KIMMO KOSKENNIEMI

SYNTACTIC METHODS IN THE STUDY OF THE INDUS SCRIPT

This paper discusses some aspects of the decipherment of unknown scripts and proposes a framework for the decipherment process. We believe that such methods enable one to propose better solutions and to formulate these for easy inspection and evaluation by other scholars, and thus enhance the possibilities to arrive at a generally accepted solution. The methods presented here can also be considered an example of a proper role for the computer in linguistic studies.

Introduction

In Indus script was created and used by the Harappan or Indus civilisation, which flourished in the plains of the Indus river about 2500 - 1800 BC. About 4000 inscriptions, mostly seals, have been preserved. However, no bilingual texts have been found yet, nor do we know any proper names or other direct clues about the script. Although we do not know what the underlying language of the inscriptions is, the growing amount of archaeological evidence puts early Dravidian as the most probable candidate for the language.

Some recent methods

A Soviet team lead by Yuri V. Knorozov has used a computer to extract and identify signs corresponding to roots, derivational and inflectional morphemes in the Indus script. We have not, however, been able to find any explicit description of the algorithms or the criteria used for making such decisions.

E. Barber has worked mostly on linear A texts and has collected a set of probabilistic criteria for segmenting continuous text into words. Using several independent criteria in parallel, she can locate the most

likely points of the boundaries. Segments with similar distribution are then grouped into categories. The segments and categories thus obtained are proposed to be compiled to form a dictionary that indicates the distributional properties of each unit. These methods based on segmentation seem to be somewhat directed towards syllabic or word syllabic writing systems, where phonological constraints are also present.

Problems with proposed decipherments

Scholars have proposed several different decipherments for the Indus script. Many of these disagree with each other even in the very basic concepts, including the direction of writing, the type of the writing system and the identity of the underlying language. General opinion among the scholars has rejected many of these decipherments, but the rest of them remain in an ambivalent state: they are neither accepted nor proven to be false.

It is not the impossibility of the proposals that prevents them from being accepted. Rather, it is the fact that no solution includes arguments that would be convincing enough. The problem seems to be that the proposals are not sufficiently explicit and open for inspection by other scholars, and that the intermediate steps of reasoning remain hidden. It is also only too common that readings for the signs are assigned in order to support some personal hypotheses on the culture rather than vice versa.

The scope of this paper is to propose some remedies to this situation by methods that would make the proposed decipherments more explicit and thus more open to evaluation by others. We would also like to start from the least subjective features, i.e. the distribution of the signs and procede carefully to more delicate questions.

Some assumptions on the Indus script

We make the explicit assumption that the signs in the Indus script usually correspond to morphemes or some larger units in the underlying language. Thus we assume that the writing system is not phonological, i.e. neither phonemic nor syllabic. This assumption cannot, of course, be verified at this stage, but some arguments in its favour can be listed. The number of distinct signs (about 400) and their relative

frequencies support this assumption, as most of the signs are rare and new signs keep turning up as more inscriptions are found. The historical context also favours this assumption, because mainly phonological writing systems were not used in other cultures at the time when the Indus script was created and stabilised.

If this assumption of the morphemic character of the script is correct, neither the similarities in the distribution nor the systematic alternations of the signs reflect any phonological similarities, e.g. common or related phonological components. All alternations and similar distributions must be interpreted in the terms of the distribution of morphemes or morpheme clusters, and this is the domain of syntax (or of syntax and morphology together).

It should be noted that we do not assume that the correspondence between signs and morphemes would be one to one. All morphemes need not be indicated in the script -- what is redundant in the context need not have been indicated explicitly.

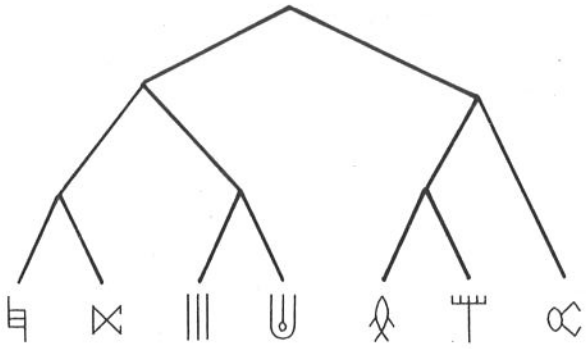
The syntactic approach

The word boundaries are not indicated in the Indus script, and this is the case in most early writing systems. Traditionally it has been felt that the continuous text should first be segmented into words and then these segments should be analysed and classified. Our opinion is that these steps should not be separated, but instead, unified into a single syntactic analysis.

If we study sentences (or utterances) as sequences of morphemes, and not as sequences of phonological units, we notice that there are no formal criteria for distinguishing between attributes and affixes, or between postpositions and inflectional endings. Therefore we think that there is no hurry in making such decisions so early. The stability of patterns and the freedom of occurrence will perhaps later provide evidence for such decisions. Furthermore, we do not know in advance whether the underlying language happens to be analytic (i.e. one morpheme per word) or synthetic (i.e. several morphemes per word). The syntactic approach can handle both cases.

An example of an analysis of a text

Figure 1 shows an analysis of one Indus inscription according to the pairwise frequencies of signs. The strength of the syntactic connection between each pair of signs has been estimated according to principles given in Koskenniemi, S. et al. (1970). The measure of this strength is computed from the number of pairs actually occurring in the corpus and the theoretically expected number. An observed number that is much higher than the expected one indicates that the signs belong to the same low level constituent. On the other hand, an observed number near the expected value probably indicates a major syntactic boundary.



The observed frequency of the pair	62	5	155	12	80	24
The expected frequency of the pair	1	1	4	6	4	1
Approximation to the strength of the joint	69	3	75	2	38	21

Figure 1

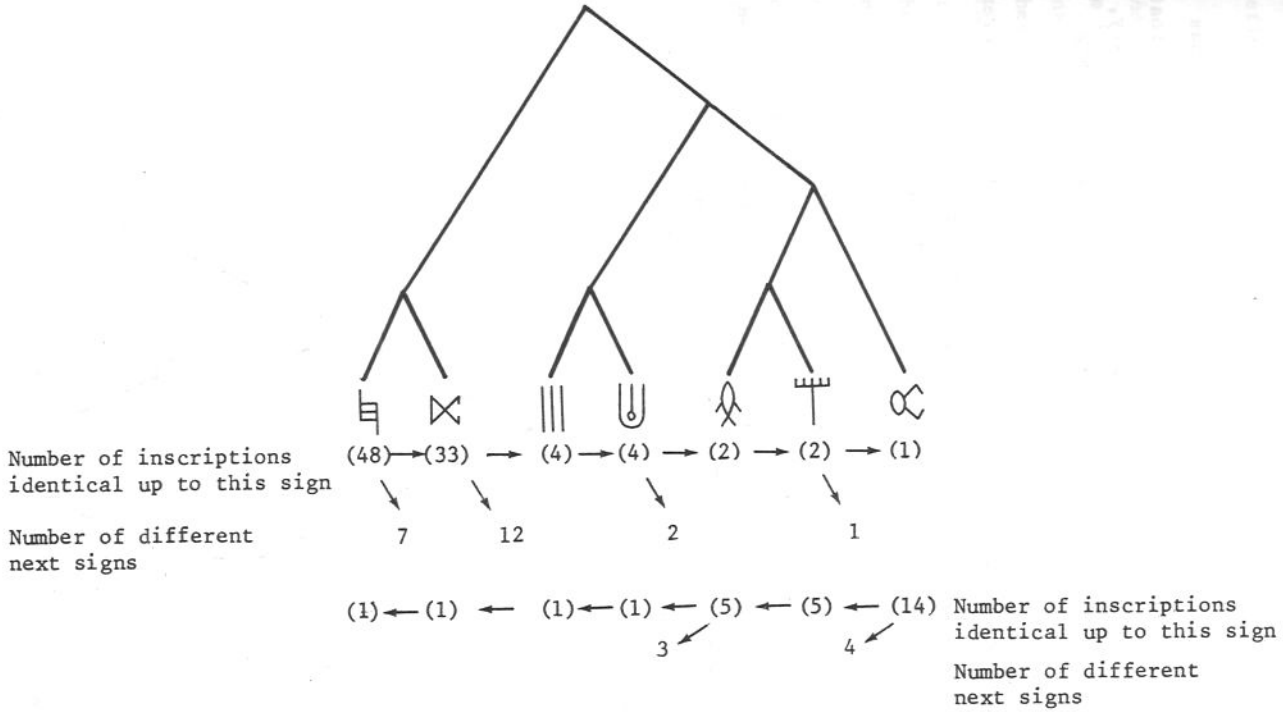

Number of inscriptions identical up to this sign
(48) → (33) → (4) → (4) → (2) → (2) → (1)
Number of different next signs
7
12
2
1
(1) ← (1) ← (1) ← (1) ← (5) ← (5) ← (14)
3
4
Number of inscriptions identical up to this sign
Number of different next signs

Figure 2

When we have the approximations for the strengths of the joints, we can draw an approximate syntax tree for the inscription by joining adjacent signs/constituents in the descending order of the joint strength.

Figure 2 gives another syntax tree for the same inscription. It has been constructed by using the number of other texts in the corpus that have an identical beginning or ending. At a major syntactic boundary, there is expected to be a rise in the number of different possible next signs.

These two syntax trees have been constructed according to two different criteria, but they show remarkable similarity. Only the two highest nodes have changed their order, all lower level constituents are identical.

The above syntax trees were unlabelled. Even a preliminary study of the paradigmatic relations shows that certain elements are optional, i.e. one can find inscriptions with the element and others, otherwise identical, without it. Nodes that govern both an obligatory and an optional element can be labeled according the obligatory constituent as in figure 3.

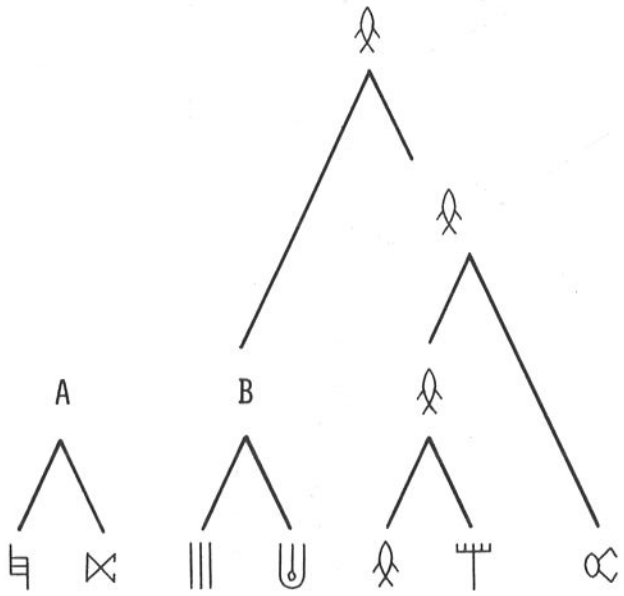


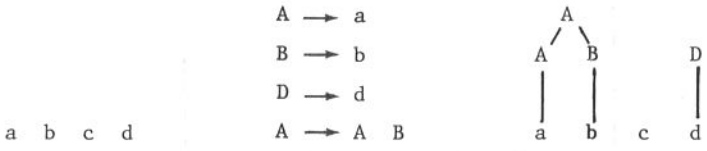
Figure 3

The problems with the segmentation can now be seen if we look at the three rightmost signs in our sample inscription. Should there be one, two or three segments? We think that none of these would adequately reflect the relations these signs have.

The proposed formalism

We hope to have now shown that it is quite feasible to study the Indus inscriptions syntactically. We propose that the findings and the hypotheses of the distributional analysis should be formulated as a *formal grammar*, which would grow gradually as the work procedes. Each addition or correction to the current version of this grammar would be based on the statistics and lists that have been computed using the information in the current grammar.

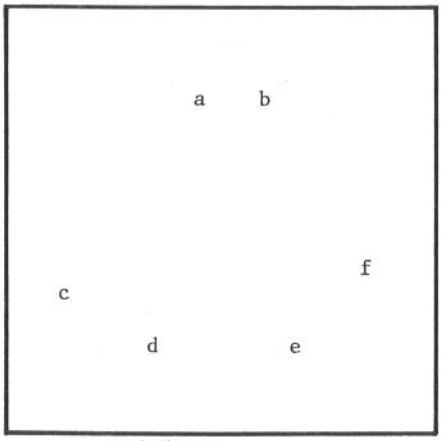
The current grammar is utilised by parsing the inscriptions partially before the statistics are calculated or lists produced. To take a very simple example consider the hypothetic sentence below left and fraction of an intermediate grammar in the middle.



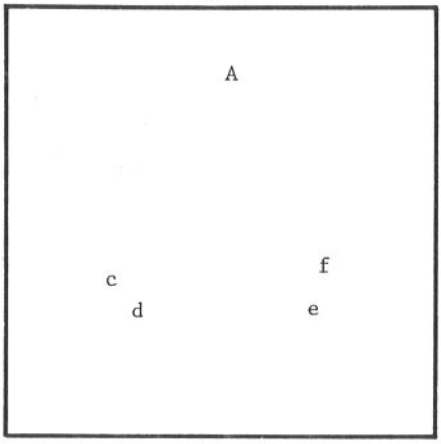
At the right is the partial parse induced by this grammar.

What should an actual formal grammar for the Indus corpus look like? It would have (at least) two levels of categories, subcategories for signs or sign pairs with very similar distribution, and broader categories for collecting similarly functioning subcategories together. These broader categories would also cover various constructions to be discovered, e.g. *attribute + noun* that functions like a single *noun*.

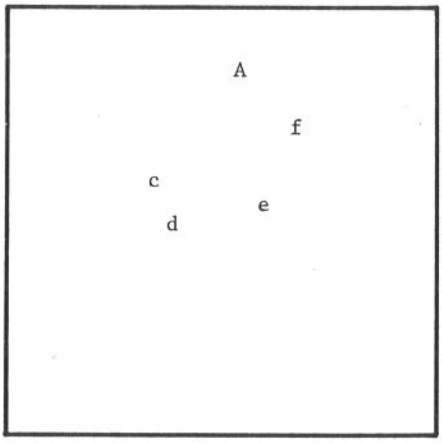
How do we then know whether we are going in the right direction as we add more and more of these productions into our current grammar? Correct productions are expected to produce grammars, where the broad categories stay well apart from each other. Incorrect productions tend



Initially



After a correct new rule



After a wrong new rule

Figure 4

to degenerate the categories into a single dummy category (*signs*) that covers anything, see figure 4.

The proposed role for the computer is twofold. Firstly it would compute the tedious statistics and sort the data over and over again. Secondly it would manage the grammar and carry out all consequences by parsing the inscriptions according to it before the statistics are calculated. In this scheme, all decisions are done by the scholar, not by the computer. The machine only provides data for the decisions and stores the evolving model or description of the corpus.

With the computer one can find all consequences of new rules rapidly and accurately. During the process of decipherment one often has to reject some earlier hypotheses. The human mind has great difficulties in forgetting all consequences of them, whereas the computer is excellent in this respect.

Homography

We must assume that Indus script, like other writing systems contemporary to it, contains homography, i.e. different meanings or functions are expressed with the same sign. The different functions associated with a sign are derived from each other either by homophony or semantic connection. It is, however, reasonable to assume that signs with many functions form only a minority, and that even the polyphonic signs usually have one function that is significantly more frequent than the others.

Context free productions in our grammar are adequate only for unambiguous signs. Signs that have more than one function can be detected with methods given by Zellig Harris (1951). Suppose we have two signs y and z which do not occur in similar contexts. If then sign x is such that there are environments $\alpha\_\_\_\beta$ where x and y occur, and environments $\gamma\_\_\_\delta$ where x and z occur, and these two environments cover most of the occurrences of x, we have a probably homographic sign x. Such ambiguous signs x can be described in the grammar with context sensitive rules of the form $A \rightarrow x/\alpha\_\_\_\beta$ and $B \rightarrow x/\gamma\_\_\_\delta$

Barber considered homography harmful for analysis and thought that it is difficult or even impossible to detect and manage it with formal methods. We think that the recognition and description of homography

is not only possible, but extremely valuable, because it provides one of the very few ways of discriminating among the various candidates for the underlying language. The pictorial shapes of the signs induce different pairs of meaning in these languages, and these can be compared with the iconography, the shapes and functions of the objects etc.

A comparison with the Linear B

The decipherers of the Indus script have often tried to reach some grids that would resemble the famous ones used by Ventris, when he deciphered the Linear B script. These grids for the Linear B indicated the identical initial segments of the signs in each row and the identical final segment in each column. The method did not give any actual values for these segments, but any value guessed for one sign would automatically imply the same first segment for the whole row, and the same final segment for the whole column. Thus the positive or the negative effect of any guess was greatly amplified, which led to a successful solution.

With the morphemic assumption of the Indus script we have rejected all hope of finding similar phonologically arranged grids. What we are hoping to have instead is a formal grammar for the corpus that indicates the paradigmatic and syntagmatic relations of the signs and sequences of signs. The stricter categories are expected to indicate both syntactic and semantic similarity, and the broader categories plain syntactic similarity. Now, a guess of a single value for a sign within a category will imply a syntactically and semantically similar value for all other signs within the same strict category, and a syntactically similar function for all signs in the broad category. In this way, we can hope to reach a similar amplification of the consequences of guesses, but the similarity is in the levels of syntax and semantics rather than in the phonological level.

Relations to linguistic theories

Different methods of decipherment have interesting relations to some of the major theories in general linguistics, especially to the American structuralist and generativist schools. Barber's methods belong clearly to the post-Bloomfieldian American structuralistic tradition. She had adopted the fundamental thesis that forbids the mixing

of levels of the description. She also seems to put more stress on the method by which the description has been obtained than to the operationality of the final result.

The generativists do not care how the final description has been found, it only has to work well. They also let such things as the syntactic arguments affect the establishment of lower level units. Many other features of the generativist theories, like a mentalistic overall approach and a transformational apparatus, however, seem to be totally inapplicable to the decipherment process.

Our line has combined elements from these two. Basically, we accept the distributional analysis of post-Bloomfieldians. We use approximately the same concepts of syntactical constructions, although we prefer the more compact generativist notation. We are prepared to mix the levels of the description by making any safe decisions on syntactic similarities before any less safe decisions on potential word boundaries. We agree that it is very important to give good justification for each step of the decipherment, but we would also like to stress the overall functioning of the description. We anticipate that some seemingly hazardous steps have to be taken in order to arrive at the best solution, and that these steps cannot perhaps be justified until afterwards.

Acknowledgements

This paper continues the work on the Indus script initiated in 1964 by a Finnish research team, which comprised Pentti Aalto, Seppo Koskenniemi, Asko Parpola and Simo Parpola.

I wish to thank Asko and Simo Parpola for their encouragement and valuable suggestions.

References

Barber, E. J. W., 1974. *Archaeological decipherment.* Princeton.

Harris, Zellig, 1951. *Methods in Structural Linguistics.* Chicago.

Koskenniemi, Kimmo & Asko Parpola, 1979. *Corpus of Texts in the Indus Script*. (Department of Asian and African Studies, University of Helsinki, Research Reports 1.) Helsinki.

Koskenniemi, Seppo & Asko Parpola & Simo Parpola, 1970. A Method to Classify Characters of Unknown Ancient Scripts. *Linguistics* 61, pp. 65-91.

Parpola, Asko, 1975. Tasks, Methods and Results in the Study of the Indus Script. *Journal of the Royal Asiatic Society* 1975:2, pp. 178-209.