ASKO PARPOLA

COMPUTER TECHNIQUES IN THE STUDY OF THE INDUS SCRIPT

Seppo Koskenniemi, Simo Parpola, Pentti Aalto and the writer published a first announcement of a breakthrough in their computer-aided studies of the Indus script in February 1969. This breakthrough meant the decipherment of the script, for the nature of the script and the language behind it were unveiled and some signs read with cross-checked phonetic and semantic values. From this basis the work has been successfully carried considerably further[1]. We very much appreciate the kind invitation of the Editor to contribute our findings on the usefulness of the computer in this type of research. We are glad of this opportunity to exchange experiences with our colleagues and to participate in this most useful project.

The rôle of the computer

The computer is a faithful servant of man, which, in carrying through quickly and efficiently his orders, can save him a lot of time and trouble. Especially when extensive mechanical work is concerned, its help, if not indispensable, is instrumental in sparing man the frustration and mental fatigue resulting from lengthy and

[1] Koskenniemi, S., and A. and S. Parpola, Computing approach to Proto-Indian, 1965: An interim report (stencilled), 1966

—, "A method to classify characters of unknown ancient scripts", submitted for publication in Linguistics (in press)

Parpola, A., S. Koskenniemi, S. Parpola and P. Aalto, Decipherment of the Proto-Dravidian Inscriptions of the Indus Civilization. A First Announcement, Copenhagen 1969. (The Scandinavian Institute of Asian Studies, Special Publications No. 1)

—, Progress in the Decipherment of the Proto-Dravidian Indus Script, Copenhagen 1969. (Ibid. No. 2)

—, Further Progress in the Indus Script Decipherment, Copenhagen 1970 (Ibid. No. 3)

Clauson, G., and J. Chadwick, "The Indus Script deciphered?" Antiquity 43, 1969, 200—208.

tiresome routine labour. One can concentrate on theoretical problems and leave the practical realization of ideas to the machine. With small-scale mechanical tasks it is of course often more appropriate and easy to complete by hand than to programme the computer to do the job. On the other hand, the computer can quickly perform complicated mechanical operations that would take ages to do by hand, and it thus opens up possibilities of study that have never before been at the disposal of scholars. But the computer cannot do any job unless man has defined exactly how it must do it. The reasoning is thus left to man, and this includes decisions as to whether and how one should use the computer, the rôle of which is confined to that of a faithful servant (at least for the time being).

The basic tools of study

An obvious routine task to be done with the help of the computer in the study of an unknown ancient script is the compilation of the basic tools of study: that is, all kinds of indexes to the existing epigraphic material that may be thought to prove useful. This was therefore our first endeavour. For the purposes of comparison with other writing systems, and of internal analysis of the Indus script itself, we wished to know:

1. how many signs there are *altogether* in the Indus inscriptions that are available for study, and what is the *average* length of an inscription;
2. how many *different* signs there are;
3. what is the *frequency of occurrence* of each individual sign;
4. what is the *order of frequency* of the signs;
5. how many times within its total frequency each sign occurs in *initial* and *final* positions in the inscriptions;
6. which other signs occur on the left-hand side, and which signs on the right-hand side, of each individual sign; and what are the total, initial and final frequencies of these *pairwise occurrences* and what is their order of frequency;
7. in the *lists of all occurrences* of all pairwise sign combinations, the full context and all relevant information about each inscription (type, provenance and iconography of the object on which the inscription is found, number of lines, and so forth).

For processing these data, we transformed the material into a form readable by machine by drawing up a sign-list, allotting to

each sign a three-digit number, and transcribing the inscriptions into numerical form. Each inscription (the average length being only five signs) was then punched on a separate IBM card, headed by a source reference in the form of a four-digit number.

Difficulties in preparing these tools

There were a number of difficulties due to the character of the material for study and, to some extent, to technical problems which hampered an ideal realization of our goals.

1. Many inscriptions were badly broken or worn, partly or totally, illegible or written indistinctly (especially graffiti on potsherds). In addition, several inscriptions have scratches that may or may not belong to the writing. In many cases, a collation with the original would enable one to overcome these difficulties, as we noticed later on in practice, but in the beginning we had to rely entirely on photographs and copies produced by scholars who had had access to the original documents.

2. There were several cases where two or more signs, resembling each other but drawn slightly differently, could be either graphic variants of one and the same sign, or else separate signs. A wrong decision in either direction would be equally detrimental.

3. Though the direction of writing is normally constant (from right to left), there are exceptions to the rule. Anomalous direction of writing (from left to right) is not uncommon, and on inscriptions with more than one line the writing may or may not run boustrophedon. It is important to remember that the relevant archaeological publications sometimes reproduce a seal instead of its impression.

4. The inscriptions are not broken into words, which made it imperative to take the inscription and not the word as the measure for initial and final occurrences of the signs. The beginning and end of an inscription of course coincide with the beginning of the initial word and the end of the final word, but, as many inscriptions are likely to contain several words, the figures of the final and initial frequencies would be more telling if the word could be used as the unit. The inscriptions containing two or more lines are here too a problem and a source of errors, for the end of a line could coincide or not with the end of the word, and in case it did not, the word could continue from either end of the next line[2].

[2] We realized that many of these problems could be solved by a comparative study of the sign combinations. If we prepared preliminary lists trying our best but

Computer-drawn signs

When our preliminary lists were processed five years ago in 1965, we could get the output in numerical form only, which meant that for easy reference the signs had to be retranscribed. For this reason, complete lists of occurrences were prepared only for selected sign combinations. (Samples of our preliminary lists were published in our First Announcement, pp. 66—68.) This was sufficient for breaking the code, but at a more advanced stage a complete concordance is necessary. In the meanwhile, computer techniques have developed considerably, and the critical edition and concordance with all relevant data which we hope to publish soon will be drawn in Indus characters by a plotter. The sample signs reproduced in Figure 1 were drawn by the plotter at the Northern Europe University Computing Centre (NEUCC) in Kongens Lyngby near Copenhagen. The programme was written by Seppo Koskenniemi, and the basic elements of the characters designed by the writer. The system involves scaling up and down in size, and moving elements to various positions.

Automatic classification of signs

We tried in several ways to use the possibilities offered by the computer for deciphering the script, as described in a cyclostyled paper which we circulated to a limited number of scholars in 1966. Among other achievements, we developed a new automatic method which may prove helpful in the decipherment of other unknown scripts. The basic idea was that writing is subject to the measurable laws governing speech and dictated by human physiology, which make only certain sound combinations possible. Therefore it should be possible to develop an automatic method of decipherment based on statistical data only, which would have universal

---

without wasting time in trying to solve all these problems, we would quickly have lists which would however contain a few errors. Statistically these errors would be meaningless, and we would be able to correct many of them and improve the text on the basis of these preliminary lists. That is why we left many judgements to a later stage, when the sign combinations could be effectively studied by comparing their frequency patterns and occurrences in different contexts. All suspected variants were coded as separate signs. Broken or unclear signs were given a common code number, and less doubtful signs were individually read with their most likely values. Each line in inscriptions of more than one line was treated as a separate unit, but all were given the same reference.

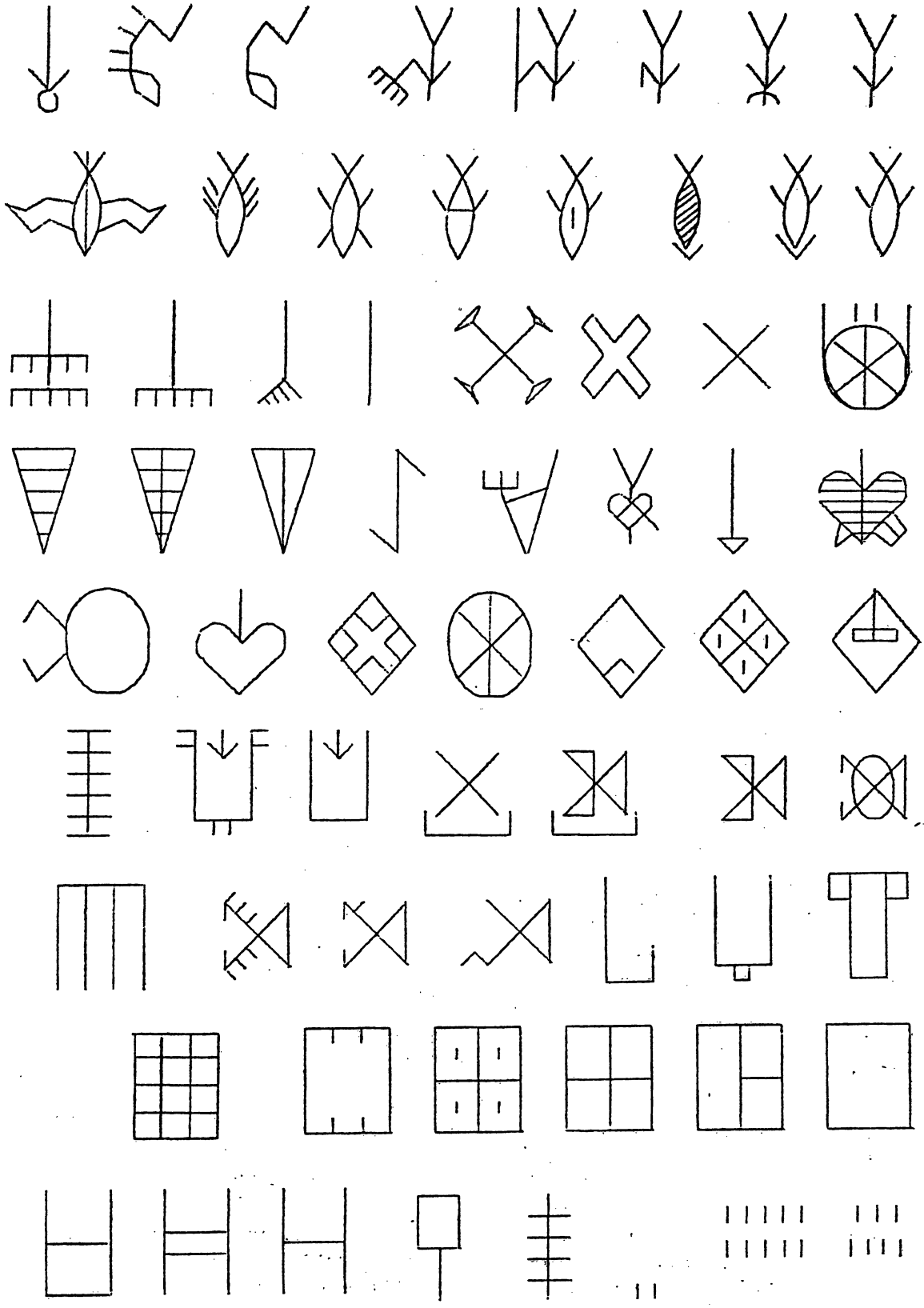

Fig. 1

validity. In order to develop and check such a method we prepared, in form readable by machine, samples of several known ancient scripts representing different types of writing and different languages (Sumerian, Neo-Assyrian, Middle-Egyptian, Linear B, and Elamite cuneiform) as well as samples of different languages in alphabetic script (Finnish, English, French, and two samples of Latin). In a paper submitted for publication in Linguistics (The Hague) in October 1968, we describe a classification of the characters in these samples based only on pairwise occurrences of each individual sign. The computer counts the other signs which occur on the right-hand side of a sign. This gives the 'behaviour pattern' of the sign, which is then compared with similar patterns in the cases of the other signs. The frequencies of sign combinations are taken into account, but normalized to eliminate the error resulting from the stereotyped nature of the material. The computer classifies the signs into groups, in such a way that signs having a similar pattern of behaviour fall within the same group. We begin with two groups, then three groups, then four, and so forth. The time it would take to do this by hand would be enormous. The signs are then classified in the same way into different groups according to their behaviour patterns with reference to the signs occurring on their left-hand side; and thereafter these two classifications were combined. The samples of syllabic scripts with signs for both open and closed syllables gave extremely promising results, and a classification into two groups of the phonetically exact Finnish sample resulted in a pure division into consonants and vowels.