Linguistics

Sign Language

Could machine learning show symbols encode language without decoding a script? BY VIRAT MARKANDEYA

It seems almost mystical: a process that fills in missing symbols—to be sure undeciphered symbols—from a 4,500-year-old advanced urban civilisation. Yet, a statistical model does precisely so for the undeciphered Indus Valley script.

"We have used this model to restore signs in damaged or illegible Indus texts and for generation of new Indus-like texts. It can predict signs with about 75 per cent accuracy," says Nisha Yadav, a scientist at the Tata Institute of Fundamental Research, Mumbai. Along with her colleagues Mayank Vahia and Hrishikesh Joglekar, she first began to test the script statistically, using a database created by Iravatham Mahadevan, an Indian civil servant who, working on punch card computers at TIFR in the 1970s, put together a corpus of Harappan symbols.

Collaborating with scientists from the University of Washington, Seattle and the Institute of Mathematical Sciences, Chennai, they claim to have found evidence that the Harappan script encodes language.

But they have no idea what it says.

"The key idea is that it is possible using statistical measures to distinguish a script as linguistic or non-linguistic without having deciphered it" says Ronojoy Adhikari of IMSc, Chennai, and a member of the team.

Led by Rajesh Rao of the University of Washington, the scientists published an article in the journal *Science*, attempting to measure the level of order in the Indus Valley script and compare it with four languages as well as other systems that encode information. In calculating the level of order—or entropy—they used Markov Models.

A Markov Model is a sequence of symbols where the choice of the next symbol depends only on the preceding symbol and not on the history of past symbols. The shift from one symbol to the next is governed by a "transition" probability for every pair of symbols. So, if one Indus sign A tends to follow another Indus sign B but the reverse rarely occurs, one would have a high transition probability for the pair AB and a low probability for BA. The researchers estimated transition probabilities between all pairs of Indus signs as well as the probability that each sign can start a text.

By such modelling, the researchers tried to capture a property of written language: that certain symbols work in tandem, preceding or succeeding each other. To take an example from English, the word 'the' can be followed by a large number of words but usually not verbs such as 'eat'. At the level of an individual word, the letter 'q' is usually succeeded by the letter 'u'.

INDUS inscriptions have an average length of 5 signs. This brevity is cited to claim that it does not encode language. Others say it is in line with Proto-Cuneiform and Proto-Sumerian, which have an average of 6 signs.

J. M. KENOYER / HARAPPA.COM

"Our work used conditional entropy to compute the similarity of the Indus script's sequential structure to four natural spoken languages (English, Sanskrit, Old Tamil, Sumerian), one artificial formally-defined language (Fortran), and two biological non-linguistic systems (DNA and proteins), in addition to two artificial datasets representing the two extremes of random and rigid symbol order," says Rao.

When conditional entropy was plotted graphically, the researchers found that the four languages and the Indus script bunched together—and the Indus script's conditional entropy was closest to Sumerian and old Tamil.

The paper prompted a sharp riposte from scientists who believe that the Indus symbols do not encode language. The main charge is that the two artificial extreme datasets make it appear that the language sets occupy a unique range, whereas it is possible to produce other data sets that fall into the range. Without this, it was surmised that Indus symbols falling into that range only shows that they display some kind of structure, which has been known for some time.

"Their criticism misses the point," counters Rao, adding, "The other data sets some people have produced in the range of languages have nowhere near the statistical regularities seen in languages—some have no correlations at all between symbols. Our paper does not claim that conditional entropy by itself is a sufficient statistic for a system to be linguistic. Rather, the paper provides evidence which, in conjunction with the rich syntactic structure in the Indus script, increases the probability that the script represents language."

In a yet-to-be published paper, the team will attempt to show that Indus texts procured from West Asia (ancient Mesopotamia) show striking statistical departures from those procured from within the Indian subcontinent. This raises the possibility that the Indus script may have been used to represent a different language or subject matter by traders living in West Asia. A tantalising prospect indeed—but one that may remain unverifiable until the discovery of a multi-lingual artefact like the Rosetta Stone to actually crack the script.