Identifying Anomalous Indus Texts from West Asia Using Markov Chain Language Models

1st Varun Venkatesh
Dublin High School
Dublin, CA, USA
varun.v9725@gmail.com

2nd Ali Farghaly
Faculty of Computers and Information
Cairo University
Cairo, Egypt
alifarghaly@yahoo.com

***Abstract*—The Indus Valley civilization thrived during its mature period between 2500 BCE and 1800 BCE. and traded with the bronze age civilizations of West Asia in Mesopotamia and the Persian Gulf region. During this period, the Indus civilization developed a logosyllabic writing system that remains undeciphered. In our research, we built various Markov chain language models from the Indus texts corpus. Using a best-fit language model, we calculated the model perplexity for each Indus text found in the West Asian region. Our results show that the model perplexity was high for a majority of West Asian Indus texts and that these texts did not fit in well with the language model built with the Indus texts from just the Indian subcontinent. From this, we surmise that West Asian Indus Texts were written in a different language or syntax than those from the Indian subcontinent.**

***Index Terms*—Data Science, Machine Learning, Natural Language Processing, Language Model, Indus Script**

I. Introduction

The Indus Valley civilization that grew around the Indus River Valley started circa 3300 BCE and reached its mature phase around 2500 BCE. This civilization was an urban civilization centered around small cities spread around some 80,000 sq. km. During its mature stage, a writing system developed and was extensively used by its population [1]. While the actual use of the writing system, now called the Indus script, could have been used mainly for documentation and trade, we don't know much about it. The Indus script disappeared from its widespread use after around 1900 BCE. Excavations conducted in the northwest of the Indian subcontinent in the past hundred-odd years have revealed an urban civilization with well-planned cities, houses laid out in grids, drainage systems, and baths. This civilization also left behind beads, seals used for trade, pottery, and other artifacts. Archeologists and linguists have collected writings from these artifacts and built an Indus script text corpora. Over the years, extensive research has taken place to understand the underlying language and what these Indus texts convey. The Indus script is now categorized as logo syllabic script that is primarily written right to left.

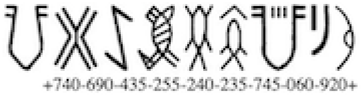


Fig. 1. Indus Text with ICIT code below for a seal from Mohenjo-Daro M-671. The text is read right to the left.

There is consensus that the language(s) could belong to one of the proto-Dravidian families, Indo-Aryan, or a proto-Munda language family [2] [3] [4]. The texts found thus far are short; the text corpus is limited, and the language it encodes is unknown. We have also not found a Rosetta stone with a multilingual inscription to help the decipherment effort. Consequently, the decipherment effort has hit significant roadblocks; the language still needs to be deciphered [2] [3] [4]. Fig. 1 shows an Indus text from Mohenjo-Daro with ICIT code at the bottom corresponding to each sign.

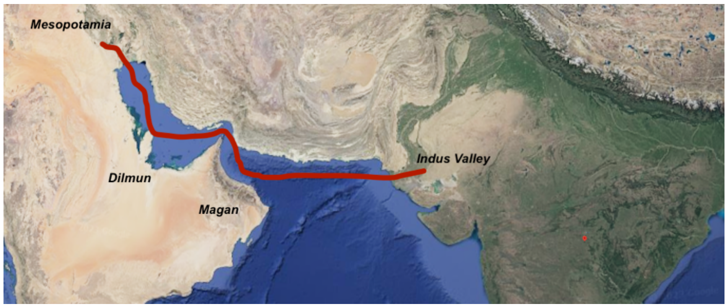


Fig. 2. Illustrative trading Routes of Indus Valley Civilization with West Asia.

While archeologists have found Indus texts in the northwest part of the Indian subcontinent, Mohenja-Daro and Harrapa being the most important sites, some artifacts and texts have also been found far from the Indian subcontinent [5]. Researchers have found that a vast trading network connected the Indus region (called Meluhha) with Mesopotamia and several areas of the Persian Gulf in the West Asia region from 2500 BCE till about 1900 BCE. Sumer, Ur, Uruk, Susa (Southern Iraq, Eastern Iran), and Dilmun (Bahrain) traded with the Indus Valley civilization during the early and mature Harappan period [6]. Indus Valley traders also traveled and lived in these far-off regions and produced Indus text seals. Archeologists have excavated these artifacts with Indus scripts from West Asia. Fig. 2 shows possible trade routes of the Indus people to these places in West Asia [6], [7]. Many of the seals with Indus texts that were found in the Persian Gulf are circular and called the Gulf-type seals [7]. While the number of Indus texts found is small, researchers generally agree that these texts are

Indus script [5] [7]. The focus of this research is to understand if there are differences in the Indus text corpus found in West Asia compared with the texts from the Indian subcontinent using statistical methods. Identifying these differences will lead to a better understanding of the language these texts encode.

II. Related Works

There are interesting statistical analysis works done on the Indus script that have contributed to a greater understanding of the linguistic nature of the texts [8] [9] [10]. Some have employed n-gram Markov models of the Indus texts to emphasize the linguistic nature of the Indus texts and use that to further fill in missing texts [11] [12] [13] [14]. Some researchers have also noted that West Asian texts often contain unusual sign combinations indicating that they were different from the Indus texts from the Indian subcontinent [15]. Researchers have also explored anomalous Indus texts from West Asia using statistical likelihood measures and have flagged the texts for further exploration [13]. They have used an older Indus text corpus with a basic language model and pairwise dependencies. We decided to research this area further using an updated Indus text corpus, the Interactive Corpus of Indus Text (ICIT), and more sophisticated n-gram language models with higher order n-grams with smoothing and interpolation [16]. Other researchers have even used language model perplexity to identify differences and distances between languages for other languages [18] that we have considered for this research. Based on these previous research works, we decided to build n-gram language models from Indus scripts to specifically understand Indus Texts from West Asia. With a more recent and complete ICIT corpus and sophisticated Markov chain n-gram language models used in this research, we can identify anomalies in the West Asian Indus texts using perplexity measures.

III. Hypothesis

We hypothesized that the West Asian Indus texts are different from the Indus texts found in the Indian subcontinent. We think the Indus traders settled in West Asia and started producing seals by writing in the local language using local names and measurements. We can identify this difference in texts by building a language model with texts from the Indian subcontinent and then calculating how well the West Asian Indus texts fit in that language model using model perplexity measures. By proving that the language or the syntax in the West Asian Indus script texts is different from that of the Indus script texts from the Indian subcontinent, we can highlight an example of long distant regional differences in the Indus script.

IV. Markov Chain Language Model and Perplexity

We used the Markov Chain and n-grams to build a statistical language model from the Indus Corpus. An n-gram, in this context, is a contiguous sequence of ‘n’ Indus signs for a given sequence of Indus text. To build up the language model, we calculate the frequencies of the n-grams and derive their probabilities. Language models then use a simplification called the Markov chain assumption. We can obtain the probability of a sign appearing from the conditional probability of nearby preceding signs without looking too far behind [17] [19].

The language model uses this Markov chain assumption to build a conditional probability matrix for the n-grams.

$$P\left(w_n \mid w_{1:n-1}\right) \approx P\left(w_n \mid w_{n-N+1:n-1}\right) \tag{1}$$

Where:

N = Size of n-gram
w_n = The sign in n th position
$w_{1:n-1}$ = Signs in position 1 through n-1

The Markov chain approximation of the conditional probability for the n-gram size of N is given by (1). We can use the resulting language model to predict the next possible sign given a set of signs before it. It can also be used to measure how well a set of signs fit in the language model.

Perplexity is a measure typically used to evaluate a language model intrinsically and is often explained as ‘how confused is the language model is for your test set”. It is commonly used in language modeling as a quality indicator for language models created using n-grams retrieved from text corpora [19].

$$PP(W) = P\left(w_1 w_2 \ldots w_N\right)^{-\frac{1}{N}} \tag{2}$$

Where:

PP = Perplexity
W = Test set
N = Number of words
w_n = The sign in n th position

The perplexity of N words (or characters) in a language is defined as the inverse probability of the test dataset normalized by the number of words or characters (2), [19].

The lower the probability of a test dataset occurring in the model, the higher the perplexity of the model. A language model returns higher perplexity for a test dataset usually because the language model itself is not optimal for the training dataset, and we need to choose a better language model, or the test dataset is so different from the training data that the model does not see a fit [17] [19]. Perplexity measures are generally specific to the model, the n value for the n-gram used, and the dataset that is used. So it is essential to use the same model, and the same n value for building the n-gram tokens to compare perplexities.

Some language models handle Out Of Vocabulary (OOV) words or characters by assigning probabilities and marking them as ’unknown’ so that perplexity is not infinite. Language models also use smoothing techniques for sequences not seen in the training model. Various models and n-gram orders must be tried on the language corpus to find the optimal language model for the training dataset.

V. Methods

A. Corpus Selection

The popular Indus texts corpus are M77 corpus, which has about 3500 lines of text [8], which most of the previous research works have used, and the ICIT corpus, which has some 4500+ artifacts with texts [16]. It is a living corpus; the corpus maintainers add newly found texts regularly, and the corpus is regularly updated. There are 417 distinct Indus signs in M77 and 695 in the ICIT corpus. Since ICIT was much more complete, latest, and digitally available, we used the ICIT corpus in this study.

B. Data Pre-Processing

As the first step for any statistical analysis, we must organize and clean the data. We first compiled the texts from the ICIT database into easily workable CSV files. To clean up the data, repeated texts from the same dig site with the same cult object in the artifact (usually an animal or an object) were considered duplicate texts to avoid the TAB effect [9], and to avoid counting some texts multiple times. We kept only one copy of such repeated texts. We discarded texts with ambiguous directionality and eliminated texts that had missing signs. Similarly, we eliminated multi-line texts, as in some of them, the directionality and continuity of texts was ambiguous. We also eliminated multipart texts. Other researchers have done similar data cleaning when dealing with the Indus text data [11] [13].

ICIT signs are 3-digit numbers; texts have these three-digit numbers separated by a '-.' Text beginning and end are marked by a '+.' Unclear signs are marked as '000'. For example, the seal from Lothal (L-95) is represented as +740-000-175-002-880+, and read right to left [16]. Since it is easier to work with left-to-right texts, we converted all texts to left-to-right text and removed the + and - signs. The original ICIT text corpus has about 4500+ texts, and after data cleaning, we built a cleaned-up CSV file with about 2200 unique texts. We had only eleven texts from West Asia without any missing signs and whose length was over three in our corpus. We removed these West Asian texts and kept them aside.

C. Model Building and Selection

We split this cleaned-up Indus texts into 80% training and 20% test datasets. We picked six popular language models, some with interpolation, some with smoothing, and some with both, for n-grams ranging from n=2 to n=4. We aimed to train our dataset and select a suitable model for the Indus script. The training involved padding each Indus text with text beginner and text ender. We used <s>and </s>for it. We needed the padding to differentiate between initial, terminal signs, and other signs. We then used NLTK library to build n-grams from each padded text in the training dataset for various 'n' values and passed the generated n-grams to train the models. We trained the MLE, KneserNey Interpolated, Lidstone, Stupid Back-off, and Witten Bell Interpolated language models for n-grams for different n values [17]. For example, n=4 means that we trained the model from all n-grams where n=1, n=2, n=3, till n=4. Earlier findings indicate that a cluster of 4 signs on the right and 3 signs on the middle seems significant in the ICIT corpus, and the significance of clusters of signs reduces after n=4 [11]. Based on this, we did this perplexity analysis for up to n=4. We then ended up with language models that understood the sign patterns in the Indus texts well. We then used the texts from the test dataset as the hold-out set to calculate perplexities. We calculated the average and median perplexity of the test dataset texts. We eliminated any model that returned an infinite perplexity for any text in the test dataset.

D. Randomized Testing of Our Selected Model

To ensure the selected language model works well, we wanted to do some tests to see the perplexities it generated when tested against two types of randomly generated texts. For the first one, we shuffled the sign positions randomly for each text in that test dataset and generated a new Randomly Shuffled Signs dataset. For example, we randomly shuffled the text +740-001-175-002-880+ to +001-880-002-175-740+. For the second one, we created a Randomly Generated Signs dataset by generating the texts with random signs from all possible Indus signs. We kept the text length distribution similar to the test dataset for both these datasets to avoid introducing perplexity variations due to text length differences.

We used the random datasets as the hold-out datasets against the trained model to compare our perplexities with perplexity numbers obtained from the test dataset. We expected that perplexities for the test dataset would be lower than those of the Randomly Shuffled Signs dataset, which would be lower than that of the Randomly Generated Signs dataset.

E. West Asian Indus Texts

We eliminated the West Asian Indus texts with at least one unclear and missing sign. We also eliminated texts with lengths less than three signs as smaller texts generally represent the general corpus well and result in low perplexity, making perplexity measures distorted. We then used the remaining West Asian Indus texts as the hold-out texts. The text length distribution of the West Asian Indus text was similar to the text length distribution of the test dataset. We passed each of the remaining West Asian Indus texts through our model and computed the perplexity values of the texts. We also reversed the direction of the West Asian texts to calculate the perplexity in the opposite direction just in case the directionality of the texts was incorrectly captured in the ICIT corpus.

VI. Results

A. Choosing a Language Model

We passed each text in the test dataset through all models and calculated perplexities. One of our goals was to eliminate language models that generate infinite perplexities. Most of these models generated infinite perplexities for at least some of the texts in the dataset, likely because they did not have a smoothing algorithm or the smoothing functions did not work well for the data, as shown in Table I. The Lidstone language model, with n=4 (trained on every n-gram from length 2 to length 4) and a gamma of 0.75, was picked for the rest of the tests.

B. *Perplexities for Random Texts*

As mentioned earlier, we created two types of random Indus texts: Randomly shuffled signs and Randomly generated signs, and we generated the perplexities. Using the language model selected (Lidstone, n=4), the test dataset produced median, mean, and 90th percentile perplexities much lesser than the perplexities of both the Randomly shuffled signs and the Randomly generated signs datasets. Table II shows that the perplexity for Randomly generated signs was even more than the Randomly shuffled signs. This observation aligns with our expectations and proves that our chosen language model worked well.

C. *Perplexities for West Asian Indus Texts*

The perplexities of West Asian Indus texts are shown below in Table III. We used the perplexity at the 90th Percentile value of 74.1 for the test dataset as the threshold and considered any text with perplexity greater than it as an outlier. West Asian Indus texts with ICIT IDs 2153, 3863, 3882, 3897, 3898, 4173 and 5229 had high perplexities and were outliers. The high perplexities indicate that their sign patterns are significantly different from the Indus texts from the Indian subcontinent.

D. *Beyond Perplexities*

Not only did the West Asian texts have high perplexity, but also specific sign patterns had a very low probability of occurrence. For example, the sign pattern 091-840 found in a text from Luristan and the sign pattern 595-278 found in Salut was absent from the rest of the Indus corpus from the Indian subcontinent. For high-perplexity texts, there were patterns found in these West Asian texts that never occurred elsewhere, indicating that the syntax or pattern of texts is different from the texts in the Indus script from the Indian subcontinent. When we reversed the West Asian Indus text and did perplexity measurement, we found that the reversed texts had higher perplexities than the perplexities of the un-reversed text, indicating that misreading the direction of the text was not the problem.

VII. Discussion

A. *Discussion*

Researchers have observed that the Indus texts exhibit a structure that encodes a language [8]. They have concluded that the Indus script exhibits a language-like behavior with sign clusters and a language model would be apt to build using it. Based on this, we built various Markov chain language models with various smoothing and interpolation techniques for various orders of n-grams to understand what language model works well for our texts.

We chose the Lidstone model with n=4 for further analysis. Our results on perplexity for the Randomly shuffled signs and Randomly generated signs indicated that our chosen language model represented the cleaned-up Indus text corpus well. We then obtained the perplexity of each of the West Asian Indus texts. High perplexities of the texts for most texts indicate that the pattern/sign clusters of West Asian Indus texts differed from the Indus texts from the Indian subcontinent. High perplexities suggest that these West Asian Indus texts were encoding a different language or using an other linguistic syntax than the Indus texts from the Indian subcontinent. If the texts represent names of persons or a trade group, quantities and measures as some suggest (as the seals with writing are typically used in trade), the names in West Asian Indus texts perhaps sound different, or the quantities and measures are somewhat different than in the Indian subcontinent. Fig. 3 shows the distribution of texts whose perplexity is an outlier for the Indus texts from the Indian subcontinent, the Persian Gulf, and Mesopotamia.

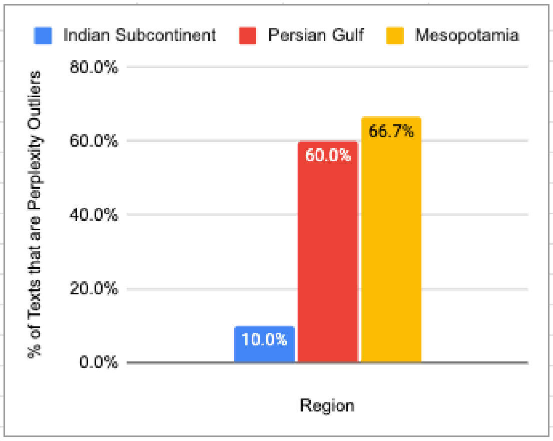


Fig. 3. Percent of Indus texts that are Perplexity outliers by Region where the texts are found

Others have used ranking basic distance to derive distances between languages [18]. In this approach, n-grams are ranked according to the frequency with which it appears in the language corpus for the two languages we want to compare. They are then pruned to retain high-frequency n-grams only, and the language distance is computed by building an 'out-of-place' measure on the n-gram ranks across the two languages. This method works well if we have enough data in both languages. Since the number of texts from West Asia is very small, we did not think we would get usable results with the ranking-based distanced method and decided not to proceed with it.

VIII. Conclusions

Researchers have suspected that some of the Indus texts in West Asia differed from those in the Indian subcontinent. This difference was because of the occurrence of certain signs which are rarely or never used in the Indian subcontinent and due to the occurrence of rare sign patterns in the West Asian Indus texts. We aimed to prove that some of the Indus texts from West Asia were different using a quantitative analysis method, and we chose perplexity as a measure to quantify

TABLE I

Mean and Median Perplexity for different language models for different maximum values of ‘n’ for the n-grams. Here the test data is the Hold-out set. We rejected any model that returned infinity as the perplexity value for any text. The Lidstone model with n=4 was picked for further use in our tests

Model Name	n=2	n=3	n=4
KneserNeyInterpolated	∞, ∞	75, ∞	53, ∞
Lidstone	65.5,87.4	44.5,55.1	**34.2,41.1**
MLE	∞, ∞	∞, ∞	∞, ∞
StupidBackoff	33.0, ∞	13.1,∞	6.5, ∞
WittenBellInterpolated	21.2, ∞	13.3, ∞	9.8, ∞

TABLE II

Perplexity statistics of Test Set, Randomly shuffled signs, Randomly generated signs using Lidstone model n=4. The perplexity of the Test set was much lower than that of the random sets, as expected

Statistics	Test	Random Shuffle	Full Random
Median	34.2	72.7	167.0
Mean	40.6	74.6	171.7
90th percentile	75.4	114.5	247.9
Std Deviation	23.8	28.8	54.0

TABLE III

Perplexities of the West Asian Indus Texts with Outliers. The texts marked Yes in the Outlier column are considered anomalous texts. A high percentage of anomalous texts indicate that West Asian Indus texts are different. Note: The Indus texts here are read left to right

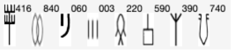
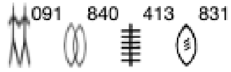
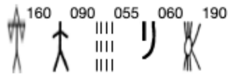
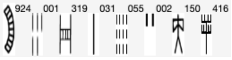
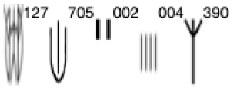
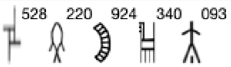
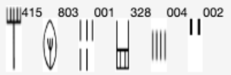
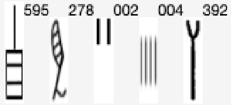
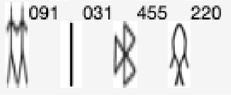
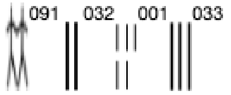
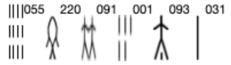
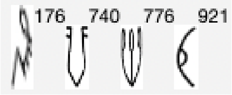
ICIT ID	Text Length	Perplexity	Site	Region	Outlier?	Text Image with ICIT Code
1971	8	41.4	Kish	Mesopotamia	No	
2153	4	89.4	Luristan	Mesopotamia	Yes	
3863	5	101.6	Qala’at al-Bahrain	Persian Gulf	Yes	
3882	8	162.4	Susa	Mesopotamia	Yes	
3884	5	49.2	Tell Umma	Mesopotamia	No	
3897	5	140.4	Ur	Mesopotamia	Yes	
3898	6	86.1	Ur	Mesopotamia	Yes	
4173	5	108.78	Salut	Persian Gulf	Yes	
5227	4	54.86	Karzakan	Persian Gulf	No	
5228	4	60.39	Karzakan	Persian Gulf	No	
5229	6	162.39	Saar	Persian Gulf	Yes	

TABLE IV
PERPLEXITIES OF A SOUTH INDIAN INDUS TEXT. NOTE: INDUS TEXT HERE IS READ LEFT TO RIGHT

ICIT ID	Text Length	Perplexity	Site	Culture	Outlier?	Text Image with ICIT code
None	4	49.4	Tamil Nadu	South India	No	

these differences. To compute the perplexities, we trained various language models from the Indus texts. To select a good language model among the models, we computed Randomly shuffled signs and Randomly generated signs datasets and computed the perplexities for them. We found that perplexities were high for both of them, as expected, indicating that our chosen Lidstone language model with n=4 and gamma =0.75 worked well. We used this language model to compute the perplexities of West Asian Indus texts. We found that 67% of the texts from Mesopotamia and 60% from the Persian Gulf had perplexities that were outliers. In short, most of these texts did not fit in well with the language model built with Indus texts from just the Indian subcontinent. We think this is because the language or the syntax in the West Asian Indus texts are different from the Indus texts from the Indian subcontinent. This is further corroborated by the fact that certain sign combinations found in texts from West Asia seldom appear in texts from the Indian subcontinent. The Indus text corpus we used to build the language models is limited, and we don't have long texts. So we may be trying to build a language model from snippets of texts, names, or measures that may not represent the entire script and the language. Furthermore, the number of texts found thus far from West Asia is small and may only partially represent some Indus texts used in that area, and could distort our results.

Results and code are shared with the Indus script research community for others to work on this further in GitHub at https://github.com/varundataquest/Indus Valley-Text-Analysis.

IX. FUTURE WORK

We can extend this research to analyze sites within the Indian subcontinent to understand if any differences existed in the script used in significant sites such as Mohenjo Daro and Harappa. We could also use this method to determine if the Indus script found in an outlying site within the Indian subcontinent fits in well with the Indus script from the rest of the Indian subcontinent.

To test this, we calculated the perplexities of an Indus text found in Southern India [22]. Sign 740 following 176 is prevalent in the Indus text. Still, the perplexity of this text was higher than the average of the test dataset as no Indus texts exist where sign 776 followed sign 740 and sign 921 followed sign 776, but not high enough to call it anomalous text. Table IV shows that this text found in deep Souther India fits well with the rest of the Indus scripts from the Indian subcontinent.

The Indus script could have changed over time and researchers have not done enough work to understand the evolution of this script. One could do a similar analysis across the Indus texts from early Harappan (3300-2800 BCE), mature Harappan (2600-1900 BCE), and late Harappan period (1900-1300 BCE) and understand how the script changed between these periods.

ACKNOWLEDGMENTS

We thank Dr. Andreas Fuls, Technische Universität Berlin, for managing, maintaining, and graciously sharing the Indus Text corpus through the ICIT website. Thanks to the Indus script research community for inspiring us to contribute to computational historical linguistics work on the Indus scripts.

REFERENCES

[1] A. Parpola, and J.P. Joshi, "Corpus of Indus Seals and Inscriptions", Collections in India Memoirs of the Archeological Survey of India. vol. 86, ASI, 1987.

[2] B. Alex, "Why We Still Can't Read the Writing of the Ancient Indus Civilization?", Discover Magazine, Sep 2021.

[3] S. Bonta, The Indus Valley Script: A New Interpretation. Book. Penn State University -Altoona College, 2010.

[4] A. Fuls. "Classifying Undeciphered Writing Systems", Journal of Historical Linguistics, vol. 128, pp. 42-58, 2015.

[5] G. Gadd, "Seals Of Ancient Indian Style Found At Ur", Proceedings of the British Academy (18) (pp 3-22, pls I-III), 1958.

[6] J. Kenoyer, "Indus and Mesopotamian trade networks: New insights from shell and carnelian artifacts. Intercultural Relations Between South and Southwest Asia", Studies in Commemoration of E.C. L. During-Caspers (1934-1996). 19-28, 2008.

[7] S.T. Laursen, "The westward transmission of Indus Valley sealing technology: Origin and development of the "Gulf Type" seal and other administrative technologies in Early Dilmun, c.2100–2000 BC. Arabian Archaeology and Epigraphy, 21(2), 96– 134, 2010.

[8] I. Mahadevan, "The Indus script: Texts, concordance, and tables", Memoirs - Archaeological Survey of India. vol. 77, Archaeological Survey of India, 1977.

[9] N. Yadav, M.N. Vahia, I. Mahadevan, H. Joglekar. A Statistical Approach For Pattern Search In Indus Writing. International Journal of Dravidian Linguistics, vol. 37, 2008, pp. 39-52.

[10] R. P. N. Rao, N. Yadav, M.N. Vahia, H. Joglekar, R. Adhikari, I. Mahadevan. "Entropic evidence for linguistic structure in the Indus script." Science, vol. 324, no. 5931, 2009.

[11] R. P. N. Rao, N. Yadav, M.N. Vahia, I. Mahadevan, "A Markov model of the Indus script", PNAS, vol. 106, no. 33., 2009.

[12] N. Yadav, H. Joglekar, R. P. N. Rao, M.N. Vahia, R. Adhikari, I. Mahadevan, "Statistical Analysis of the Indus Script Using n-Grams", PLOS One, vol. 5(3), no. e9506, 2010.

[13] V. Venkatesh, A. Farghaly, "Statistical Models for Identifying Missing and Unclear Signs of the Indus Script", Journal of Emerging Investigators, In press, 2023.

[14] K. Papavassileiou, D. Kosmopoulos, G. Owens, "A generative model for the Mycenaean Linear B script and its application in infilling text from ancient tablets", Journal on Computing and Cultural Heritage, 2023.

[15] A. Parpola, Deciphering the Indus script (Cambridge Univ Press, Cambridge, UK), 1994.

[16] Wells, Bryan, and Fuls, Andreas: Online Indus Writing Database. Berlin 2010, http://www.indus.epigraphica.de/ , Accessed 30 Nov 2022.

[17] D. Jurafsky,, and J. H. Martin. Speech and Language Processing, 2nd Edition. Pearson Prentice Hall, 2008.

[18] P. Gamallo, P.J Ramon, and A. Iñaki, ”Measuring language distance among historical varieties using perplexity. application to European Portuguese”, Proceedings of Fourth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2018), 2018.

[19] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[20] P. Gamallo, J.R. Pichel, and Alegria Iñaki. ”From language identification to language distance”, Physica A, 483:162–172, 2017b.

[21] B. Wells. Epigraphic Approaches to Indus Writing. Oxbow Books, 2011.

[22] Significance of Mayiladuthurai find, The Hindu Newspaper, http://www.hinduonnet.com/2006/05/01/stories/2006050101992000.htm