Indus script encodes language, reveals new study of ancient symbols

Hannah Hickey
UW News

Courtesy of J. M. Kenoyer / Harappa.com
Examples of the Indus script. The four square artifacts with animal and human iconography are stamp seals that measure one or two inches per side. On the top right are three elongated seals that have no iconography, as well as three miniature tablets (one twisted). The tablets measure about 1.25 inches long by 0.5 inches wide.

The Rosetta Stone allowed 19th century scholars to translate symbols left by an ancient civilization and thus decipher the meaning of Egyptian hieroglyphics.

But the symbols found on many other ancient artifacts remain a mystery, including those of a people that inhabited the Indus valley on the present-day border between Pakistan and India. Some experts question whether the symbols represent a language at all, or are merely pictograms that bear no relation to the language spoken by their creators.

A University of Washington computer scientist has led a statistical study of the Indus script, comparing the pattern of symbols to various linguistic scripts and nonlinguistic systems, including DNA and a computer programming language. The results, published online Thursday by the journal Science, found the Indus script's pattern is closer to that of spoken words, supporting the hypothesis that it codes for an as-yet-unknown language.

"We applied techniques of computer science, specifically machine learning, to an ancient problem," said Rajesh Rao, a UW associate professor of computer science and engineering and lead author of the study. "At this point we can say that the Indus script seems to have statistical regularities that are in line with natural languages."

Co-authors are Nisha Yadav and Mayank Vahia at the Tata Institute of Fundamental Research in Mumbai, India; Hrishikesh Joglekar, a software engineer from Mumbai; R. Adhikari at the Institute of Mathematical Sciences in Chennai, India; and Iravatham Mahadevan at the Indus Research Center in Chennai. The research was supported by the Packard Foundation and the Sir Jamsetji Tata Trust.

The Indus people were contemporaries of the Egyptian and Mesopotamian civilizations, inhabiting the Indus river valley in present-day eastern Pakistan and northwestern India from about 2600 to 1900 B.C. This was an advanced,

urbanized civilization that left written symbols on tiny stamp seals, amulets, ceramic objects and small tablets.

“The Indus script has been known for almost 130 years,” said Rao, an Indian native with a longtime personal interest in the subject. “Despite more than 100 attempts, it has not yet been deciphered. The underlying assumption has always been that the script encodes language.”

In 2004 a provocative paper titled The Collapse of the Indus-Script Thesis claimed that the short inscriptions have no linguistic content and are merely brief pictograms depicting religious or political symbols. That paper’s lead author offered a $10,000 reward to anybody who could produce an Indus artifact with more than 50 symbols.

Taking a scientific approach, the U.S.-Indian team of computer scientists and mathematicians looked at the statistical patterns in sequences of Indus symbols. They calculated the amount of randomness allowed in choosing the next symbol in a sequence. Some nonlinguistic systems display a random pattern, while others, such as pictures that represent deities, follow a strict order that reflects some underlying hierarchy. Spoken languages tend to fall between the two extremes, incorporating some order as well as some flexibility.

The new study compared a well-known compilation of Indus texts with linguistic and nonlinguistic samples. The researchers performed calculations on present-day texts of English; texts of the Sumerian language spoken in Mesopotamia during the time of the Indus civilization; texts in Old Tamil, a Dravidian language originating in southern India that some scholars have hypothesized is related to the Indus script; and ancient Sanskrit, one of the earliest members of the Indo-European language family. In each case the authors calculated the conditional entropy, or randomness, of the symbols’ order.

They then repeated the calculations for samples of symbols that are not spoken languages: one in which the placement of symbols was completely random; another in which the placement of symbols followed a strict hierarchy; DNA sequences from the human genome; bacterial protein

sequences; and an artificially created linguistic system, the computer programming language Fortran.

Results showed that the Indus inscriptions fell in the middle of the spoken languages and differed from any of the nonlinguistic systems.

If the Indus symbols are a spoken language, then deciphering them would open a window onto a civilization that lived more than 4,000 years ago. The researchers hope to continue their international collaboration, using a mathematical approach to delve further into the Indus script.

“We would like to make as much headway as possible and ideally, yes, we’d like to crack the code,” Rao said. “For now we want to analyze the structure and syntax of the script and infer its grammatical rules. Someday we could leverage this information to get to a decipherment, if, for example, an Indus equivalent of the Rosetta Stone is unearthed in the future.”