Indian Journal of History of Science, 51.2.1 (2016) 292-302 DOI: 10.16943/ijhs/2016/v51i2/48441

Project Reports

Study of the Indus Valley Scripts through Linguistic and Markov Chain Methods*

Niladri Sarkar**

1. Introduction

The main aim of this project was to study and apply statistical and linguistic methods in order to understand the nature and the structure of the Indus Valley Scripts. The focus was on the application of the Markov Chain Method on the Indus Scripts. A new method based on Metropolis Algorithm has also been used for deciphering the scripts. Indus Valley Civilization was one of the most technologically advanced civilizations of the ancient world. It existed roughly between 6000 BC-1800 BC and was at its peak between 2550 BC-1900 BC. This civilization was contemporary with the ancient river valley civilizations of Nile, Euphrates and Tigris. It had well established trading relations with its contemporary civilizations. It was an urban civilization with very well planned cities and ports. There were magnificent drainage systems, well built roads, facilities of water storage, great public baths and very well planned houses made of burnt bricks. One feature that is quite unusual from its other contemporary civilizations is the relative even distribution of their wealth. There is no evidence of any extravagant monument building by any king or any influential person of that era to placate any supernatural and superstitious beliefs if at all existed(Kenoyer, 1998). These salient features described above makes it more superior civilization than its contemporaries. However, unlike its counterparts in Egypt, Persia and Mesopotamia, the Indus Civilization did not survive until the historical times. A gap of about 1200 years between the sudden demise of the Indus valley civilization and the golden period of Ashoka remains. Many scholars and archaeologists have shown some similarity between the religion and culture of this civilization and those of later known times. Some of the symbols and motifs discovered during various excavations are still seen in present times. The most common example is the symbol known as the '*svastika'*, incised on hundreds of Harappan tablets, generally single but occasionally double, with no preferred direction.

This symbol continues to be depicted on pottery at several early sites, on punch-marked coins, on some of Ashoka's edicts and other early inscriptions. Another typical symbol is the 'endless knot' which appears on several inscriptions of the ninth century AD in Gujarat and can still be seen today in some of the *rangolis* drawn by Indian woman in front of their houses during auspicious occasions. The motif of 'intersecting circles' is a frequent one on Harrapan pottery. This type of design occurs on top of the Boddhi throne at Bodh Gaya, which dates from the third century BC. Apart from these, there are lots of other symbols like the hollow cross, the tree on a raised platform, the fish, the peacock, the antelope, the Pipal leaf and many other symbols which are common to Indus seals or pottery and punch-marked coins or

* The project was accomplished under the sponsorship of Indian National Commission for History of Science between August 2012 and July 2015.

** Assistant Professor, Department of Physics, Birla Institute of Technology & Science, Pilani-333031; Email: niladri_2001@yahoo.com

Fig. 1. Typical Indus Valley Seals which are painted for more clarity. The seals depict a one horned bull and an elephant with inscriptions at the top. There are similar seals of one horned bull with different inscriptions

other artifacts of later historical times. Pipal tree is still regarded as one of the most sacred of trees in India today. All the above mentioned facts point to a very strong link between people and culture of the Indus Valley civilizations and those of known historical times, which roughly begin from 500 BC. We can identify ourselves with the people and culture of historical times due to the continuity of the communication between the past and present times through various written and engraved records of the past, which have been deciphered successfully till date. However, the moment we try to connect ourselves with the people of Indus Valley, a missing link appears. This is because we haven't completely understood the Indus Scripts properly. This is the greatest obstacle towards understanding this civilization, and a successful decipherment of this script may bridge the gap between the Indus Valley Civilization and historical times. Only written evidences that we see today are the Indus Scripts which is found on the thousands of terracotta seals excavated from various archeological sites of Harappan and Mohenjodaro(Kenoyer, 1998, Kak, 1987).

This was not a prehistoric settlement where people were hunters and fisherman. The concept of wheels was known to them. People had developed various skills like pottery and had the knowledge of metallurgy. There were magnificent drainage system, well built roads, facilities for water storage, great baths and planned houses made of burned bricks of optimized dimensions to sustain extreme environmental conditions. They also made toys for children. If we try to reconstruct a model out of the ruins of Indus Valley Civilization, it will very well resemble with any of the ancient urban civilization with planned cities and ports. One can see such models in the Harrapan gallery of the National Museum in New-Delhi. Hence forth, so many things should have developed during that era along with the development of a language.

In order to correctly interpret the Indus Scripts, following objectives and methodologies were adopted. These are:

(i) Linguistic analysis of the Indus Script to determine the grammatical structure of the texts and its affinity to documented natural languages.

(ii) Determination of the Entropic evidences for the linguistic structure in the Indus script using Markov chain model.

(iii) Syntactic connections between the signs of the Indus Scripts should be established.

(iv) Application of conditional entropy on Indus Scripts and comparing it with other known scripts and non-scripts will establish a strong evidence for the linguistic structure of the Indus Scripts.

The project was accomplished under the following chapters:

I. Traces of the lost civilization and its scripts.
II. Scientific & linguistic approaches towards understanding of the Indus Scripts.
III. Scientific Evidence of a Script and a Language by Markov Chain Model.
IV. Conclusions and future work to be done.

2. Traces of the lost civilization and its scripts

This section starts with a concept of sudden end of an era. The reason for this is due to the fact there exist a gap of about 1200 years of missing information between the Indus Period to the Indian historical times. If we look into other ancient civilizations, there exists a strong link between the ancient times and modern times. One can see an analogous example in the catastrophic end of the Dinosaurs after the Jurassic era. However, many other species which existed during that era continue to survive till the present times. Hence, the only things which we find today after the sudden extinction are the fossils in the case of dinosaurs and the traces of the ancient cities and its scripts in the case of Indus Valley Civilization although both the events may have occurred during very different eras separated by millions of years. The dawn of the mankind occurred much later after the catastrophic end of the Jurassic Age. It evolved very slowly with man learned to hunt and tool making for various purposes. With time and ages, the humans very slowly learnt the art of speaking which created very early languages with a limited stock of words. Such primitive languages even survive today in the present times with the tribes of Andaman Islands and aboriginal people of Australia, New-Zealand and other native people small Pacific Islands. Many of these primitive languages are agglutinative where the complex words are strings formed by joining the morphemes together without changing the spellings and the phonetics of the individual units. With the evolution and enrichment of the language, the ideas were exchanged between persons and from hunting and cave dwelling evolved early settlements which finally resulted into small villages which finally integrated to form towns and cities of the early civilizations.

The cities of the ancient times like Mohenjodaro and Harrapa at sometimes housed some more than 40,000 inhabitants. These were the important and greatest cities and the centers of trades during that time.

Fig. 2. (a) The Sudden and Catastrophic End of an Era and (b) the only evidence of the life that existed during the Jurassic era can be obtained from the fossil records

Fig. 3. From an Early Settlement to a well established Port Cities which had trade links with Western Asia

Apart from these, from the evidences of the types of the Indus Seals and artifacts which were found in many parts of Modern North-Western India and the parts of Western Pakistan, one can infer that there existed a regular network of towns and cities which had spread over an area of about 1.25 Million Sq. Kms. There are very strong evidences that there existed Harrapan contacts with the Western Asia. Similar type of signs of the Indus seals have been found in some round seals which were excavated at Ur. There foreign countries have been identified as participants in the trade from 2400 BC to 1900 BC. They were Dilmun, Magan and Meluhha. Dilmun is the Islands of Tarut, Baharin and Failaka. Magan is the Oman peninsula and the Iranian Markran across the Gulf. Meluhha is the Indo (undivided India)-Iranian borderlands and the Indus Civilization. A very important picture which is found as drawing on the pots in far off places from Indus valley is that of peacock. Woods for making chairs and Ivory were some of the most important products that were imported from the port cities of the Indus Valley. There exist evidences of irrigation channels. The potters mastered the techniques of air reduction to produce high quality ceramics. The ceramic style which appeared in the Indus Valley is *Kot Dijjian ware*. Most of the globular vessels had a wide band painted around its neck. The remains of typical Harrapan large storage jars were discovered at Wadi Asimah in the Oman Mountains in 1987. The correct interpretation of the Indus Valley Script is a challenging problem. We are still unsure about its origin and know very little about it. We don't know whether these signs represent a well developed language. We can guess it by looking into the developments that had occurred during that period. From the ruins of Indus Valley, one can infer that it was an advanced civilization which flourished along with other contemporary river valley civilizations of Nile, Euphrates and Tigris. It had trade links with Western Asia. One thing is certain that there existed very well defined language because without it they wouldn't have established trade relations.

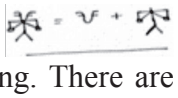
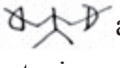
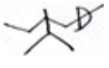
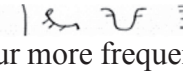
3. Scientific & Linguistic Approaches Towards understanding of the Indus Scripts

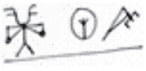
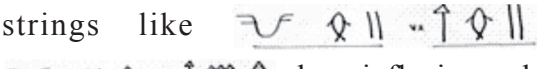
This chapter deals with the Scientific and Linguistic approaches made for understanding the nature of the Indus Scripts (Kak, 1987; Jayaswal, 1933; Mahadevan, 1977; Fairservis,1992; Parpola, 1994). Over 4000 inscriptions mostly on seals, tablets and pottery have been excavated from various Harrappan sites. About 400 distinct signs have been identified of which only 200 have been used more than five times. Various methodical attempts have been made in the past ninety years. Sir John Marshall in 1924 was the first to suggest

that the Indus civilization was non-Aryan and that its language probably close to the Dravidian family. Piero Meriggi considered the Dravidian language of Brahui, spoken in Baluchistan as the only reasonable contender. The Spanish Jesuit Henry Heras who spent most of his time in India, had numerous publications on the Indus script that culminated in 1953 in a book called *Studies in Proto-Indo-Mediterranean culture*. Experts have long pointed to important structural features of the Indus script that resemble with the features in Brahmi script. The use of composite signs and modifiers, respectively call to mind the use of composite letters and diacritical marks in Brahmi to denote vowels as in later Indian scripts. Subhash Kak, Stephen Langdon and G.R. Hunter have independently proposed a parallel between the Indus and Brahmi scripts. The mysterious inscription in the Vikramkhol caves in Odisha has been studied by eminent epigraphist K.P. Jayaswal and has concluded that the inscription provides a link between the letter forms of the Indus Script and the Brahmi Script. Soviet scholars, headed by Yurij V. Knorozov, who have published extensively on the Indus script, have based their work on the number of different signs they could distinguish. They have come to the conclusion that the script is logo-syllabic rather than syllabic or alphabetic. Based on the statistical analysis of the distribution of various signs, they have concluded that signs which occur more frequently represented grammatical markers, while those occurring less so, represented root morphemes. Iravatham Mahadevan has contributed greatly towards the documentation of the Indus scripts and its technical aspects. He has stressed on the pictorial meaning of the Indus signs and tried to work out its interpretations from the symbolic uses of the depicted objects in later Indian traditions. Since 1976, Walter A. Fairservis, one of the leading archaeologists specializing in the Harrappan culture has elaborated on general assumptions and principles relevant to the decipherment of the Indus script. In recent times, a Finnish team lead by Professor Parpola has presented a decipherment of a small number of interlocking Indus signs and has proposed a method which may permit further progress in decipherment. In spite of all this, none of the attempts at deciphering the Indus script has gained wide acceptance, although there have been solid contributions on various technical aspects of the script. First and the most important thing in the decipherment is the collection of the material into a corpus. Accuracy is vital in such a corpus. It is extremely important to generate various kinds of statistical data namely, frequencies of individual signs and sign combinations, of text with different lengths, of object types, of iconographic motifs etc, and to tabulate the distributions of signs and other variables in relation to different criteria.

Fig. 4. The Inscriptions of Vikramkhol Caves in Odisha

For the linguistic analysis of the script, one has to first know the direction of the writing. Asko Parpola has proposed the following criteria for determining the direction of the writing. When signs are incised on pots, the lines drawn first could become partially obliterated by those drawn later. One criterion proposed for the determination of the direction is based on the idea that the animals depicted on the Indus seals, and the pictograms in the accompanying inscriptions should face in the direction in which the writing was to be read. More dependable evidence is supplied by the spacing, both in the individual

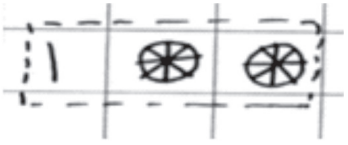
signs and in inscriptions. Sometimes the inscription is shorter than the allotted space. Texts starting from the right edge and leaving an unused space to the left may reasonably be expected to run from right to left. Numerous tests have established right to left as the most probable direction of writing in the Indus inscriptions. This was already clear in the early 1930s to Gadd and Smith, Marshall and Hunter. However, a few uncertain cases still remain. Iravatham Mahadevan, who carefully recorded the direction of the original in each of his 3,573 lines, distinguished 2,974 lines running from right to left and 235 going from left to right, in addition to ambiguous sequences, such as 190 single-sign lines, 12 symmetrical sequences and 155 cases that are doubtful on account of damaged or illegible lines. A top-to-bottom sequence is recorded for seven lines. After the completion of the preliminary tasks of collecting and establishing the direction of writing, several attempts have been made for a structural analysis of the Indus script by Asko Papola, Iravatham Mahadevan and others. There exist evidences of patterns which reflects an intrinsic structure in the Indus Scripts. There are evidences on the direction of the writing. It has been established by various methods that the direction of the writing is from right to left. Also, the scripts show diacritical modification of basic signs and ligaturing. For example,  shows an evidence of this ligaturing. There are evidences of double ligaturing as well. For example,  shows the evidence of double ligature. The example  and  show double and single ligaturing. There are mirror images like  . There are some signs which occur more frequently than the others. For example, the sign  occurs as a terminal in many strings. This sign shows diacritical modification as well. This modification occurs as respectively.

Also, the fish sign  shows the same modification as  . There exist mirror images of sign pairs like  and  where the axis of reflection lies horizontally. Sometimes signs occur with two, three and four fold repetitions. Some example of such repetitions are  , , , and respectively. There are evidences of two different words in a string. For example,  and  have the word  which is common in both strings. One can split a string into three segments as  . There are also evidences of inflexion in the strings. Some pair of strings like  and  show inflexions where the terminals and act as inflexion suffixes.

4. The inscription from the city Dholavira in Kutch, Gujarat

A very important discovery by Dr R. S. Bisht is an inscription which may have indicated very important information about the city, possibly its name. These are ten signs made from the pieces of crystalline rock. The following figure shows the inscription and two other inscriptions that bear a same pattern of three signs.

Here, we identify the sign sequence,



a possible Trigram.

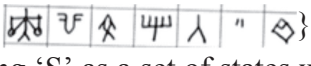
5. Scientific Evidence of a Script and a Language by Markov Chain Model

The problem here is that both the language as well as the script is unknown to us. Only one thing is certain that these are not randomly generated string of signs. These were used by people of that era for trade purposes as one can guess from the seals as similar type of seals are found in Western Asia. However the Indus seals were rectangular in shape, whereas the foreign seals where mostly circular in shape. The figures and the scripts had similarities. Various tests based on the concepts of entropy in the communication theory have been applied on these scripts. The results have shown that there is an intrinsic statistical structure in these scripts. One can construct crosswords out of these strings which indicate non zero redundancy.

6. Concepts of Entropy in Communication theory

How to analyze these seals is one of the most challenging problems. The first thing which comes to our mind is whether these are a part of some scripts which lead to a language or just signs which may convey some ideas like the modern day traffic sign boards. For this, one may take help from the concepts of entropy in communication theory. According to Shannon's information theory, there is entropy associated with every spoken language (Shannon, 1948). One can understand it in terms of the existence of 2-D and 3-D cross-word puzzles. The ratio of the entropy of a source to the maximum value it can have is called the relative entropy. Entropy in a language is a measure of the number of degrees of freedom to construct meaningful words from a set of limited units which may be alphabets in case of English language or some set of signs in the case of Indus Scripts. There exists in every language whether it is a natural spoken language or a computer language, an intrinsic statistical structure which invokes a constraint in the usage of the alphabets or building blocks. For example, the occurrence of "u" after "q" in English language is highly probable than the occurrence of "q" after "u". Hence, when we write in any language, a fraction '*F*' of the letters chosen is determined by the internal structure of the language and the rest '1-*F*' letters can be chosen freely. This '*F*' is related to the redundancy of a language. If redundancy is zero, then we will not have a language as all combinations are possible.

7. What is a Markov Chain?

We will now describe the Markov chain in terms of a language (Shannon, 1948; Rao & Rajesh, 2009; Rao 2009). Suppose we have a string of symbols of a language which may be known or unknown. For examples, let's take a string from an Indus seal, S={  }. Now, we define the string 'S' as a set of states where each symbol is an element of the set. Now this set 'S' would be a sub-set of the language which may contain a very large collection of strings which are sequences of the signs. Hence, the language, L={ $S_1, S_2, \ldots, S_N, \ldots$} will be the set of all the possible meaningful sequences of the signs. Now, if we take a particular string 'S' from 'L'. Now, one can only move from the first sign of this string to the next sign and so on with a transition probability, Pij. Now, the transition probability can be obtained from the transition matrix of the particular language 'L'. One can understand it from the transition matrix of signs which is shown in fig. 5. From a matrix like this, we can construct the relevant strings which may belong to the language

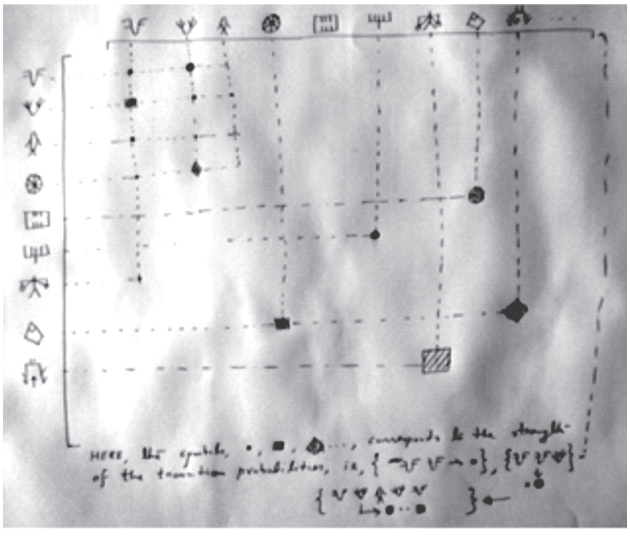


Fig. 5. A Matrix of Indus Signs for First Order Transition

set L. Here, we find that corresponding to every sign in the outer column there is an associated strength assigned for the occurrence of the next sign. Now, the construction of such a matrix for the Indus script is a challenging problem. For example for a known language like English, it is very easy to construct such a matrix by writing a computer code in MATLAB or C. All that is required for this is a book or a novel in a digital form. Now, the signs which will be placed at the other column and the row of the matrix will be the alphabets and the space. The computer program will assign the strengths for the first order transition. A simple observation that one sees here is the strength of the occurrence of the alphabet 'u' after 'q' than the strength of the occurrence of the alphabet 'q' after 'u'. It will observed that the former's strength will be very large compared to the latter's strength. Now with the help of the transition matrix, one can construct a relevant string which may belong to the language set. The algorithm for generating a first order sequence of strings is shown below.

(i) Choose a sign randomly from the outer column of the Matrix.

(ii) Now, generate a random number between '0' to '1'.

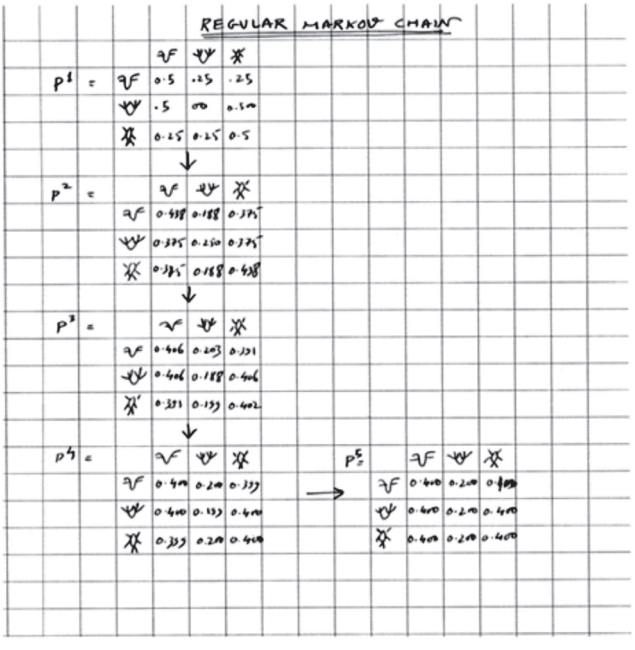
(iii) Now, the Next Sign in the sequence would be according to the value of the random number generated and this will correspond to the first order transition strength.

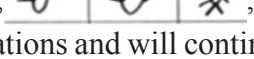
(iv) Once, the sign is assigned, repeat the process till a terminal sign comes in the string.

In this way, one can generate lot of strings of which few may belong to the language set 'L'.

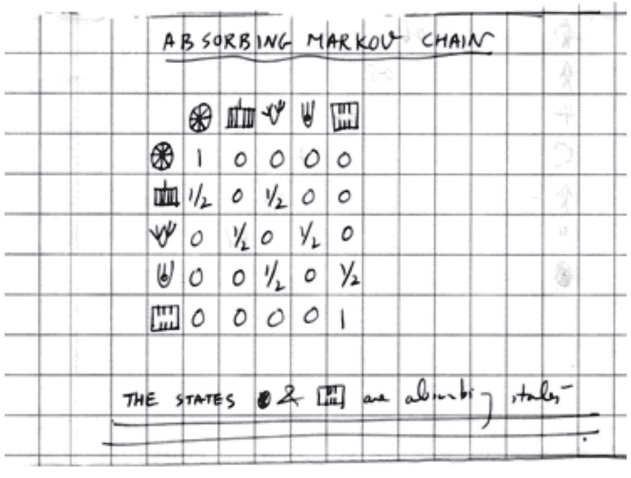
8. Sequences of Strings Generated by Markov Chain Method

The following figure shows the saturation of the first order transition matrix for three Indus signs. The elements of the matrices will not change after 5 iterations.

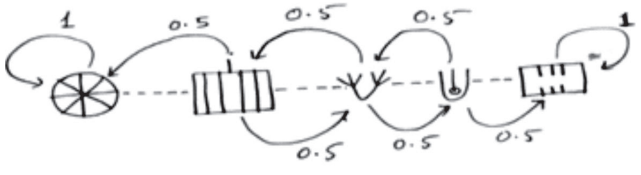


The probabilities of the three states,  , are 0.4,0.2 and 0.4 after 5 iterations and will continue to be that for higher iterations. The states will be independent of the starting states. This type of chain is called "Regular Markov Chain". The following figure shows a transition matrix for an Absorbing Markov Chain. The states and are the absorbing states. It

is impossible to leave that state. However, it is possible that one can go from all the non-absorbing states to the final absorbing state.



The following figure shows the transitions from transient states to final absorbing states:

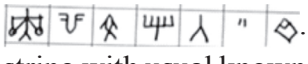


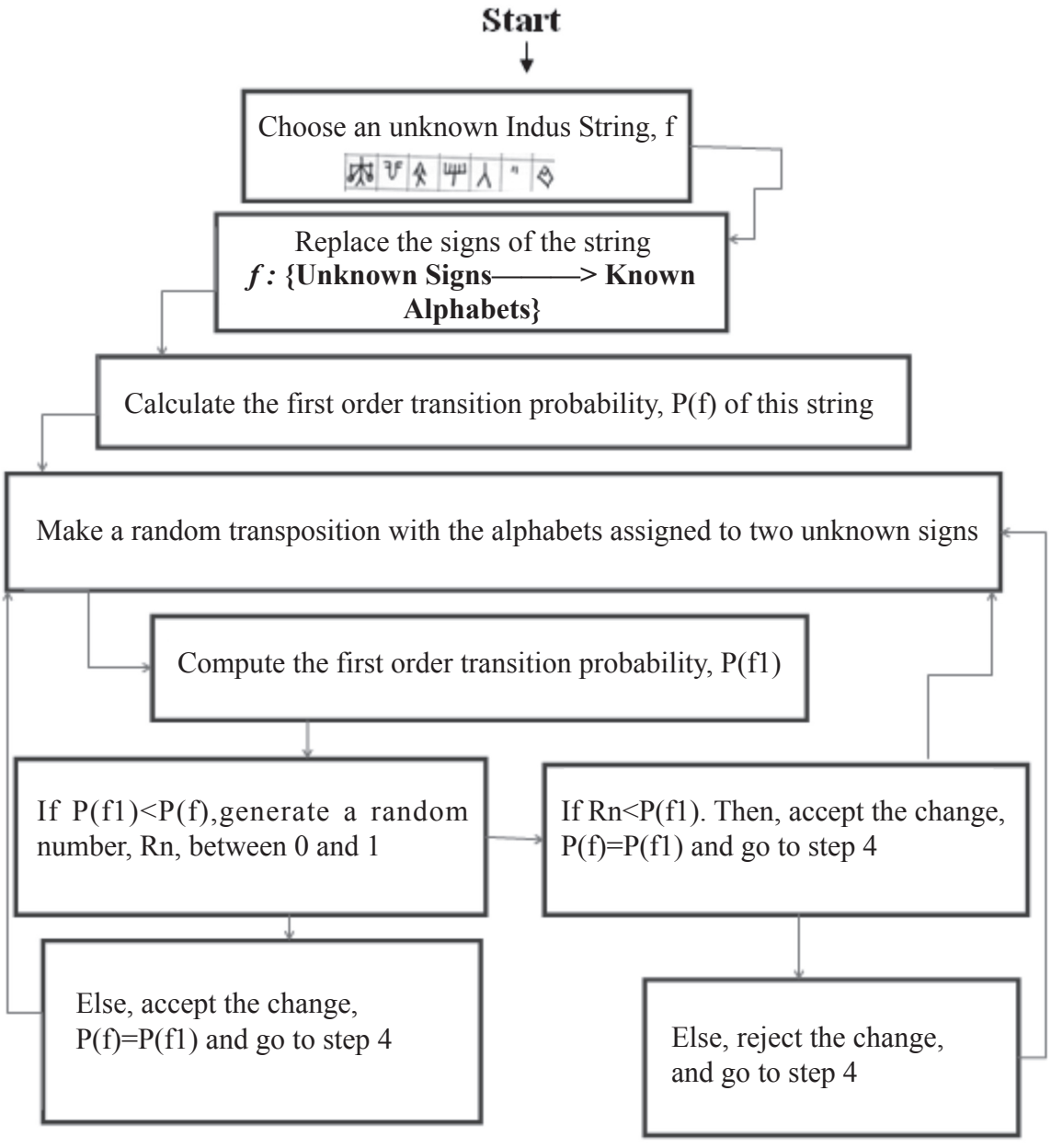
The following figure shows some possible strings of Indus scripts obtained from the above transition matrix.

Possible strings

Having, made such logical observations, one needs a scientific methodology to decipher these scripts. The technique which can be successfully applied here is based on Markov Chain and Monte Carlo Method using Metropolis algorithm(Harvey). Under this an intelligent guess about the language is to be made by narrowing down the known ancient languages which were spoken during that particular era in Western Asia. Then a statistics of that particular language has to be recorded by recording the first order transitions, proportion of the consecutive signs. A matrix of consecutive signs will have the transition probabilities as its elements. Then, we have to replace properly the unknown signs of the Indus Scripts by the known signs of that particular language by comparing the statistics of the Indus scripts with the known signs. Then, we have to compute the first order transition probability of this new assembly. Then, we have to make a random transposition of the assigned signs. After this, again the first order transition probability is to be calculated. We have to keep on making random transpositions till the program converges to the first order transition probability of the known ancient language. This type of technique will work, provided we get some texts of known ancient languages of that era.

9. Monte Carlo Metropolis Algorithm

Start with a known text of ancient language. Calculate the probability of first order transition, i.e, the probability of the occurrence of the consecutive symbols. Call it $P(f_0)$. Now, choose an unknown Indus String, ***f***=  . Replace the signs of the string with usual known alphabets, ***f*****: {Unknown Signs———> Known Alphabets}.** Calculate the first order transition probability, P(f) of this string. Now, make a random transposition with the alphabets assigned to two unknown signs. Again, compute the first order transition probability, P(f1). If P(f1)>P(f), then accept the change. If P(f1)<P(f),generate a random number, Rn, between 0 and 1. If Rn<P(f1). Then, accept the change. Else, the transposition is rejected. Keep on repeating the above procedure till P(f) saturates i.e, no more transposition is allowed.



Flowchart for the Monte Carlo Metropolis Algorithm

10. Conclusions and Future Work to be done

An attempt based on Shannon's information theory has been attempted on Indus Scripts. The construction of the crosswords gave us some weak indication of a language in the script. This can be said because the strings have been chosen random from the corpus. However, it doesn't give us any clue about the origin of the language and its connection with the present closely related language. This is very important since we cannot proceed without knowing the language. So it is important to narrow down our choices on ancient languages that were spoken during that era. Now, by targeting a particular language, the Monte-Carlo method with Metropolis algorithm should be used on the unknown Indus Scripts. The Metropolis algorithm

will insure the convergence of the problem. The most ambitious work that can be attempted is the testing of the Markov Chain Monte Carlo Method using Metropolis algorithms.

Bibliography

Fairservis, Walter A, *The Harappan Civilization and its Writing: a Model for the Decipherment of the Indus Script*, New Delhi, 1992.

Harvey, Gould; Jan, Tobochnik and Wolfgang, Christian. *An Introduction to Computer Simulation Methods: Applications to Physical Systems* (3rd Edition).

Jayaswal, K. P., 'The Vikramkhol Inscription', *The Indian Antiquary*, 1933, vol. LXII, p.60.

Kak, Subhash. 'On the Decipherment of the Indus Script: A Preliminary Study of its Connection with Brahmi', *IJHS*. 22(1987): 51-62.

Kenoyer, J. M. *Ancient Cities of the Indus Valley Civilization*, Oxford University Press & American Institute of Pakistan Studies, Karachi & Islamabad, 1998.

Mahadevan, I , *The Indus Script: Text, Concordance and Tables*, Archeological Survey of India, New Delhi, 1977.

Parpola, A, *Deciphering the Indus Script*, Cambridge University Press, 1994.

Rao, P. N., Rajesh et. al., A Markov Model of the Indus Script, *PNAS*, 5 August, 2009.

Rao, RPN, et. al., Entropic Evidence for Linguistic Structure in the Indus Script, *Science* 324 (2009): 1165.

Shannon, C. E. *The Bell System Technical Journal*, 27, (1948): 379.