# Deciphering Ancient Lost Languages With Machine Learning + Phonetic Correlation

## Cracking the Code of Linear A Using Linear B

A HIST 1056 Research Project
By Jonas Hansen | Harvard University

## Research Question + Abstract

**How can machine learning be combined with existing archeological knowledge of linguistics to decipher ancient languages?**

Linear A remains the holy grail of ancient lost languages yet to be deciphered. Here, we present a possible method of reconstructing Linear A by **training a machine learning model to recognize inscription patterns**, and then comparing these results with existing knowledge of Linear B phonetics to develop a Linear A lexicon. The frequency correlation between Linear A and Linear B was strong, with a **correlation coefficient of ~0.97**. This derived lexicon was then compared to ten other ancient languages, with the highest percentage match being **25.8% with Proto-Basque**.

## Introduction

- In 1886, a British Archaeologist named Arthur Evans happened upon numerous stone tablets dating back to 1400 BCE with inscriptions in two unknown languages. The first and oldest was dubbed Linear A, which was determined to originate from the Mediterranean island of Crete when it was dominated by the Ancient Minoan civilization. The second and more recent language, Linear B, appeared after 1400 BCE when the Mycenaeans from the Greek mainland conquered Crete and adopted Minoan linguistics.[4]
- For years, these languages remained a mystery, until 1952 when linguist Michael Ventris successfully deciphered Linear B. By assuming that Linear B was directly related to ancient Greek and that many of the repeated words found on the tablets were names of locations on the island of Crete, Ventris was able to construct a complete model for Linear B.[4] While this took years to decipher, recent machine learning methods have demonstrated a successful deciphering of Linear B with the click of a button.[5]
- Linear A has remained quite elusive. The problem is that it does not seem to be related to any known languages, and all attempts at trying to find correlations with ancient scripts have been unsuccessful.[6]
- However, recent research has provided insight into how phonetics and natural language processing (NLP) can be used in conjunction with machine learning models to automatically decipher languages without any prior knowledge of their relationships to other languages.[3] Such a method would help fill in the missing gaps of languages once thought lost and demonstrate the validity of using machine learning to uncover elements of the missing past.

## Methods

- First, images of Linear A tablets were analyzed to train a deep learning classifier to recognize similar symbols.

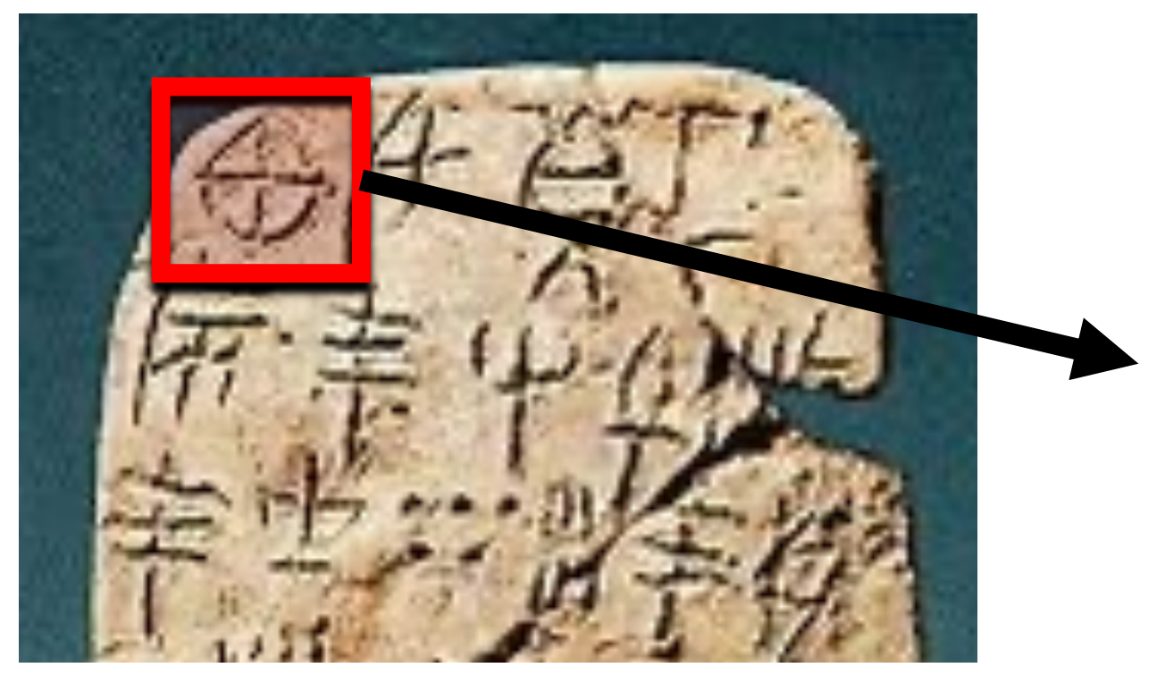

**Figure 1.** Sample Image of Linear A Tablet.[7]

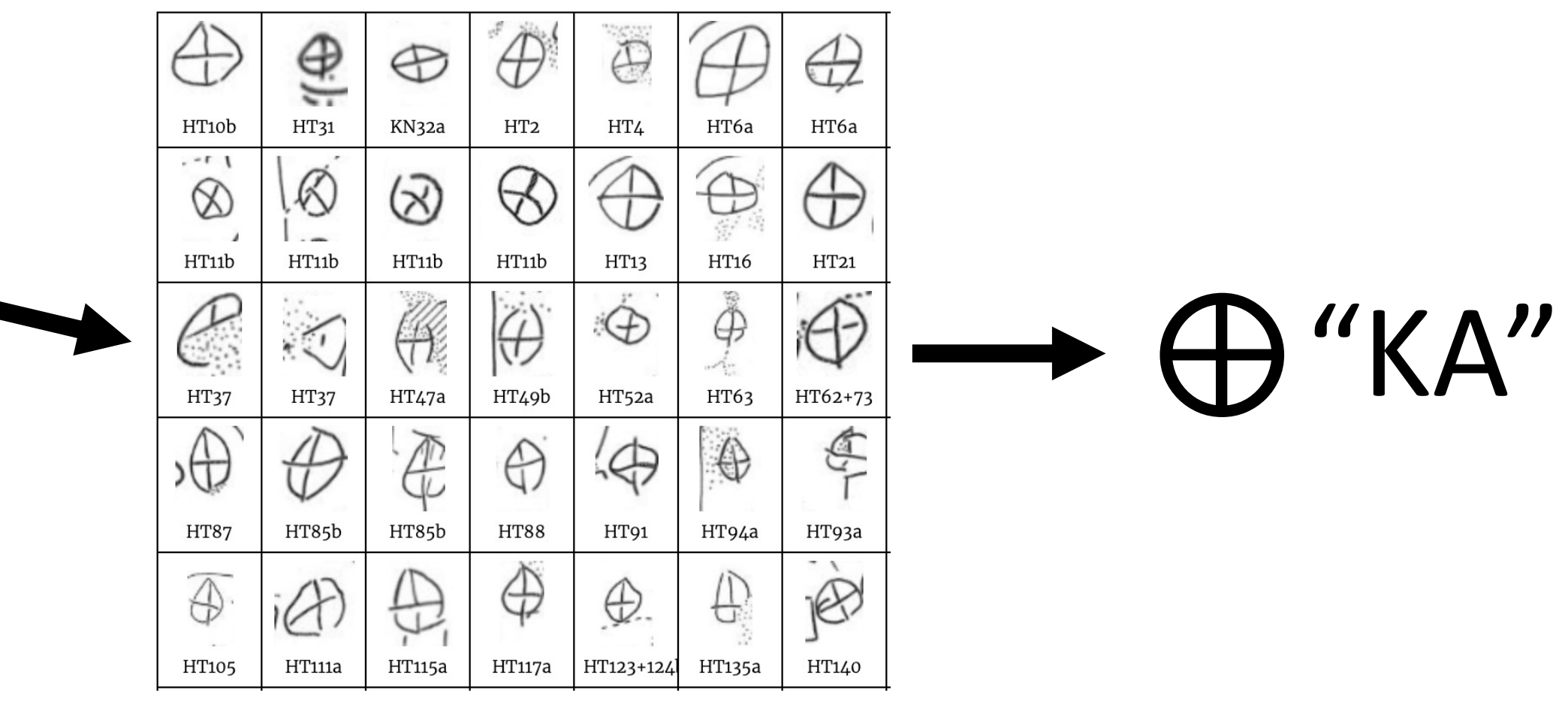

**Figure 2.** Dataset of Symbol Images.[1]

- Using a phonetic comparison between syllables of Linear A and those of Linear B, a frequency correlation plot was developed (see Figure 3 below). This correlation was then extended to find identical word matches between Linear A and ten other ancient languages (see Figure 4 below).

## Results

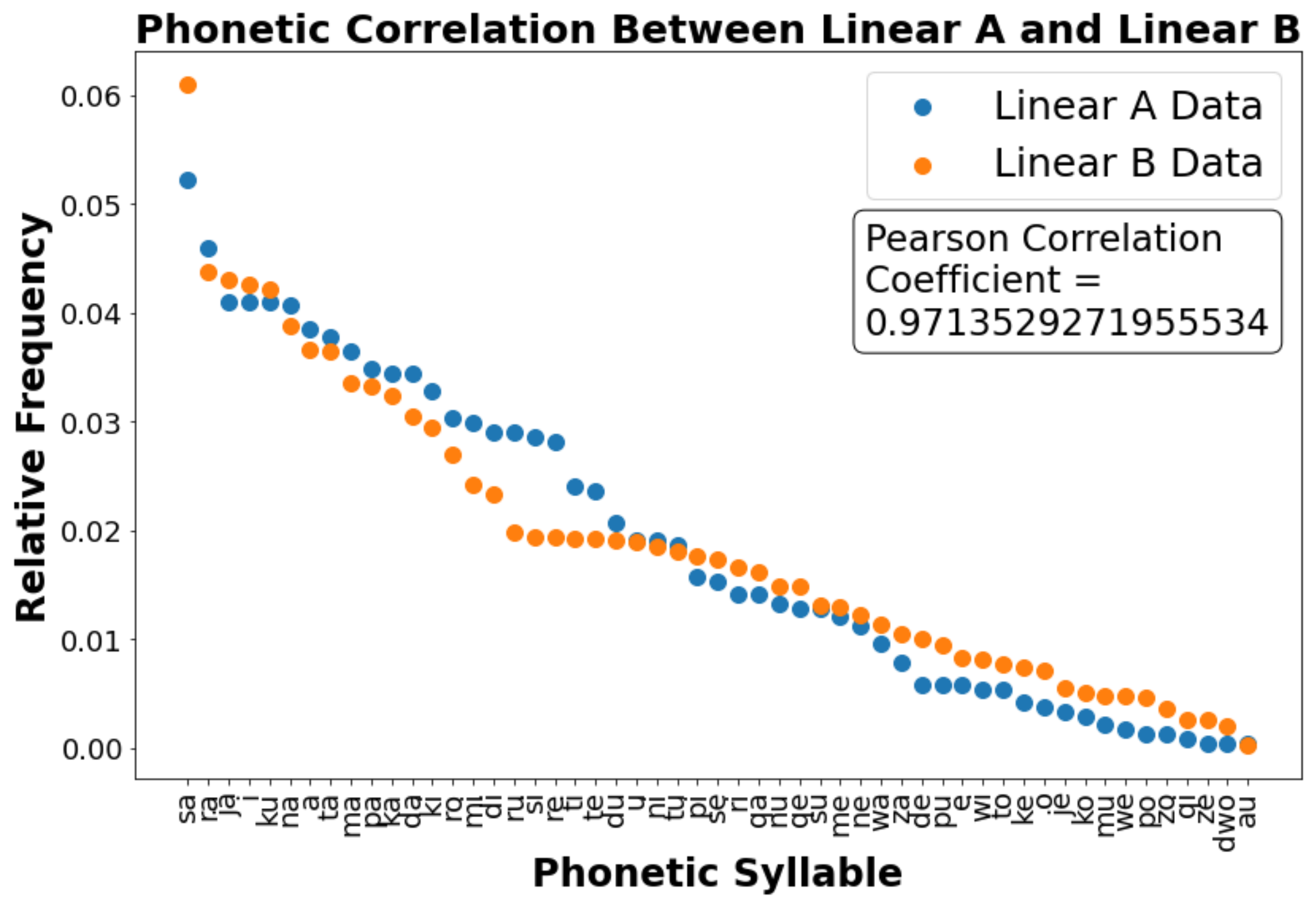

**Figure 3.** Frequency correlation of the trained Linear A syllable set with known relevant syllables of Linear B.[9] The syllables are shown in order of descending frequency. The resulting Pearson correlation coefficient was found to be ~0.97.

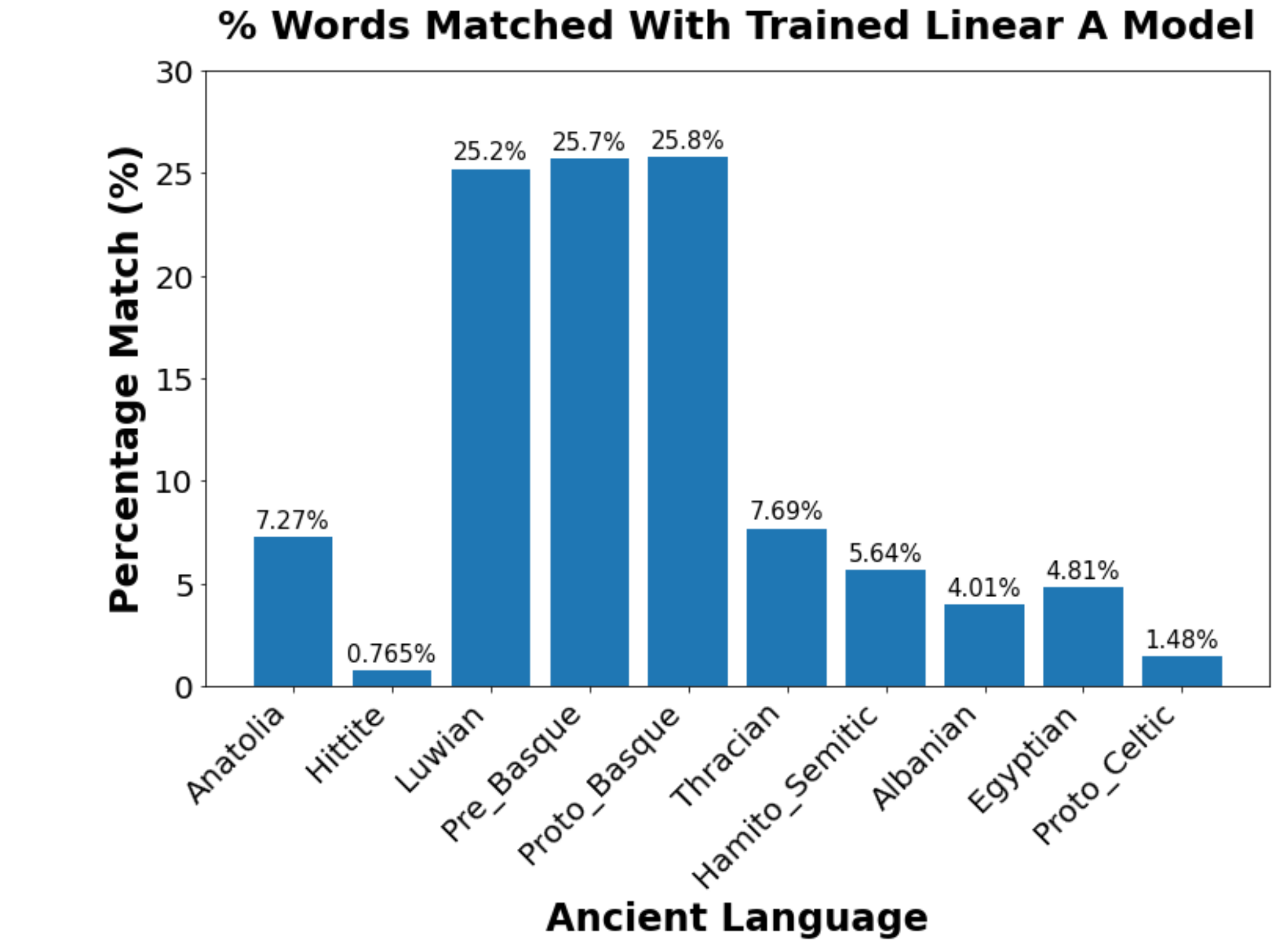

**Figure 4.** Histogram showing relative percentage match of words from the Linear A model with words in ten other ancient languages.[8] The words from the Linear A model were inferred via the Linear B phonetic correlation.

## Discussion

- From the trained deep learning model of Linear A, we have demonstrated how it is possible to **construct a lexicon based on character recognition** from thousands of available inscription images. Deriving the meaning of characters in this lexicon, however, is what remains undetermined. Therefore, we present a possible derivation based on **comparing the phonetics between Linear A and Linear B**. The frequency correlation between the proposed Linear A phonetics and known Linear B phonetics is strong, with a **calculated Pearson coefficient of ~0.97** (Figure 3).
- Nearly **26% of Pre-Basque and Proto-Basque words matched** the derived Linear A words (Figure 4).

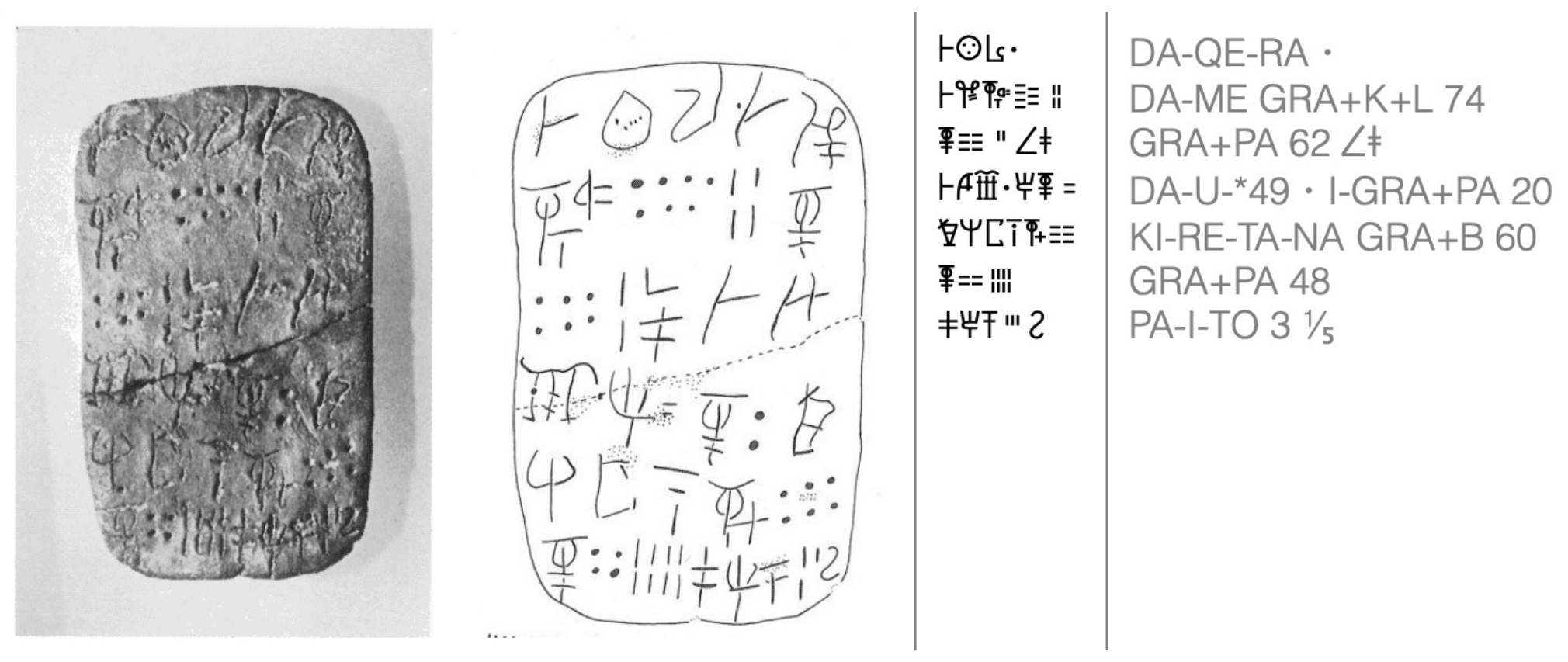

**Figure 5.** Our correlation model closely aligned with another possible phonetic derivation of Linear A developed by Robert Hogan.[1]

## Future Work

- A next step for this project would be to corroborate this data with other attempts at deciphering Linear A to try and fill in any gaps and discrepancies that exist between similar sounding words. For example, Figure 6 shows an intriguing attempt by Janke et al. to translate Linear A based on phonetic similarities with Linear B.
- Another improvement would be to employ a deep learning model that could automatically infer linguistic similarity based on phonetics, geography, time period, historical events where civilizations may have interacted, etc., to estimate the meaning of words in Linear A.

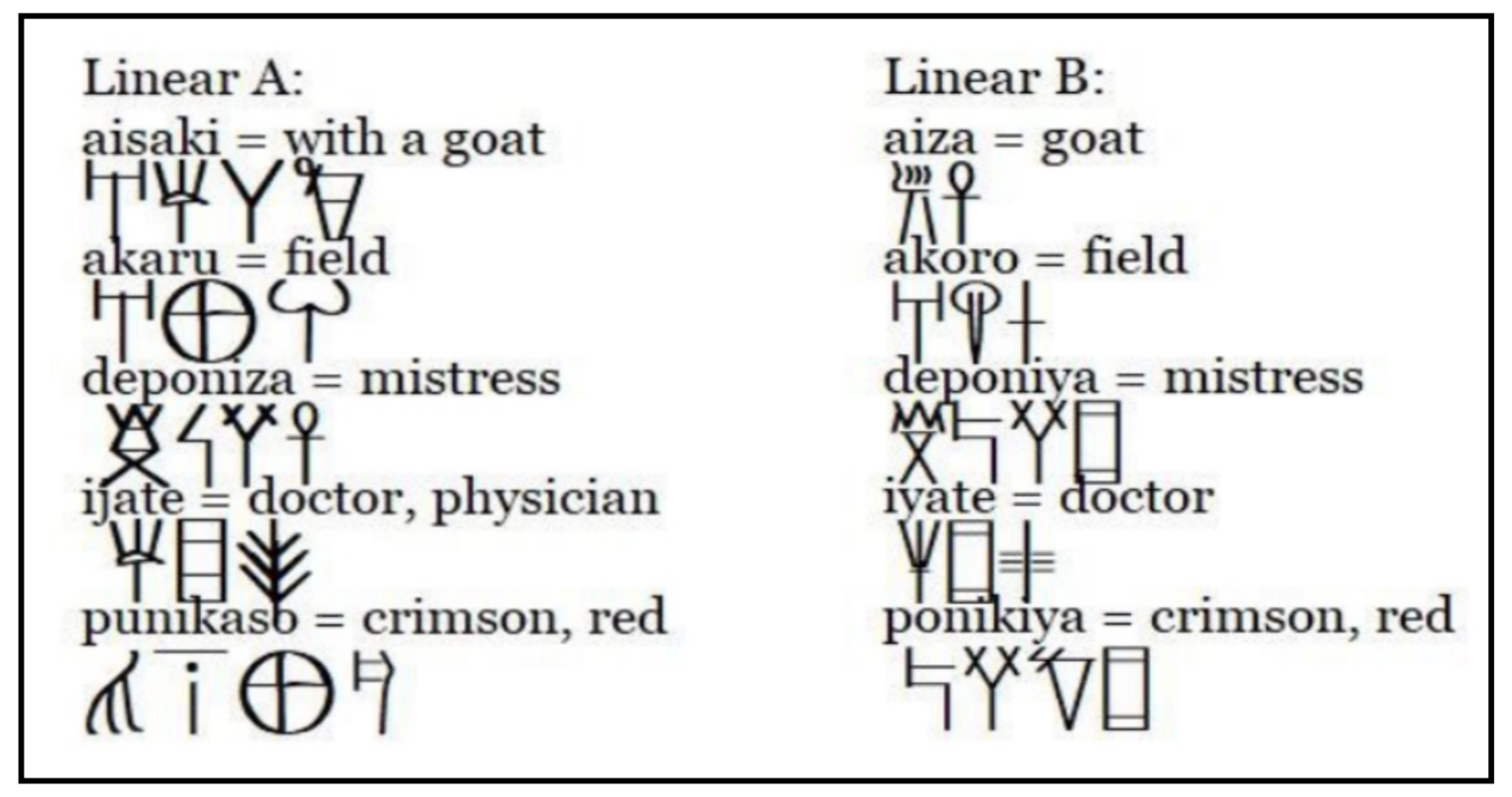

**Figure 6.** A proposed method of translating Linear A vocabulary by Janke et el., based on phonetic similarities with Linear B (e.g. "akaru" sounds like "akoro," therefore "akaru" means "field").[2]

## References/Acknowledgments

1. Hogan, R. (2021). *LinearA Explorer* [JavaScript]. https://github.com/mwenge/lineara.xyz (Original work published 2019)
2. Janke, R., Solcà, A., Bengtson, J., Binnberg, J., Gingras, J.-P., Canada, R., & Koslenko, R. (2018, December 31). *High Correlation Linear A-Linear B vocabulary, grammar and orthography in Linear A Board of Editors/Conseil des rédacteurs.*
3. Karajgikar, J. R., Al-Khulaidy, A., & Berea, A. (2021). *Computational Pattern Recognition in Linear A.* https://hal.archives-ouvertes.fr/hal-03207615
4. *Linear B.* (n.d.). Archaeologies of the Greek Past. Retrieved November 8, 2021, from https://www.brown.edu/Departments/Joukowsky_Institute/courses/greekpast/4690.html
5. Luo, J., Cao, Y., & Barzilay, R. (2019). Neural Decipherment via Minimum-Cost Flow: From Ugaritic to Linear B. *ArXiv:1906.06718 [Cs].* http://arxiv.org/abs/1906.06718
6. Luo, J., Hartmann, F., Santus, E., Barzilay, R., & Cao, Y. (2020). Deciphering Undersegmented Ancient Scripts Using Phonetic Prior. *Transactions of the Association for Computational Linguistics, 9,* 69–81. https://doi.org/10.1162/tacl_a_00354
7. *Mnamon: Ancient Writing Systems in the Mediterranean.* Scuola Normale Superiore. http://mnamon.sns.it/index.php?page=Esempi&id=19&lang=en. Accessed 8 Nov. 2021.
8. PERONO CACCIAFOCO, F., & Loh, C. (2021, March 1). *A New Approach to the Decipherment of Linear A, Stage 2 - Cryptanalysis and Language Deciphering: A "Brute Force Attack" on an Undeciphered Writing System*. https://doi.org/10.36824/2020-graf-cacc
9. Tselentis, C. (n.d.). *Linear B Lexicon, by Chris Tselentis (Greece).* Retrieved November 8, 2021, from https://www.academia.edu/15310428/Linear_B_Lexicon_by_Chris_Tselentis_Greece_

**With A Special Thanks to Professor Michael McCormick, Reed Morgan, and Yingxue Wang.**