

Network analysis reveals structure indicative of syntax in the corpus of undeciphered Indus civilization inscriptions

Sitabhra Sinha

Institute of Mathematical Sciences
Taramani, Chennai 600113, India
sitabhra@imsc.res.in

Raj Kumar Pan

Institute of Mathematical Sciences
Taramani, Chennai 600113
rajkp@imsc.res.in

Nisha Yadav

Tata Institute of Fundamental Research,
Homi Bhabha Road, Mumbai
400005, India
Y_nisha@tifr.res.in

Mayank Vahia

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

Iravatham Mahadevan

Indus Research Center,
Taramani, Chennai 600113, India
iravatham@vsnl.net

Abstract

Archaeological excavations in the sites of the Indus Valley civilization (2500-1900 BCE) in Pakistan and northwestern India have unearthed a large number of artifacts with inscriptions made up of hundreds of distinct signs. To date, there is no generally accepted decipherment of these sign sequences, and there have been suggestions that the signs could be non-linguistic. Here we apply complex network analysis techniques on the database of available Indus inscriptions, with the aim of detecting patterns indicative of syntactic structure in this sign system. Our results show the presence of regularities, e.g., in the segmentation trees of the sequences, that suggest the existence of a grammar underlying the construction of the sequences.

1 Introduction

The recent interest in complex networks among physicists over the past few years has meant that the graph theoretic paradigm has been applied to many different areas, including networks defined in corpora of textual units (Mehler, 2008), and has often revealed hitherto unsuspected patterns. While graph-based representation of texts had been used for some time in natural language processing tasks, such as, text parsing, disam-

biguation and clustering (Radev and Mihalcea, 2008), the approach based on the new physics of complex networks often asks questions from a different perspective that can shed new light on the organization of linguistic structure. For example, networks constructed on the basis of co-occurrence of words in sentences have been seen to exhibit the small-world effect, i.e., a small average distance between any pair of arbitrarily chosen words, and, a scale-free distribution of the number of words a given word is connected to (i.e., its degree) (Ferrer i Cancho and Sole, 2001). These properties have been proposed to reflect the evolutionary history of lexicons and the origin of their flexibility and combinatorial nature. Note that, a recent study of a lexical network of words that are phonological neighbors has found that the degree distribution might be better fit by an exponential rather than a power-law function (Vitevitch, 2008). A theoretical model for such word co-occurrence network, which treats language as a self-organizing network of interacting words, has led to the suggestion that languages may have a core (the “kernel lexicon”) that does not vary as the language evolves (Dorogovtsev and Mendes, 2001). However, even though text and speech are sequential, the local correlation between immediately consecutive words may not describe natural languages well – due to the presence of non-local relations between words that occur apart from each other in a sentence. Therefore, network



Fig. 1: A typical example of Indus sign sequence (having 8 distinct signs) occurring at the top of a seal, with the picture of a “unicorn” in the foreground (i.e., the field symbol), one of the common animal motifs observed in such artifacts. Note that, on seals, the conventional order in which the signs are read (right to left) is reversed.

analysis has been extended to syntactic dependency networks, where two words are connected if they have been related syntactically in a number of sentences (Ferrer i Cancho et al, 2003). The theory of complex networks has also been used to investigate the structure of meaningful concepts in the written texts of individual authors, which have been seen to have small-world as well as scale-free characteristics (Caldeira et al, 2006). The conceptual network of a language has been explored by using the semantic relatedness of words as defined by a thesaurus, and this network too is seen to have small-world nature with scale-free degree distribution (Motter et al, 2002).

In this article, we look at a corpus of inscriptions obtained through archaeological excavations carried out in the ruins of the Indus valley civilization. The sequences comprise signs, of which there are more than four hundred unique and distinct types. Since discovery in the early part of the 20th century, there have been attempts at deciphering them. However, to date there is no generally accepted method of interpreting these inscriptions. We analyze a representative database of these sequences using techniques inspired by complex network theory. Our aim is to see whether such methods can reveal the existence of patterns suggesting syntactic organiza-

tion in the sign sequences. In the next section, we briefly introduce the historical context of the Indus inscriptions, while in Section 3, we discuss the dataset on which analysis has been carried out. Our results are reported in Section 4, and we finally conclude with a discussion of unresolved questions and further work that needs to be carried out.

2 The Indus inscriptions

The Indus civilization, also known as the Mature Harappan civilization (2500-1900 BCE), was geographically spread over what is now Pakistan and northwestern India, covering approximately a million square kilometers (Possehl, 2002). It was marked by urbanization centered around large planned cities, as evidenced in the ruins of Harappa and Mohenjo-daro. Craft specialization and long-distance trade with Mesopotamia and Central Asia have been well-demonstrated. This civilization came to an end early in the 2nd millennium BC. There were no historical records of its existence until archaeological excavations in the late 19th and early 20th century uncovered artifacts, and some of the ruined urban centers (Marshall, 1931).

Among the artifacts uncovered during these discoveries were a variety of objects (especially seals) that were inscribed with a variety of signs arranged in sequences (Fig. 1). Although found primarily on seals and their impressions (sealings), inscriptions with similar signs have also been discovered on miniature tablets, pottery, copper tablets, bronze implements, etc. Unsurprisingly, given the high sophistication of the civilization and the level of social complexity it implies, with the concomitant requirements of coordination and communication, these inscriptions have been interpreted as corresponding to writing. However, despite periodic claims about decipherment of this writing system, there has as yet been no generally accepted interpretation of the signs. The failure of decipherment is partly due to lack of knowledge about the language which the signs encode and the lack of any bilingual texts such as the Rosetta stone which was crucial in deciphering Egyptian hieroglyphs. While there is disagreement on the exact number of unique and distinct signs that occur in the inscriptions, there is overall agreement that they lie in the range of a few hundred. This rules out the possibility of the signs belonging either to an alphabetic system, which contains on average

about 25 letters (such as English) or an ideographic (e.g., Chinese) writing system, comprising more than 50,000 characters. The brevity of the inscriptions (the longest that occurs on a single line has 14 signs) and the existence of a large number of signs that occur with very low frequency have led to some alternative suggestions regarding the meaning of the sign sequences. These include the possibilities that, e.g., (i) the signs correspond to a label specifying an individual and his belongings, in the manner of heraldic badges (Fairservis, 1971) and (ii) the signs are ritual or religious symbols which do not encode speech nor serve as mnemonic devices, much as the Vinca signs or emblems in Near Eastern artifacts (Farmer et al, 2004). The latter possibility implies the absence of any syntactic structure in the Indus inscriptions, a possibility that can be tested without making any *a priori* assumptions about the meaning of the signs.

3 Description of dataset

The database for Indus sign sequences that we have used is based on the electronic concordance constructed by Mahadevan (1977), referred here as M77. This is based on the study of a total of 3573 sequences recorded from 2906 distinct artifacts and it identifies 417 unique signs. In the following we identify each sign in a sign sequence by its corresponding identification number (1, ..., 417) in M77. Most of the inscriptions seem to have been written from right to left. However, according to the convention we use, the sequence of numbers representing each text is read from left to right (i.e., the leftmost number in the sequence is read as the first sign in the inscription). Yadav et al (2008) have constructed

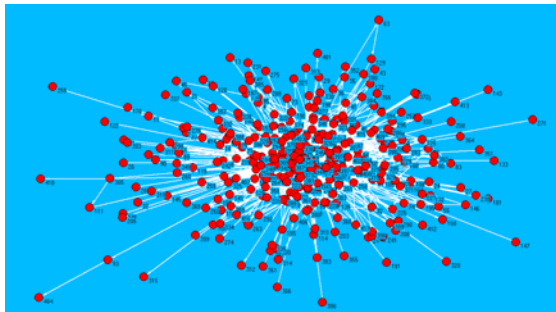


Fig. 2: The directed network of 377 distinct Indus signs in EBUDS, with arrows pointing from a preceding sign to a sign that follows it in the corpus of empirically observed sign sequences. Links are weighted by the frequency of occurrence of that particular sign pair.

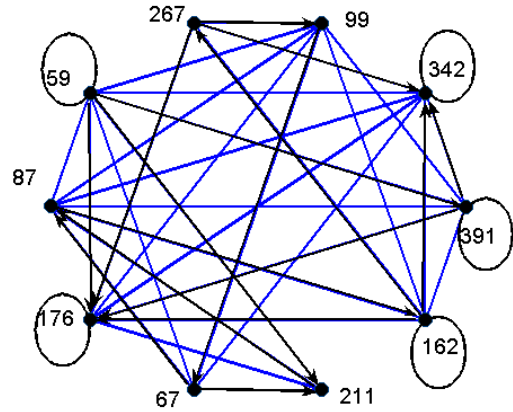


Fig. 3: The subnetwork of connections between the 10 highest frequency signs in EBUDS. Different colors are used to represent the two different orientations possible for arrows between a nodal pair (e.g., the pairs 342-162 and 162-342 are both possible and are indicated by a blue and a black arrow, respectively). Loops indicate successive occurrences of the same sign.

an Extended Basic Unique Data Set (EBUDS) by removing from M77 all sequences that are incomplete, because of the presence of signs that are damaged, lost, illegible or not unambiguously identified. Further, only sequences which are written on a single line are retained. This is to remove the ambiguity regarding the interpretation of sequences with multiple lines, namely, whether the different lines should be considered as independent sequences or whether they form one continuous sequence. Moreover, if the same sequence is found in multiple artifacts, it is represented only once in EBUDS. Following these operations, the original number of 3573 sequences in M77 is reduced to 1548 sequences in EBUDS. Moreover, 40 of the 417 unique signs in M77, which occur with relatively very low frequency, do not have any representation in EBUDS – so that latter dataset comprises 377 unique signs. However, it has been verified by Yadav et al (2008) that the frequency distributions of signs in the two datasets are qualitatively similar.

4 Results

In the following sections we report the results of applying network analysis techniques to the sign sequences in EBUDS. We should note at this point that, the distributions of the in- and out- strengths of all the nodes (i.e., the sum of the weights of the incoming and outgoing links,

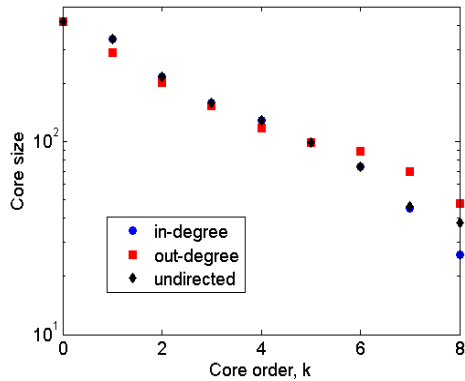


Fig. 4: Core-decomposition of the undirected and directed networks of Indus signs. For the latter, both the in-degree (circles) and out-degree (squares) cores are shown, while the undirected cores are represented with diamonds. All three core decompositions show an innermost core of order 8.

respectively) do not show a scale-free distribution.

4.1 The directed network of signs

To have a graphical view of the entire sign system, we first construct the directed network of Indus signs (Fig. 2). This has 377 nodes corresponding to the distinct, uniquely identified signs. Node i has a directed link to node j , if sign j immediately follows sign i in any of the inscriptions included in EBUDS. The link between i and j is weighted by the frequency of occurrence of $*ij*$ in the corpus (“*” is a wildcard character that may be substituted by any of the 377 signs or blank space).

We note immediately that only 1.5% (=2178) of the 377×377 possible directed pairs are seen to occur in the actual inscriptions. Furthermore, most of the signs are connected to only one or two other signs. The connectance (probability of link between any given pair) is only around 0.01, implying that the network is extremely sparse. However, if we plot the sub-network of connections between nodes corresponding to the 10 most common signs in EBUDS (i.e., the ones occurring with the highest frequency), we note that they are strongly inter-connected (Fig. 3). Therefore the adjacency matrix of the sign network is far from homogeneous, with patches of dense connectivity in certain regions.

As the above evidence indicates that there exists a core set of signs which occur very frequently as pairs, a natural question is whether the network

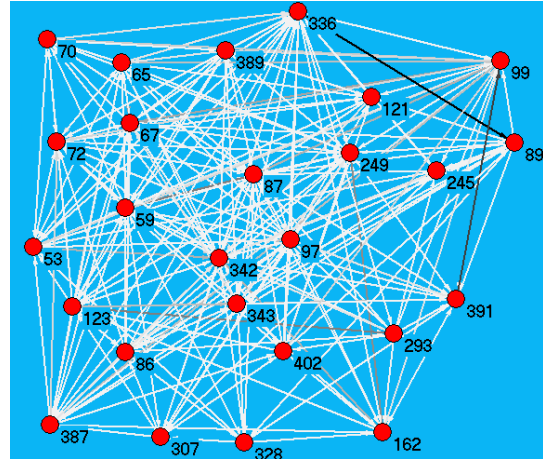


Fig. 5: The innermost (order 8) in-degree core of the Indus sign network with 26 signs. Grayscale color of each link corresponds to the frequency of occurrence of a particular pair (e.g., 391-99 and 336-89 are the commonest pairs).

generated from EBUDS has a core-periphery organization. This would imply the existence of a densely connected central core (central in terms of graph distance) and a larger class of sparsely connected peripheral nodes, like that seen in the case of geographically embedded transportation networks (Holme, 2005). To obtain such a decomposition of the network we use a pruning algorithm that successively peels away layers of a given core-order of the network. The k -core of a network is defined as the subnetwork containing all nodes that have degree *at least* equal to k . Thus, to obtain it, we have to iteratively remove all nodes having degree less than k . In particular, the 2-core of a network is obtained by eliminating all nodes that do not form part of a loop (i.e., a closed path through a subset of the connected nodes). For a k -core, there exist at least k paths between any pair of nodes belonging to it. It is obvious that for any network, there exists an innermost core of maximum order which cannot exceed the highest degree of the network.

In a directed network, one can define a k -core either in terms of the in-degree (number of connections arriving at the node) or the out-degree (number of connections sent from the node). For the EBUDS network, the innermost core turns out to have order 8, regardless of the type of network considered (Fig. 4). Fig. 5 shows the innermost core for the in-degree network. Even a casual inspection shows that many of the common sign pairs in the database belong to this subnetwork. Thus, a large part of the corpus can

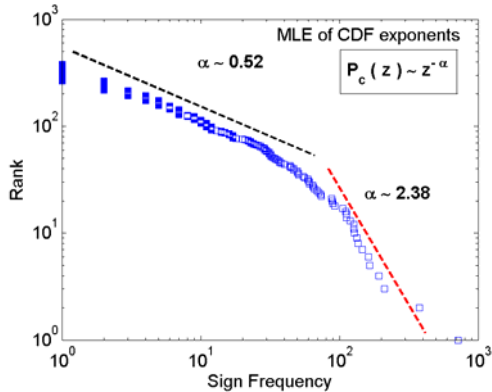


Fig. 6: Rank-frequency distribution of Indus sign occurrences, shown on a double logarithmic scale. The two lines indicate power law fits to different regions of the distribution, with distinct exponents. The latter are calculated using Maximum Likelihood Estimation (MLE). Neither equal 1, as would have been the case for a simple Zipf distribution.

be generated by using members of these “kernel lexicons”.

4.2 Modularity of the network

Many networks that we see in nature are modular, i.e., they comprise several subnetworks (often called *communities*) whose members are more densely connected to each other than with the rest of the network. In several systems, such structural modules are often associated with functional modularity, with each community being identified as being responsible for certain specific functions (e.g., in the protein interaction network). In the EBUDS network, existence of modules will imply that certain sets of signs occur together far more often than would be expected had their frequencies of appearance in the corpus been statistically independent.

The unambiguous identification of communities in a network is a problem that still has not been solved to complete satisfaction. However, several near-optimal algorithms exist. The technique we use was proposed in Newman and Girvan (2004) and involves calculating the following measure of modularity of a network:

$$Q = \sum_s \left[\frac{L_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right],$$

where, L is the total number of links in the network, L_s is the number of links between nodes within a module s and d_s is the sum of the degrees for nodes belonging to module s . By defi-

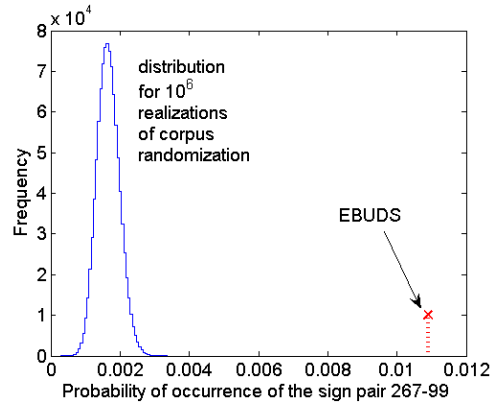


Fig. 7: The probability of occurrence of the sign pair 267-99 in EBUDS compared against the corresponding distribution for the randomized corpus (obtained by considering a million realizations). The large deviation of the empirical value of the pair occurrence probability from the randomized corpus indicates that this is a statistically significant sign pair.

inition, this gives a large value when the network has an unambiguous partition into several communities. Thus, the method for finding the modules involves carrying out many trial divisions of the network into modules and calculating the corresponding Q . The partition for which Q is maximum will correspond to the true modular partition of the system. Needless to say, a brute force method for finding the best partition is impossible for modest sized networks, and we use an extremal optimization technique for this purpose. We obtain 8 communities whose sizes range from 6 to 87 nodes.

Having identified the communities in the sign network, the obvious question is whether they correspond to sign groups that occur in a particular context, e.g., commercial or religious. With this aim in view we have examined the correlation between the modules and (i) artifact types, (ii) nature of the field symbols and (iii) site of excavation. None of them show any significant correlation with the modules, implying that the signs are not simple attributes of either artifact or symbol portrayed in a seal, nor were the use of certain sign subsets confined exclusively to certain regions. The latter point underlines the surprising homogeneity of sign usage over the vast area that the Indus civilization covered. Let us stress that we are looking at correlation between *groups* of signs (that have a strong probability of co-occurrence) and specific contexts, rather than the significant frequency of occurrence of an *individual sign* in a specific context, of which there

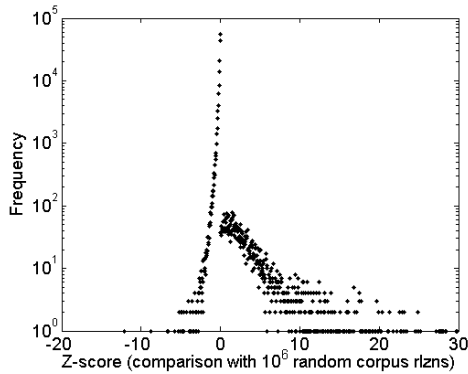


Fig. 8: Distribution of z-scores for all 377×377 possible sign pairs. Note that many potential sign pairs are not observed in EBUDS at all, which are responsible for the negative z-score values. The randomization is over the entire corpus, and the mean and standard deviation are calculated over a million random realizations.

are several examples. By focusing on correlation patterns at the level of groups of signs, rather than individual signs, we aim to arrive at results that are robust with respect to fluctuations in individual sign frequencies occurring as a result of further archaeological excavations.

4.3 Network of “significant” links

So far we had placed all sign pairs that occur in EBUDS on an equal footing. However, certain pairs may occur with high probability simply because the individual signs that make up the pair occur with high frequency. Fig. 6 shows that the frequency distribution of sign occurrence in EBUDS has an approximately power law distribution. This implies that the few commonest signs will occur over a very large number of cases (the most frequent sign appearing as many as 715 times, which is 10% of the total of 7000 occurrences of the 377 signs in EBUDS). By using the information about the probability of occurrence for individual signs in EBUDS we can investigate significant sign relations, i.e., sign combinations that occur far more frequently than that expected from the individual probabilities of the component signs.

Thus, if sign i occurs with a probability $p(i)$ and j with $p(j)$, then the pair ij is significant if $p(ij) \gg p(i)p(j)$. To measure by how much $p(ij)$ has to be larger than the product of $p(i)$ and $p(j)$ in order to be “significant”, we need to compare the empirical joint occurrence probability against the corresponding value for randomized surrogates. The

randomized corpus is generated by shuffling the sign sequences in the dataset so that, while the individual sign frequencies are unchanged, all pair correlations in the original inscriptions are lost. The shuffling can be done over either (i) the entire corpus, or (ii) over each individual seal.

Fig. 7 shows a comparison between the empirical probability of a certain significant sign pair, and the corresponding probability distribution obtained upon corpus randomization. It is evident that the pair would never have been observed with the actual EBUDS frequency had the two signs been independent, i.e., had there been no dependency relation between them. This deviation can be quantified by computing the z-score, which is the difference between the empirical sign pair probability and the mean of the randomized cases, divided by the standard deviation for the randomizations. The distribution of z-scores for all 377×377 possible pairs are shown in Fig. 8. We note that there are 284 sign pairs with z-score larger than 10, while 46 pairs have z-score more than 20.

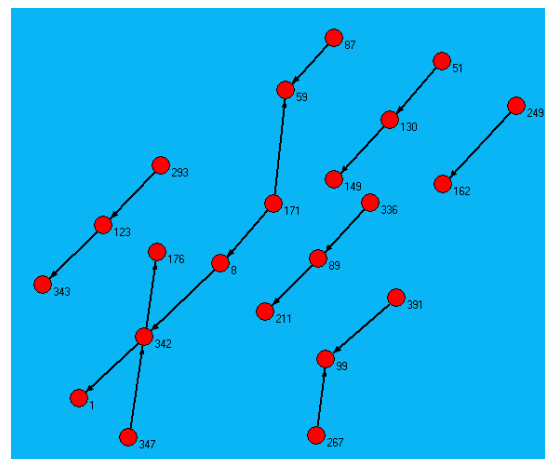


Fig. 9: The network of significant sign pairs as obtained after comparison with the randomized corpus constructed by shuffling signs in each seal.

As an example, using the individual seal shuffled randomization we obtain significant relations among 22 signs for a specified z-score threshold (Fig. 9). There are 6 isolated clusters in this sub-network, with the longest cluster containing 8 signs including sign 342 (the commonest sign in the corpus). Out of the 16 most frequently appearing signs in the database, 13 appear in this group, indicating that some of the common signs have significant relations with each other. While most pair relations are between such common signs, one exception is the cluster of signs no. 51 (35th most common sign), no. 149 (50th most

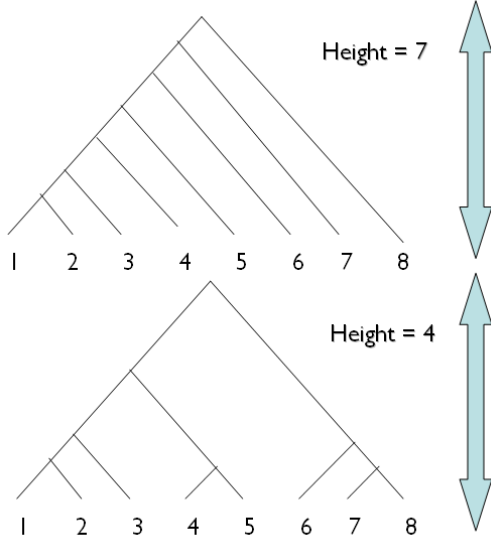


Fig. 10: Schematic segmentation trees for a sign sequence of length 8, representing two alternative possibilities. The top example is a relatively unstructured sign sequence, with the tree height being almost identical to the sequence length. The bottom example shows significant recursive structure and a corresponding lower tree height.

common sign) and no. 130 (59th most common sign). As the individual signs are themselves not very common, the observed sign relation is indeed quite intriguing and possibly has some functional significance in terms of interpreting the sign sequences.

4.4 “Syntactic” tree generation

We will finally attempt to reveal structure indicative of syntactic trees by “parsing” the longest sign sequences. We do this by generating segmentation trees of the sign sequences based on the relative frequency of sign combination occurrences. Given an inscription of length n , sign pairs are iteratively merged to form meta-signs, with the first merge being done for the sign pair with the highest z-score among all pairs in that sequence. This merged sign is then included as a meta-sign and assigned a new number. The reduced sequence of length $n-1$ is now again scanned for the pair of signs or meta-signs that is most “significant” and merged together. This process continues until the entire sign sequence reduces to a single meta-sign. In case of a tie between two or more pairs at any stage, one pair is randomly chosen. The resulting segmentation tree of the sign sequence is shown schematically in Fig. 10. The height of the tree is an indicator of the presence of significant recursive structure

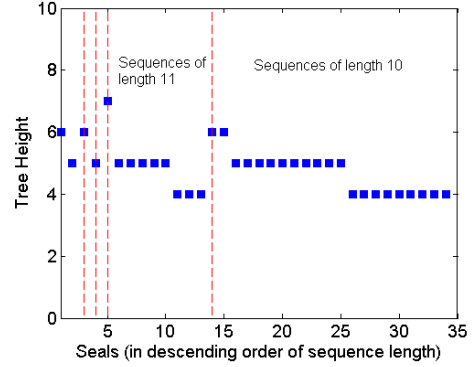


Fig. 11: Segmentation tree height for all sequences (of length 10 or more) in EBUDS arranged in descending order.

in the sign sequence. In particular, if the signs are all independent of each other, then the segmentation tree has essentially the same height as the length of the sequence (Fig. 10, top). On the other hand, if for long sequences, there exists subsequences that also appear in the corpus as separate sequences in their own right, this is indicative of recursion. The corresponding tree height is substantially reduced as compared to the sequence length (Fig. 10, bottom).

We use this criterion to seek signature of recursive, and hence syntactic, structure in the EBUDS network. For shorter length sequences, it becomes difficult to obtain subsequences that also appear as sequences in the database. We have thus confined our attention to inscriptions having 10 or more signs. Arranging the heights of the segmentation trees of these sequences in descending order (for seals of each specific length), we see that the average tree height is around 5 (Fig. 11). Such a characteristic length scale indicates that the longer sequences may actually be composed of multiple smaller sequences, each of which has a particular syntactic relation among its constituent signs.

5 Discussion

In this paper we have used complex network analysis techniques on the sign network constructed from a subset of the corpus of inscriptions obtained in Indus civilization excavations. Our results suggest that though these sign sequences are yet to be deciphered, they have a highly structured arrangement which is suggestive of the existence of syntax. The inference of a set of rules (i.e., the grammar) for arranging these signs in a particular order, so as to be able

to create pseudotexts that are indistinguishable from the excavated ones is the eventual aim of the analysis described here. However, prior to this several open problems need to be addressed. One of the extensions of the present work has to do with looking beyond sign pairs to sign triplets, quadruplets, etc. Preliminary analysis of networks of meta-signs by us indicates that, combinations beyond four signs may not have statistical significance. A detailed comparison between the sign network described here and the meta-sign network may provide clues about the possible hierarchical arrangement of subsequences in the longer sequences. Evidence of this is already seen from the construction of segmentation trees of individual sequences based on relative pair frequencies.

It is also possible that there are non-local correlations between signs in a given inscription. To analyze this, we need to redefine the links in the network as being connections between all signs that occur in the same inscription. Again, preliminary analysis seems to suggest that this does not provide substantially new results from those reported here.

Based on the number of distinct signs (more than 400) there have been several suggestions that, as the number is too high to be an alphabetic system but too small to be an ideographic system, the inscriptions could well be written in a logosyllabic system. Such a writing system combines both logograms (morphemic signs) and syllabic (or phonetic) signs without inherent meaning. In future work, we plan to investigate the differences that arise in the network structure of languages belonging to these very different systems, in order to make an inference on the nature of the writing system used in the Indus inscriptions.

One of the most controversial aspects of Indus decipherment is the question of how many distinct signs are there. M77 identified 417 signs, but other researchers have come up with a wide range of different numbers. Therefore, an important open issue that needs to be settled in the future is the robustness of these results, with respect to analysis based on another sign list, e.g., that created by B. K. Wells (Wells, 2006).

Our analysis of correlations, or rather, the lack of it, between the modules of the network (i.e., groups of signs that have a high probability of co-occurrence) and contexts such as site of exca-

vation, artifact types and field symbols, indicates that the patterns seen in the sequence organization are intrinsic to the sign usage system and not so much dependent on the context in which they arise. This supports the long-held belief that the signs encode writing, or, at least, proto-writing.

Acknowledgments

We would like to thank Bryan K. Wells for carefully reading through the manuscript and making numerous suggestions for improvement. We are grateful to the anonymous referees for several helpful suggestions. Ashraf M. Izhar helped in preparing Figures 7 and 8. We would also like to thank P. P. Divakaran for his constant encouragement.

References

- S. M. G. Caldeira, T. C. Petit Lobao, R. F. S. Andrade A. Neme and J. G. V. Miranda. 2006. The network of concepts in written texts. *European Physical Journal B*, 49(4):523-529.
- S. N. Dorogovtsev and J. F. Mendes. 2001. Language as an evolving word web. *Proceedings of the Royal Society of London B*, 268(1485):2603-2606.
- W. A. Fairervis. 1971. *The Roots of Ancient India*, George Allen and Unwin, London.
- S. Farmer, R. Sproat and M. Witzel. 2004. The collapse of the Indus-script thesis. *Electronic Journal of Vedic Studies*, 11(2):19-57.
- R. Ferrer i Cancho, and R. V. Sole. 2001. The small world of human language. *Proceedings of the Royal Society of London B*, 268(1482): 2261-2265.
- R. Ferrer i Cancho, R. V. Sole and R. Kohler. 2003. Patterns in syntactic dependency networks. *Physical Review E*, 69(5): 051915 (1-8).
- P. Holme. 2005. Core-periphery organization of complex networks. *Physical Review E*, 72(4): 046111 (1-4).
- I. Mahadevan. 1977. *The Indus Script: Texts, Concordances and Tables*, Archaeological Survey of India, New Delhi.
- J. Marshall. 1931. *Mohenjo-daro and the Indus Civilization*, Arthur Probsthain, London.
- A. Mehler. 2008. Large text networks as an object of corpus linguistic studies. In A. Ludeling and K. Marja (Eds) *Corpus Linguistics: An Interna-*

- tional Handbook*, Mouton de Gruyter, Berlin, 328-382.
- A. E. Motter, A. P. S. De Moura, Y-C. Lai and P. Dasgupta. 2002. Topology of the conceptual network of language. *Physical Review E*, 65(6):065102 (1-4).
- M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113 (1-15).
- A. Parpola. 1994. *Deciphering the Indus Script*, Cambridge University Press, Cambridge.
- G. Possehl. 2002. *The Indus Civilization: A Contemporary Perspective*, AltaMira Press, Lanham, MD.
- D. R. Radev and R. Mihalcea. 2008. Networks and natural language processing. *AI Magazine*, 29(3):16-28.
- M. S. Vitevitch. 2008. What can graph theory tell us about word learning and lexical retrieval. *Journal of Speech, Language and Hearing Research*, 51(2):408-422.
- B. K. Wells. 2006. *Epigraphic Approaches to Indus Writing*, PhD thesis, Harvard University, Cambridge, Mass.
- N. Yadav, M. N. Vahia, I. Mahadevan and H. Jogelkar. 2008. A statistical approach for pattern search in Indus writing. *International Journal of Dravidian Linguistics*, 37(1):39-52.