SCHOOL OF
ELECTRICAL AND
ELECTRONIC ENGINEERING

THE UNIVERSITY
of ADELAIDE

Statistical Analysis of Unknown Written Language: The Voynich Manuscript

Project Group 31

Andrew McInnes (a1211832)

Lifei Wang (a1614410)

ELEC ENG 4068 A/B HONOURS PROJECT

B.E. in Avionics & Electronic Systems Engineering
B.E. in Computer Systems Engineering
B.E. in Electrical and Electronic Engineering
B.E. in Electrical & Sustainable Energy
B.E. in Telecommunications Engineering

Date submitted:	17/04/2015
Supervisors:	Prof. Derek Abbot Maryam Ebrahimpour Dr. Brian Ng.

Abstract

The Voynich Manuscript is a document written in an unknown language or cipher. This research proposal presents an idea into determining possible relationships within the Voynich. This is to be performed through known statistical methods relating to linguistics. The document reviews previous research carried out by other researchers. The proposed method is given and shows the current results obtained by the project team at this time. The project management is briefly outlined.

Contents

1.0 Introduction

1.1 Background

The Voynich Manuscript is a document written in an unknown script that has been carbon dated back to the early 15th century [1] and believed to be created within Europe [2]. Named after Wilfrid Voynich, whom purchased the folio in 1912, the manuscript has become a well-known mystery within linguistics and cryptology. It is divided into several different section based on the nature of the drawings [3]. These sections are:

- Herbal
- Astronomical
- Biological
- Cosmological
- Pharmaceutical
- Recipes

The folio numbers and examples of each section are outlined in appendix *section A.2*.

In general, the Voynich Manuscript has fallen into three particular hypotheses [4]. These are as follows:

- Cipher Text: The text is encrypted.
- Plain Text: The text is in a plain, natural language that is currently unidentified.
- Hoax: The text has no meaningful information.

Note that the manuscript may fall into more than one of these hypotheses [4]. It may be that the manuscript is written through steganography, the concealing of the true meaning within the possibly meaningless text.

1.2 Aim

The aim of the project is to determine possible features and relationships of the Voynich Manuscript using statistical methods that can be used to aid in the investigation of unknown languages and linguistics. It is not to fully decode or understand the Voynich Manuscript itself. This outcome would be beyond excellent but is unreasonable to expect in a single year project.

1.3 Motivation

The research project, that is to be carried out, shall attempt to find relationships and patterns within unknown text through the usage of known statistical methods on languages and linguistics. The Voynich Manuscript is a prime candidate for this as there is no known accepted translations of any part within the document. The relationships found can be used to verify the statistical methods and also be used to conclude on specific features of the unknown language(s)[1] within the Voynich Manuscript.

Knowledge produced from the relationships and patterns of languages and linguistics can be used to further the current linguistic computation and encryption/decryption technologies of today [5].

1.4 Significance

There are many computational linguistic and encryption/decryption technologies that are in use today. As mentioned in *section 1.3*, knowledge produced from this research can help advance these technologies in a range of different applications [5]. These include, but are not limited to, information retrieval systems, search engines, machine translators, automatic summarizers, and social networks [5].

Particular technologies, that are widely used today, that can benefit from the research, include:

- Turn-It-In (Authorship/Plagiarism Detection)
- Google (Search Engines)
- Google Translate (Machine Runnable Language Translations)

1.5 Technical Background

The vast majority of the project relies on a technique known as data mining. Data mining is the process of taking and analysing a large data set in order to uncover particular patterns and correlations within said data thus creating useful knowledge [6]. In terms of the project, data shall be acquired from the Interlinear Archive, a digital archive of transcriptions from the Voynich Manuscript, and other sources of digital texts in known languages. Data mined from the Interlinear Archive will be tested and analysed for specific linguistic properties using varying statistical methods.

The Interlinear Archive, as mentioned, will be the main source of data in regards to the Voynich Manuscript. It has been compiled to be a machine readable version of the Voynich Manuscript

[1] Currier (1976) suggests that there may be multiple ‘languages’ within the Voynich Manuscript

based on transcriptions from various transcribers. Each transcription has been translated into the European Voynich Alphabet (EVA). An example of the archive in EVA and the corresponding text within the Voynich Manuscript can be seen within the appendix *section A.3.* The EVA itself can be seen within appendix *section A.4*.

1.6 Technical Challenges

Due to the difficulty of transcribing a hand-written 15th century document, no transcriptions within the Archive are completed, nor do they all agree with each other. Many tokens within the Voynich Manuscript have been considered as a different token, or even multiple tokens. Spacing between word tokens has also been a key ambiguity as one transcription may consider one word token to be multiple word tokens or vice-versa. It is also believed that the manuscript is missing 14 pages [7]. These uncertainties will make it difficult to effectively conclude on any linguistic analyses.

The statistical methods relating to linguistics are numerous, leading to many different possible approaches that can be used upon the Voynich Manuscript. However many of the intricate techniques require some form of knowledge of the language itself. This limits the possible linguistic analysis techniques that can be used. Despite previous research on the Voynich Manuscript, no current conclusion has been widely accepted [3]. Due to this the research will be focussed on the basics of linguistics.

1.7 Knowledge Gaps

The project requires a large amount of software code using various statistical techniques. No project team members are particularly knowledgeable in these areas. As such, all members within the project team shall be developing skills in software programming and knowledge within these statistical techniques as the project develops.

From a broader view, knowledge from statistical methods used on the Voynich Manuscript is plentiful but, so far, none have shown any conclusive, widely-accepted understanding of the text [3]. Throughout the project life, the team hopes to show possible relationships within the Voynich Manuscript through the investigation of different linguistic properties.

2.0 Requirements

It is not expected that the project fully decodes, or even partially decodes, the Voynich Manuscript. Nonetheless the project must show the following:

- A logical approach to investigating the Voynich Manuscript

- Critical evaluation of any and all results
- Testing on all code
- Hypotheses based on results

3.0 Literature Review

Over the years, the Voynich Manuscript has been investigated by numerous scholars and professionals. This has given rise to many possible hypotheses [4] through many different forms of analysis based on its linguistic properties [2]. These properties range from the character tokens to word tokens, to the syntax and pages themselves. The currently reviewed literature, which is of interest to the project, is summarized below.

A broad, albeit brief, summary of linguistic analyses that have been completed over the previous years is given by Reddy and Knight [2] and include some of their own tests. They perform multiple analyses on the letter, the word, syntax, pages, and the manuscript itself while giving reference to other works on the same property. Their work on the letter and word are of a particular interest of this project. They suggest that vowels may not be represented within the Voynich Manuscript and that Abjad languages have the closest similarities [2]. This is concluded through two-state hidden Markov models and word length distributions respectively. Reddy and Knight also suggest that there are some particular structural similarities within the words when using a minimum description length based algorithm [2].

Gabriel Landini's [3] paper "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis" looks into the statistical characteristics of the manuscript and natural languages. Characterising the text through Zipf's Law and performing analysis on entropy and character token correlation, Landini suggests that there is some form of linguistic structure behind the Voynich Manuscript [3]. In particular, the work reveals long range correlation, a modal token length, and periodic structures within the text.

Andreas Schinner [4] takes a different approach in the paper "The Voynich Manuscript: Evidence of the Hoax Hypothesis". Schinner performs a random walk model and tests token repetition distances through the Levenshtein distance metric. It is concluded that while the results seem to support the hoax hypothesis more so than the others, it cannot rule out any of them [4].

Diego R. Amancio, Eduardo G. Altmann, Diego Rybski, Osvaldo N. Oliveira Jr., and Luciano da F. Costa [5] investigate the statistical properties of unknown texts. They apply various techniques to the Voynich Manuscript looking at vocabulary size, distinct word frequency, selectivity of words, network characterization, and intermittency of words. Their techniques were aimed at determining useful statistic properties with no prior knowledge of the meaning of the text. Although not aimed specifically at deciphering the Voynich Manuscript, they do conclude that the Voynich Manuscript is compatible with natural languages [5].

Jorge Stolfi's [8] website "Voynich Manuscript stuff" gave multiple views and analyses of the Voynich Manuscript. Stolfi's work on word length distributions and morphological structure [8] are of particular interest to the project. He displays a remarkable similarity in word length distributions between the Voynich Manuscript and Eastern Asian languages [8]. He also shows evidence of morphological structure, displaying prefix-midfix-suffix structure [9], and later displaying a crust-mantle-core paradigm [10].

In regards to research on the Voynich Manuscript carried out at the University of Adelaide. This is the second year that this project has been undertaken by students. Bryce Shi and Peter Roush provide a report on their results [11]. They carry out a multitude of tests on the Voynich Manuscript including:

- Zipf's Law
- Word Length Distribution
- Word and Picture Association
- Word Recurrence Intervals
- Entropy
- N-Grams
- Punctuation
- Authorship

Shi and Roush give short conclusions to each of these tests but realise that further research is required for any to be considered conclusive [11].

4.0 Proposed Method

The proposed approach to the project has been broken down into multiple phases and is briefly shown in *Figure 1*.

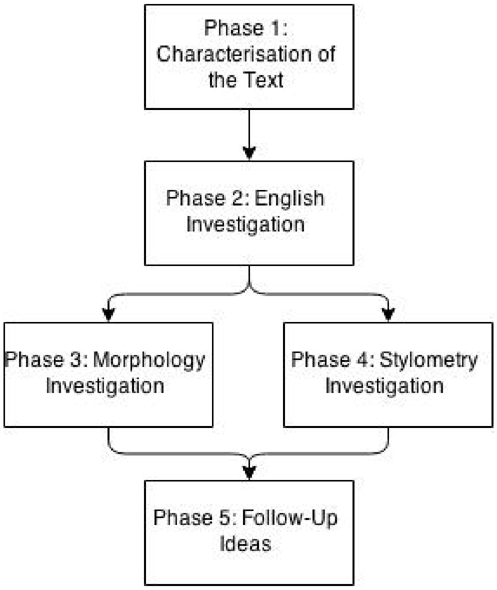

Phase 1: Characterisation of the Text
Phase 2: English Investigation
Phase 3: Morphology Investigation
Phase 4: Stylometry Investigation
Phase 5: Follow-Up Ideas

Figure 1: Basic Work Breakdown Overview

Each phase will be considering a specific feature of the Voynich Manuscript and linguistics while building onto what was learned in the previous phase(s). Many techniques may replicate previous research outlined in section 3.0. The results within these documents will be used to compare and complement results obtained throughout the life of the project.

All phases will be coded and will therefore include testing as all code must be verified for results to be considered accurate. Code will also be well commented and documented within the project wiki.

Code shall be written in C++ and MATLAB languages as the project members have experience using these programming languages. MATLAB, in particular, is chosen as it provides a simple, easy to use mathematical toolbox that is readily available on the University systems. Other programming languages may be used if it is found to be more suitable.

Completion of each phase is considered a milestone, as referenced in section 6.1.

4.1 Phase 1 – Characterization of the Text

Characterization of the text involves determining the first-order statistics of the Voynich Manuscript. This first involves pre-processing the Interlinear Archive into a simpler machine-readable format.

The pre-processed files are then characterized through MATLAB code by finding and determining:

- Unique word tokens
- Unique character tokens
- Frequency of word tokens
- Frequency of character tokens
- Word token length frequency
- Character tokens that only appear at the start, end, or middle of word tokens

Resulting statistics can then be compared with other known languages through using the same code on the various translations of the Universal Declaration of Human Rights.

Unfortunately the Universal Declaration of Human Rights is, by comparison, a small document which will limit results. However it will give a basis for follow-up research into the languages that have a possible relationship to the Voynich Manuscript based on the first-order statistics.

Further research can be carried out using any languages that appear to have a relationship to the manuscript through the compilation of much a larger corpus.

4.2 Phase 2 – English Investigation

The English investigation looks into the elementary structure of English text. It specifically examines the representation of the English alphabet and how the alphabetical tokens can be extracted from an English text using statistics.

Initially, a corpus of English texts shall be passed through the characterisation code of phase 1 to determine the first-order statistics of each text. These will be compared to grasp a basic understanding of how each of the tokens can be statistically represented and how these statistics differ between texts. These tokens include alphabetical, numerical, and punctuation tokens.

The characterization code will then be expanded upon to include character token bigrams to further define the differences between character tokens. Bigrams give the conditional probability, P, of a token, T_n, given the proceeding token, T_{n-1}. This is given in the following formula:

$$P(T_n|T_{n-1}) = \frac{P(T_{n-1}, T_n)}{P(T_{n-1})}$$

It is expected that the probability of the different tokens along with the first-order statistics, obtained through the phase 1 code, will show definitive differences between alphabetical, numerical, and punctuation tokens.

Code will be written that takes these statistical findings into account to attempt to extract the English alphabet from any given English text with no prior knowledge of English itself. This will be used to examine the Voynich Manuscript to search for any character token relationships.

4.3 Phase 3 – Morphology Investigation

Morphology deals with the structure of the words. Specifically, phase 3 will be looking into the possibility of affixes within the Voynich Manuscript.

As described in section 2, previous research has found the possibility of morphological structure within the Voynich Manuscript [2]. A Minimum Description Length model[2] [12] will be used to attempt to segment word tokens into possible affix models.

The basis of the code will be examining word tokens within the Interlinear Archive and attempting to find all similar tokens. Following the Minimum Description Length model, the code will then attempt to find the most compact representation of the word token and any pre or post word tokens.

Coding this model into MATLAB will allow for use on the Interlinear Archive. The code will also be used on English texts to provide a qualitative comparison on the effectiveness and limitations of the algorithm.

4.4 Phase 4 – Stylometry Investigation

Stylometry examines the linguistic style of written language. Phase 4 will be investigating the authorship of the Voynich Manuscript through stylometry.

[2] Minimum Description Length attempts to find the most compact representation of the data [12]

As with phase 2, phase 4 will involve the creation of a corpus of texts from known authors of the 15th century and any authors that are suspected of writing the manuscript. Statistics will be lifted from this corpus and analysed against those of the manuscript.

Due to the unknown nature of the Voynich Manuscript, the stylometry investigation will, again, be focussing on the basic statistics of the script. These include:

- Average word lengths
- Distribution of word lengths
- Lexical Richness

However, it should be noted that the manuscript may have been written by multiple authors and in multiple languages[3] [13]. This means that any stylometry results from the manuscript as a whole may be corrupted if this is true. Sections of the manuscript will need to be investigated separately, particularly those written in different languages, along with the manuscript as a whole.

4.5 Phase 5 – Other Ideas

This phase will essentially be determining follow-up investigations based on current findings from the other phases. The empirical data found may lead to possible investigations that can be followed up during this phase. It is also quite possible that a phase, particularly phases 3 and 4, may not provide a definitive conclusion or may lead to an impasse. Due to this, phase 5 has been left much more open than the other phases.

Some other particular investigations that may be completed during this phase include:

- Keywords and co-occurrence within the manuscript [14]
- Vowel and consonant representation [2]
- Word order [2]
- Hidden Markov Modelling [11]
- 15th Century Cipher Analysis [11]

It is expected that this phase will eventually be split up into multiple separate phases. At this time it is unknown as to which follow-up investigations will be completed and, as such, has been left for discussion at a later date as previous phases become completed.

[3] The usage of the word 'language' here denotes a statistical difference between two sets of text [13]. 'Language' is used again here for consistency between cited documentation.

5.0 Preliminary Results

The project team has begun research by pre-processing the Interlinear Archive into separate simple files containing the transcriptions of each unique transcriber. All unnecessary data, such as comments, were removed from each of these transcriptions. In-line formatting was also converted to follow a simpler, machine-readable standard (see appendix *section A.5* for an example).

To get the most accurate results the team must look into which transcriptions are the most complete. Shi and Roush (2014) suggest that the Takahashi transcription was the most complete transcription by checking the total number of lines transcribed [11]. A test on the amount of transcribed lines per transcriber is performed again giving the results within *figure 2* (see appendix *section A.6* for a complete list of transcriber codes).

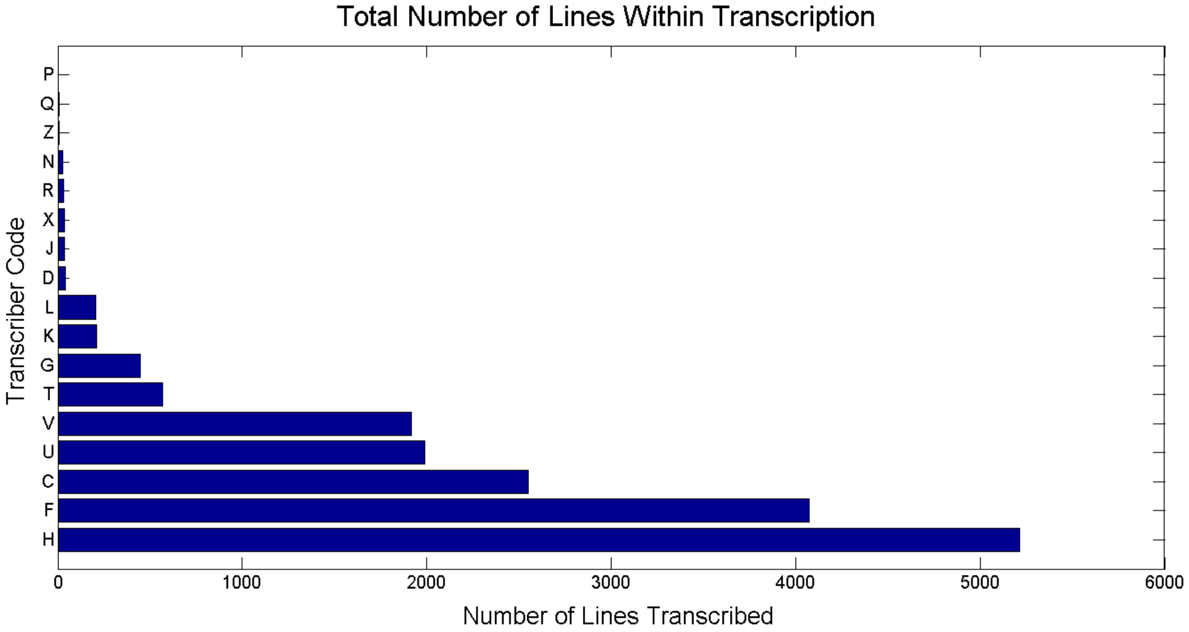


Figure 2: Total Number of Lines Transcribed by Transcriber

This follows the same conclusion of Shi and Roush (2014).

A comparison of the top five most completed transcriptions word-length distribution was then carried out. Takahashi’s transcription showed an unusual peculiarity with a single word token of length 35 with the next highest being of length 15. However, this word token[4] was composed of mainly unknown ‘*’ characters and was therefore removed from our data set. This resulted in the following word-length distribution plot in *figure 3*.

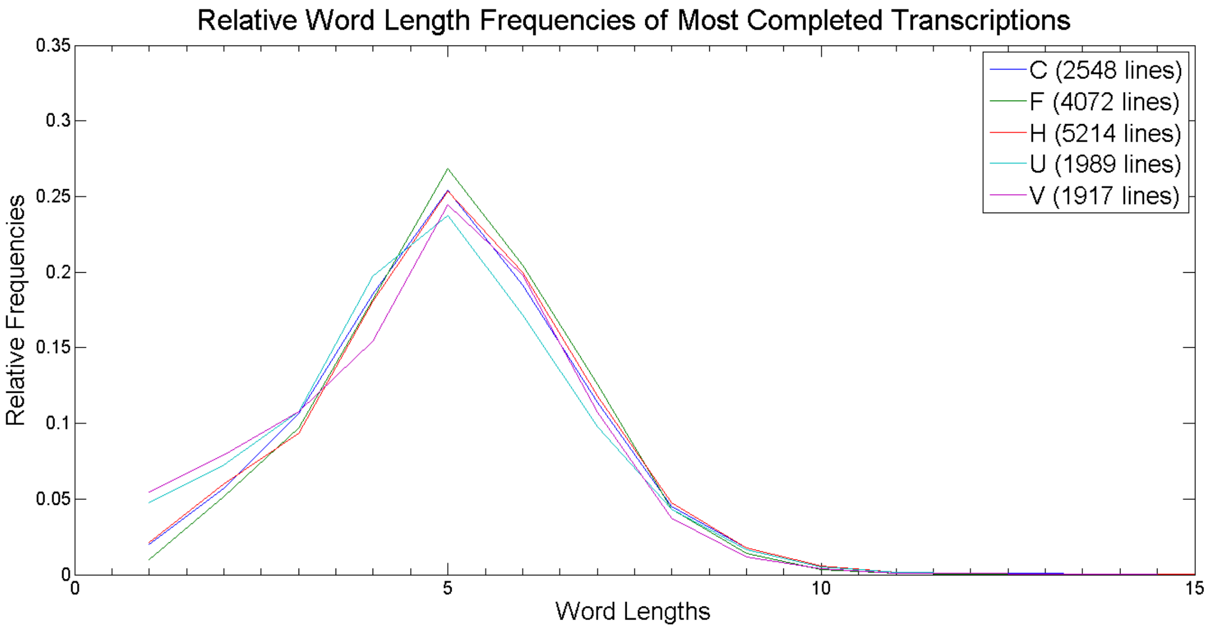


Figure 3: Relative Word Length Frequencies of Most Completed Transcriptions

This result, again, conforms to the results found by Shi and Roush (2014), showing a peak word length distribution of 5 and giving a binomial distribution. This can also be seen in Reddy and Knight (2011). However Reddy and Knight specifically investigated the word lengths of language B[5] within the Voynich Manuscript.

The Universal Declaration of Human Rights was also mined for relative word length distributions. This is, unfortunately, limited to a much smaller amount of tokens than that of the Voynich Manuscript but shall give a good indication as to which languages to investigate further.

[4] Actual word token ‘************ooooooooolar*****s**r**’

[5] As discussed in section 4, phase 4. ‘Languages’ here refers to a distinct statistic difference within the text. Currier (1976) describes two ‘languages’ within the Voynich Manuscript as language A and language B.

As it is believed that the Voynich originated from Europe [2], European languages were initially compared with the results found above. Using the Takahashi transcription, as it is the most complete, resulted in the following word-length distribution plot in *figure 4.*

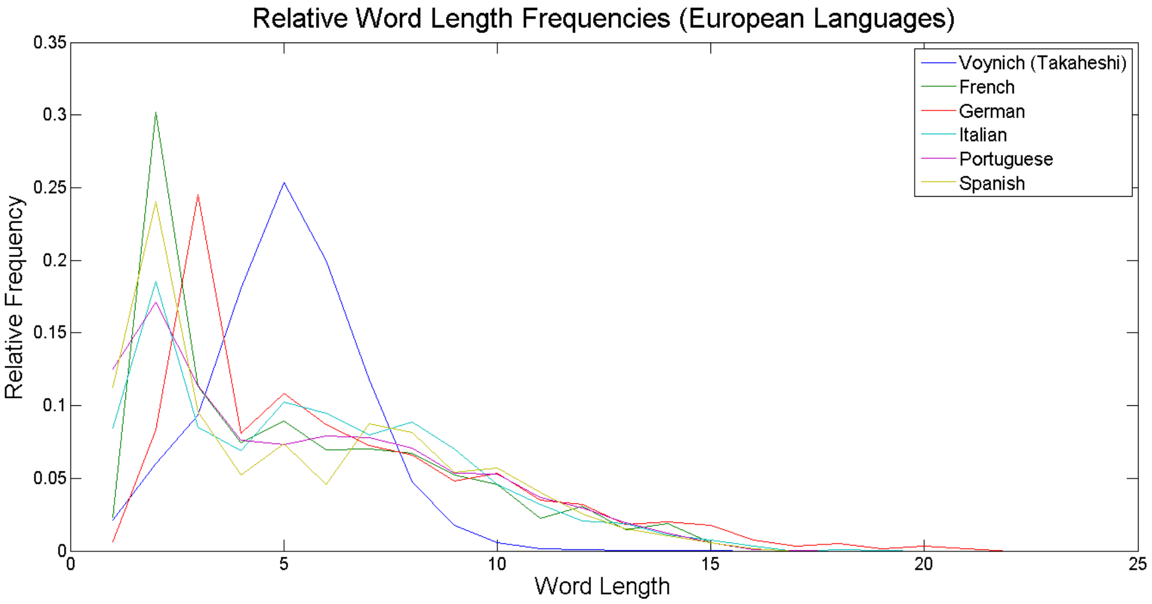


Figure 4: Relative Word Length Frequencies of European Languages

Many European languages were removed from the plot to make it more readable. Regardless, the resulting conclusion was the same as no tested European language appeared to fit the peak word length and binomial distribution of the Voynich Manuscript. Shi and Roush [11] found similar results, but also showed that the language within the manuscript had a closer resemblance to Hebrew. Reddy and Knight [2] tested Buckwalter Arabic, Pinyin, and 'de-voweled' English, resulting in much closer relationships. All gave the appearance of a binomial distribution much like the manuscript, with Buckwalter Arabic being very similar to Voynich Language B. This leads to the hypothesis that the manuscript may be in the form of Abjad[6] [2].

[6] Abjad is a writing system that essentially leaves out vowels and only uses consonants.

Looking specifically at the Takahashi transcription, the following first-order statistics were found (as shown in *Table 1*).

Total Word Tokens	37919
Total Unique Word Tokens	8151
Total Character Tokens	191825
Total Unique Character Tokens	23
Longest Word Token	15

Table 1: First-Order Statistics (Takahashi)

All word and character tokens of each transcription have been recorded along with the frequency that each occur. Note that the character tokens are currently limited to the basic EVA characters (see appendix *section A.4)* but is currently being expanded to include the extended EVA characters. The code itself is also being extended to determine all character tokens that only appear at the start or end of a word token and also replicated for the characterisation of English.

6.0 Project Management

6.1 Timeline

As shown in section 4, the project has currently been split up into 5 distinct phases with the expectation of the 5th phase to be split off into much smaller phases. Each of these are to be worked on and completed within a given time to keep the project on schedule. The current project schedule has been graphically organized and displayed on a Gantt chart viewable in the appendix *section A.1*.

6.2 Deliverables

The deliverables of the project are summarized below in *table 2*, detailing the deliverable and the respective deadline[7]. The deliverable work schedule can also be viewed within the Gantt chart of the appendix *section A.1*.

Deliverable	Deadline
Proposal Seminar	31st of March, 2015

[7] Some exact deadlines are currently not known, these have been approximated with week and semester numbers.

Research Proposal Draft	17th of April, 2015
Research Proposal	Week 12, Semester 1
Progress Report	Week 12, Semester 1
Final Seminar	Week 10, Semester 2
Thesis	Week 11, Semester 2
Expo Poster	Week 11, Semester 2
Expo Presentation	Week 12, Semester 2
YouTube Video	Week 12, Semester 2
USB Flash Drive of all Code and Work	Week 12, Semester 2

Table 2: Deliverables

6.3 Task Allocation

Tasks have been allocated to the project team members through the phases in section 4. Collaboration between members will occur during phases 1, 2, and 5. However, it is expected that there will be a considerable amount of collaboration throughout all phases. The current allocations are summarized in *table 3* below.

Task	Phase	Allocated Member
Pre-processing of Interlinear Archive	1	Andrew McInnes
Writing and Testing Voynich First-Order Statistics Code	1	Andrew McInnes
Writing and Testing Character Token Code	1	Lifei Wang
Expanding First-Order Statistics Code	2	Andrew McInnes
Expanding Character Token Code	2	Lifei Wang
Writing and Test English Alphabet Extraction Code	2	Andrew McInnes
Writing and Testing Morphology Code	3	Andrew McInnes
Writing and Testing Stylometry Code	4	Lifei Wang
Discussing and Determining Follow-Up Investigations	5	Andrew McInnes, Lifei Wang

Table 3: Task Allocation

6.4 Management Strategy

The project team will be managed through a minimum of one internal meeting between members outside of scheduled workshop and project time, and a minimum of one fortnightly meeting with supervisors. Each meeting will involve:

- Current phase progress
- Issue(s) encountered
- Display of any relevant result(s) or research finding(s)

Feedback can then be gathered through both team members and supervisors.

All working copies of code and documents shall also be kept on a group Google Drive[8]. These will be updated as necessary and are available for all necessary members.

6.5 Budget

The project team has been assigned a budget of $500. However the project is heavily computer-based where all currently required programs are freely available on the University systems. Therefore it is expected that none of the budget will need to be used.

It is possible that works may be found that are unavailable within the University. Should it be found that these works would further the research then the budget may be used on these works. This will be discussed with supervisors.

6.6 Risk Analysis

Multiple risks have been identified by the project team and ranked according to the likelihood of occurring, and the consequence of an occurrence. The likelihood and consequence was given a number ranking as denoted by the brackets '[ ]'. The main risks are summarised within *table 4* below.

No.	Risk	Likelihood	Consequence	Risk Level
1	Underestimation and/or mismanagement of time and resources	High [8]	High [7]	56
2	Health related issues from long periods at computers	High [7]	Moderate [6]	42

[8] The Google Drive is available at https://drive.google.com/a/student.adelaide.edu.au/folderview?id=0B3xk_r8iaE_IYURhTEhLd1dyeVk&usp=sharing

3	Team member illness or injury	Very High [9]	Moderate [4]	36
4	Issues with communication between team and/or supervisors	Low [3]	High [7]	21
5	Loss of software code	Low [2]	Very High [10]	20

Table 4: Risk Analysis (Summary)

The risk level was calculated by multiplying the likelihood rank and the consequence rank. This risk level corresponds to the overall risk that is posed to the project.

Short descriptions along with mitigation and continuity plans for each risk are detailed below.

6.6.1 Underestimation and/or mismanagement of time and resources

As the team members undertaking the project have no experience with such a large, software focused project, the likelihood of underestimation or mismanagement of time and/or resources is high.

Mitigation of this risk shall be through continual meeting within the project team and with supervisors. A minimum of a weekly internal meeting within the team and a fortnightly progress meeting with supervisors shall occur. Phase 5 of the project has also been left deliberately long for this purpose.

Should the risk still occur, the project schedule shall be discussed and reworked to allow for successful completion within the project time frame.

6.6.2 Health related issues from long periods at computers

Due to project being mostly computer-based, team members will be in front of computers for large quantities of time.

To mitigate any possible issues due to long periods in front of computers, team members will take periodic breaks from the computer during their work.

6.6.3 Team member illness or injury

The project shall be occurring over two semesters. There is a high likelihood that one, if not both, team members may fall ill or become injured within that time.

Should any health related issues arise, the member shall inform the rest of the project team and supervisors. Depending on the illness or injury, the other member may take over work from the ill or injured member. As the majority of work is computer-based it is expected that team members will be able to work to some extent during most illnesses or injuries.

6.6.4 Issues with communication between team and/or supervisors

Team members and supervisors are very busy throughout the project lifetime. As the main form of communication shall be through emails it is possible, although not likely, that miscommunication of information may occur.

Communication issues shall be mitigated through thoughtful, concise messages. Emails should not contain any ambiguities where possible and any questions that may be raised should be asked immediately.

6.6.5 Loss of software code

As explained multiple times, the majority of the project is software based. It is possible that through some errors or malicious intent that software code(s) may be lost. While unlikely, the consequences of this occurring are severe.

All code will therefore be kept in multiple backups. This includes the use of external flash drives, the University system, and Google Drive.

7.0 Conclusion

Due to issues with the basic characterisation code and the extended EVA characters, the team is slightly behind schedule. As discussed within section 6.6, estimation of time was the most probable risk. The team has split the members such that progress has begun on the second phase while the first phase is being rectified. It is expected that this will not affect the overall schedule of the following phases.

Despite this, some interesting basic results have been found that conform to both previous research carried out at the University of Adelaide by Shi and Roush [11] and by other researchers.

8.0 Citations and References

[1] D. Stolte, “Experts determine age of book 'nobody can read',” 10 February 2011. [Online]. Available: http://phys.org/news/2011-02-experts-age.html. [Accessed 12 March 2015].

[2] S. Reddy and K. Knight, “What We Know About The Voynich Manuscript,” *LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 78-86, 2011.

[3] G. Landini, “Evidence Of Linguistic Structure In The Voynich Manuscript Using Spectral Analysis,” *Cryptologia*, pp. 275-295, 2001.

[4] A. Schinner, “The Voynich Manuscript: Evidence of the Hoax Hypothesis,” *Cryptologia*, pp. 95-107, 2007.

[5] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr. and L. d. F. Costa, “Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript,” *PLoS ONE 8(7)*, vol. 8, no. 7, pp. 1-10, 2013.

[6] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro and W. Wang, “Data Mining Curriculum: A Proposal (Version 1.0),” 12 April 2015. [Online]. Available: http://www.kdd.org/curriculum/index.html.

[7] R. Zandbergen, “Description of the Manuscript,” 25 March 2015. [Online]. Available: http://voynich.nu/descr.html.

[8] J. Stolfi, “Voynich Manuscript stuff,” 23 May 2005. [Online]. Available: http://www.ic.unicamp.br/~stolfi/EXPORT/projects/voynich/Welcome.html.

[9] J. Stolfi, “A prefix-midfix-suffix decomposition of Voynichese words,” 10 12 1997. [Online]. Available: http://www.ic.unicamp.br/~stolfi/voynich/97-11-12-pms/.

[10] J. Stolfi, “A Grammar for Voynichese Words,” 14 June 2000. [Online]. Available: http://www.ic.unicamp.br/~stolfi/EXPORT/projects/voynich/00-06-07-word-grammar/.

[11] B. Shi and P. Roush, "Semester B Final Report 2014 - Cracking the Voynich code," University of Adelaide, Adelaide, 2014.

[12] J. Goldsmith, "Unsupervised Learning of the Morphology of a Natural Language," *Computational Linguistics*, pp. 153-198, 2001.

[13] P. Currier, "New Research on the Voynich Manuscript: Proceedings of a Seminar," 30 November 1976. [Online]. Available: http://www.voynich.nu/extra/curr_main.html.

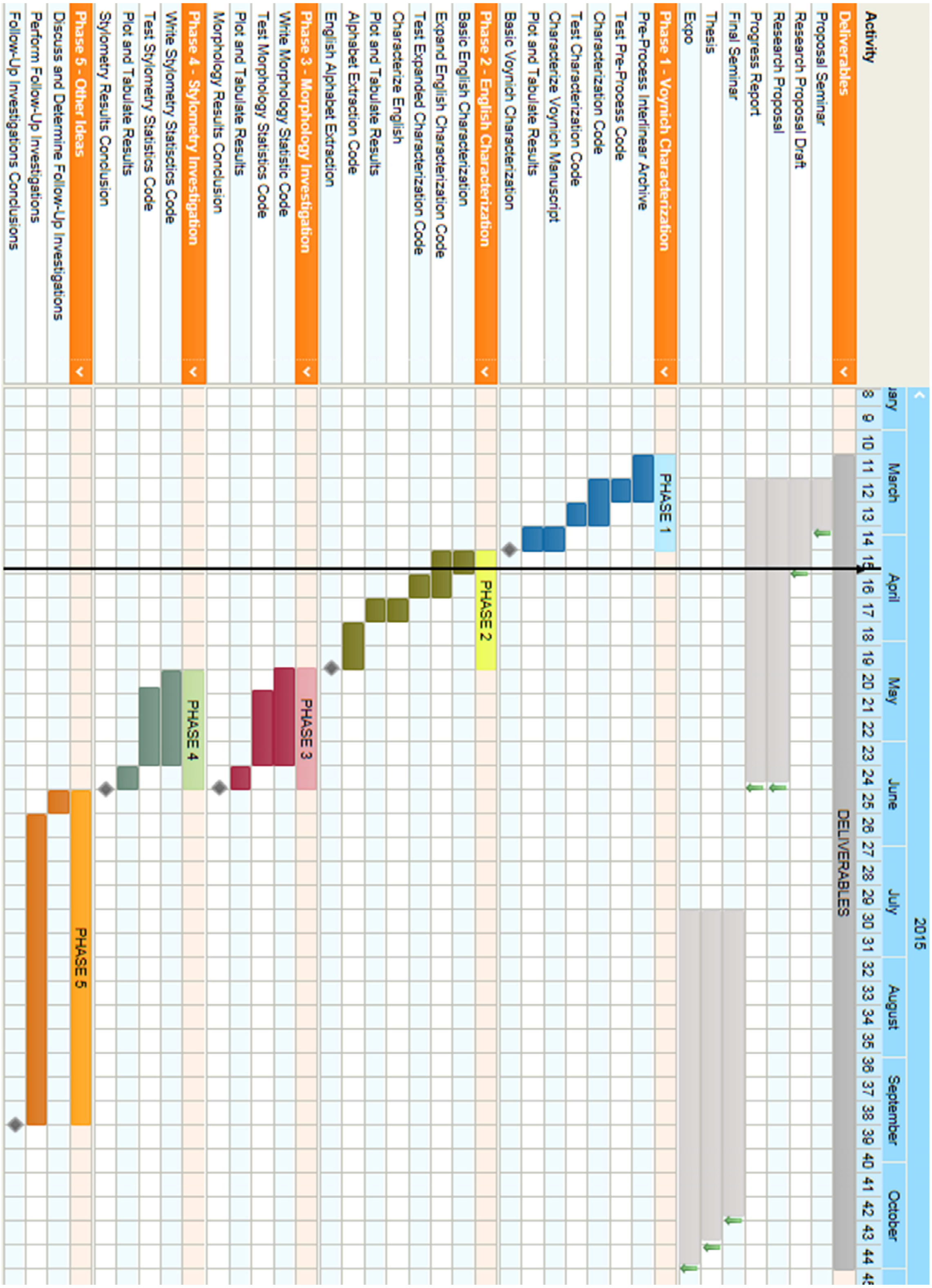
[14] M. A. Montemurro and D. H. Zanette, "Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis," *PLoS ONE,* vol. 8, no. 6, pp. 1-9, 2013.

[15] "The Voynich Manuscript," 22 March 2015. [Online]. Available: https://archive.org/details/TheVoynichManuscript.

[16] R. Zandbergen, "Analysis of the text," 13 April 2015. [Online]. Available: http://www.voynich.nu/analysis.html.

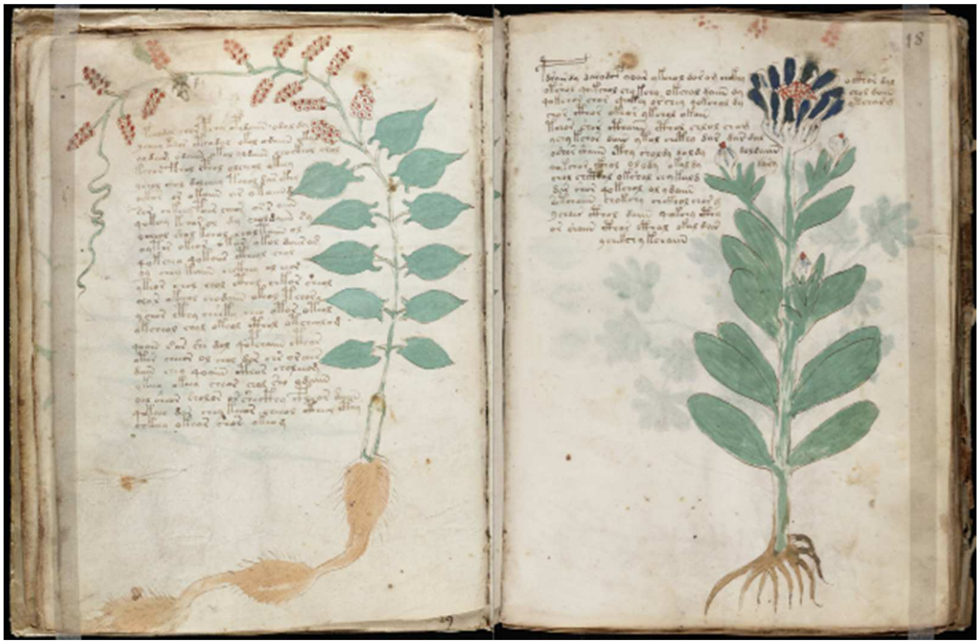
A. Appendix

A.1 Project Gantt Chart

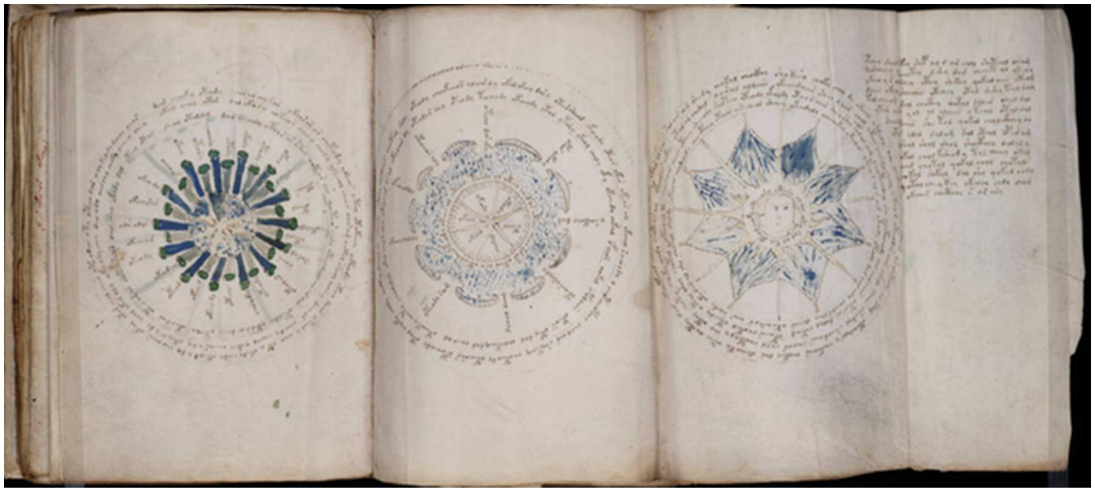


A.2 The Voynich Manuscript

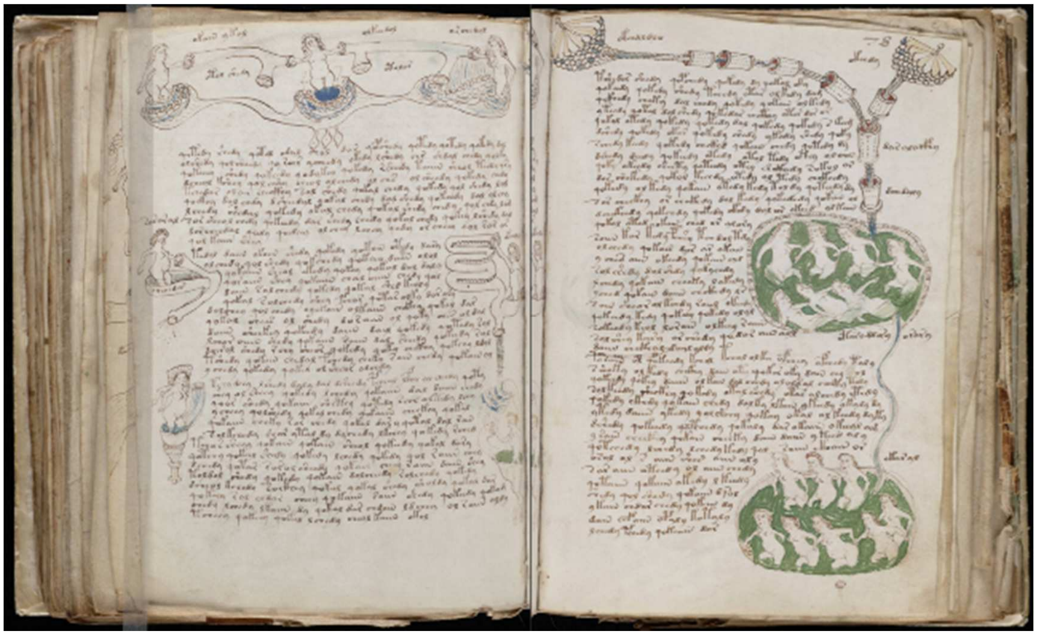
The following images are of the Voynich Manuscript. These images have been reproduced from the Internet Archive [15]. Note that ‘v’ denotes verso, and ‘r’ denotes recto.



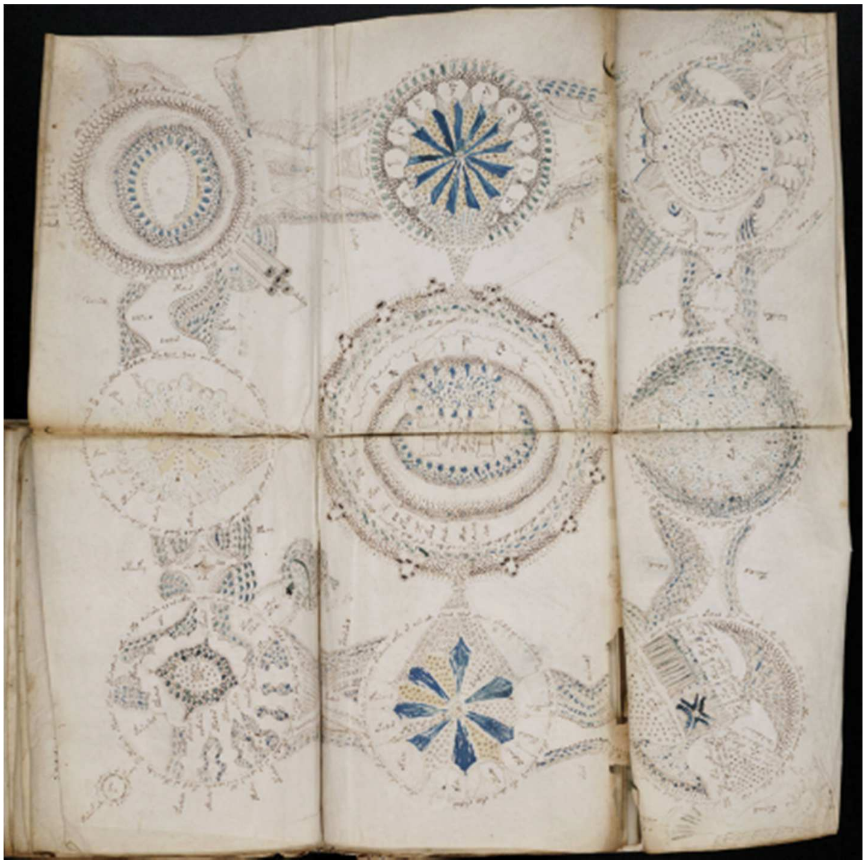
The herbal section, folios 1r – 66v.



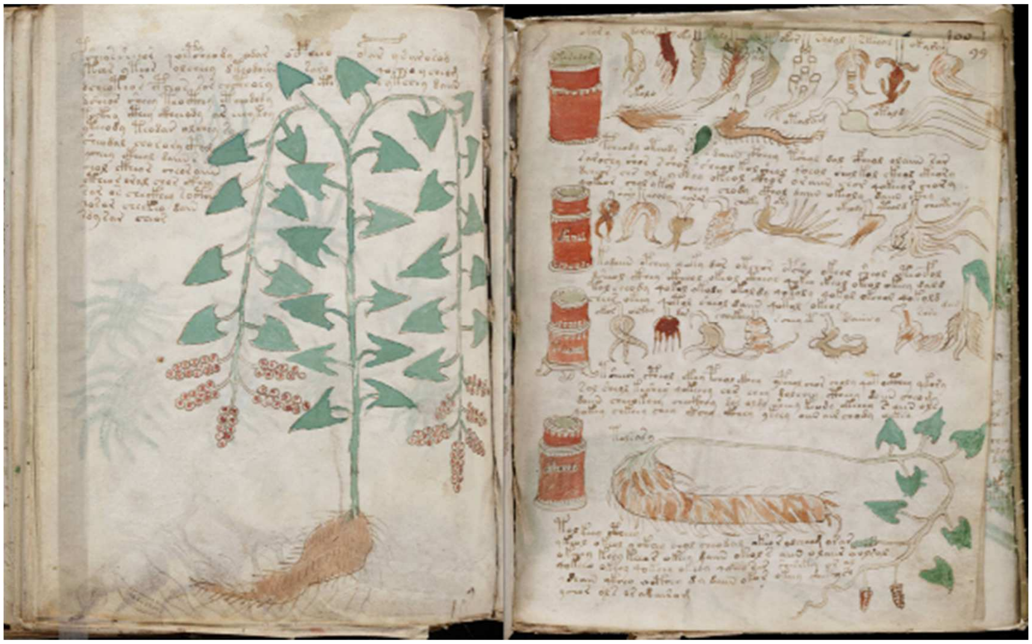
The astronomical section, folios 67r – 73v.



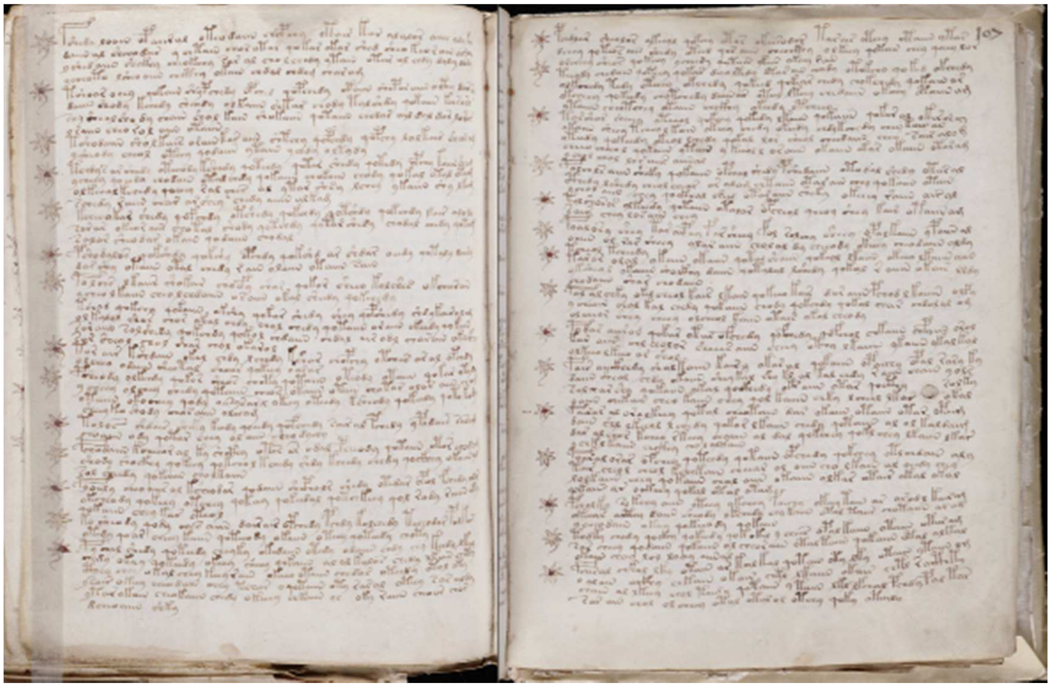
The biological section, folios 75r - 84v.



The cosmological section, folios 85r – 86v



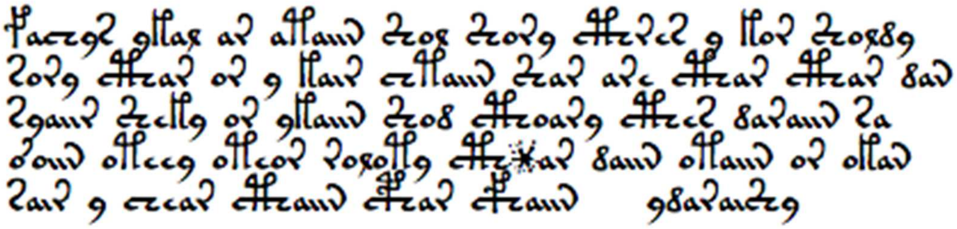
The pharmaceutical section, folios 87r – 102v.

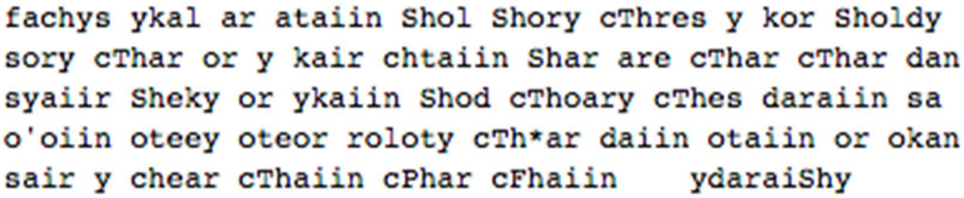


The recipes section, folios 103r – 116v.

A.3 The Interlinear Archive Example

The following images are an example of the text found within the Voynich Manuscript and its corresponding translation into the machine-readable EVA. These images have been reproduced from René Zandbergen’s website [16].

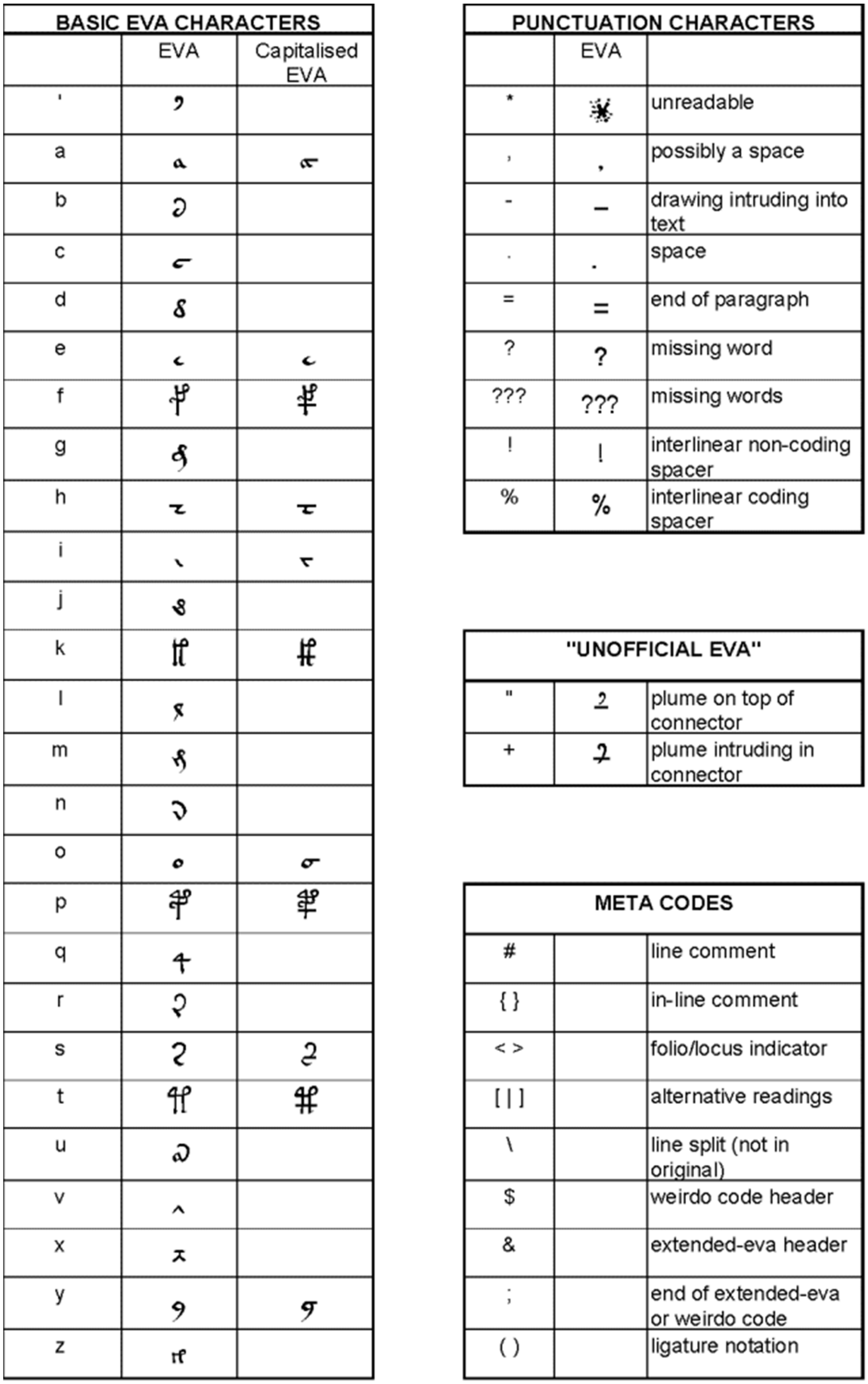




```
fachys ykal ar ataiin Shol Shory cThres y kor Sholdy
sory cThar or y kair chtaiin Shar are cThar cThar dan
syaiir Sheky or ykaiin Shod cThoary cThes daraiin sa
o'oiin oteey oteor roloty cTh*ar daiin otaiin or okan
sair y chear cThaiin cPhar cFhaiin    ydaraiShy
```

A.4 The European Voynich Alphabet

The EVA as shown on René Zandbergen’s website [16].

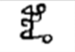
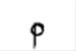
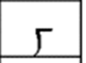
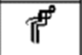
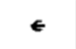
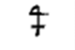
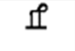
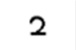
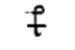
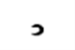
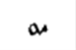
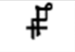
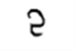
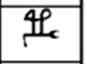
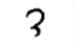
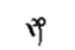
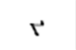
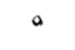
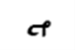
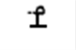
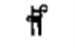
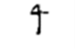
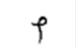
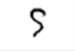
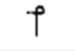
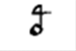
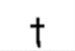
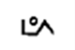
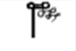
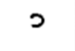
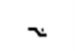
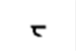
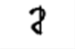
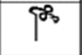
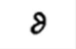
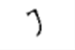
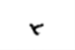
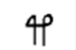
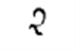
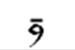
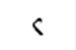
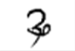
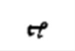
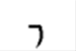
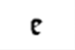
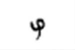
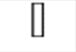
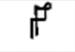
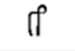
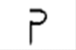
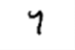
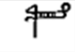
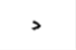
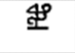
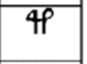
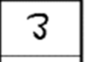
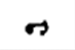
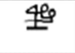
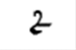
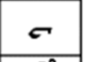
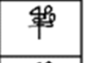
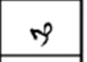
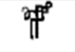
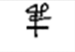
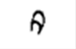
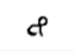
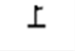
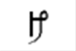
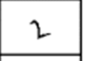
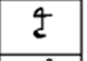
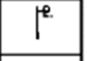
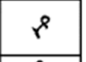
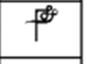
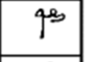
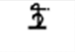
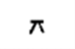
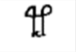
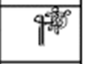
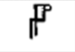
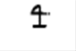
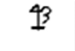
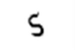
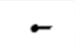
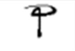
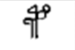
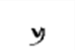


BASIC EVA CHARACTERS		
	EVA	Capitalised EVA
'		
a		
b		
c		
d		
e		
f		
g		
h		
i		
j		
k		
l		
m		
n		
o		
p		
q		
r		
s		
t		
u		
v		
x		
y		
z		

PUNCTUATION CHARACTERS		
	EVA	
*	*	unreadable
,	,	possibly a space
-	–	drawing intruding into text
.	.	space
=	=	end of paragraph
?	?	missing word
???	???	missing words
!	!	interlinear non-coding spacer
%	%	interlinear coding spacer

"UNOFFICIAL EVA"		
"		plume on top of connector
+		plume intruding in connector

META CODES		
#		line comment
{ }		in-line comment
< >		folio/locus indicator
[|]		alternative readings
\		line split (not in original)
$		weirdo code header
&		extended-eva header
;		end of extended-eva or weirdo code
( )		ligature notation

EXTENDED EVA CHARACTERS										
& code	Arial	EVA		& code	Arial	EVA		& code	Arial	EVA
130	‚			-	-	-		190	¾	
131	ƒ			161	¡			191	¿	
132	„			162	¢			192	À	
133	…			163	£			193	Á	
134	†			164	¤			194	Â	
135	‡			165	¥			195	Ã	
136	ˆ			166	¦			196	Ä	
137	‰			167	§			197	Å	
138	Š			168	¨			198	Æ	
139	‹			169	©			199	Ç	
140	Œ			170	ª			200	È	
141	□			171	«			201	É	
142	□			172	¬			202	Ê	
143	□			173	-			203	Ë	
144	□			174	®			204	Ì	
145	‘			175	¯			205	Í	
146	’			176	°			206	Î	
147	“			177	±			207	Ï	
148	”			178	²			208	Ð	
149	•			179	³			209	Ñ	
150	–			180	´			210	Ò	
151	—			181	µ			211	Ó	
152	˜			182	¶			212	Ô	
153	™			183	·			213	Õ	
154	š			184	¸			214	Ö	
155	›			185	¹			215	×	
156	œ			186	º			216	Ø	
157	□			187	»					
158	□			188	¼					
159	Ÿ			189	½					

A.5 Pre-Processing Example

The following gives an example of the pre-processing that is completed during the initial stages of phase 1.

Unprocessed Interlinear Archive Example

```
## <f17v.P> {}
# text
# Last edited on 1998-12-06 20:57:24 by stolfi
#
<f17v.P.1;H>      pchodol.chor.fchy.opydaiin.odaldy-{plant}
<f17v.P.1;C>      pchodol.chor.pchy.opydaiin.odaldy-{plant}
<f17v.P.1;F>      pchodol.chor.fchy.opydaiin.odaldy-{plant}
#
<f17v.P.2;H>      ycheey.keeor.ctho!dal.okol.odaiin.okal-{plant}
<f17v.P.2;C>      ycheey.kshor.ctho!dal.okol.odaiin.okal-{plant}
<f17v.P.2;F>      ycheey.keeor.ctho.dal.okol.odaiin.okal-{plant}
#
<f17v.P.3;H>      oldaim.odaiin.okal.oldaiin.chockhol.olol-{plant}
<f17v.P.3;C>      oldaim.odaiin.okal.oldaiin.chockhol.olol-{plant}
<f17v.P.3;F>      oldaim.odaiin.okal.oldaiin.chockhol.olol-{plant}
#
<f17v.P.4;H>      kchor.fchol.cphol.olcheol.okeeey-{plant}
<f17v.P.4;C>      kchor.fchol.cphol.olcheol.okee!y-{plant}
<f17v.P.4;F>      kchor.fchol.cphol.olcheol.okeeey-{plant}
```

Processed File for H

```
pchodol chor fchy opydaiin odaldy
ycheey keeor cthodal okol odaiin okal
oldaim odaiin okal oldaiin chockhol olol
kchor fchol cphol olcheol okeeey
```

A.6 Transcriber Codes

The following is a list of the transcriber codes and their respective transcriber

Transcriber Code	Transcriber
C	Currier
F	Friedman (First Study Group)
T	John Tiltman
L	Don Latham
R	Mike Roe
K	Karl Kluge
J	Jim Reed
D	Currier Alternative
G	Friedman Alternative
I	Jim Reed Alternative
Q	Karl Kluge Alternative
M	Don Latham Alternative
H	Takeshi Takahashi
N	Gabriel Landini
U	Jorge Stolfi
V	John Grove
P	Father Th. Petersen
X	Denis V. Mardle
Z	René Zandbergen