# GMU CDS490 AC

**Aron Cottman**

**May-August 2020**

**Agnostic Analysis of Symbolic Writing Systems: Linear A**

## 1 Introduction

The data that will be worked with in this project is the Linear A corpus. I would like to be able to be more comfortable with image analysis, natural language processing and statistical analysis methods relating to NLP and image processing. The idea of contributing to new methods that could provide insight into deciphering ancient languages is very interesting. =

Developing algorithms for analyzing symbolic data can also help us generalize current status quo NLP methods in a way that is transferable and reproducible to other languages and other forms of communication as well. Data science, artificial intelligence and machine learning techniques are constantly evolving, and have not been exhaustively applied to indecipherable languages yes. this

Linear A is an undeciphered writing system used by the inhabitants of Crete from 1800 - 1450 BC. It was followed by the Linear B writing system that was deciphered in the 1950's. Linear A consists of about 350 unique symbols, across 1400 documents containing about 7400 signs. In contrast, linear B has over 4600 documents, and 57,400 signs. [7]. Added to this lack of data, a large number of symbols in the Linear A corpus are unreadable. If the unreadable symbols were readable in the set of discovered documents, the corpus would add about 2500 more signs.

## 2 Executive Summary

Decryption of ancient languages has a large amount of work being done, and is an active area of research. There is work that is using modern machine learning and AI techniques to try and gain further insight into various languages. [1][3] While deciphering a previously indecipherable language would be an amazing outcome of this project, a realistic goal is to uncover some previously unknown underlying structure and statistical properties of the writing system.

## 2.1 What is the problem you are trying to solve? What are you trying to do?

The purpose of this project is to develop methods and solutions for analysing symbolic data. The symbols of a writing system have a structure that imparts information. If this structure can be learned, it can be used to uncover previously unknown patterns and relationships within the data. a writing system is intended to convey meaning, and the structure of the writing system is to support the conveyance of the intended meaning. Simply changing the structure of a language can drastically change how the words are processed. Take the sentence "Data scientists not only are adept at working with data, but appreciate data itself as a first-class product." written by Hillary Mason, founder of Fast Forward Labs. A change to the structure that still conveys the same meaning is "Adept at working with data, data scientists not only are, but appreciate data itself as a first-class product." This syntax can be recognized as the format that Yoda from the Star Wars movies uses to speak. This observation is significant in that the relationships between words are important, and can influence how a passage is understood. A step further is a completely random shuffle of the same words "a with product Data adept data as at first-class appreciate itself are scientists only data working but not" In this case, the entire meaning of the passage is lost.

Entropy is a measure of disorder, and as part of this project, I will be calculating the entropy for Linear A and possibly other languages. The entropy for many different languages has been calculated, but it seems that Linear A and other ancient languages have not been done. The purpose is to give further insight into the structure of this ancient undeciphered language and further efforts to understand it.

## 2.2 Main field of study or studied phenomenon

The main fields applicable to this project are cryptography, linguistics and natural language processing.

## 2.3 Who has done something similar and how?

Entropy for many different languages has been calculated and it seems that Linear A and other ancient texts have not been done. A leading expert that has published numerous papers in this area is Francesco Perono Cacciafoco. [4], [6]

## 2.4 How will you solve it?

(I think this will be solved by developing a new algorithm that is using some NLP and image processing methods, as well as methods from information theory (entropy); and while we are not necessarily looking at deciphering ancient texts and manuscripts, we are looking at creating general methods and algorithms to work with unknown data.)

The entropy will be calculated using Shannon's entropy and applied to different combinations of symbols within the corpus. Entropy for individual characters as well as pairs and trios will be conducted. If possible, whole "word" and document entropy can be calculated.

The first steps will be to process the corpus into individual symbols and documents. Then it can be converted into a document term matrix. Once the data in in these two forms, the analysis can be conducted by performing any number of analysis on it.

## 2.5 Methodology used in this project

In this project, we will be utilizing ML and AI techniques to explore the Linear A dataset. Information gained based on the structure of the symbols, using techniques from information theory, including single symbol entropy, symbol pairs and trio entropy will be calculated. The idea is to develop methods and algorithms that can work with unknown data. Various forms of unsupervised clustering will be conducted on the data including k-means, topic modeling using Latent Dirichlet Allocation (LDA), and word networks.

### 2.5.1 Agnostic classification using topic modeling

In 2011, Barbara Montecchi classified 147 of the Linear A tablets from the Haghia Triada site [5]. In this stage of the project, we attempt to model categories using topic modeling and compare them to the results that Montecchi published. This was done in two different stages. The first was conducting topic modeling on the entire corpus of Linear A, and the second was modeled only using the Haghia Triada tablets. in each case, the documents that were classified into a topic were compared to the documents that were classified by Montecchi.

Montecchi used 15 different categories to classify the Haghia Triada documents. We used these same groupings as well as a consolidated list of categories. The consolidated categories combined classes A, B, and M these were general commodities. Next, livestock and sheep, wool and cloth classes C, D, L, and O were consolidated. Cloth and wool were grouped since they could be considered a textile group and wool and sheep are closely related. Classes E, F, G, K, N, and O were grouped since barley, olive oil, wheat, figs and wine were all consumable items. Class K was kept with the food/ consumable group since the food is probably kept in a vessel and closely related to food and consumables. Classes U and X were consolidated since they were either unreadable, or not classifiable. Class, V was kept by itself since it seemed like a catch all group of things that didn't fit the other groups.

## 2.6 Datasets in this project

The data sets that will be used will be the corpus of linear A to begin with. This contains all known inscriptions of Linear A, and is organized by the document

number and the line that it appears on the piece. [2]

## 2.7 Why is it important?

This work is important, as it can help provide insight into the nature of the Linear A language, and the nature of communication in general. AI has not been exhaustively used to decipher this particular writing system, and there is opportunity to uncover some new underlying structure that may have eluded anthropologists.

## 2.8 Validation

Validation of this project will consist of using the analysis techniques used on the unknown language on similar known language. in the case of topic modeling, Barbara Montecchi proposed a categorization for all of the tablets discovered by Haghia Triada [5]. This will be used to compare the results of the topic modeling. If the results of the topic modeling are close to the categories proposed by archaeologists and anthropologists, then the topic models can be extended to the tablets that have not been categorized by anthropologists. (yuck, rewrite this)

## 2.9 Presentation of the Project

The code will be hosted publicly hosted on git hub to allow anyone access to the methods and data used. The full write up will be hosted there as well. The project, including visualizations, analysis, justifications and reasoning will most likely be contained in an R shiny or plotly dashboard.

The dashboard will have a page for each major phase of the project, and flow from one concept to the next, taking the reader on a tour of the data, allowing them to experience our findings and understand why we came to the conclusions we did. It will tell a story and showcase he findings appropriately.

## 2.10 Symbol Mining Linear A

The linear A corpus has a few problems with it. One is that the entire known corpus only has about 10000 total symbols. Of this, approximately 20% of the symbols are unreadable. Even with these problems, an analysis of the frequencies can be conducted, and the entropy can be calculated. An initial plot of symbol frequency is done, and the distribution is compared to itself with and without the unknown symbol included. The curve significantly flattens when the unknown symbol is removed from the distribution. The next distribution to look at is the pairs of symbols and trios of symbols that occur in the data. There are 2300 unique pairs of symbols in the data, and a portion of them contain the unknown symbol, in fact, the most common symbol pair is two unknown symbols, and the most common trio is three unknown symbols.

I attempted to see if there was a method of potentially predicting what symbol could be filled into some of the unknown positions. The frequency of all symbol pairs was sorted and then recompiled after filtering out the unknown character. Then the list of known symbol pairs was compared to two different lists. the first list is the symbol pairs that had the pattern "unknown, known", and the second is the pattern "known, unknown". The list of known trios was compared to the list that had unknown symbols in it as well. The entropy's for each of these symbol combinations were calculated and other information can be visualized in the shiny app here

The next step consisted of creating a table of known pairs from the most frequent unknown/known pair. Snippets of these tables can be seen in [figure and figure] It is interesting to note how much the entropy changes with the different pairings of characters. The entropy of all of the pairs of symbols both with and without the unknown symbol is very high, above 10. While the entropy of the "unknown, known" pair is lower right around 6, and then the "known, unknown" pair is much lower around 4.5.

### 2.11 Pivot

The main purpose of the project was to use symbolic data to uncover structure and process non-textual data using NLP techniques. The initial results provided a framework that appear to be promising for use in the ultimate point of the project which is to analyze luminosity data of stars supporting a team for NASA's Frontier Development Lab (FDL). The processes uncovered during the analysis of Linear A were presented to the mentor of the NASA FDL team. The response was very positive, and wanted to see more results using their data. The results of this new project will be available in another repository and write up.

## References


[1] Hauer, B., and Kondrak, G. Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computational Linguistics 4* (2016), 75–86.

[2] Hogan, R. Linear a explorer. `https://github.com/mwenge/lineara.xyz`, 2019.

[3] Luo, J., Cao, Y., and Barzilay, R. Neural decipherment via minimum-cost flow: from ugaritic to linear b. *arXiv preprint arXiv:1906.06718* (2019).

[4] Min Eu, N. C., Xu, D. D., and Cacciafoco, F. P. Coding to decipher linear a. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)* (2019), pp. 1–6.

[5] Montecchi, B. A classification proposal of linear a tablets from haghia triada in classes and series. *Kadmos 49*, 1 (2011), 11 – 38.

[6] Petrolito, T., Petrolito, R., Perono Cacciafoco, F., and Winterstein, G. Minoan linguistic resources: The linear a digital corpus. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 95–104.

[7] Younger, J. Table of contents, 2000.