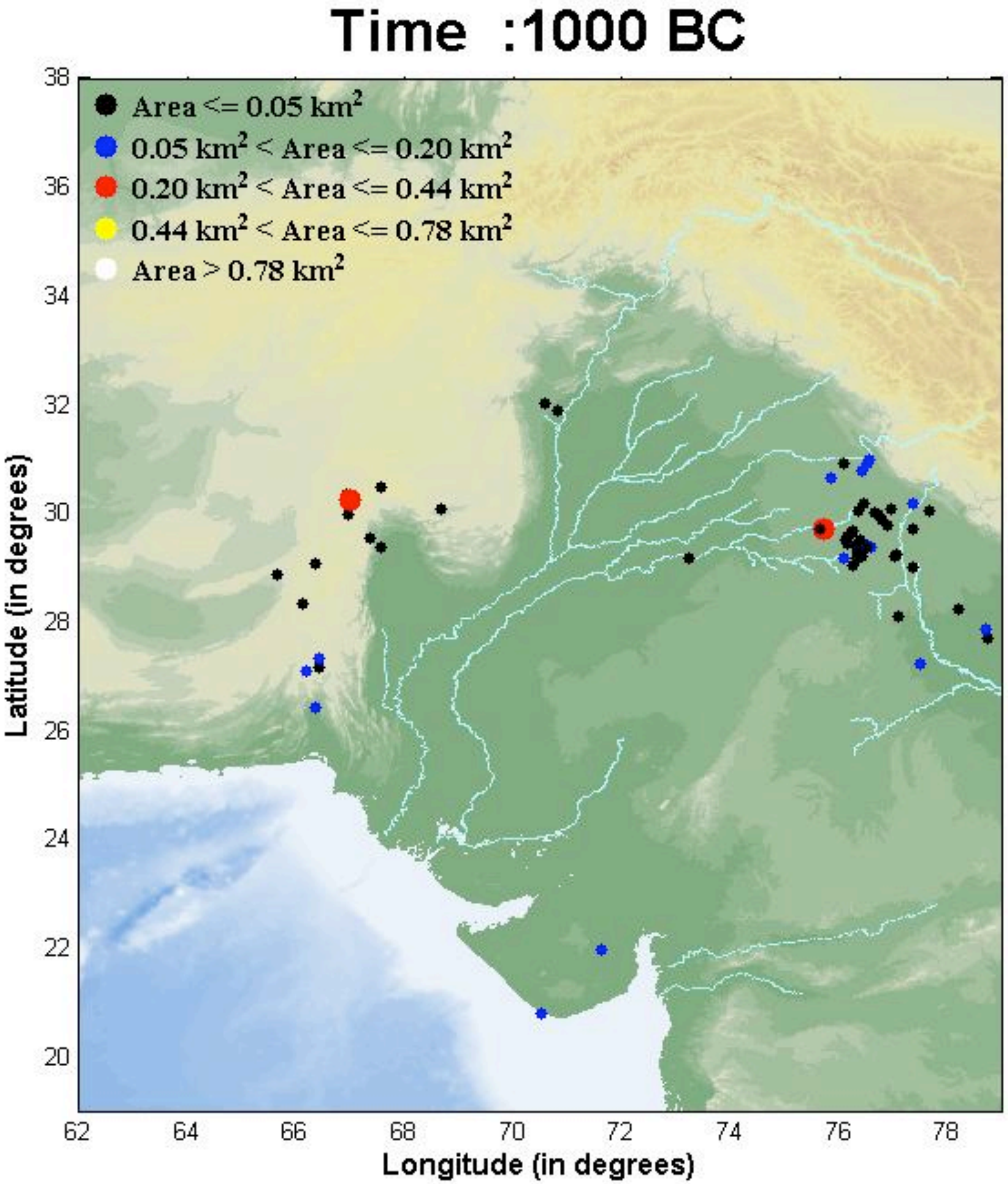
The Indus civilisation : spatio-temporal growth



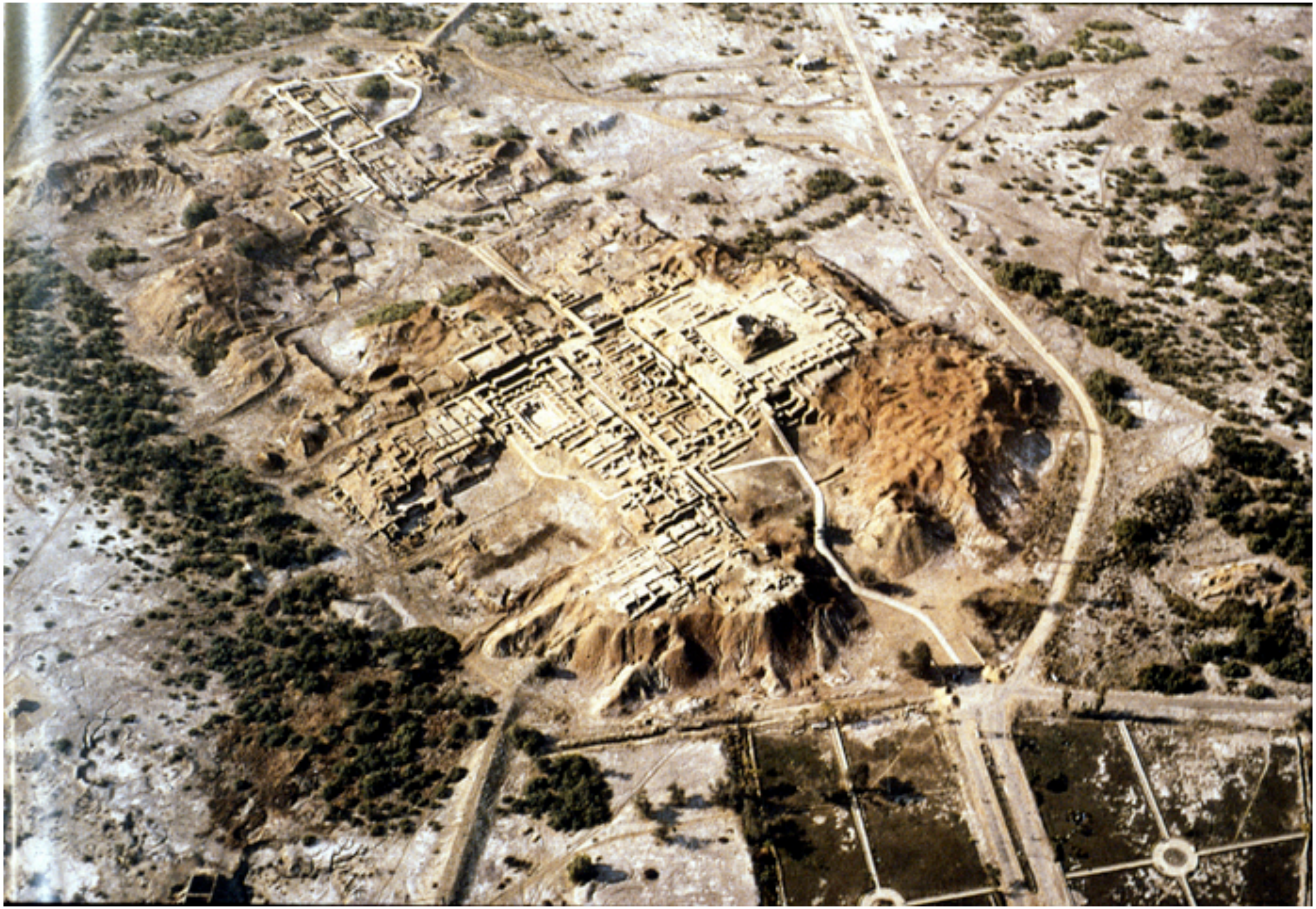
The Indus civilisation : spatio-temporal growth



The Indus civilisation : spatio-temporal growth

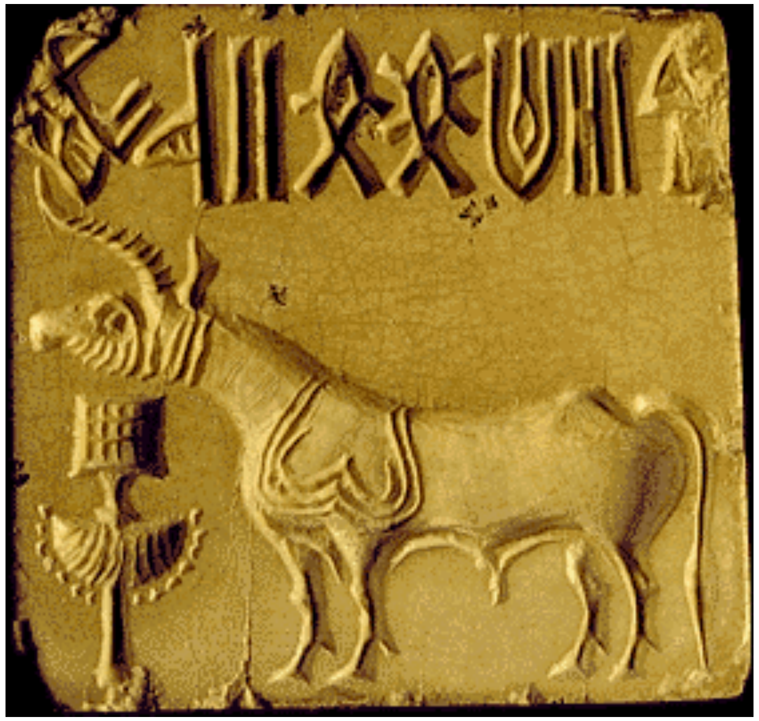


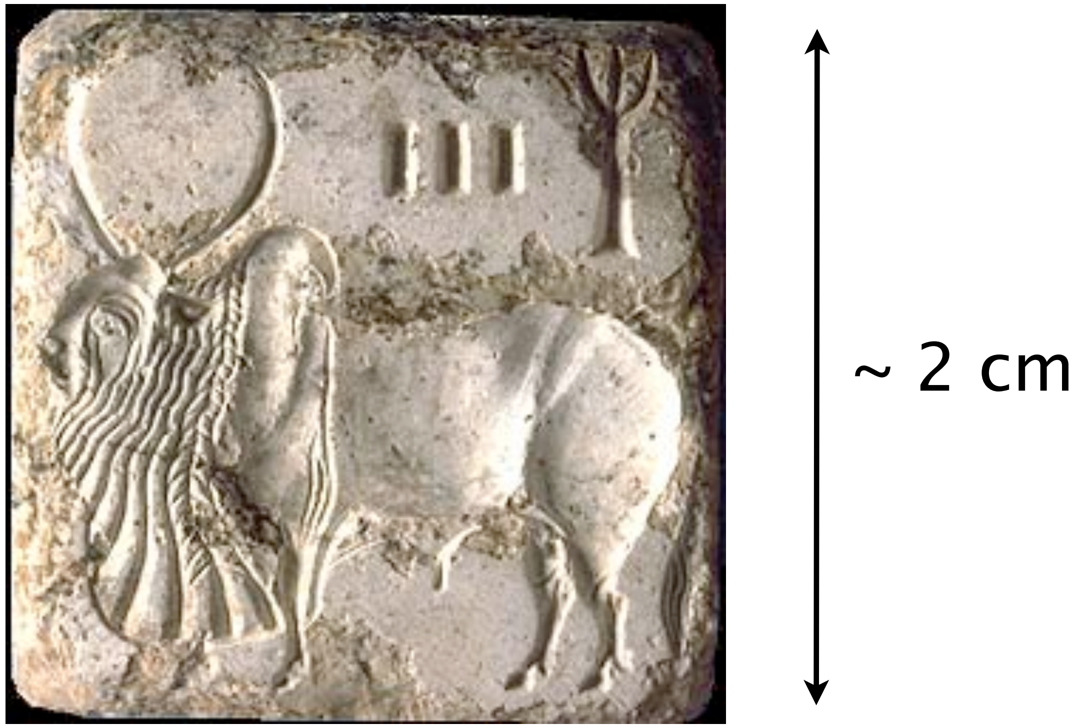
An urban civilisation : Mohenjo Daro

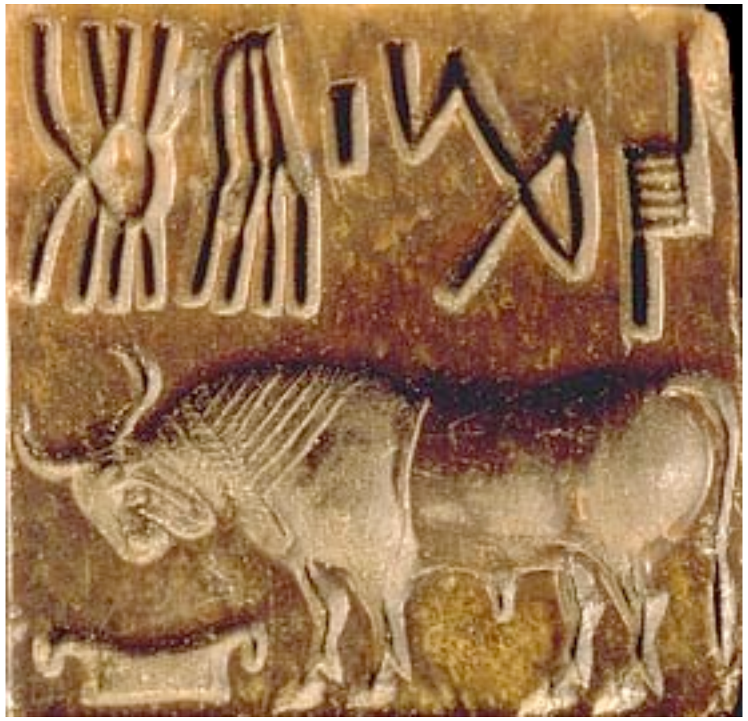


Acknowledgement : Bryan Wells

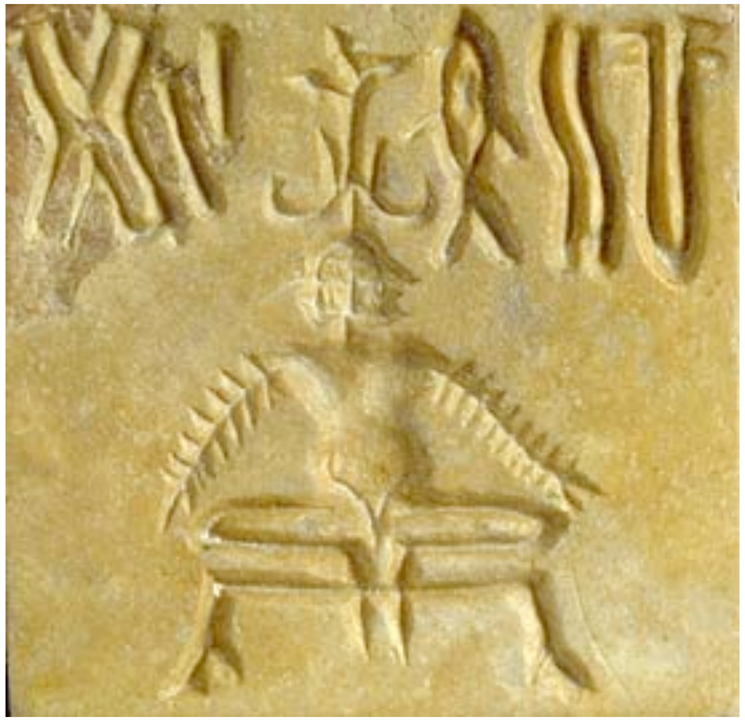
The Indus script : seals







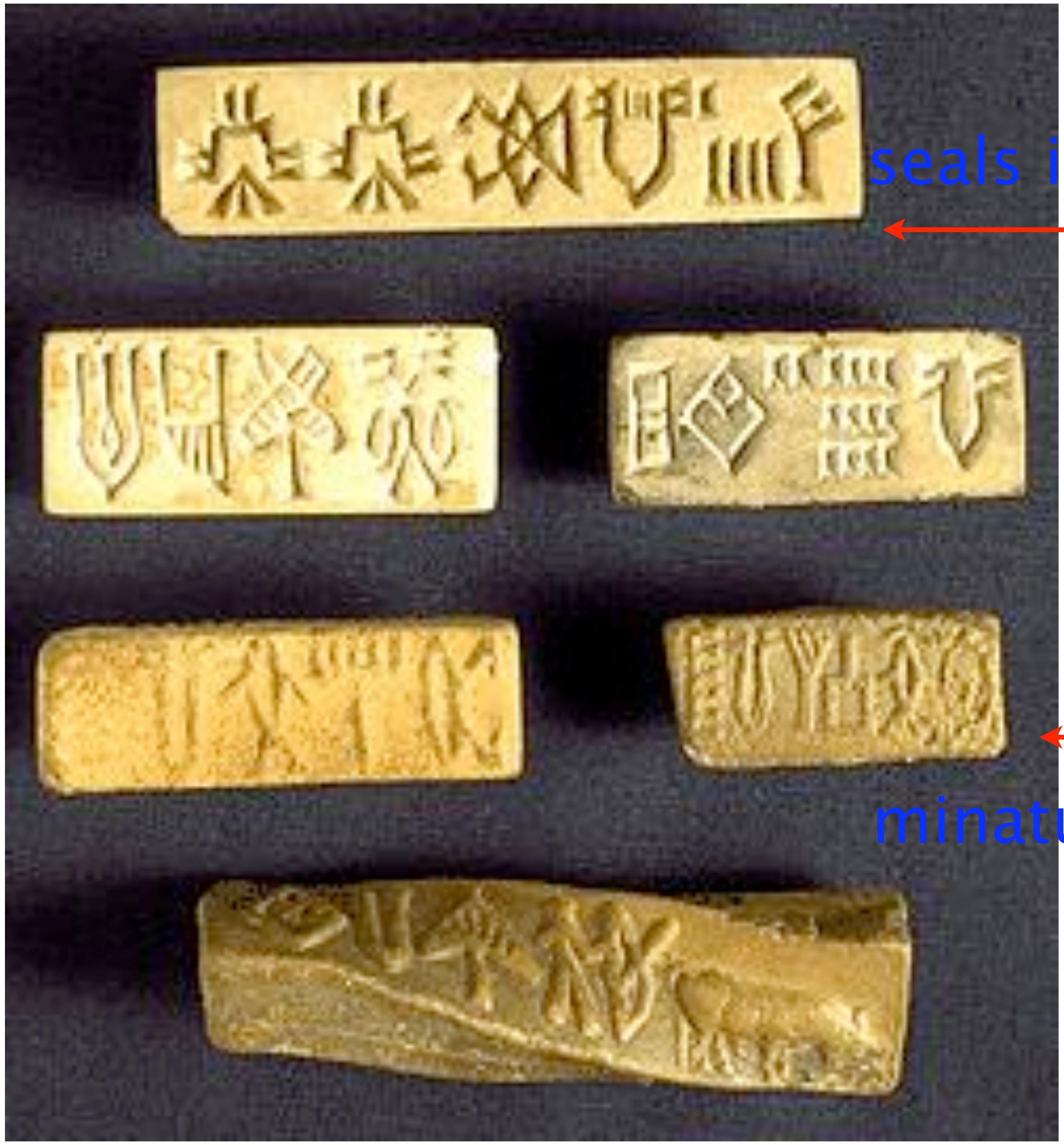
copyright : J. M. Kenoyer



source : harappa.com

The Indus script : tablets

The script is read from right to left.



seals in intaglio

minature tablet

The Indus people wrote on steatite, carnelian, ivory and bone, pottery, stoneware, faience, copper and gold, and inlays on wooden boards.

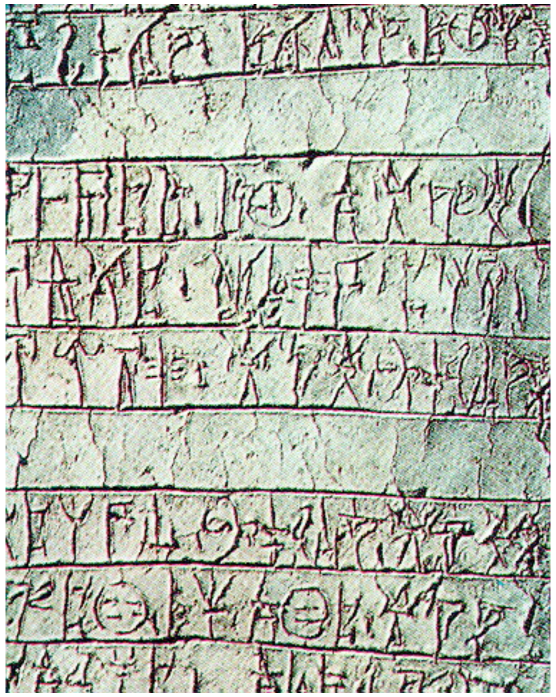
Inspite of almost a century of effort, the script is still undeciphered.

copyright : J. M. Kenoyer source : harappa.com

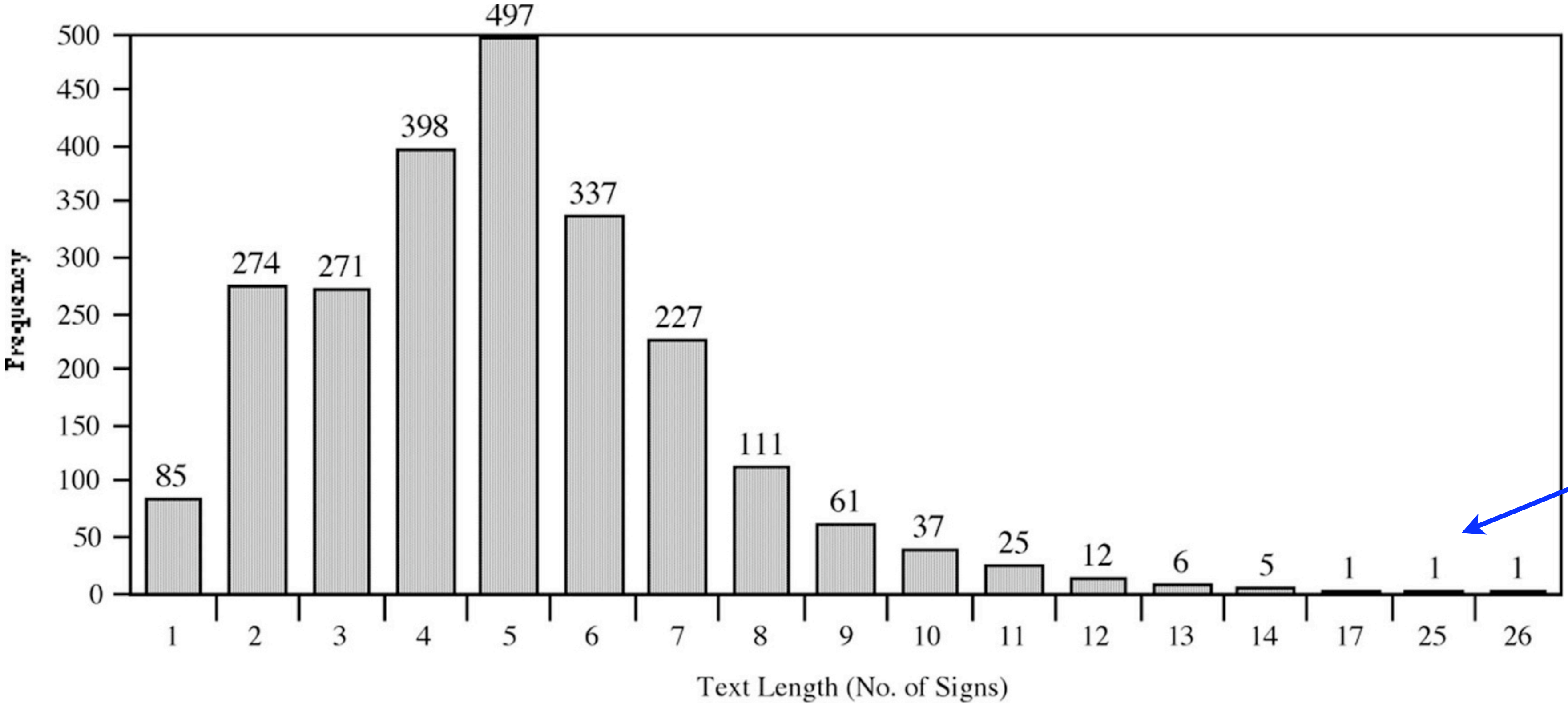
Why is the script still undeciphered ?

Short texts and small corpus

Indus

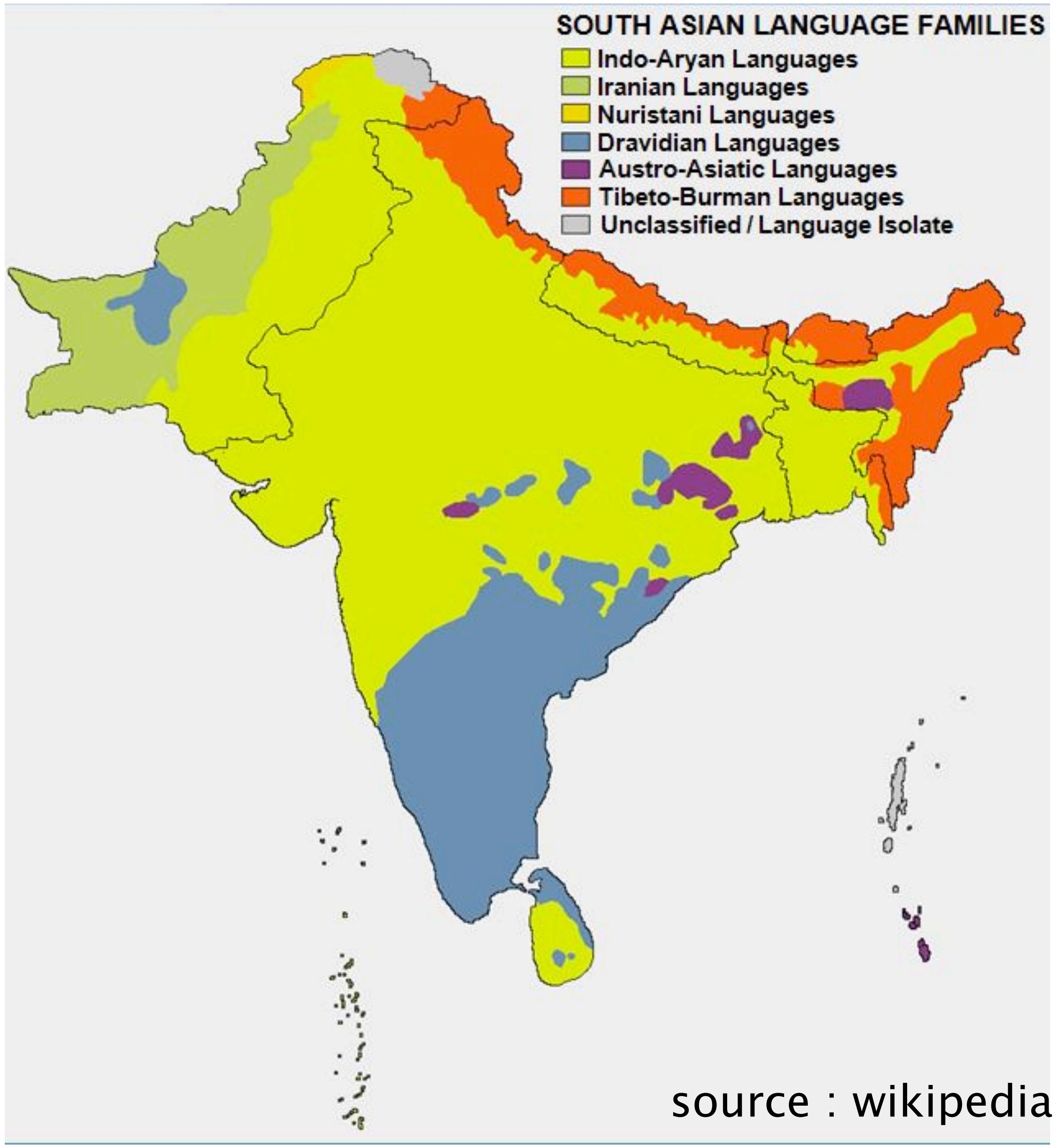


Linear B
source : wikipedia



on multiple faces

Language unknown



source : wikipedia

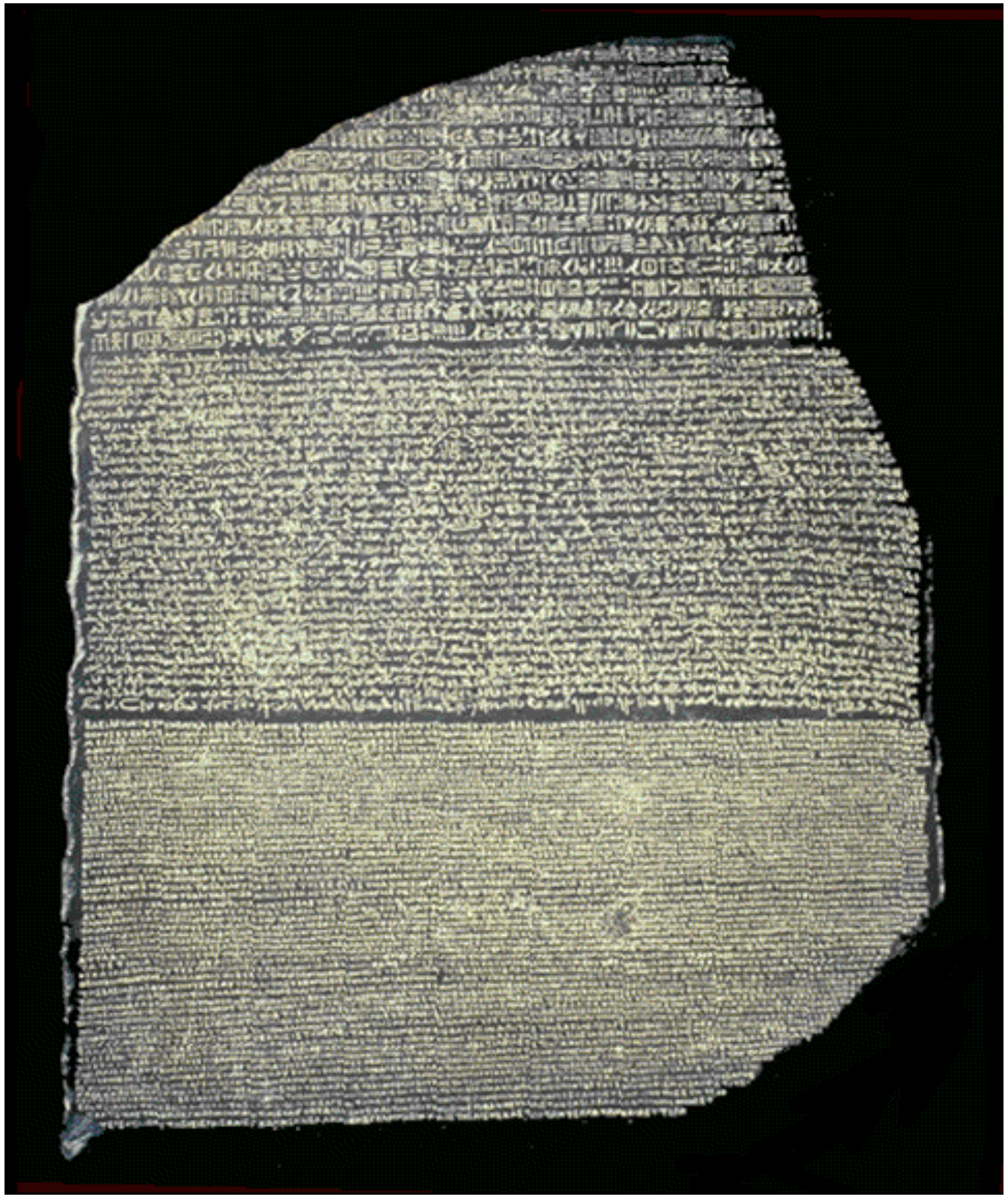
100
RESERVE BANK OF INDIA
GUARANTEED BY THE CENTRAL GOVERNMENT
6FB 444444
भारतीय रिज़र्व बैंक
I PROMISE TO PAY THE BEARER THE SUM OF ONE HUNDRED RUPEES
100
6FB 444444
एक सौ रुपये

The subcontinent is a very linguistically diverse region.

1576 classified mother tongues, 29 language with more than a 1 million speakers. (Indian Census, 1991).

Current geographical distributions may not reflect historical distributions.

No multilingual texts

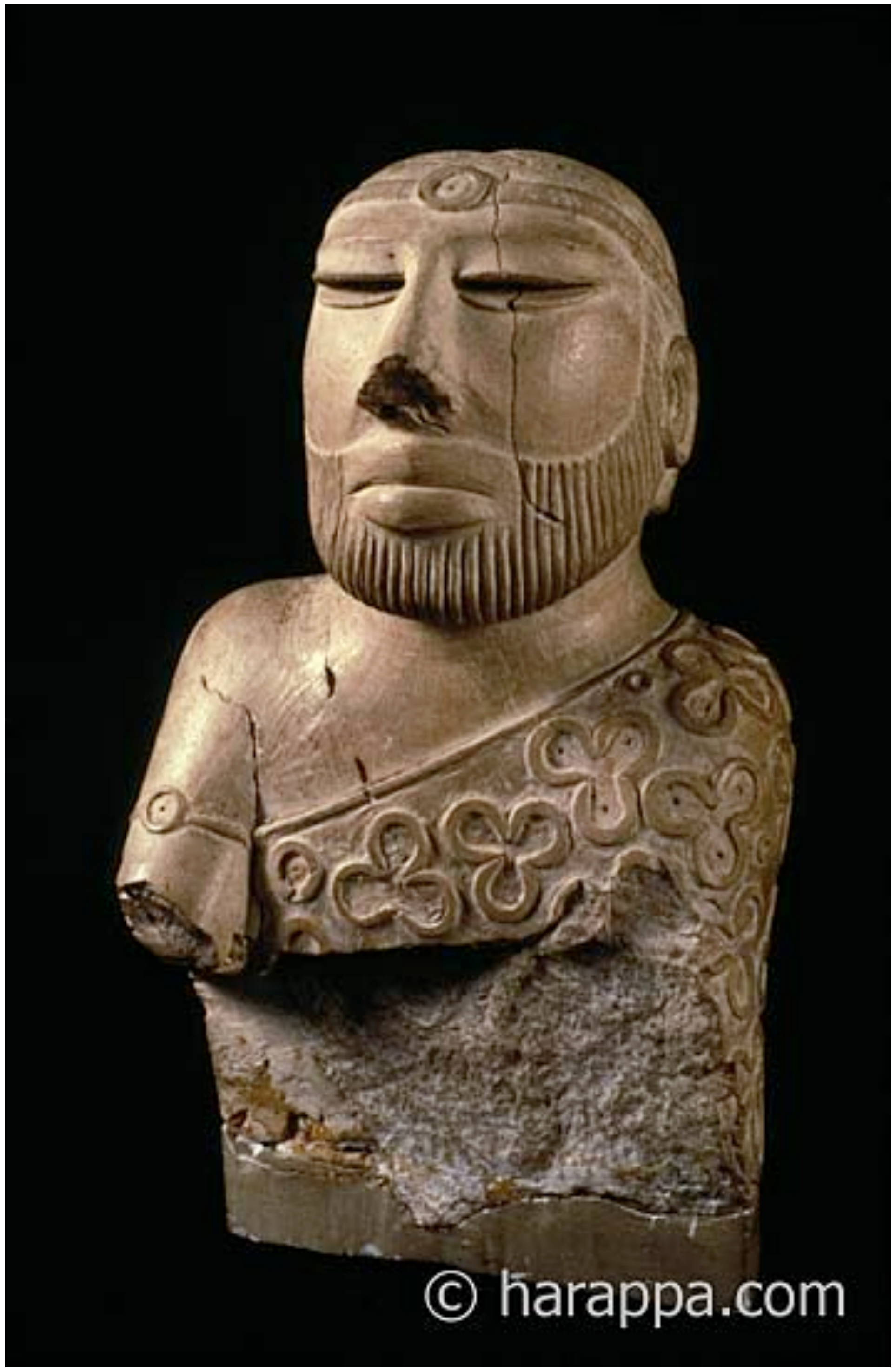


The Rosetta stone has a single text written in hieroglyphic, Demotic, and Greek.

This helped Thomas Young and Jean-Francois Champollion to decipher the hieroglyphics.

source : wikipedia

No contexts



?

No place names, or names of kings, or dynasties or rulers.

Attempts at decipherment

Proto-Dravidian

Indo-European

Proto-Munda

No consensus on any of these readings.

Ideographic ? Syllabic ? Logo-syllabic ?

“I shall pass over in silence many other attempts based on intuition rather than on analysis.’’

The non-linguistic hypothesis

S. Farmer, R. Sproat, M. Witzel, EJVS, 2004

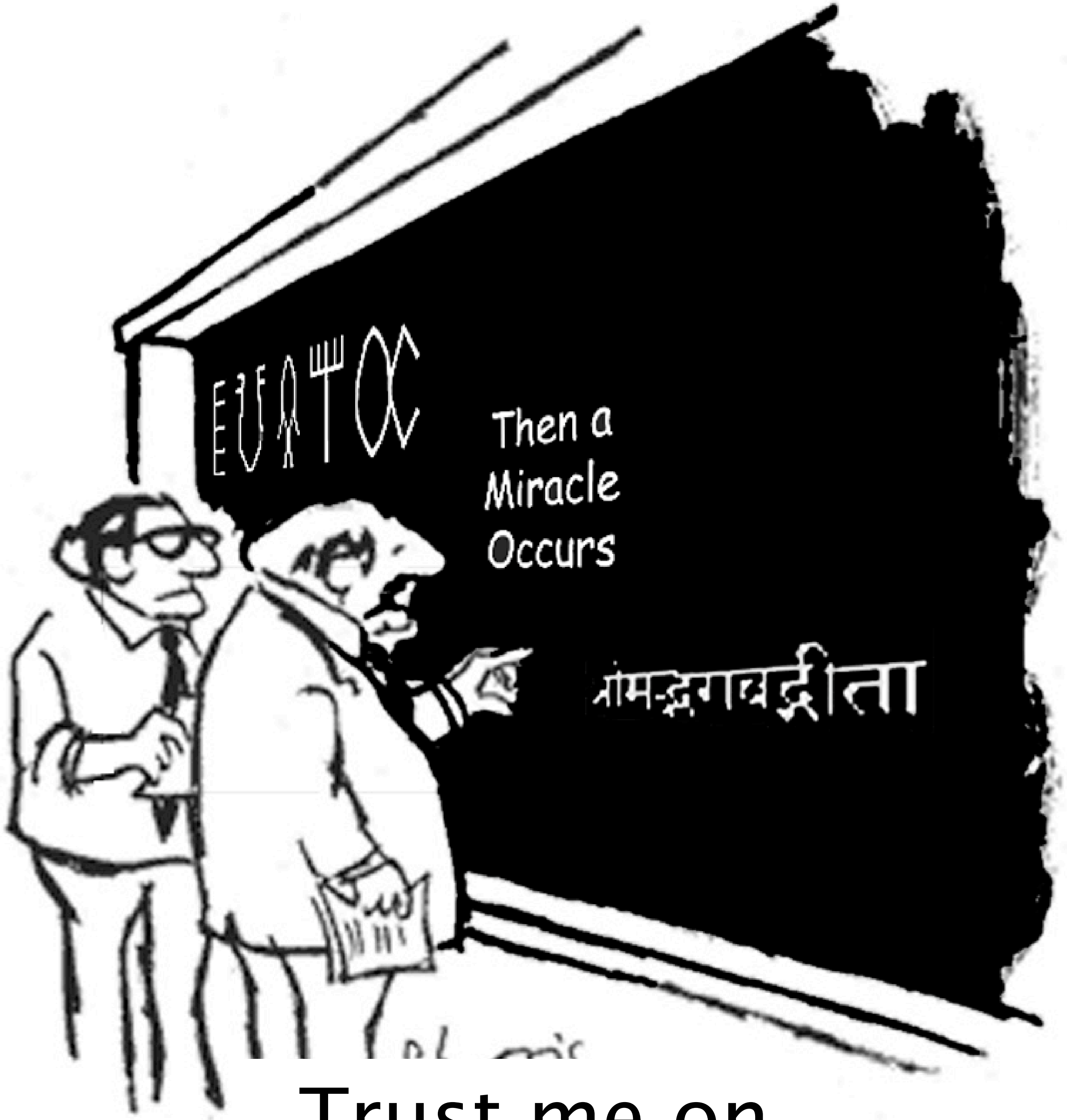
The collapse of the Indus script hypothesis : the myth of a literate Harappan civilisation.

No long texts.
'Unusual' frequency distributions.
'Unusual' archaeological features.

Massimo Vidale, East and West, 2007

The collapse melts down : a reply to Farmer, Sproat and Witzel

"Their way of handling archaeological information on the Indus civilisation (my field of expertise) is sometimes so poor, outdated and factious that I feel fully authorised to answer on my own terms."



Trust me on this!

Acknowledgement : Bryan Wells

Syntax implies statistical regularities

Power-law frequency distribution

Ranked word frequencies have a power-law distribution. This empirical result is called the Zipf-Mandelbrot law. All tested languages show this feature.

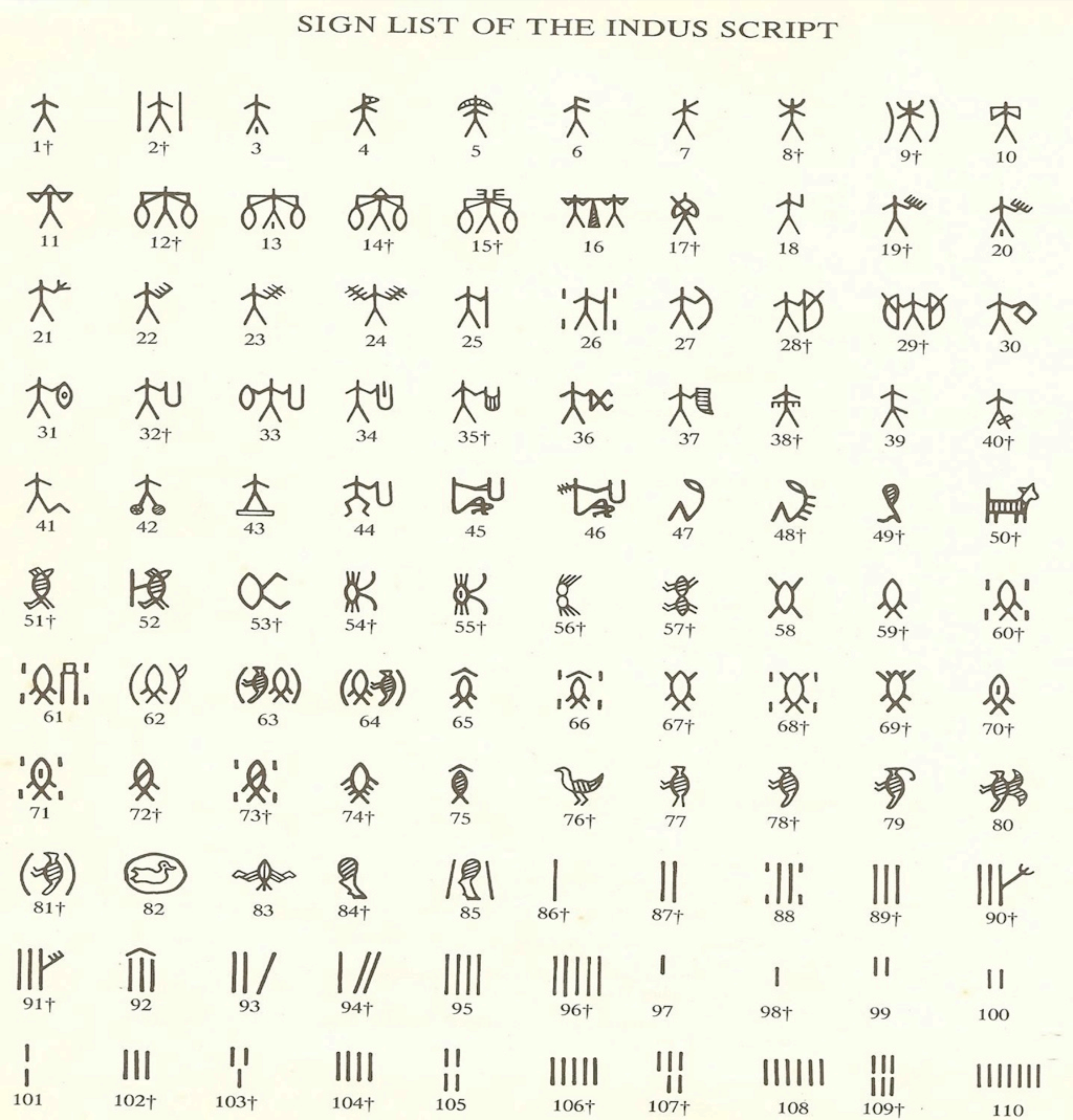
Beginner-ender asymmetry :

Languages have preferred order in Subject Object and Verb. Articles like ‘a’ or ‘the’ never end sentences.

Correlations between tokens :

In English, ‘u’ follows ‘q’ with overwhelming probability.
SVO order has to be maintained in sentences.
Prescriptive grammar : infinitives are not to be split.

From corpus to concordance


SIGN LIST OF THE INDUS SCRIPT

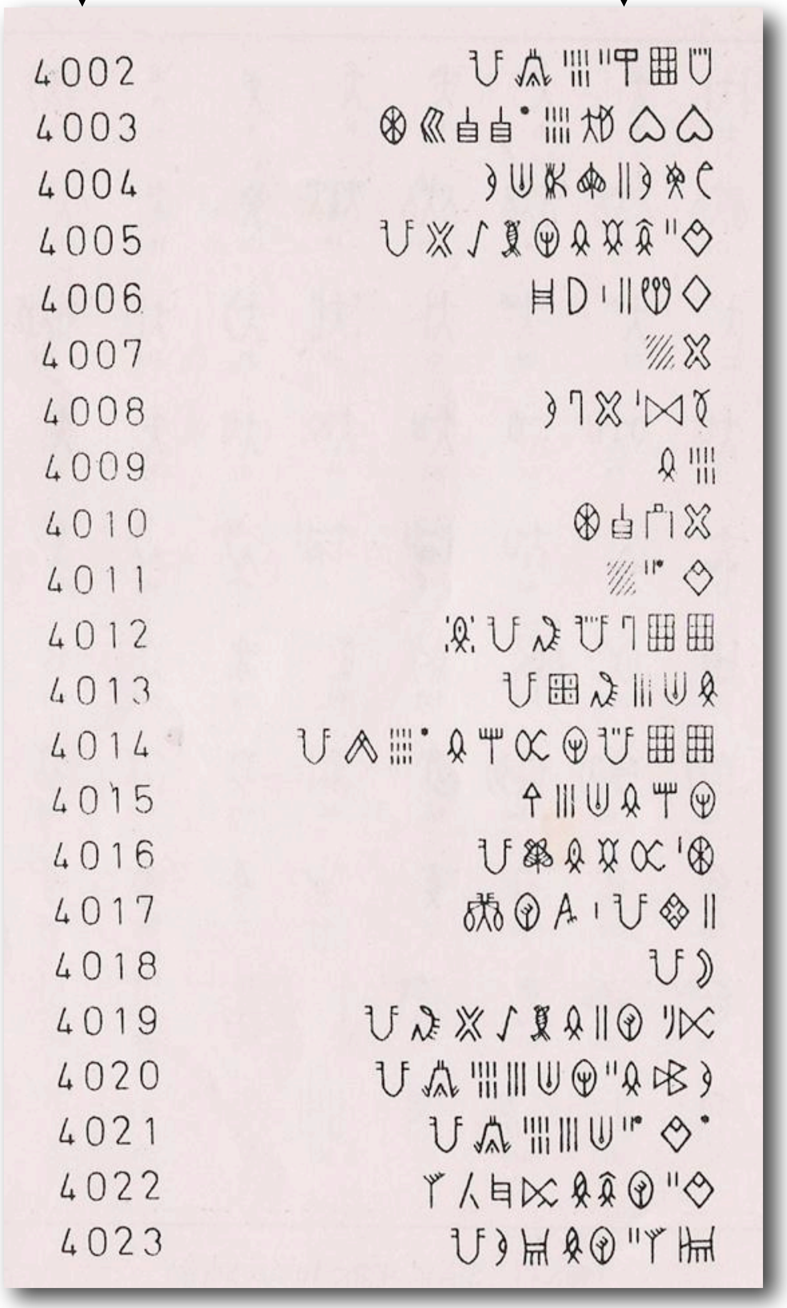
1† 2† 3 4 5 6 7 8† 9† 10
11 12† 13 14† 15† 16 17† 18 19† 20
21 22 23 24 25 26 27 28† 29† 30
31 32† 33 34 35† 36 37 38† 39 40†
41 42 43 44 45 46 47 48† 49† 50†
51† 52 53† 54† 55† 56† 57† 58 59† 60†
61 62 63 64 65 66 67† 68† 69† 70†
71 72† 73† 74† 75 76† 77 78† 79 80
81† 82 83 84† 85 86† 87† 88 89† 90†
91† 92 93 94† 95 96† 97 98† 99 100
101 102† 103† 104† 105 106† 107† 108 109† 110

Compiled by Iravatham Mahadevan in 1977 at the Tata Institute of Fundamental Research. Punch cards were used for the data processing.

← 417 unique signs.

Mahadevan concordance : our data set

text identifier Indus text


4002
4003
4004
4005
4006
4007
4008
4009
4010
4011
4012
4013
4014
4015
4016
4017
4018
4019
4020
4021
4022
4023

2906 texts.
3573 lines.

Signs are mapped to numbers in our analysis.

101-220-59-67-119-23
-97

Probabilities are assigned on the basis of data, with smoothing for unseen n-grams. Technical, but straightforward.

Estimating the probabilities of unseen events

HHHHHH : 6 heads in 6 throws. $\xrightarrow{?}$ $P(H) = 1$

$P(T) = 0$

maximum likelihood estimate

$$\boxed{P(i) = \frac{n_i}{N}}$$

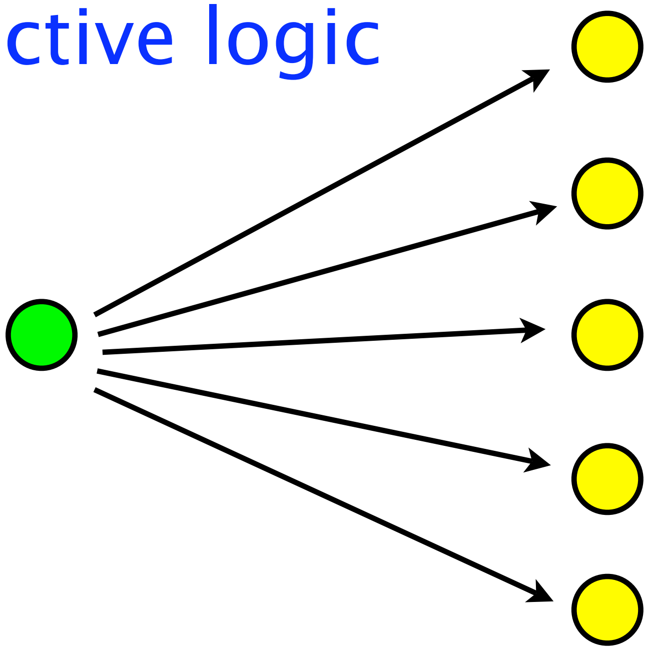
Laplace's rule of succession

$$\boxed{P(i) = \frac{n_i + 1}{N + 2}}$$

Not a deductive problem,
but an inductive problem!

Scientific inference and Bayesian probability

Deductive logic

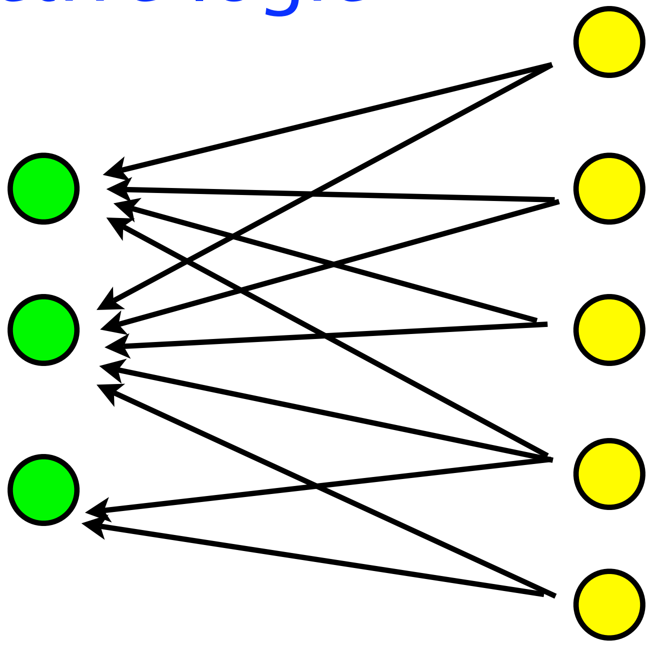


Cause

Effects
or
Outcomes

Mathematical derivation.

Inductive logic



Possible
Causes

Effects
or
Observations

$$P(H|D) = P(D|H)P(H)/P(D)$$

posterior = likelihood x prior / evidence

after D. Sivia in Data Analysis : A Bayesian Tutorial

Inference with uniform prior for binomial distribution

$$P(n_1|\theta, N) = \frac{N!}{n_1!(N-n_1)!}\theta^{n_1}(1-\theta)^{N-n_1}$$ P(D|H) – likelihood

$$P(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$ P(H) = prior

$$\langle\theta\rangle = \frac{a}{a+b}$$

$$P(\theta|n_1, N) \sim \theta^{n_1+a-1}(1-\theta)^{n-n_1+b-1}$$ P(H|D) = posterior

Posterior estimates

$$\theta_{mode} = \frac{n_1 + a - 1}{N + a + b - 2}$$

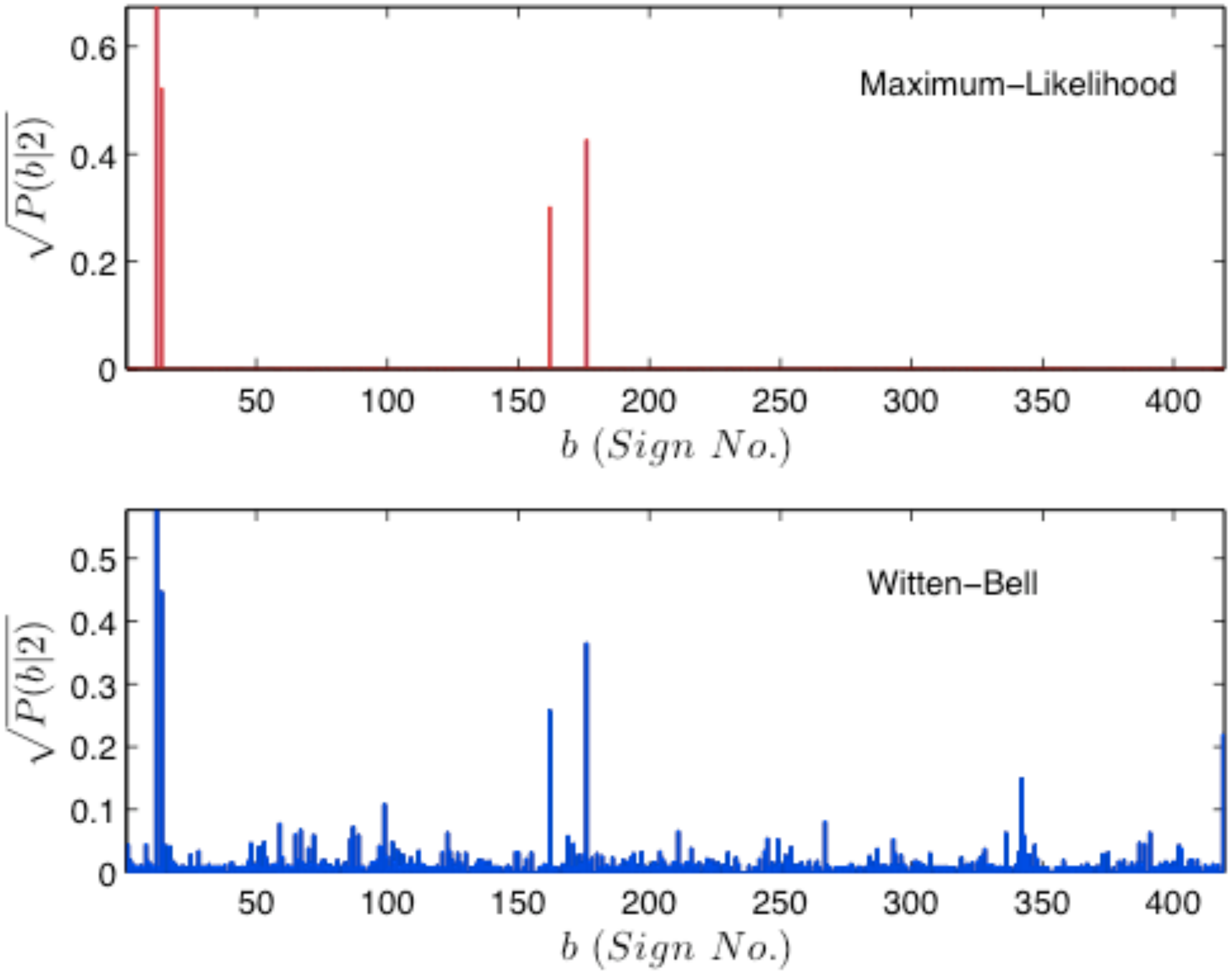
a = 1, b = 1
Estimate using mode. Gives MLE.
Like doing mean-field theory.

$$\langle \theta \rangle_{posterior} = \frac{n_1 + a}{N + a + b}$$

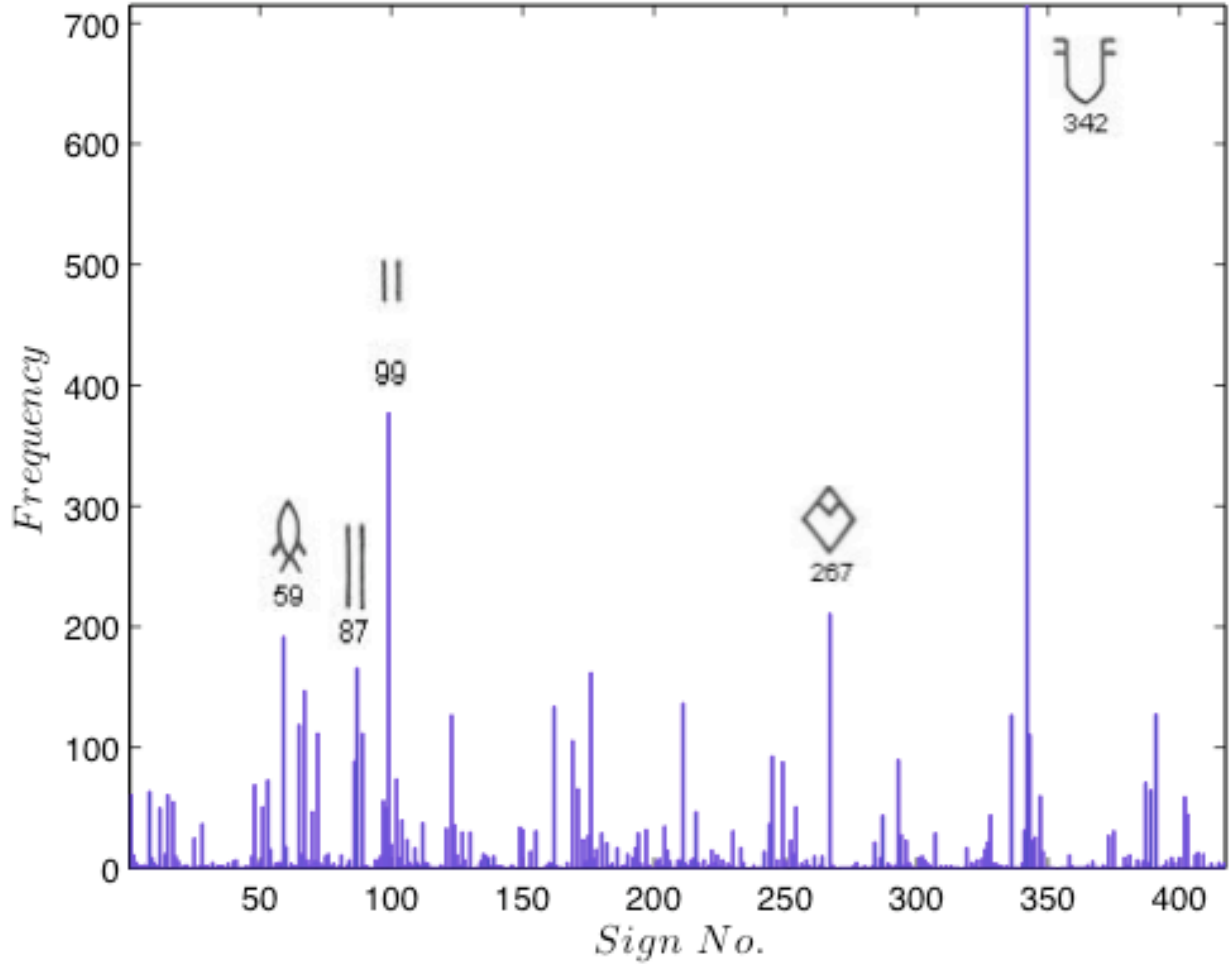
a = 1, b = 1
Estimate using mean. Gives LRS.
Like retaining fluctuations.

Generalising this to multinomial distributions is straightforward but tedious.

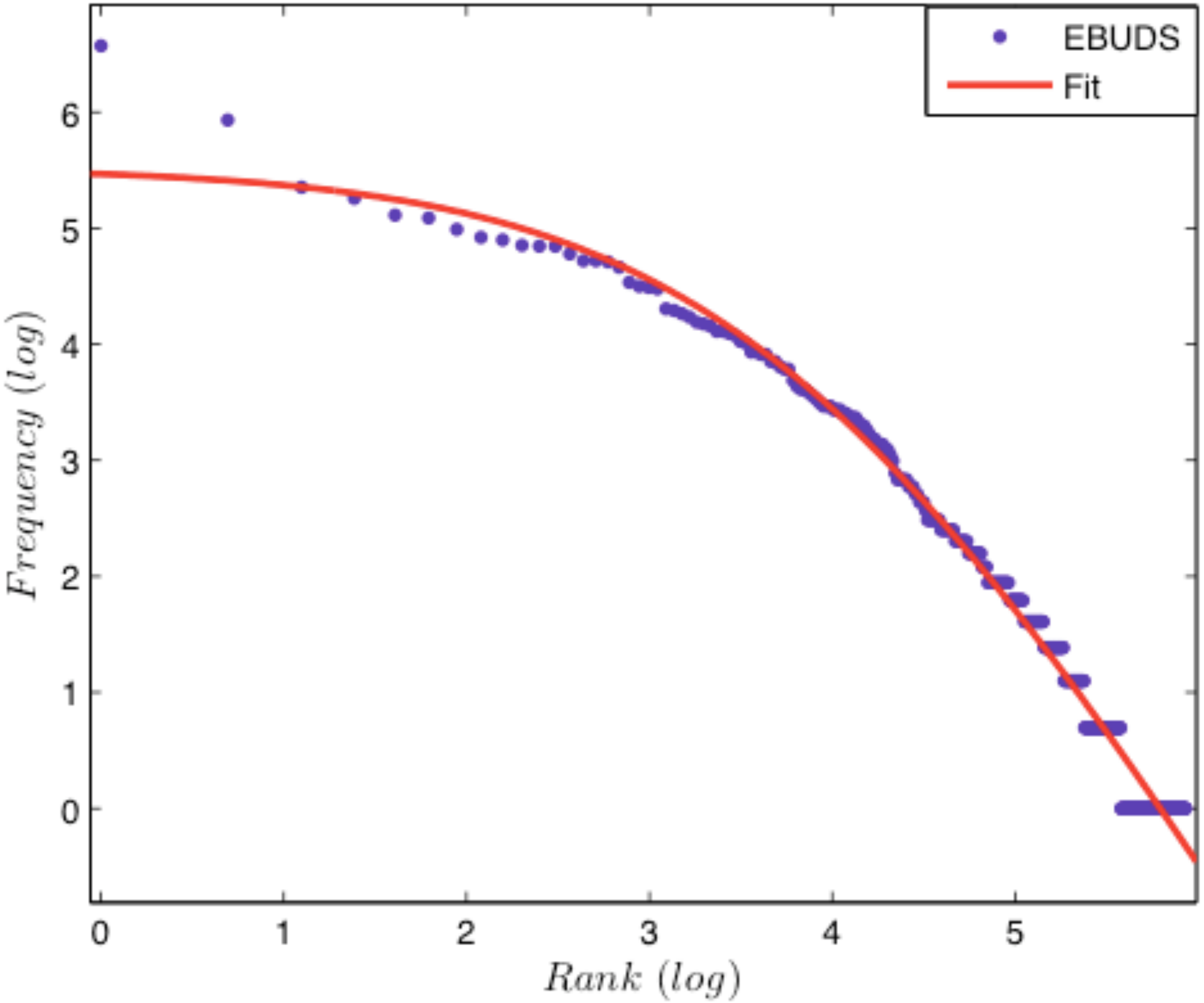
Smoothing of n-grams



Results from the Markov chain : unigrams



Unigrams follow the Zipf-Mandelbrot law

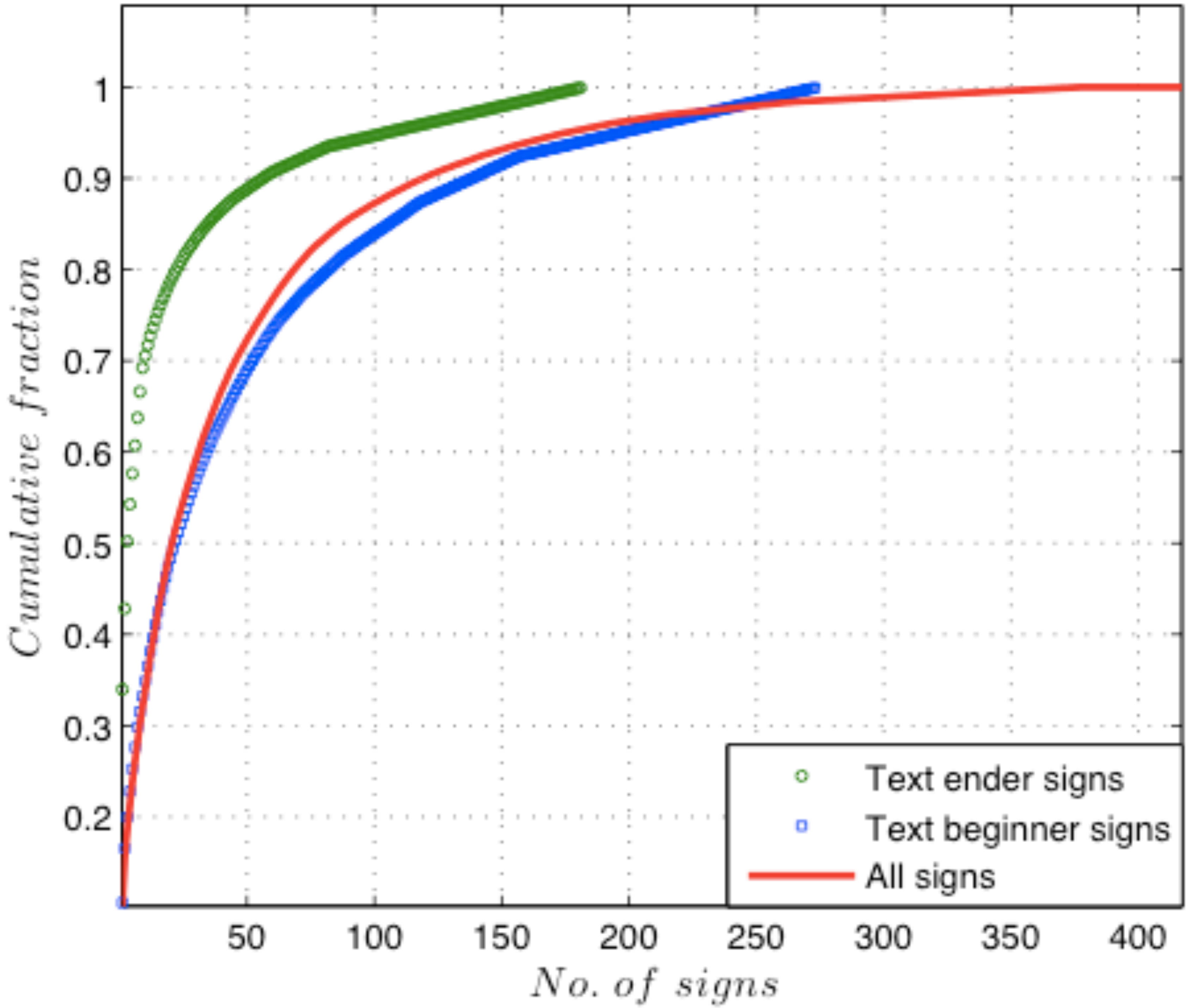


$$\log f_r = a - b\log(r+c)$$

	Indus	English
a	15.39	12.43
b	2.59	1.15
c	44.47	100.00

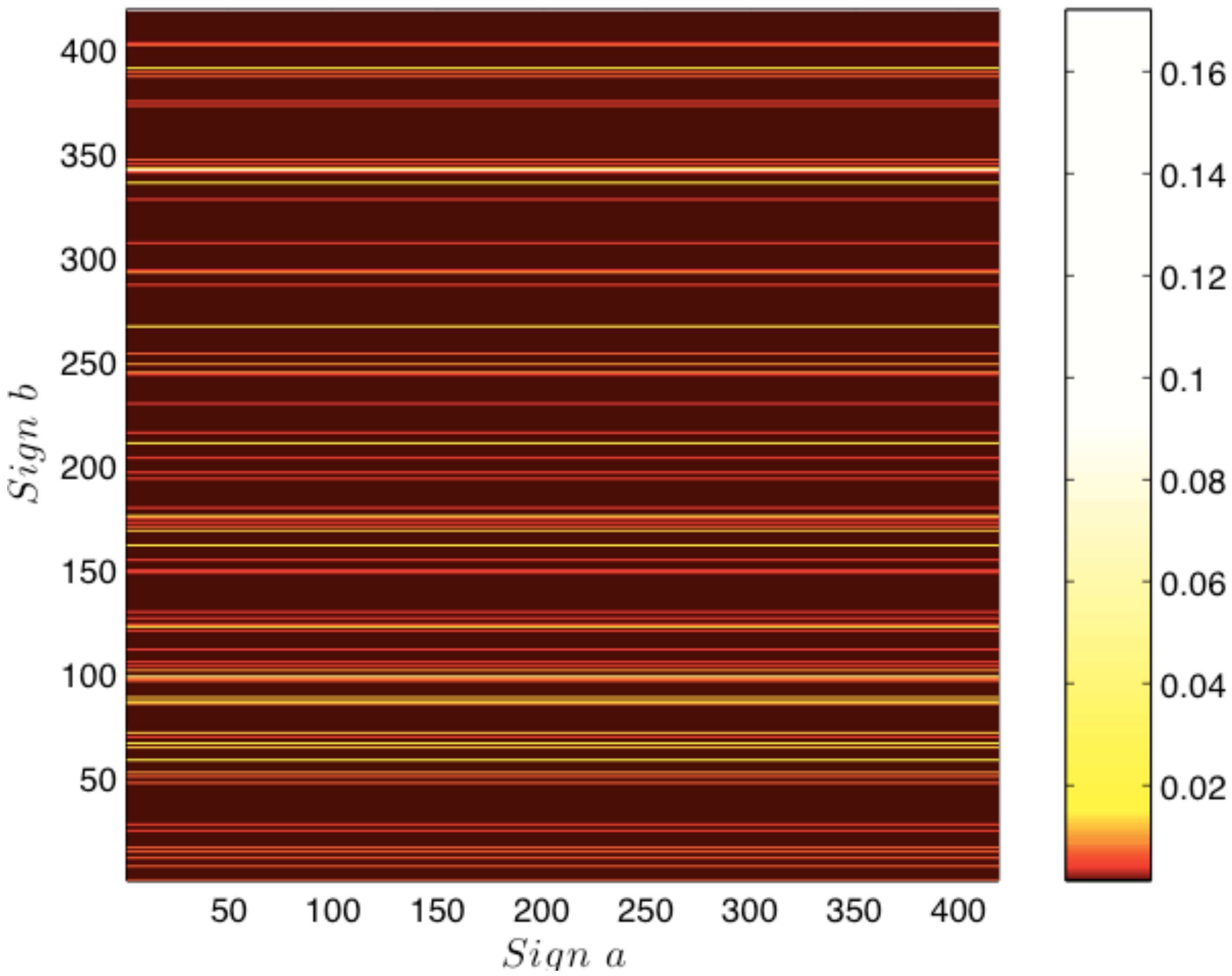
Do the signs encode words ?

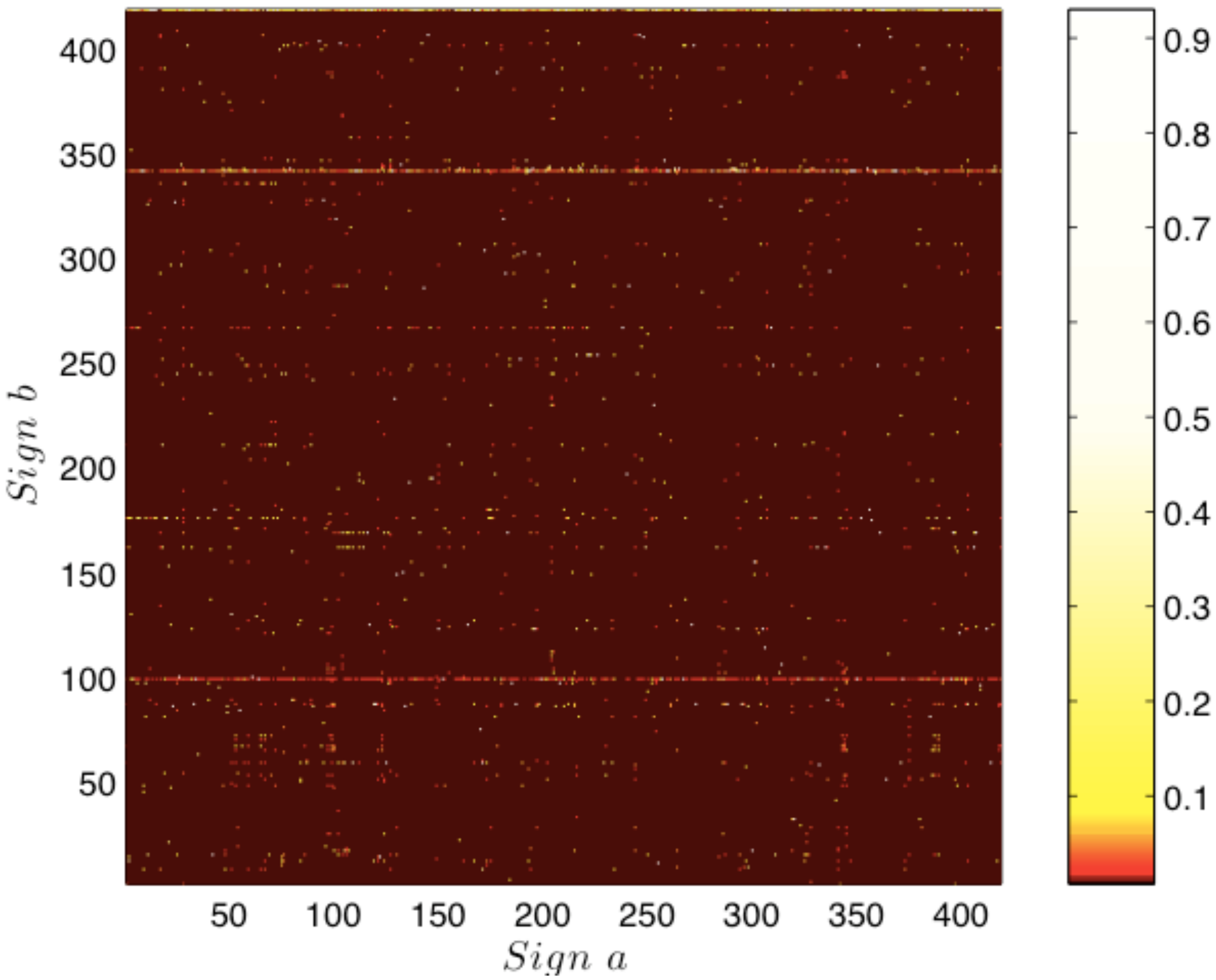
Beginners, enders and unigrams



Does this indicate SOV order ?

Results from the Markov chains : bigrams





Independent sequence

Indus script

Information content of n-grams

unigram entropy

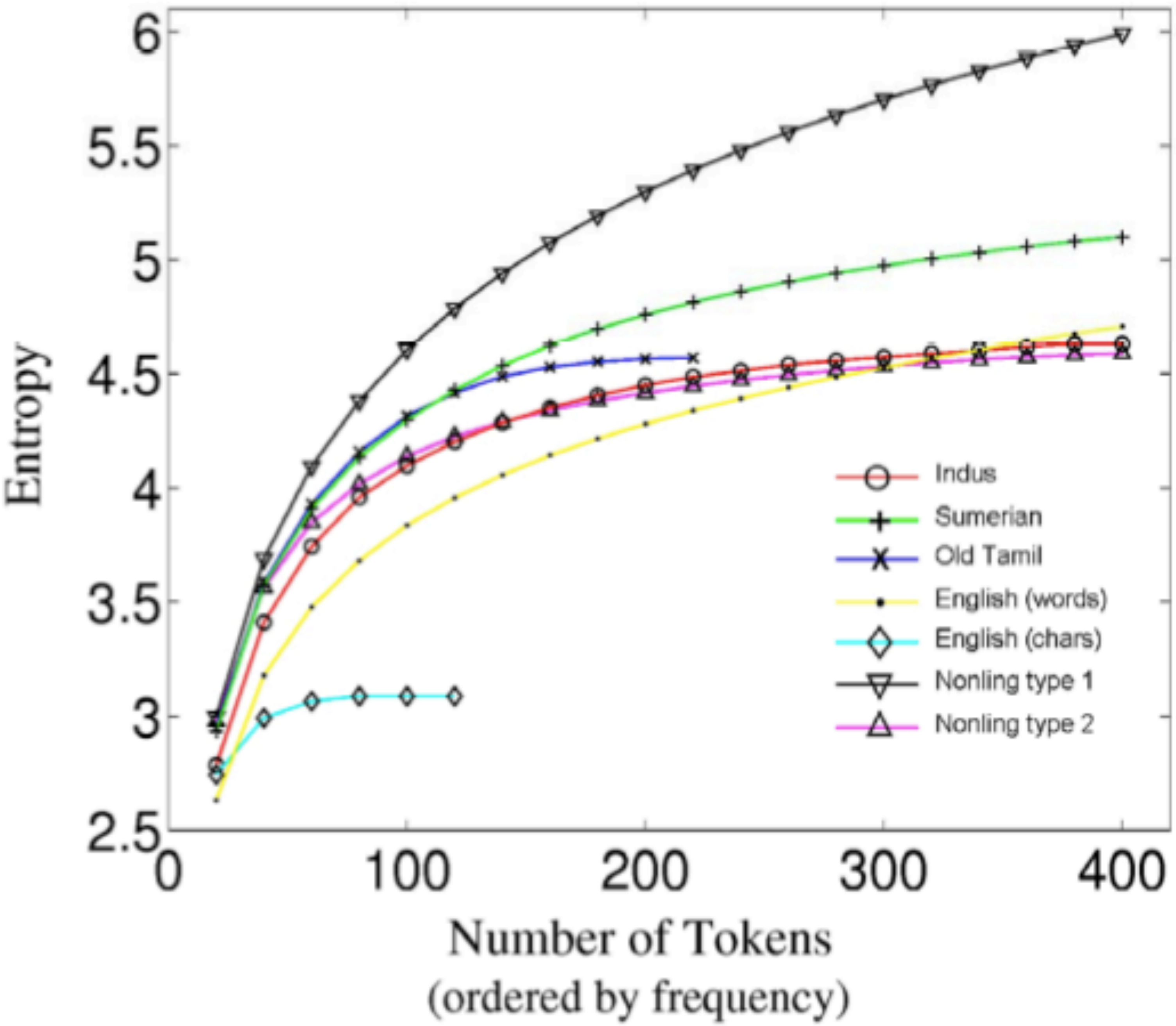
$$H_1 = -\sum_a P(a) \ln P(a)$$

bigram conditional entropy

$$H_{1|1} = -\sum_a P(a) \sum_b P(b|a) \ln P(b|a)$$

We calculate the entropy as a function of the number of tokens, where tokens are ranked by frequency. We compare linguistic and non-linguistic systems using these measures. Two artificial sets of data, representing minimum and maximum conditional entropies, are generated as controls.

Unigram entropies



Indus : Mahadevan Corpus

English : Brown Corpus

Sanskrit : Rig Veda

Old Tamil : Ettuthokai

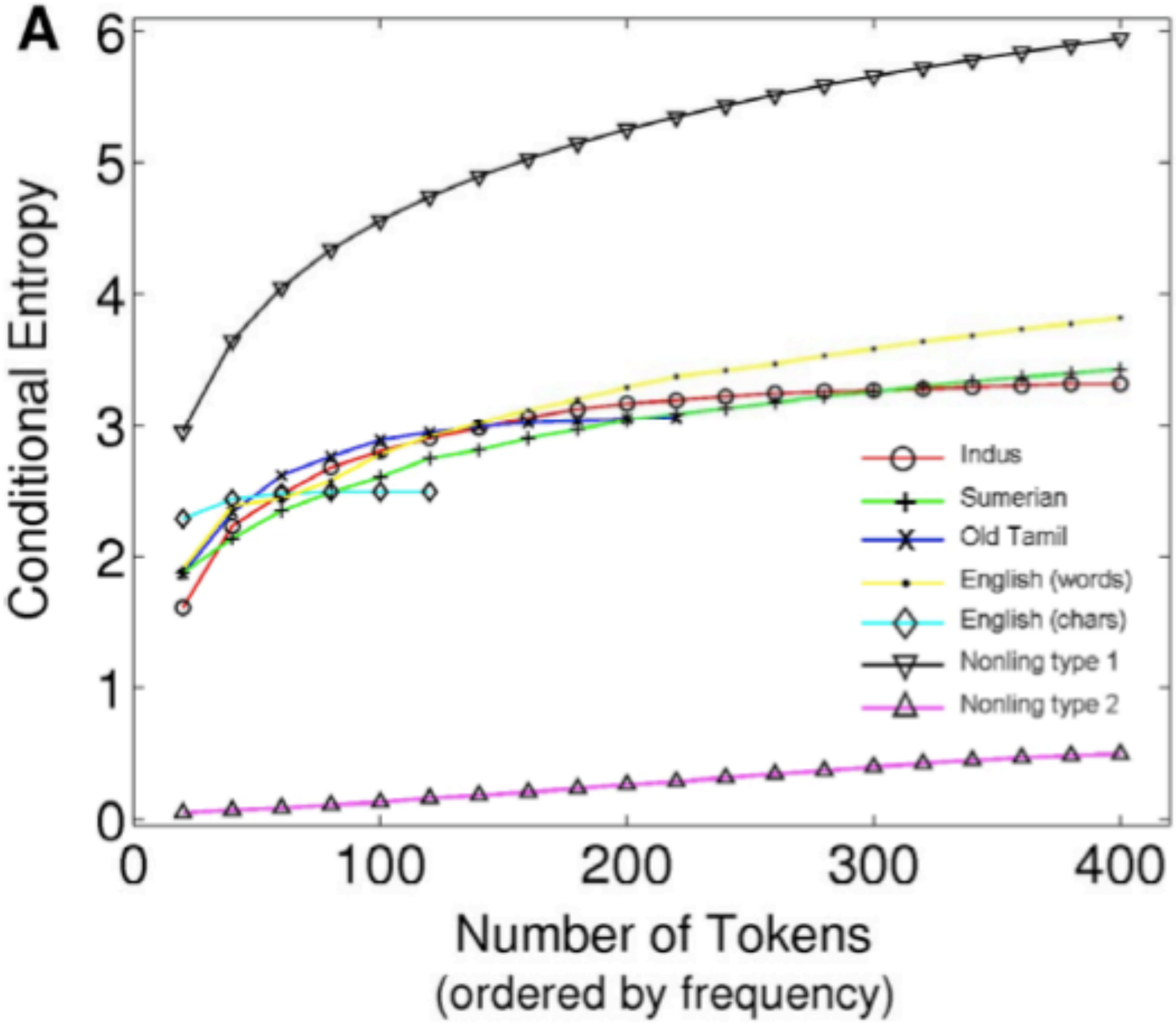
Sumerian : Oxford Corpus

DNA : Human Genome

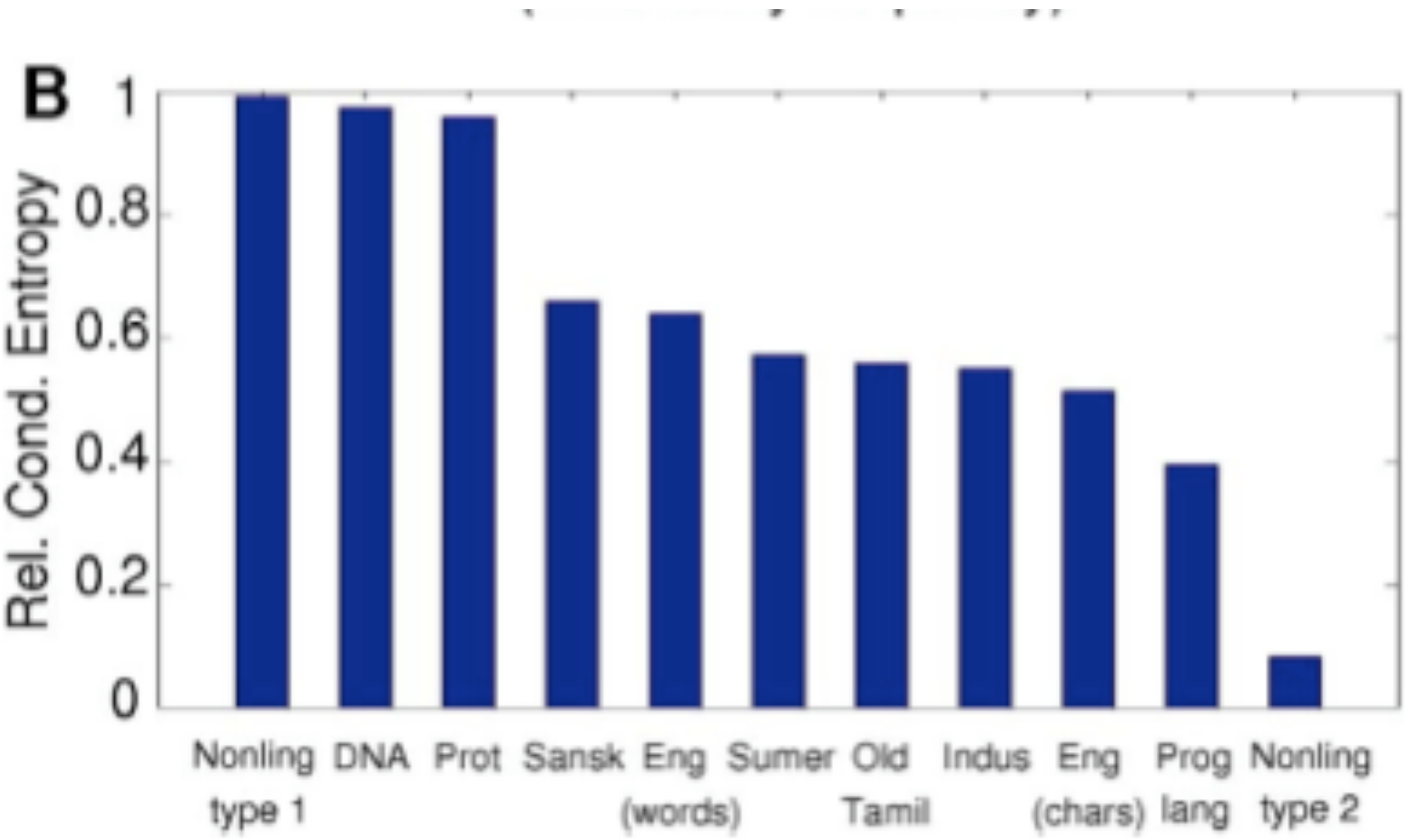
Protein : E. Coli

Fortran : CFD code

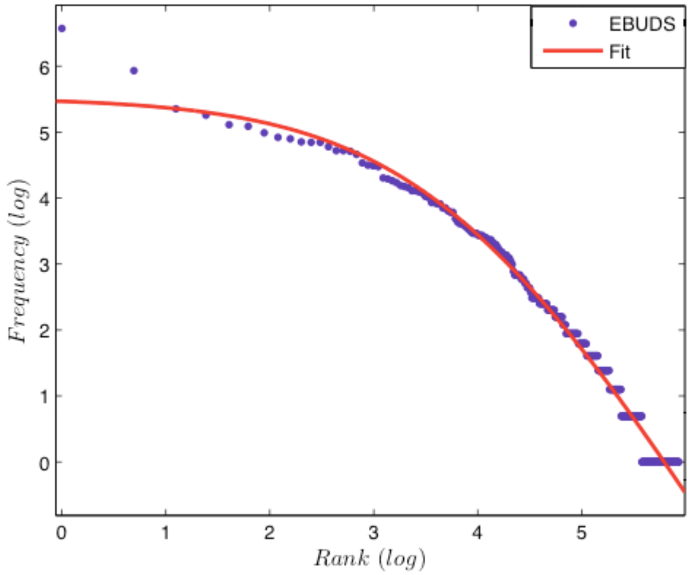
Bigram conditional entropies



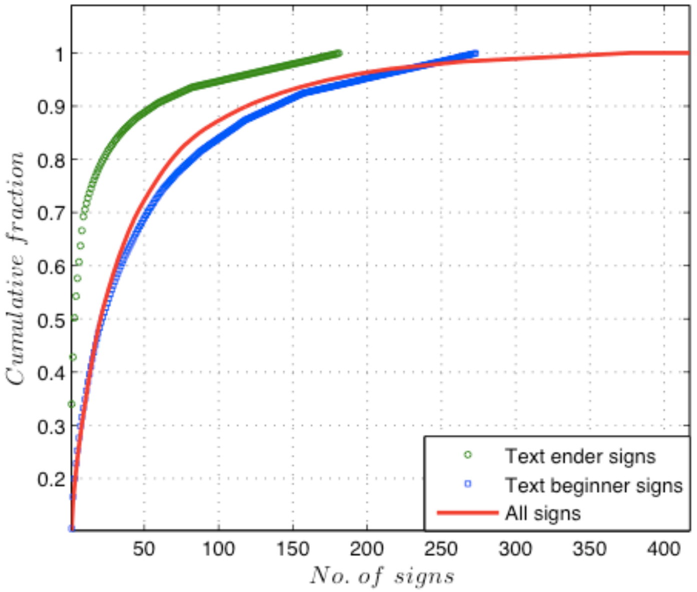
Comparing conditional entropies



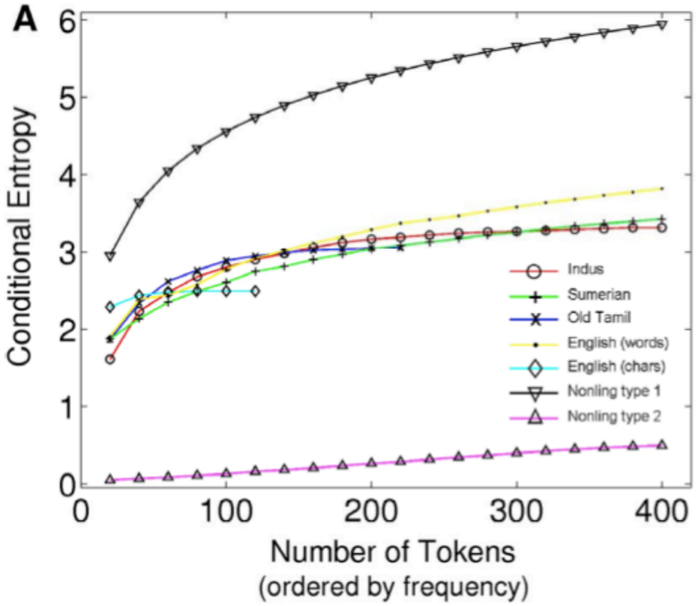
Evidence for language



Unigrams follows the Zipf–Mandelbrot law.



Clear presence of beginners and enders.



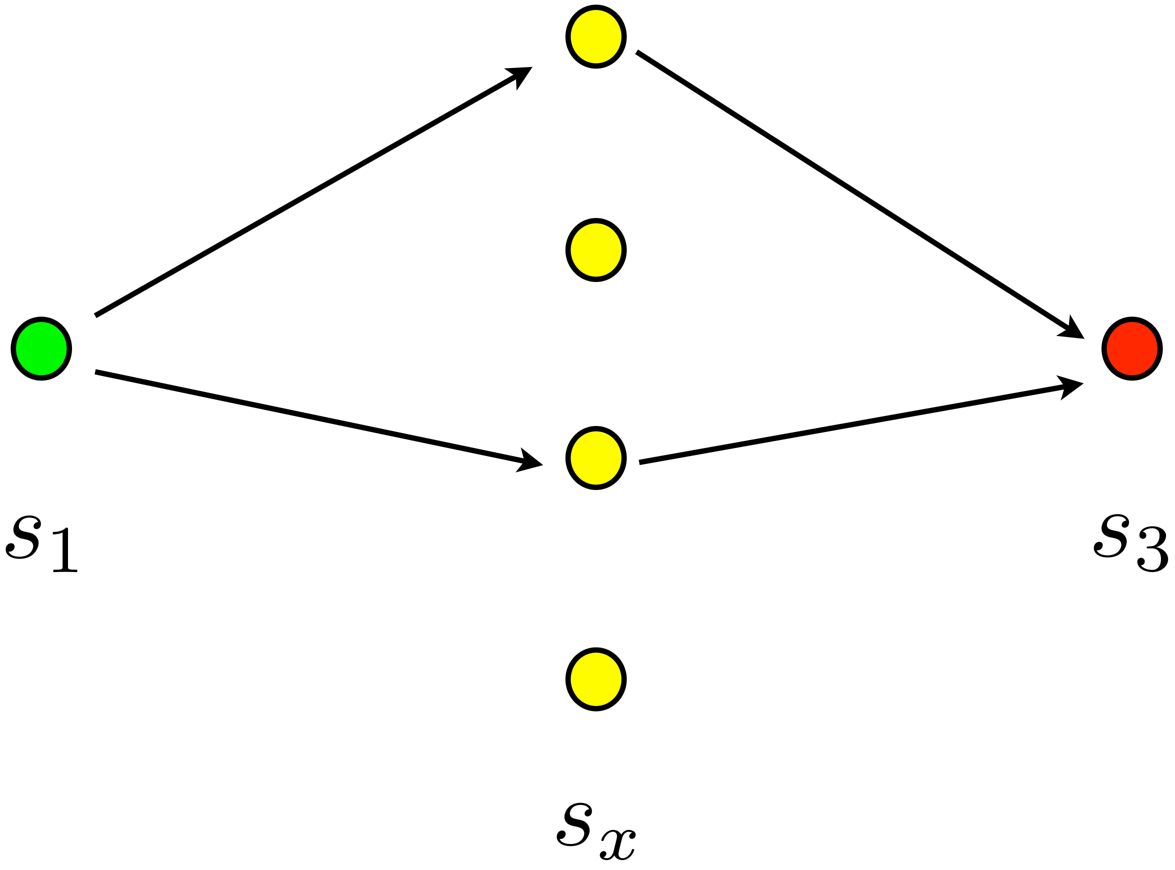
Conditional entropy is like natural language.

Conclusion : evidence in favour of language is greater than against.

An application : restoring illegible signs.

Fill in the blanks problem : c ? t

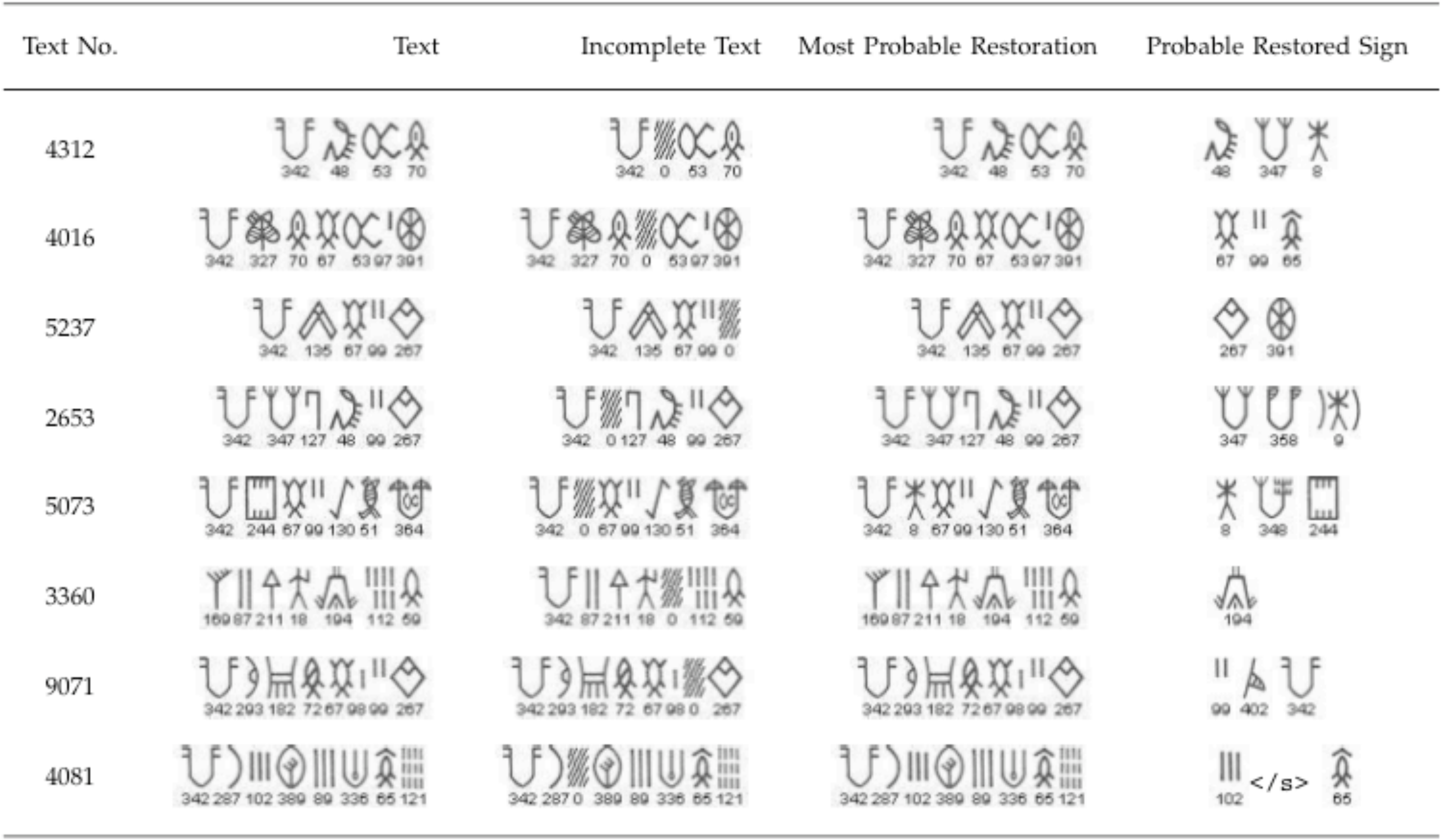
$$P(s_1 x s_3) = P(s_3|x)P(x|s_1)P(s_1)$$



Most probable path in state-space gives the best estimate of missing sign. For large spaces, we use the Viterbi algorithm.

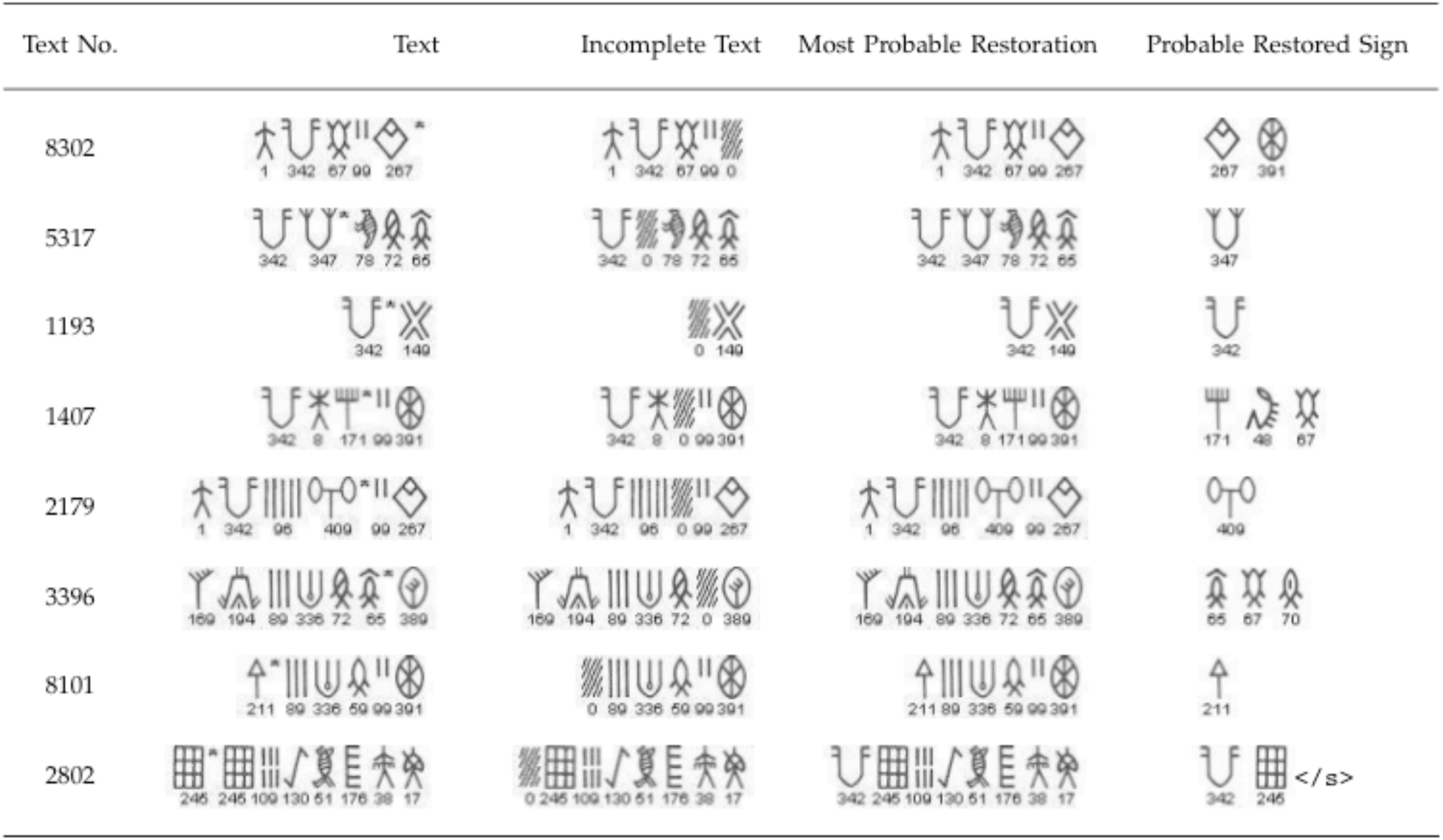
Benchmarking the restoration algorithm

Success rate on simulated examples is greater than 75% for most probable sign.



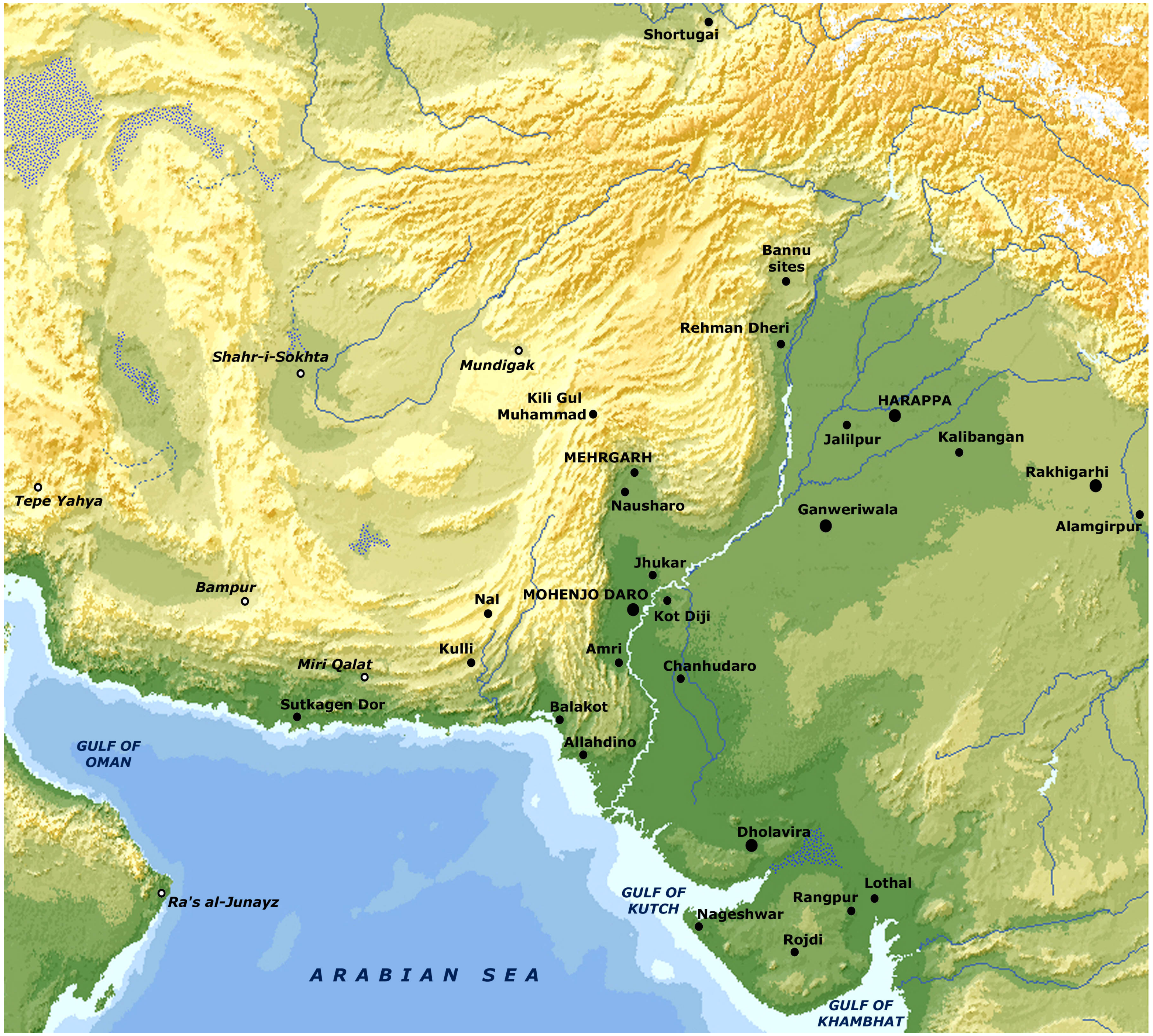
Text No.	Text	Incomplete Text	Most Probable Restoration	Probable Restored Sign
4312	342 48 53 70	342 0 53 70	342 48 53 70	48 347 8
4016	342 327 70 67 53 97 391	342 327 70 0 53 97 391	342 327 70 67 53 97 391	67 99 65
5237	342 135 67 99 267	342 135 67 99 0	342 135 67 99 267	267 391
2653	342 347 127 48 99 267	342 0 127 48 99 267	342 347 127 48 99 267	347 358 9
5073	342 244 67 99 130 51 364	342 0 67 99 130 51 364	342 8 67 99 130 51 364	8 348 244
3360	169 87 211 18 194 112 59	342 87 211 18 0 112 59	169 87 211 18 194 112 59	194
9071	342 293 182 72 67 98 99 267	342 293 182 72 67 98 0 267	342 293 182 72 67 98 99 267	99 402 342
4081	342 287 102 389 89 336 65 121	342 287 0 389 89 336 65 121	342 287 102 389 89 336 65 121	102 </s> 65

Restoring damaged signs in Mahadevan corpus

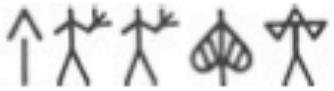
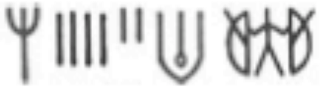
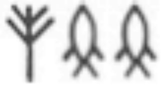
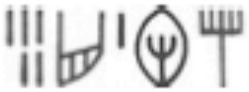
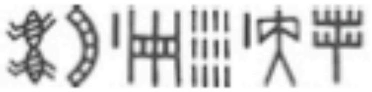
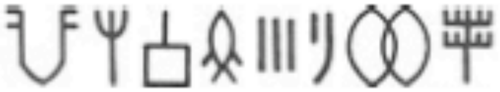
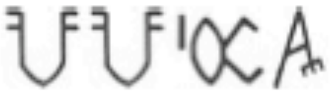
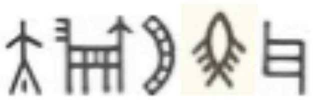


Text No.	Text	Incomplete Text	Most Probable Restoration	Probable Restored Sign
8302	1 342 67 99 267*	1 342 67 99 0	1 342 67 99 267	267 391
5317	342 347* 78 72 65	342 0 78 72 65	342 347 78 72 65	347
1193	342* 149	0 149	342 149	342
1407	342 8 171* 99 391	342 8 0 99 391	342 8 171 99 391	171 48 67
2179	1 342 96 409* 99 267	1 342 96 0 99 267	1 342 96 409 99 267	409
3396	169 104 89 336 72 65* 389	169 104 89 336 72 0 389	169 104 89 336 72 65 389	65 67 70
8101	211* 89 336 59 99 391	0 89 336 59 99 391	211 89 336 59 99 391	211
2802	246* 245 109 130 51 176 38 17	0 245 109 130 51 176 38 17	342 245 109 130 51 176 38 17	342 246 </s>

West Asian seals



Another useful application : different ‘languages’ ?

West Asian Text (from [11])	Likelihood
	0
	2.71×10^{-10}
	6.32×10^{-8}
	4.66×10^{-14}
	0
	8.82×10^{-12}
	1.20×10^{-12}
	2.22×10^{-17}
Indus valley held-out texts (median)	6.85×10^{-8}

Likelihood = P(D|H)
= P(T|M)

$$
\begin{aligned}
P(s_1 s_2 \ldots s_N) &= P(s_N | s_{N-1}) \\
&\times P(s_{N-1} | s_{N-2}) \\
&\vdots \\
&\times P(s_2 | s_1) \\
&\times P(s_1)
\end{aligned}
$$

Conclusion : West Asian texts are structurally different from the Indus texts.
Speculation : Different language ? Different names ?

Future work

- Enlarge the space of instances : more linguistic and non-linguistic systems. Enlarge the metrics used : entropy of n-grams.

- Induce classes from the Markov chain. This may help uncover parts of speech.

- Use algorithmic complexity (Kolmogorov entropy) to distinguish language from non-language.

- Borrow techniques from bio-informatics, e.g. motif-recognition in DNA to help recognise motifs.

Thanks to Vikram for inviting me to speak.

Thank you for your attention.

Epigraphist’s view of Markov chains

Markov

chains

© harappa.com