INDUS VALLEY GIRLS SCRIPT

Good evening everyone, my name is Keerthana.

And I'm Sonia. And we are the Indus Valley Girls!!

We're here to talk about our attempt to decipher the Indus valley script.

Slide 1: Background and History of Indus Valley Script – Sonia

The Indus valley or Harappan civilization is India's oldest civilization. Locationally, It spanned over modern day Pakistan. It is also known for its highly developed water management system and a script which hasn't been deciphered yet.

Slide 2: Why it hasn't been deciphered yet - Sonia

Indus valley script has been notoriously known for being difficult to decipher. Why is this? For starters, Indus texts are very short. The mean length of an Indus text is 4.5 signs while the longest text is 14 signs in a single line. For context: other civilizations have left examples that are hundreds of characters long.

This brevity also further supports the theory that the Indus script is logosyllabic. This means that signs represented full words as well as syllabic sounds. Along with this :

- The corpus (body of text) is treated as one unit - with no regards to changes over time and location
- Lack of context and usage information
- Lack of bilingual inscription/tablet, E.g. Rosetta stone which recorded two Egyptian scripts and ancient Greek offering more context
- Debate over identity of Indus language – E.g. Dravidian (southern India)

All of which add to the complexity of deciphering this script.

Slide 3: Breakthroughs/Research done on the script - Sonia

- Prior to 2013, we didn't know who the Harappans were or where they came from. But there have been recent breakthroughs in population genetics. Due to this, we could map population migrations over time.

Show arrows while saying the following

- So we now know that the Harappans were a mixture of First Indians (these were the original out of africa migrants) and Iranian agriculturalists. We also know through a mixture of archaeological and linguistic research that they spoke proto dravidian with influences from proto elamite.
- This gives us strong evidence that the root language of the indus valley script is proto Dravidian

Slide 5: Purpose - Sonia
Our goal is to decipher ancient Indus Valley (Harappan) script by comparing it to a Dravidian language such as modern Tamil

Slide 6: Preprocessing - Keetu
The first step is to convert Tamil to a script that is easy to compare with the Indus script.
The Indus script is Logosyllabic which means that signs can either be syllables or morphemes
What's a morpheme?
It's the smallest unit of a word that changes its meaning. For example in these words - cats and playing, cat, the plural marker s, play and ing are all morphemes. what this means is that The indus script has separate signs for some of these morphemes as well as for syllables. So we Converted Tamil from Syllabic to Logosyllabic by converting grammatical morphemes to signs.

This is an example of a tamil word with all the morphemes separated and here is the conversion to a logosyllabic word. As you can see, the morphemes have been replaced with their own sign.

We converted a tagged part of speech tamil database to logosyllabic using pandas for preprocessing and a postgreSQl instance on amazon AWS for database storage. The Tamil database had separated some morphemes but not all of them so we completed the separation using functional programming and regex to filter the Tamil grammar tags. Then we combined the words into sentences to perform the analysis.

Slide 8: Analysis - Keetu
Once we converted the script, we ran a statistical analysis on it. We created positional histograms and built segmentation trees. The positional histograms show the positional distribution of each sign within a sentence and the segmentation trees show connectivity between signs. This lets us identify groups of signs that appear next to each other.

Slides – Results - Keetu
So what are our results so far? Firstly we may have identified the plural marker. The most frequent sign in the Tamil data is the plural marker 'kal' After analyzing the positional distributions of the most frequent signs in the Indus database, we found one that was very similar to the distribution shown by the plural marker. Sign 220. Note that the Indus script reads from right to left but Tamil reads from left to right like English. Just for context, here are the distributions of the other frequent Indus signs.

NEXT SLIDE - Keetu
Sign 220 is a fish sign and fish signs are known for being associated with numerals. They occur with numerals so often that they have been classified as units of measurement (ex - kg or meters)
They also only occur in the middle of a sentence. Remember that the Indus script reads from right to left.
But sign 220 is different. Like the Tamil plural marker, it can occur in the initial, medial or terminal part of a sentence.
This behaviour was previously attributed to polyvalence. A polyvalent sign is a sign that means different things in different contexts. It was proposed that sign 220 functioned as a unit of measurement in the middle of a sentence but had another function in the initial and terminal parts of a sentence.
There is a possibility however that its range is because it is the plural marker.
We also looked at sign pairings and we can see that sign 220 occurs next to many different signs, not just the numerals and that matches the behaviour of

the plural marker.

NEXT SLIDE - Keetu
We also identified the possible verb signs in the Indus script. Tamil has a verb final syntax and so does the Indus script. So the next step is to analyze sign groups based on the connectivities and match the Indus signs with Tamil verb morphemes.

Slide 9: Future Steps - ML - Sonia
To automate the conversion of Tamil from syllabic to logosyllabic, we created a Support Vector Machine algorithm to perform morpheme extraction.

Through our research, we referenced two papers focusing on morpheme extraction for Tamil using Machine Learning. They reported a high accuracy using SVM which largely influenced us to opt for a SVM supervised ml model as well.

We preprocessed the data and selected the features: the index, the current letter in the word, whether or not the letter in the word was a vowel/consonant, word was a noun or verb, current prefix. The Output label for our model was the morpheme boundary. **CLICK**

- We opted for a 75% training and 25% testing data split, rbf kernel with a gamma of 0.8 to achieve an accuracy of 76%.

The tools we used were sklearn and imblearn to program the ML model and Pandas for preprocessing and feature extraction.

The plan is to use this Machine learning model on future statistical analyses as well which I will further elaborate on shortly.

Slide 10: Next Steps - Sonia
Our next steps include:
A Continued Comparison of the two scripts - Tamil and Indus script
Using the SVM machine learning algorithm to extract morphemes from an even bigger Tamil dataset, convert those morphemes to LogoSyllabic, and run our statistical analysis on that to see how our results vary
We also want to use the SVM ML algorithm to convert the Old (Sangam Tamil

and Tamil Brahmi Scripts) to LogoSyllabic and run our statistical analysis with indus valley script to see the results of this as well

We also would like to Develop a deep neural network to automatically compare similar positional histograms and connectivities

Lastly, we would like to a Create a feature in Visualization that will allow a user to:
Run the SVM algorithm on an inputted Tamil csv file/text to segment morphemes
Automatically convert the segmented script to logosyllabic
Automatically run statistical analysis and show results to the user