



Cracking the Voynich Manuscript: Using basic statistics and analyses to determine linguistic relationships

Andrew McInnes (a1211832)

ELEC ENG 4068 A/B HONOURS PROJECT

B.E. in Electrical and Electronic Engineering

Date submitted: 21st October 2015

Supervisor: Professor Derek Abbott
Co-Supervisors: Maryam Ebrahimpour
Brian Ng.

Acknowledgments

I would like to extend my deepest gratitude to my supervisor, Prof. Derek Abbott, and co-supervisors, Dr. Brian Ng and Maryam Ebrahimpour, for their continual support and guidance throughout the research project. The advice given throughout helped drive the project forward and allowed for basic investigations on a very interesting topic.

I would also like to thank my project partner, Lifei Wang, who continually contributed to the overall project as well as helping with my own sections. The project would have been as efficient without him.

Abstract

The Voynich Manuscript is a 15th century document written in an unknown language or cipher. This thesis presents that basic statistics can be used to show indications of possible linguistic relationships between the Voynich and other languages or hypotheses. Previous research is investigated before tests are carried out through data-mining a digital transcription of the Voynich. Basic features such as word and character frequencies, bigrams, affix frequencies, and word pairs are analysed against other languages and possible hypotheses. The results are then discussed and concluded upon.

Contents

Acknowledgments	2
Abstract.....	3
1 Introduction	6
1.1 Background	6
1.2 Motivation	7
1.3 Objectives	7
1.4 Pre-processing of the Interlinear Archive.....	7
1.5 Choice of Transcription.....	7
1.6 Comparison Texts and Corpora	9
2 Topic 1: Basic Statistical Characterisation of the Voynich Manuscript.....	11
2.1 Introduction	11
2.2 Literature Review	11
2.3 Zipf’s Law Theory	12
2.4 Method	12
2.5 Results	12
2.6 Discussion.....	15
2.7 Conclusion	15
3 Topic 2: English Investigation	17
3.1 Introduction	17
3.2 Literature Review	17
3.3 Method	18
3.4 Results	19
3.5 Discussion.....	20
3.6 Conclusion	21
4 Topic 3: Morphology (Affix) Investigation	22
4.1 Introduction	22
4.2 Literature Review	22
4.3 Method	24
4.4 Results	24
4.5 Discussion.....	28
4.6 Conclusion	28
5 Topic 4: Collocation Investigation.....	29
5.1 Introduction	29
5.2 Literature Review	29
5.3 Method	30
5.4 Results	31

5.5	Discussion.....	34
5.6	Conclusion	35
6	Discussion	36
7	Conclusion	37
8	References.....	38

1 Introduction

Linguistics, or the study of language, has been around for centuries and is continuing to evolve even today. With the invention of computers, linguistics can now be studied through computational linguistics. Through data-mining, statistics on written texts can be found much faster than traditional means but requires knowledge of linguistics to correctly analyse.

Using simple data-mining techniques to determine basic statistics within the written texts, indications of linguistic relationships between Voynich Manuscript and other known languages can be found. These relationships may not be definitive but will give suggestions for further research into particular linguistic properties or languages for future projects.

1.1 Background

The Voynich Manuscript is an undeciphered folio written in an unknown script that has been carbon dated back to the early 15th century [1] and is believed to have been created in 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 has been studied by both professionals and amateurs alike but, even with the aid of modern computer-based analysis techniques, neither have come to a definitive conclusion. It is divided into several different section based on the nature of the drawings [3]. These sections are:

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

Examples of these sections can be seen in Appendix A.

Many possible interpretations and hypotheses have been given [4] but these generally fall into three possibilities.

- 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 Motivation

The project attempts to find relationships and patterns within unknown text through the usage of basic linguistic properties and analyses. The Voynich Manuscript is a prime candidate for analyses as there is no known accepted translations of any part within the document. The relationships found can be used help narrow future research and to conclude on specific features of the unknown language 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].

While some may question as to why an unknown text is of any importance to Engineering, a more general view of the research project shows that it deals with data acquisition and analyses. This is integral to a wide array of businesses, including engineering, which can involve a basic service, such as survey analysis, to more complex automated system.

1.3 Objectives

The aim of the research project is to determine possible features and relationships of the Voynich Manuscript through the analyses of basic linguistic features and to gain knowledge of these linguistic features. These features can be used to aid in the future investigation of unknown languages and linguistics.

The project does not aim 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 from a small team of student engineers with very little initial knowledge on linguistics.

1.4 Pre-processing of the Interlinear Archive

The Voynich Interlinear Archive contains digital ASCII representations, see Appendix B, of the Voynich Manuscript from various different transcribers in the European Voynich Alphabet (EVA), see Appendix C. The archive contains 19 different transcriptions of the Voynich Manuscript and is formatted to allow for software code to extract each of the different transcriptions. Each page contains the transcribed lines by each transcriber with each appropriately tagged to show the line number and the transcriber. A basic example of the unprocessed file and the output after processing is shown in Appendix D.

The Interlinear Archive also included inline formatting that can be used to align the texts of each transcription and show where any extended EVA characters or illustrations within the physical book could be found.

Pre-processing the archive allows for simplification of any software processing in the future by keeping all the transcriptions separate. All unnecessary data can also be removed.

1.5 Choice of Transcription

A difficulty in data-mining the text was to determine which of the various transcriptions to use as a base for any comparisons with other texts in the following experiments. Unfortunately no transcription is complete and each varied in alphabet size and, correspondingly, vocabulary size. As the original text is hand-written dissimilarities could be attributed to the interpretations of each character

by each transcriber. It has been stated that some character tokens are very ambiguous and could be interpreted as a single, distinct character or multiple characters [2].

With any statistical research, the sample size is an important factor [6]. A larger sample size will give a broader range of the possible data and hence form a better representation for analysis. To determine the best transcription to be used, the total lines and word tokens contained by each different transcription was determined. These are shown in Figures 1-1 and 1-2 below.

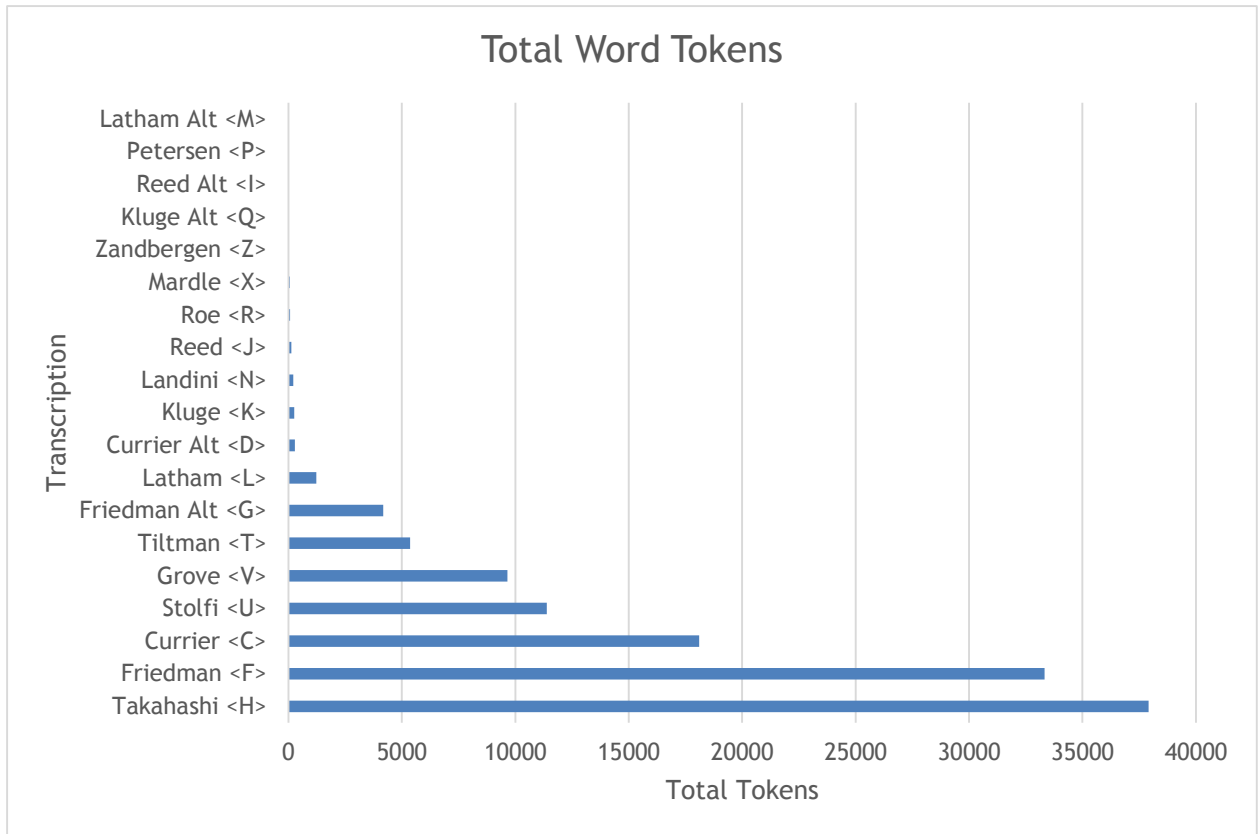


Figure 1-1: Transcription Total Word Token Comparison

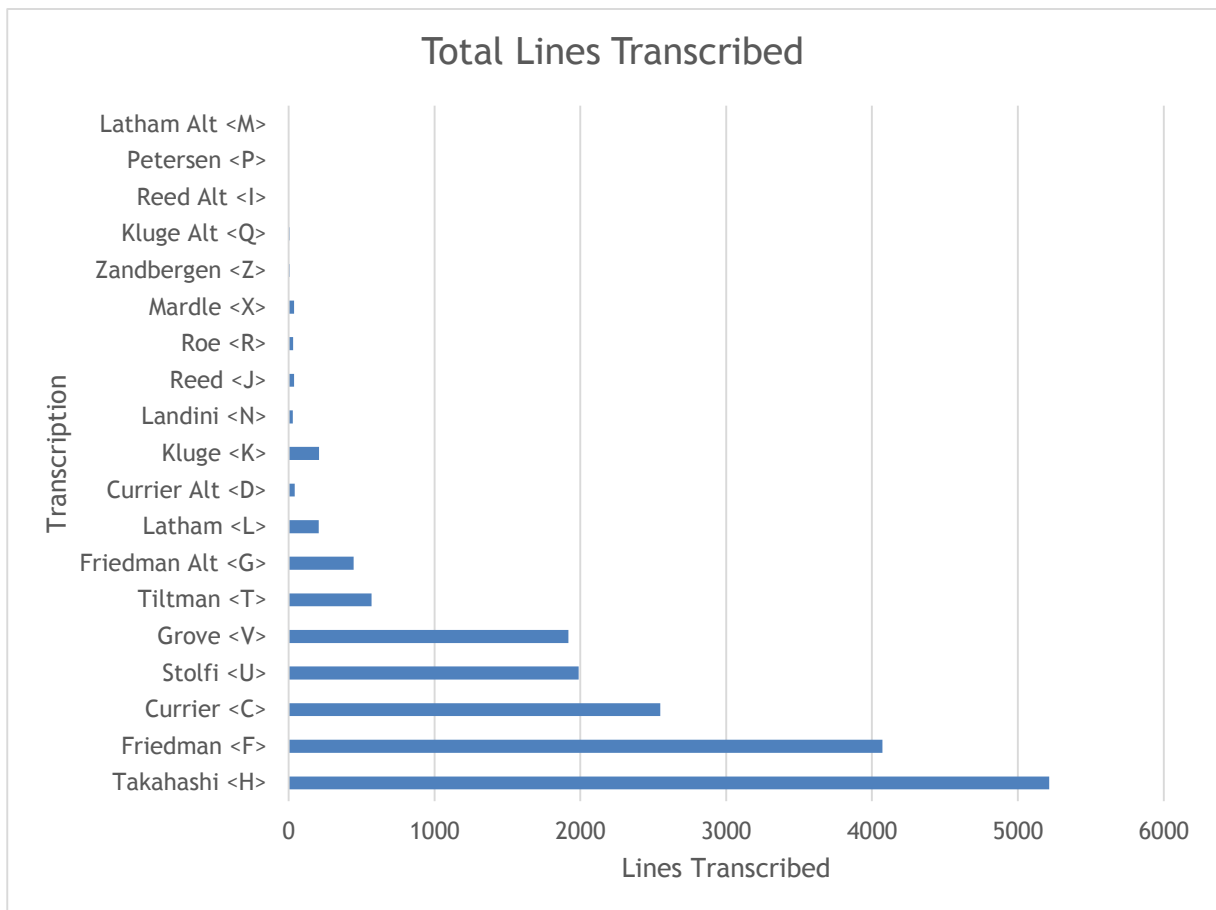


Figure 1-2: Total Lines Transcribed Comparison

From these plots in Figures 1-1 and 1-2 above, it can be easily seen that the Takahashi transcription contains the largest sample size available, having the most lines transcribed and containing the most word tokens. Based on these two metrics the Takahashi transcription was concluded as the most complete. As stated in [6] a larger sample size should give a better representation for analysis. Hence the Takahashi transcription was chosen to be used throughout the experimental study.

1.6 Comparison Texts and Corpora

Initially the Universal Declaration of Human Rights (UDHR) was used for comparisons. These were used to give basic indications of languages to use for any language comparative tests. However, with such a small word token count, the UDHR would not allow for accurate quantitative results. Therefore the UDHR was only used for the initial word length distribution testing.

An investigation into the various character token statistics of English utilized a small corpus of various different English texts. These were used specifically for investigating the statistical representation of English characters and how these could be used to determine if a specific character was either an alphabet character or a non-alphabet character. Different writing styles of texts were used to examine how the statistics could differ despite being of the same language.

To keep language comparison results coherent, a corpus of various different languages was compiled using various translations of the Old Testament. It is important to keep any texts, within a corpus, in the same domain [7] and writing style [8] as different domains and writing styles can give different statistics even

within the same language. The total word tokens within each text is also reduced to 38000 to keep the sample sizes similar to that of the 37919 word tokens of the Takahashi transcription. The majority of the languages are focused around Europe due to the belief that Voynich was created in Europe [2].

Both the English and language comparison corpora can be seen in Appendix E.

2 Basic Statistical Characterisation of the Voynich Manuscript

2.1 Introduction

Statistical characterisation of text can be handled through multiple different methods [9]. Characterisation of the Voynich Manuscript was handled through the identification of the basic statistics within the text. These included:

- Total Word Token Count
- Vocabulary Size
- Word Length Distribution
- Total Character Token Count
- Alphabet Size
- Longest Word Token
- Word Frequency Distribution

These statistics were used to examine the general size of the alphabet and words, and to determine if the data followed Zip's Law.

The various translations of the UDHR was also used to compare the word length distributions of other known languages to that of the Voynich.

2.2 Literature Review

Many previous researchers have characterized the Voynich. Reddy and Knight [2] perform various different statistical measurements to characterise the Voynich Manuscript. They determine that some character tokens mainly appear at the beginning of paragraphs and paragraphs themselves do not span multiple pages. The text appears to be written from left to right in a fully justified manner. They summarize that the Voynich is comprised of 225 pages containing a total of 8114 different words and 37919 word tokens. Word frequency and word length distribution is also investigated. It is found that the Voynich follows Zipf's Law, showing linguistic plausibility, and that the word lengths appear to have a narrow binomial distribution suggesting the Voynich is not a natural language or a form of *abjad*, a writing system that leaves out vowels and only uses consonants.

Diego R. Amancio, Eduardo G. Altmann, Diego Rybski, Osvaldo N. Oliveira Jr., and Luciano da F. Costa [10] 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 statistical properties with no prior knowledge of the meaning of the text. They also conclude that the Voynich Manuscript is compatible with natural languages [10].

Shi and Roush [11] also perform a basic statistic characterisation of the Voynich Manuscript. They give the statistics on each section and the full manuscript detailing similar statistics as found within this paper. They also include determining the primary Currier language of each section. It is again found that the Voynich

appears to follow Zipf’s Law and that the word length distribution of the Voynich appears to have a narrow binomial distribution centered on the word length of five.

2.3 Zipf’s Law Theory

Zipf’s Law is a power law that states the ‘rth’ most frequent word has a frequency that scales according to:

$$f(r) = \frac{1}{r^\alpha}$$

Where r is the “frequency rank” of a word, f(r) is its corresponding frequency, and $\alpha = 1$ [12]. In other words the frequency of a given word is inversely proportional to its rank in frequency. As human language generally follows this type of distribution [12], this law can be used to given an initial indication of whether a text can be considered a natural language.

2.4 Method

The method for characterisation of the text was simple, a MATLAB code was written and executed over the text that tracked the relevant statistics detailed in Section 2.1 through simple arrays and totaling algorithms. These could then be used to create the relevant tables and plots.

2.5 Results

The following results in the following tables detail the basic data obtained from the Takahashi transcription of the Voynich Manuscript. Table 2-1 shows the basic first-order statistics, while Tables 2-2 and 2-3 show these statistics based on the proposed sections of the Voynich Manuscript. In this paper the vocabulary size is defined as the total unique word tokens and the alphabet size is defined as the total unique character tokens. Alphabet size does not distinguish between ‘regular’ alphabet, numerical and punctuation characters.

	Excluding EVA Characters	Including EVA Characters
Total Word Tokens	37919	37919
Vocabulary Size	8151	8172
Total Character Tokens	191825	191921
Alphabet Size	23	48
Longest Word Token	15	15

Table 2-1: Basic First-Order Statistics of the Takahashi Transcription

Section	Total Word Tokens	Vocabulary Size	Total Character Tokens	Alphabet Size	Longest Word Token
Herbal	11475	3423	54977	23	13
Astronomical	3057	1630	15777	20	14
Biological	6915	1550	34681	20	11
Cosmological	1818	834	9289	21	13
Pharmaceutical	3972	1668	20168	21	15
Recipes	10682	3102	56933	21	14

Table 2-2: First-Order Statistics based on Section (excluding extended EVA)

Section	Total Word Tokens	Vocabulary Size	Total Character Tokens	Alphabet Size	Longest Word Token
Herbal	11475	3441	55040	44	13
Astronomical	3057	1630	15781	23	14
Biological	6915	1550	34684	22	11
Cosmological	1818	834	9290	22	13
Pharmaceutical	3972	1668	20180	24	15
Recipes	10682	3102	56946	29	14

Table 2-3: First-Order Statistics based on Section (including extended EVA)

The word length distribution of the various transcriptions with a significant sample size was also taken. This is given in Figure 2-1 below.

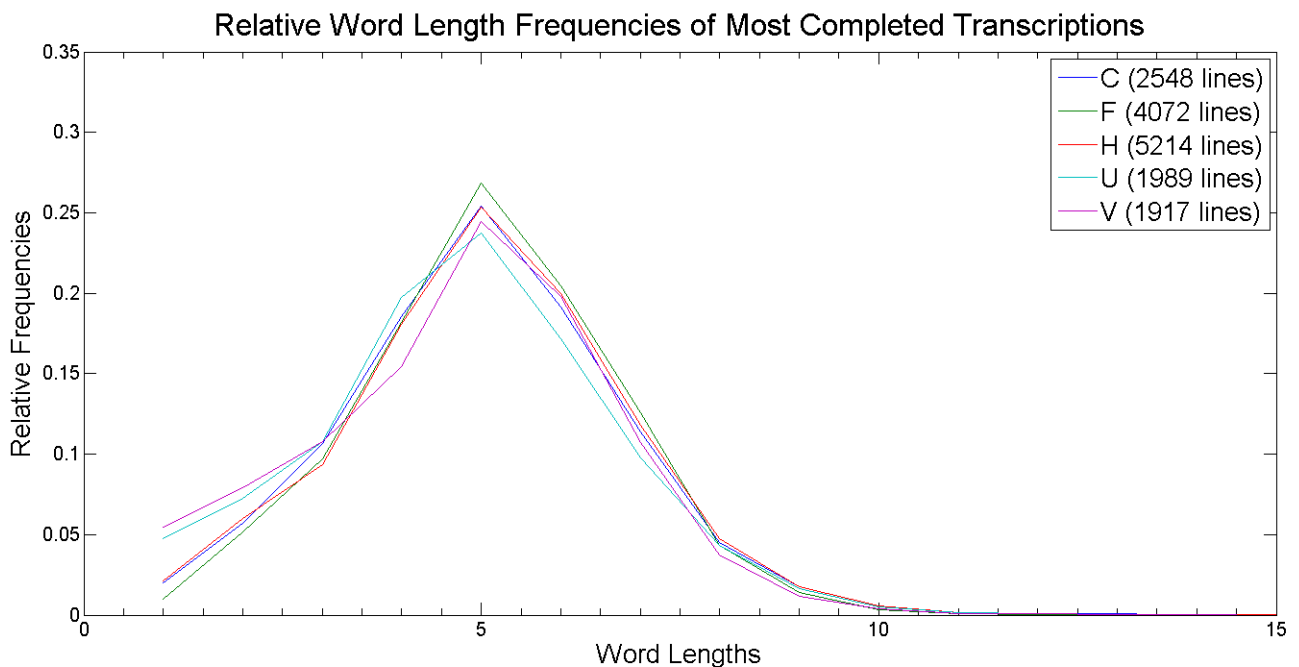


Figure 2-1: Word Frequency Distribution of Most Completed Voynich Transcriptions

The word length distribution of the Takahashi transcription against a small selection of European languages is given in Figure 2-2 below.

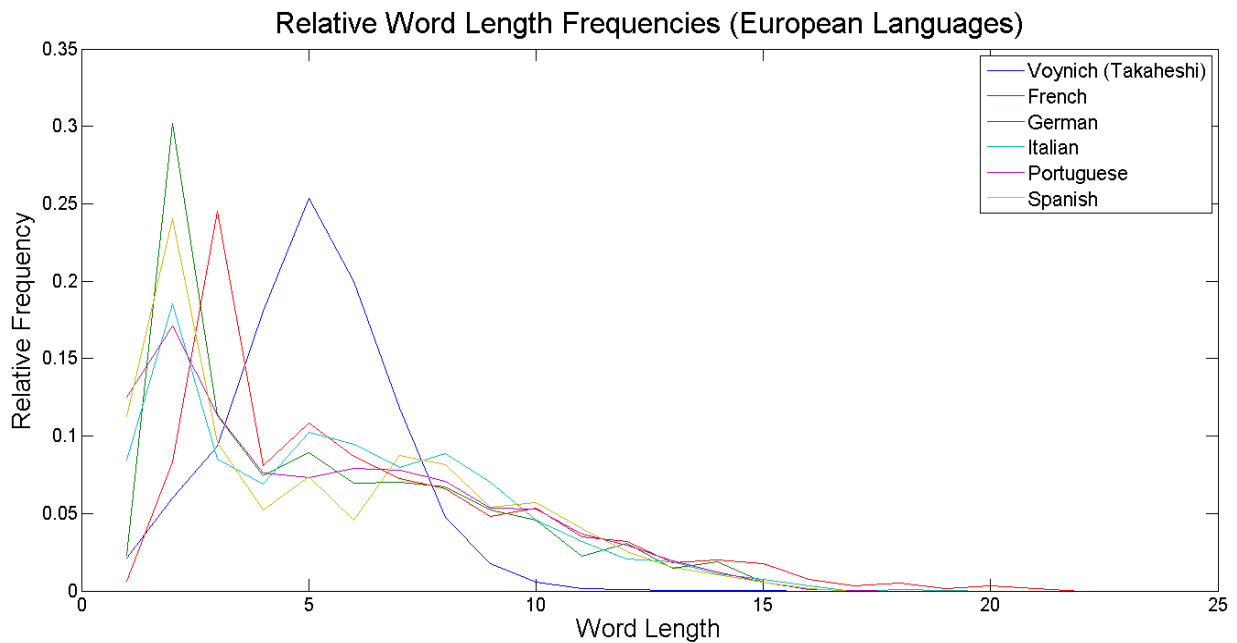


Figure 2-2: Word Length Distribution of Voynich and Various European Languages

This final graph in Figure 2-3 below shows the word frequency distribution, ranked from the highest frequency to the lowest, of the Voynich against that of English.

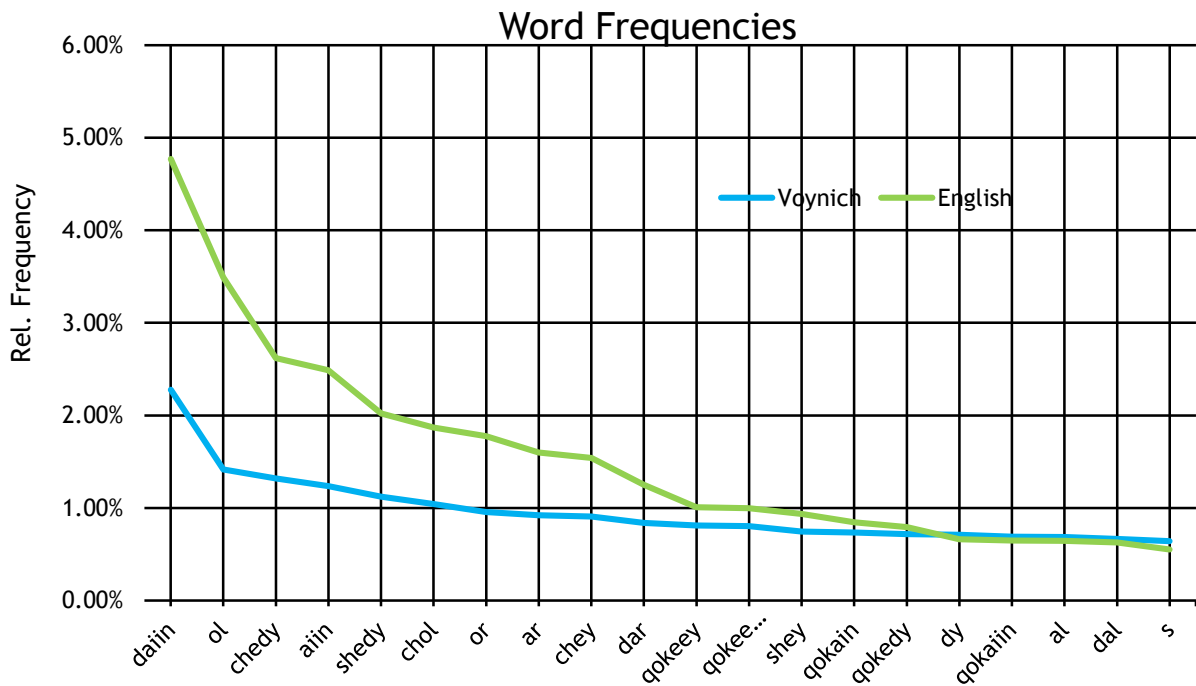


Figure 2-3: Word Frequency Distribution

2.6 Discussion

These results in Table 2-1 give a very basic impression of the Voynich Manuscript, showing that the Takahashi transcription contains 37919 words in total comprised of 8151 different words, or 8172 if including the extended alphabet characters. This is very similar to the data found by Reddy and Knight [2], and Shi and Roush [11] with minor differences in the total different words. These differences may be attributed to the choice of transcription or differences in pre-processing the archive. By including the extended EVA characters, it can also be seen that the alphabet size increases from 23 to 48 but does not result in any significant increases to the vocabulary size nor total character tokens.

By further separating the Voynich Manuscript into the proposed sections, the majority of the extended EVA characters seem to appear within the herbal section, showing an increase in alphabet size from 23 to 44. Note that, again, the vocabulary size here has a minor increase but does not increase at all within the following sections despite the increase in alphabet size.

Comparing the word length distributions of the most complete transcriptions shows that the majority of the word lengths are generally the same between the different transcriptions. All show that the word length distribution peaks at a length of 5 with a binomial distribution. This is also found in other research by Reddy and Knight [2], and Shi and Roush [11] and may suggest a form of code or cipher.

When comparing the word length distribution of the Takahashi transcription with other European languages it can be clearly seen that the other languages have peak distributions much earlier than the Voynich and do not show such a distinguishable binomial distribution. Note that the European languages are of a limited size as the data is based off of the UDHR.

The word frequencies graph in Figure 2-3 shows that the Voynich follows a similar decaying curve to that of English but has much lower frequencies at the higher ranks. However it does appear to abide to Zipf's Law.

2.7 Conclusion

The data here does not allow for any significant conclusions. However it could be speculated that the basic EVA characters do not uniquely identify any numerical or punctuation characters, in a similar fashion to English, due to the relatively small alphabet size. This does not mean that they are not represented, the numerical representations in particular may be represented using combinations of these basic EVA tokens much like Roman or Greek numerals.

The inclusion of the extended EVA characters does not present any more significant conclusions either, showing that their inclusion has very little effect on the other basic statistics. They are rare characters that may be similar to rare alphabetical tokens within the English language, such as q, x or z for example. They may also be rare punctuation tokens or even errors made by the transcribers. Some particular character tokens within the hand written Voynich are hard to distinguish [2] meaning these extended EVA characters may be errors made by the original author. Further testing is required but due to the limited data available it may be difficult to find any definitive conclusions.

The binomial distribution of the word lengths within the Voynich Manuscript suggests that the text is not a natural language and is, instead, some form of code

or cipher. As also stated in previous research, it may also be some form of abjad [2].

Zipf's Law also appears to be followed as shown in Figure 2-3. The decaying curve is not as pronounced as English but does indicate that the text may be in a form of natural language.

3 English Investigation: Character Categorisation

3.1 Introduction

Characters within a text can be divided into various different categories. Within the English language, characters can be broadly divided into:

- Alphabet Tokens
- Numerical Tokens
- Punctuation Tokens

This experiment aimed to expand on the basic character statistics found in Section 2. By incorporating character bigrams, the data could be used to attempt to categorise the characters from texts into possible alphabet and non-alphabet tokens. Utilizing MATLAB code written to determine the basic character frequencies and character bigrams, English text would be passed into MATLAB and categorised into the two different categories.

The statistics and extraction code could then be executed over the Voynich Manuscript to determine if any possible characters within the Voynich that may fall into the possible non-alphabet character category. Note that the extended EVA characters were ignored as they are characters tokens which rarely appear, hence not enough data would be available to be properly categorised.

3.2 Literature Review

Previous research did not reveal any methods used to categorise English characters as either alphabet or non-alphabet tokens. However many papers did reveal possible statistics that could be used to perform said categorisation and also highlighted possible difficulties.

Solso and Juel [13] provided a count of bigram frequencies and suggest that they may be useful in the assessment of the regularity of any word, non-word, or letter identification. Unfortunately the paper is very outdated and what they consider as comprehensive is now far below what is possible using computational methods available today. It does however show that letter identification may be possible using bigrams.

Jones and Mewhort [14] investigated the upper and lowercase letter frequency and non-alphabet characters of English over a very large (~183 million word) corpora. They find that there is no equivalence between the relative frequencies between the lowercase and corresponding uppercase characters, noting that there is a low mean correlation between upper and lower case characters. Their non-alphabet character results show that particular non-alphabet characters have much larger frequencies than some regular alphabet characters but also note that these frequencies can vary widely. The non-alphabet characters are generally found as a successor to an alphabet character but also find that on rare occasions a non-alphabet character, which regularly appears as a successor to an alphabet character, may appear before an alphabet character. It is concluded that different writing styles can affect the statistics of bigram frequencies and that both letter and bigram frequencies can have an effect on corresponding analyses.

Church and Gale [8] investigate different methods of determining the probabilities of word bigrams by initially considering a basic maximum likelihood estimator. This gives the probability of an n-gram by counting the frequency of each n-gram and dividing it by the size of the sample. Unfortunately this is very determinant on the sample but also state that these bigram frequencies could be used for the disambiguation of the output of a character recognizer. They therefore investigate two other methods, good-turning and deleted estimation methods, and compare with the results obtained from using the maximum likelihood estimator over a large corpora of 44 million words. The results show that these different methods for determining probability provide possible strengths over basic methods but note that their corpora may not be a balanced sample of English. They also state that the writing style of the texts can affect the results so particular care must be taken when selecting text for a corpus.

In terms of the Voynich Manuscript, Reddy and Knight [2] use an unsupervised algorithm, Linguistica, which returns two possible characters, K and L, as possible non-alphabet characters. The algorithm shows that these character tokens seem to only appear at the end of words, however the removal of these character tokens results in new words. Using a traditional definition of punctuation, which is punctuation only occurs at word edges, the removal of these character tokens should result in words already found within the Voynich. They therefore suggest that there is most likely no punctuation in the Voynich.

3.3 Method

The Alphabet extractor has gone through multiple different attempts to improve the performance and reliability. In general, the extractor used simple rules to determine if the character token is of a specific category. These include:

1. *Does the character token only (or the vast majority) appear at the end of a word token?*

Tokens that only appear at the end of a word token are generally only punctuation characters when using a large sample text or corpus. However, depending on the type of text, some punctuation characters may appear before another punctuation character, hence majority was taken into account.

2. *Does the character token only appear at the start of a word token?*

- *Does this character have a high relative frequency when compared to others only appearing at the start of a word token?*

In English, character tokens that only appear at the start of a word token are generally upper-case alphabet characters. Some punctuation characters may also only appear at the start of a word token, hence the relative frequencies were also taken into account.

3. *Does the character token have a high relative frequency?*

Tokens with a high relative frequency are generally alphabet characters, with the highest consisting of the vowels and commonly used consonants.

4. Does the character token have a high bigram ‘validity’?

Over a large English corpus, alphabetical characters generally appear alongside many more other tokens than non-alphabetical characters. Validity is defined as a bigram that occurs with a frequency greater than zero. Low validity suggests the character token is probably a non-alphabet character.

An English text is initially passed through a MATLAB code which finds the bigram and token frequencies which are then checked if they fit any of the rules and categorised accordingly. Note that a character may fall into multiple rules, hence multiple conditionals are given to help categorise a given character token. Any tokens that could not be categorised were considered to be alphabet tokens.

To determine if a character token only appears at the start or end of a word token, the bigrams were examined by the MATLAB code. The initial creation of the bigrams is completed by taking every unique character token within a given text and storing every possible character combination within a cell array and assigning each a frequency of zero. The MATLAB code would then read over the text and find every occurrence of the bigram, incrementing the corresponding bigram in the cell array. Using this frequency data, MATLAB checks every bigram for a specific character token appearing at the start or end of said bigram. If the character token never appears at the end of a bigram then it can be concluded that the character token never appears at the end nor middle of a word token. The same can be done to determine whether a character never appears at the start or middle of a word token.

3.4 Results

An example output of the Alphabet extractor code is given below. For example statistics on bigram frequencies and character frequencies see Appendix F and G respectively.

Possible Alphabet Characters

(A B C D E F G H I J K L M N O P R S T U W Y Z a b c d e f g h i k l m n o p q r s t u w x y z

Possible Non-Alphabet Characters

!), . : ; ? ' 0 1 2 3 4 5 6 7 8 9 V j v

The error rate of the worst-case tested text at various different word counts is shown in Figure 3-1 below. The word count of the Voynich is also shown.

English Alphabet Extraction Worst Case Error Rate

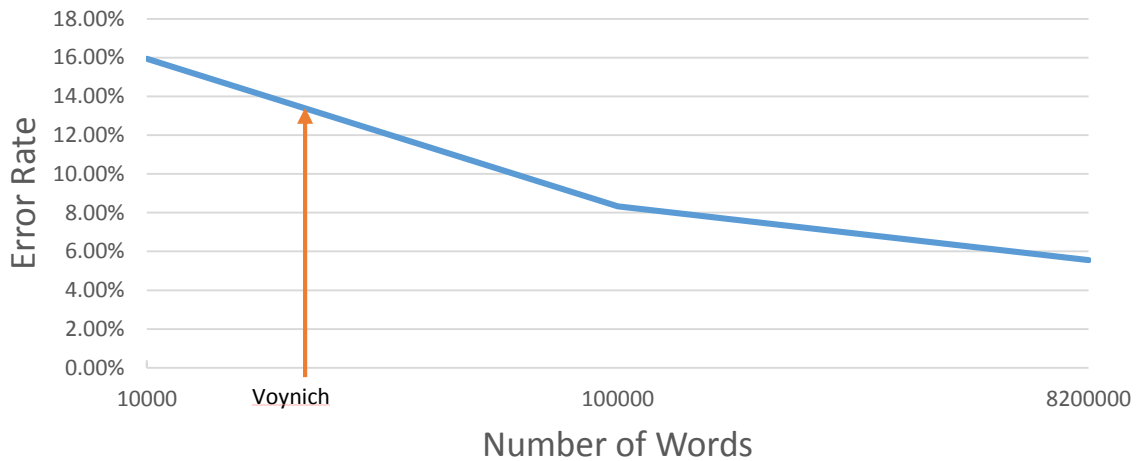


Figure 3-1: Worst-Case English Extraction Error Rate

The output of the Alphabet extractor code when ran over the Voynich is given below.

Possible Alphabet Characters

a c d e f g h i k l m n o p q r s t x y

Possible Non-Alphabet Characters

v z

3.5 Discussion

Based on the worst-case error rate graph, it can be clearly seen that the Alphabet extractor does not function with an acceptable error rate when using texts of small sample sizes. This can be attributed to higher variability in both the frequencies and bigrams. As mentioned by Church and Gale, bigrams can be very determinant on the sample size [8]. This means that bigrams that may not appear in one sample may appear in another which, based on the rules being used, can affect the results. For example, examining the character data example given in Appendix G, the character 'X' never appears within the tested New Testament meaning that all bigrams that were generated never included the character 'X'. Using smaller sample sizes may result in other missing characters.

It was also found that the Alphabet extractor seemed to give less error to basic novels (Robin Hood) than that of the New Testament, the worst cast tested. A simple comparison of the regular lower-case alphabet and numerical tokens show a significant difference in statistics. Hence the different writing styles having an effect on the extractor by varying in corresponding relative statistics [14]. The graph of the comparison can be seen in Appendix H.

With this knowledge, the results given from the Voynich are likely to be error prone, as shown by the error rate graph in Figure 3-1. When examined, the two reported possible non-alphabet characters had very low frequencies within the text which and neither character token only appeared at the start or end of a word

token. Therefore there is not enough data to reliably say that either character token are non-alphabet characters. Based on these results and the small alphabet size, it is highly likely that the basic EVA characters do not represent any non-alphabet characters. As briefly mentioned in Section 2.7, this does not mean non-alphabet characters are not represented in the text. They may be represented using combinations or sequences of character tokens similar to numerical representations in Roman and Greek.

As the rules used to determine the categorisation of a character token are generalisations made from data obtained using relatively small English texts, the categorisation of the possible alphabet characters may also have errors. Any categorisation of other languages needs further analysis.

Further modifications to the categorisation rules or the addition of more statistical measures may lead to better results in the future. However as writing style can affect the results, any alphabet extraction on different languages will be based on the data obtained from English and may cause further errors. Note that other investigations into character frequencies and bigrams used significantly larger corpora [8] [14] which should give more accurate statistics than those obtained here.

3.6 Conclusion

The output of the Alphabet extractor shows that it is possible to distinguish between alphabet and non-alphabet characters based on simple character and bigram frequencies to a certain extent. However the current implementation does require a large sample size to give an acceptable error rate and did show significant variations depending on the writing style of the input text. This was highlighted in other research papers as an area to be aware of [8] [14].

The extraction of the Voynich characters does show that it is likely that the Voynich does not contain any distinct punctuation characters as no character tokens appeared only at the start or end of a word token. As the extractor is biased towards English these results only give a very basic indication of this and relies on the Voynich character and bigram data following similar relationships to that of English.

4 Morphology Investigation: Naïve Affix Frequencies

4.1 Introduction

Linguistic morphology, broadly speaking, deals with the study of the internal structure of words [15]. It can be divided into several different categories, depending on the grammar, with the most basic being between *inflection*, the changes to the tense, gender, number, case, etc. of a word, and *word-formation*, the derivation and compounding of separate words [16].

English and many other languages contain many words that have some form of internal structure [17] that can fall into these categories. These internal structures can have multiple different forms, depending on the language itself, with the most common structural units as suffixes and prefixes [18]. This is also known as concatenative morphology.

Within this small experiment, the most common affixes in English are found and compared with those found within the Voynich Manuscript. Due to the unknown word structure and small relative size of the Voynich, this experiment defines an affix as a sequence of characters that appear at the word edges. Using this basic affix definition, a simple ranking of the affixes of various lengths could reveal potential relationships between the Voynich and other known languages through the use of the language comparison corpus.

4.2 Literature Review

Over the past years, many researchers have examined multiple different techniques of extracting different forms of linguistic morphology from various different languages [19]. Both unsupervised and supervised techniques have been used. Hammarström presents a particularly simple unsupervised algorithm for the extraction of salient affixes from an unlabelled corpus of a language [20]. This is particularly of interest as the Voynich Manuscript does not have any universally accepted morphological structure [2]. Hammarström's algorithm assumes salient affixes have to be frequent and that words are simply variable length sequences of characters. This is a naïve approach to handling the complex nature of morphology by restricting itself to concatenated morphology of which do not necessarily need to be frequent [20]. His results show that it includes many affixes that would be considered junk affixes where a junk affix is defined as a sequence of characters that, once affixed to a word, do not change the word in any meaningful way. He states that his results can only give guiding experimental data and did find that the writing-style, even in the same language, could give significant differences. More informed segmentation and peeling of affix layers was beyond the scope of the paper.

Eryiğit and Adalı offer two different approaches by using a large Turkish lexicon [21]. One approach was to initially determine the root words which allows for these to be stripped from other words, leaving the possible affixes. The other approach used the reverse order by initially determining the affixes which could then be stripped from the words leaving only root words. Both approaches used rule-based finite state machines as Turkish is a fully concatenative language that only contains suffixes [21]. This approach would not work with the Voynich Manuscript as there is no known lexicon that can be used with the Voynich. However the paper does give

evidence on how rule-based approaches can be utilised to determine morphological structure.

Minnen, Carol and Pearce show a method for analysing the inflectional morphology within English [22]. This did not use any explicit lexicon or word-base but did require knowledge of the English language as it used a set of morphological generalisations and a list of exceptions to these. This method is available as software modules which could be used in future experiments to compare with other possible methods to determine inflectional morphological structure.

Snover and Brent present an unsupervised system for the extraction of stems and suffixes with no prior knowledge of the language [18]. The system is designed to be entirely probabilistic that attempts to identify the final stems and suffixes for a given list of words. They state that the results and analysis are conservative, showing only a number of possible suffixes but, due to this, appears to be more precise than other morphology extraction algorithms. However this system requires a large corpus to determine a list of common words to use. In particular, when testing English Snover and Brent use the Hansard corpus which contains approximately 1.6 billion words. Other tests show that it has particular issues with languages that use more complex morphology.

Another paper shows extraction of morphology through the extension of the Morfessor Baseline, a tool for unsupervised morphological segmentation. Kohonen, Virpioja and Lagus state that the number of unique word formed from morphology can be very large in a given corpus [19]. They show that by adding the use of labelled data, which is data that is known as its corresponding morphological category, to unlabeled data the results of the extraction significantly improve. However this means that knowledge of the language is required to give such labelled data. They note that by using labelled data they can bias the system to a particular language or task and that it is difficult to avoid biasing across different languages. The morphemes themselves may be higher or lower depending on the language.

Morphology tests and experiments have also been carried out previously on the Voynich Manuscript. Several hypothesis of the basic structure have been given [2], these include:

- Roots and Suffixes model
- Prefix-Stem-Suffix model
- Crust-Mantel-Core model

Reddy and Knight perform a test on the Voynich Manuscript by running Linguistica, an unsupervised morphological segmentation algorithm, to segment the words into possible prefixes, stems and suffixes [2]. They conclude that the results suggest there is some form of morphological structure within the Voynich Manuscript.

Jorge Stolfi's [23] website "Voynich Manuscript stuff" gave multiple views and analyses of the morphological structure within the Voynich Manuscript. He also shows evidence of a possible prefix-midfix-suffix structure [24], and later displaying a crust-mantle-core paradigm [25].

4.3 Method

The affix extraction method exploits the simple definition given to affixes in this paper. That is, an affix is a sequences of characters that appear at the word edges. Text is read into a MATLAB code which is set to find all character sequences that begin at the start or end of a word, of a set length, and compute their relative frequencies. Any word that contains the same amount or less than the set length value is ignored. An example of extracted suffixes of character length 3 is given below.

Word Token: example Extracted Suffix: ple

Word Token: testing Extracted Suffix: ing

The extracted affixes are then ranked by frequency, with the most frequent ranked as 1 to the least frequent, and kept in their corresponding character lengths. These are plotted and compared with those found in the Voynich Manuscript and other languages. The expectation of the comparisons is to determine if any of the languages within the corpus show any similarities in affix frequency.

All punctuation within the any of the texts was also removed. The extraction of Voynich Manuscript used the simplified Takahashi transcription with the extended EVA characters removed. As the results from the previous investigations suggested that there was no punctuation within the Voynich Manuscript any texts used had punctuation removed through a simple C++ code.

4.4 Results

The initial results compared the affixes of character lengths from two to five from a section of the English text Robin Hood and those found with the Voynich. The results of the prefix extraction and ranking can be seen in Table 4-1 below.

Prefix Rank	Length 2 Relative Frequency (%)		Length 3 Relative Frequency (%)		Length 4 Relative Frequency (%)		Length 5 Relative Frequency (%)	
	Voynich	English	Voynich	English	Voynich	English	Voynich	English
1	0.1346	0.0323	0.0536	0.0107	0.0184	0.0054	0.0074	0.0032
2	0.0918	0.0215	0.0399	0.0083	0.0140	0.0035	0.0063	0.0025
3	0.0675	0.0215	0.0314	0.0062	0.0102	0.0029	0.0060	0.0022
4	0.0563	0.0195	0.0312	0.0062	0.0090	0.0029	0.0060	0.0017
5	0.0496	0.0186	0.0208	0.0055	0.0089	0.0027	0.0056	0.0017
6	0.0365	0.0162	0.0201	0.0055	0.0088	0.0025	0.0053	0.0017
7	0.0280	0.0149	0.0199	0.0053	0.0083	0.0025	0.0048	0.0017
8	0.0234	0.0145	0.0194	0.0050	0.0083	0.0025	0.0046	0.0017
9	0.0184	0.0141	0.0129	0.0048	0.0083	0.0025	0.0044	0.0017

10	0.0178	0.0138	0.0123	0.0048	0.0082	0.0023	0.0044	0.0015
-----------	--------	--------	--------	--------	--------	--------	--------	--------

Table 4-1: Prefix Frequency Comparisons

The results of the suffix extraction and ranking can be seen in Table 4-2 below.

Suffix Rank	Length 2 Relative Frequency (%)		Length 3 Relative Frequency (%)		Length 4 Relative Frequency (%)		Length 5 Relative Frequency (%)	
	Voynich	English	Voynich	English	Voynich	English	Voynich	English
1	0.1342	0.1158	0.0742	0.1039	0.0603	0.0151	0.0227	0.0092
2	0.1111	0.1041	0.0487	0.0225	0.0239	0.0143	0.0137	0.0050
3	0.0815	0.0405	0.0352	0.0146	0.0172	0.0135	0.0076	0.0047
4	0.0723	0.0377	0.0345	0.0143	0.0165	0.0118	0.0073	0.0045
5	0.0623	0.0375	0.0343	0.0140	0.0158	0.0106	0.0064	0.0042
6	0.0548	0.0319	0.0322	0.0136	0.0135	0.0101	0.0064	0.0040
7	0.0497	0.0213	0.0308	0.0117	0.0117	0.0095	0.0058	0.0037
8	0.0485	0.0202	0.0225	0.0115	0.0105	0.0091	0.0057	0.0037
9	0.0283	0.0200	0.0184	0.0105	0.0085	0.0089	0.0057	0.0035
10	0.0258	0.0188	0.0165	0.0100	0.0084	0.0073	0.0057	0.0035

Table 4-2: Suffix Frequency Comparisons

Further testing was completed over the suffixes of character length 3 and 4 as these showed the greatest variability between the Voynich Manuscript and English. In particular, there are significant differences between the top two ranked suffixes, hence the top two ranked frequencies of the various languages within the corpus were found and compared. The results of character length 3 suffixes are shown in Figures 4-1 and 4-2 below while the results of character length 4 suffixes are shown in Figures 4-3 and 4-4 below.

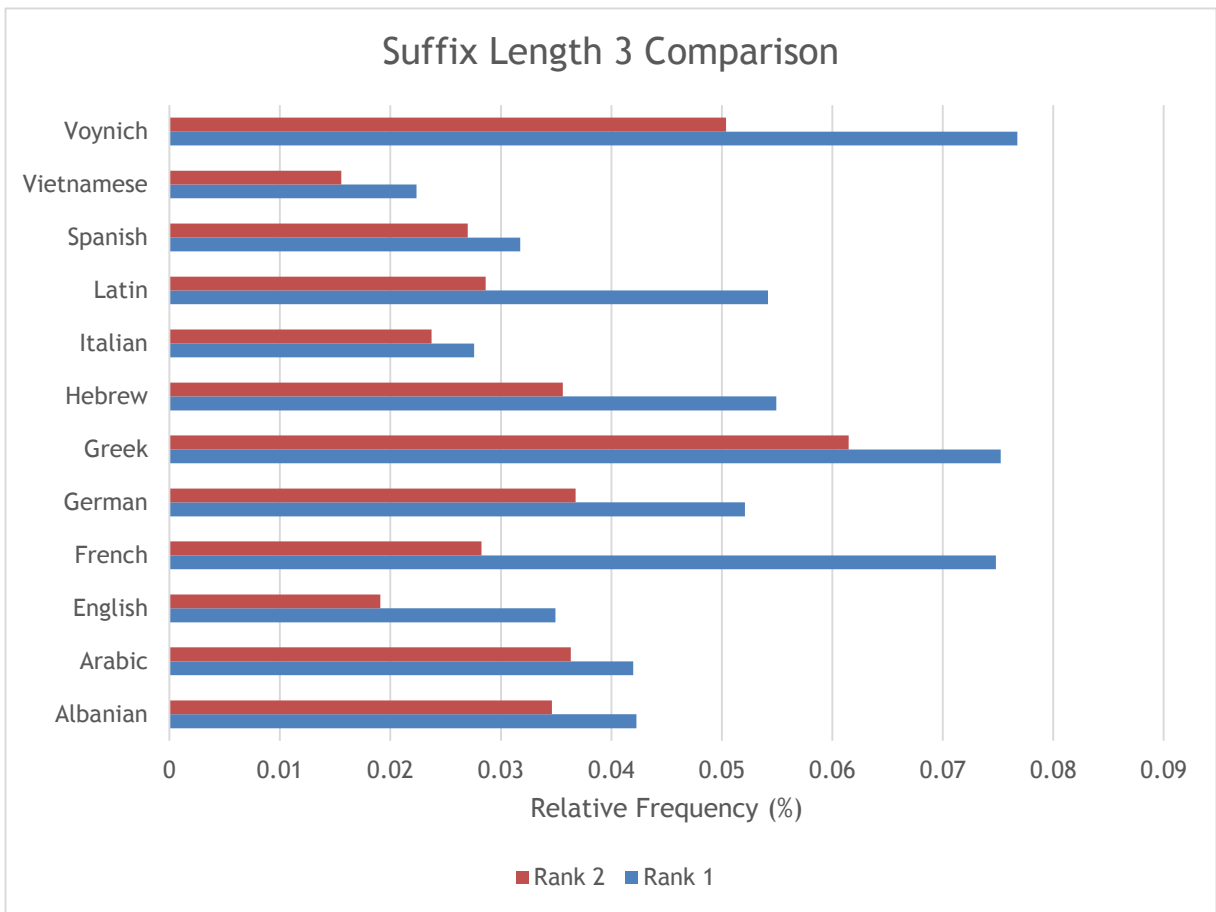


Figure 4-2: Top 2 Ranked Suffix Comparison (Character Length 3)

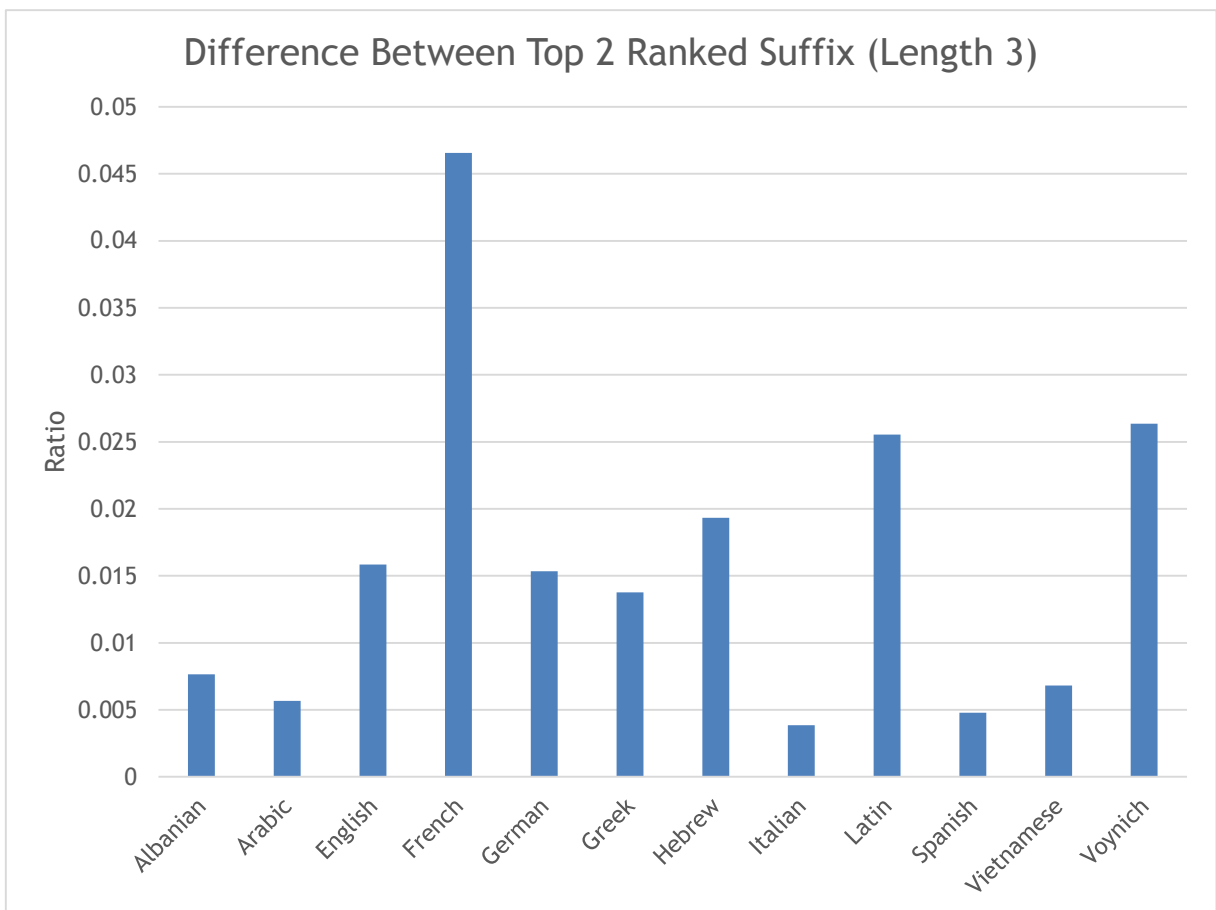


Figure 4-1: Difference Ratio of Top 2 Ranked Suffix (Character Length 3)

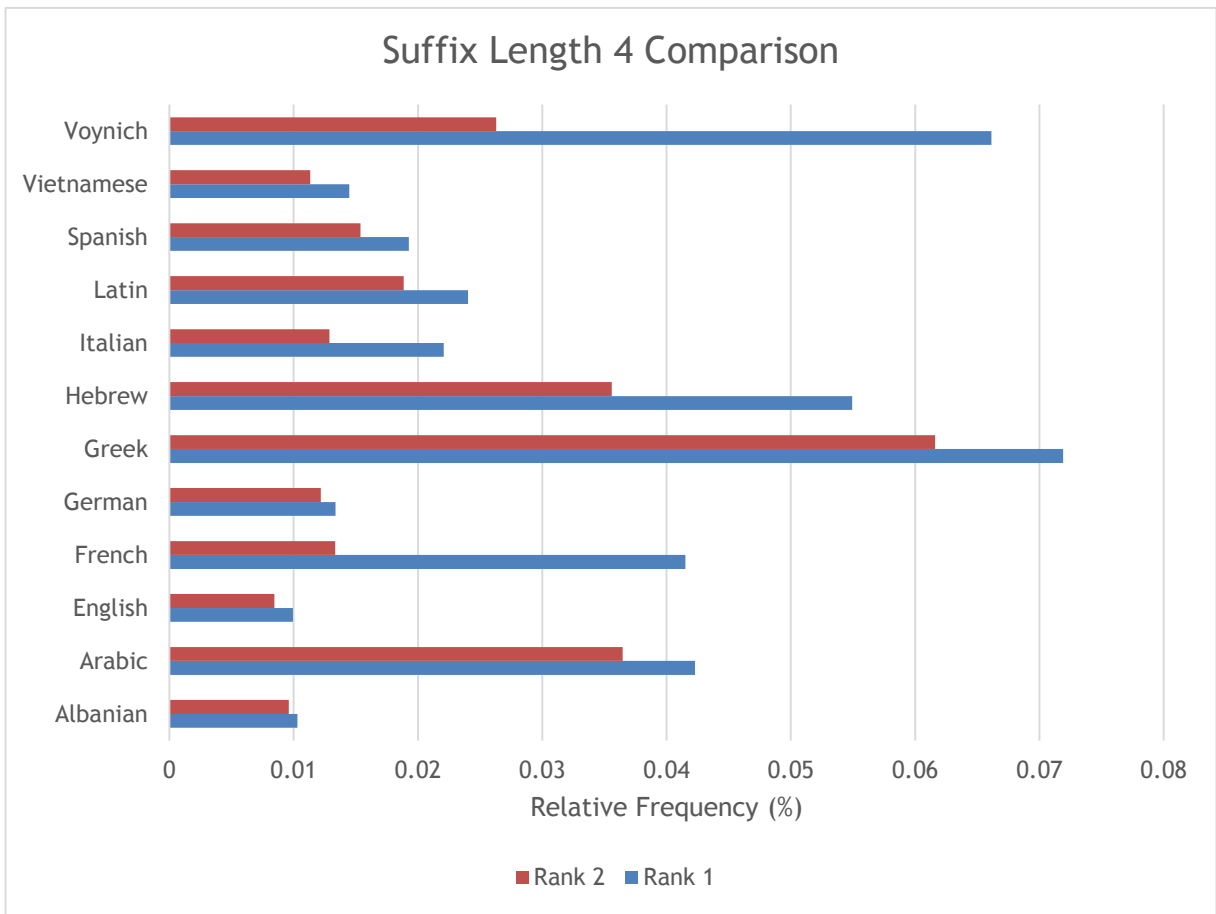


Figure 4-3: Top 2 Ranked Suffix Comparison (Character Length 4)

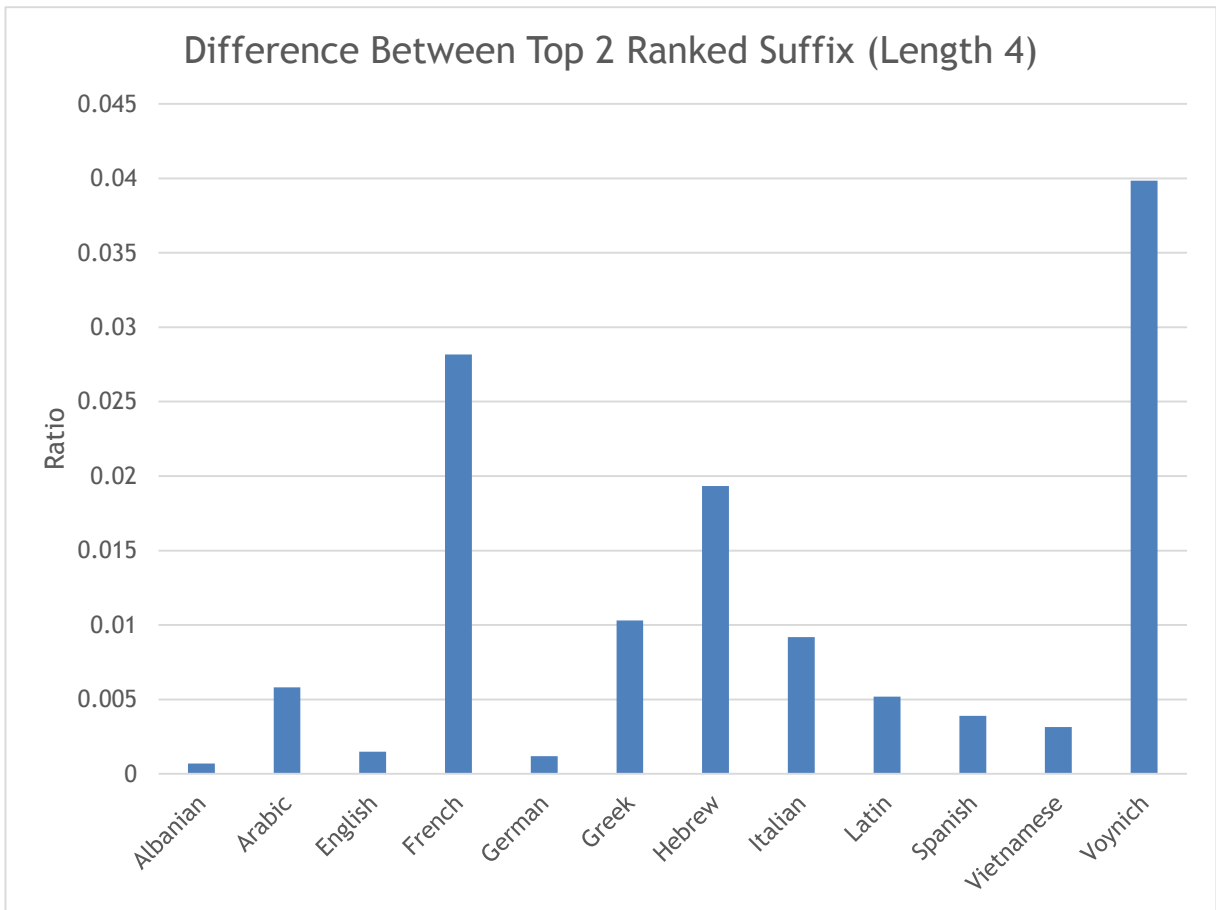


Figure 4-4: Difference Ratio of Top 2 Ranked Suffix (Character Length 4)

4.5 Discussion

The initial findings of the prefix and suffix comparisons between English and the Voynich do not appear to give any definitive relationships. From the prefix data in Table 4-1, it can be clearly seen that the Voynich contains many more frequent prefixes over the entire range when only considering the top 10 frequency ranked prefixes. This relationship does not appear to change significantly as the prefix character length is increased.

The suffix data provides a more significant difference in that English begins with very similar values to that of the Voynich at length 2, a higher ranked 1 suffix at length 3, and much lower values at lengths 4 and 5. The range of values given at length 4 for English even show an almost linear relationship while the Voynich shows an exponential decay. With such a significant difference between the length 3 and 4 suffixes the team decided to focus on these two lengths over the various languages within the corpus.

Examining the results of the length 3 and 4 comparisons of the corpus show that, at a relative frequency comparative level only Greek appears to have a relationship with the Voynich Manuscript showing a similar rank 1 suffix relative frequency but having a larger rank 2 suffix relative frequency. French also appears to have a similar rank 1 suffix but has a significantly lower rank 2. When comparing the difference between the two ranked suffixes, Latin has a very similar difference ratio to that of the Voynich.

At length 4 Greek, again, appears to have a possible relationship to that of the Voynich, showing a similar rank 1 but a much higher rank 2. However, when comparing the difference ratio between the two ranked suffixes, French has the closest relationship but is, still, significantly different.

It should be noted that morphology is full of different ambiguities [18] and are dependent on the language. These findings only give very baseline experimental data that assumes a very basic definition of an affix and is completely restricted to concatenative morphology. Much more precise data may be obtained by using a stronger definition for an affix that takes into account proper word stems or roots. Unfortunately the small sample size available from the Voynich does not allow for accurate morphological extraction based on previous research on known languages requiring significantly large sample sizes.

4.6 Conclusion

The data here does not give any conclusive findings. It does show that there may be possible morphological relationships, albeit weak, in the Voynich to other languages, in particular Greek, Latin and French, when using a naïve definition of an affix as a sequence of characters at the edge of a word.

It may also suggest that there is some form of morphological structure within the Voynich Manuscript but these results are unable to definitively conclude on that also. Further research is required.

5 Collocation Investigation: Word-Pair Association

5.1 Introduction

Collocations have no universally accepted formal definition [26] but deals with the words within a language that co-occur more often than would be expected by chance [27]. Natural languages are full of collocations [27] and can vary significantly depending on the metric, such as length or pattern, used to define a collocation [7].

In this research experiment, the definition used for a collocation is that of two words occurring directly next to each other. As collocations have varying significance within different languages, by extracting and comparing all possible collocations within the Voynich Manuscript and the corpus, a relationship based on word association could be found or provide evidence of the possibility of a hoax.

5.2 Literature Review

Similarly to the results found in the English Investigation in Section 3, collocation statistics are domain and language dependent [7]. Therefore texts within a corpus should be of the same domain to be able to compare results between languages. This also does not mean that the statistics will be the same as the recurrent property of words are typical to different types of languages [27]. This makes them difficult to translate across languages but, by using word association metrics, may show if a text in a similar domain has any relationship between different languages.

There are multiple different types of collocations which range from basic phrases to strict word-pair collocations [27] such as the collocation defined within this investigation. The word association metrics can also vary significantly and have a range of different statistical methods to assign a metric [7]. These include but are not limited to:

- T-Score
- Pearson's Chi-Square Test
- Log-Likelihood Ratio
- Pointwise Mutual Information

Thanopoulos, Fakotakis and Kokkinakis compare these various word association metrics, defining their collocations as strict word-pairs [7]. Their results show that the values of the metrics can vary significantly and that, depending on the choice of association metric, will rank the same collocations in different orders. However, despite these differences, the resulting curve from the metrics are generally quite similar.

Wermter and Hahn also investigate different word association metrics while making comparisons to a simple frequency based metric [28]. While it is generally assumed that using a statistical association measure will produce more viable results [28], Wermter and Hahn argue that this type of association may not necessarily produce better results than a simple frequency association if not including additional linguistic knowledge. Like Thanopoulos, Fakotakis and Kokkinakis, Wermter and Hahn also show that using different metrics can return similar output assuming the

metric ranks the most-likely collocations at the higher ranks while non-collocations are ranked last.

Pearce states that with no widely accepted definition on the exact nature of linguistic collocations there is a lack of any consistent evaluation methodology [29]. Many proposed computer based collocation definitions are based around the use of N-Gram statistics. An issue with this is that a dependency in a collocation may span many words, giving an example of French where a collocation may span up to 30 words. He shows different methods of giving a metric for word association and states that pointwise mutual information has so far been widely used as a basis. It is also stated that despite a universally accepted definition for a collocation, comparative evaluation is still useful.

Reddy and Knight [2] show summarized information on the word correlations of the Voynich. In particular they show that the word association of word-pairs within the Voynich at varying distances do not show any significant long-distance correlations and suggest that this may arise from scrambling of the text, generation from a unigram model, or the interleaving of words.

Shi and Roush [11] of the previous final year project group also carry out a collocation investigation using word-pairs and again found that the Voynich displayed a weak word association measure when compared to other languages. They suggest this could indicate that the manuscript is a hoax or some type of code, further stating that ciphers are designed to have weak word order.

5.3 Method

The extraction of the collocations utilized a simple MATLAB code that read a text file, determining every collocation within the text file and corresponding statistics.

Collocations were extracted by initially tokenizing all the words within the text file and pairing each adjacent word token in a separate cell array. The frequency of each collocation was tracked as the pairing of word tokens occurred.

To determine the strength of a word association two different metrics were used and each collocation was ranked based on this metric. Initially each collocation was ranked based on their corresponding relative frequencies, where the most frequent collocations would be ranked higher than the less frequent collocations. This can be defined as the probability of the word-pair occurring within the text.

The second word association metric used was pointwise mutual information which is considered as a widely accepted method to quantify the strength of word association [7] [28] [11]. This method incorporates the probabilities of each word occurring within the text as well as the two word appear coincidentally. This is defined mathematically as:

$$PMI(x, y) = \log_2\left(\frac{P(x, y)}{P(x)P(y)}\right)$$

Where $P(x)$ and $P(y)$ are the respective probabilities of each word occurring within the text and $P(x,y)$ is the probability of the words occurring coincidentally.

Both methods allowed for a plot to be generated of the corpus and the Voynich Manuscript such that comparisons could be made between the different languages.

A simple scrambling code was also written to scramble the word placements within a text. It was expected that this would uncorrelate the majority of word pairings and give comparative results to the same texts without the scrambling.

5.4 Results

The following graphs in Figures 5-1 and 5-2 show the results obtained from the initial collocation ranking using their basic frequencies.

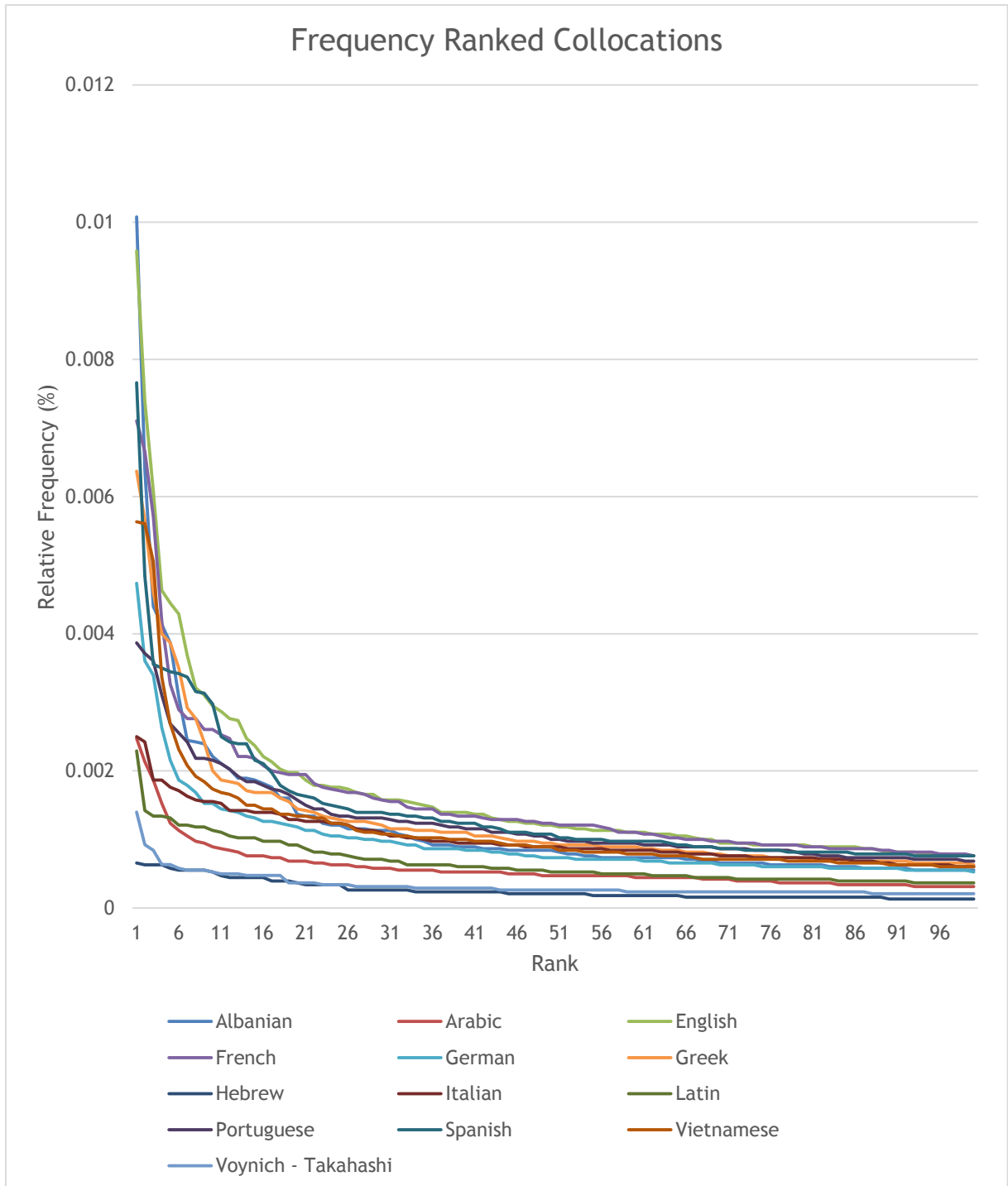


Figure 5-1: Frequency Ranked (1-100) Collocations

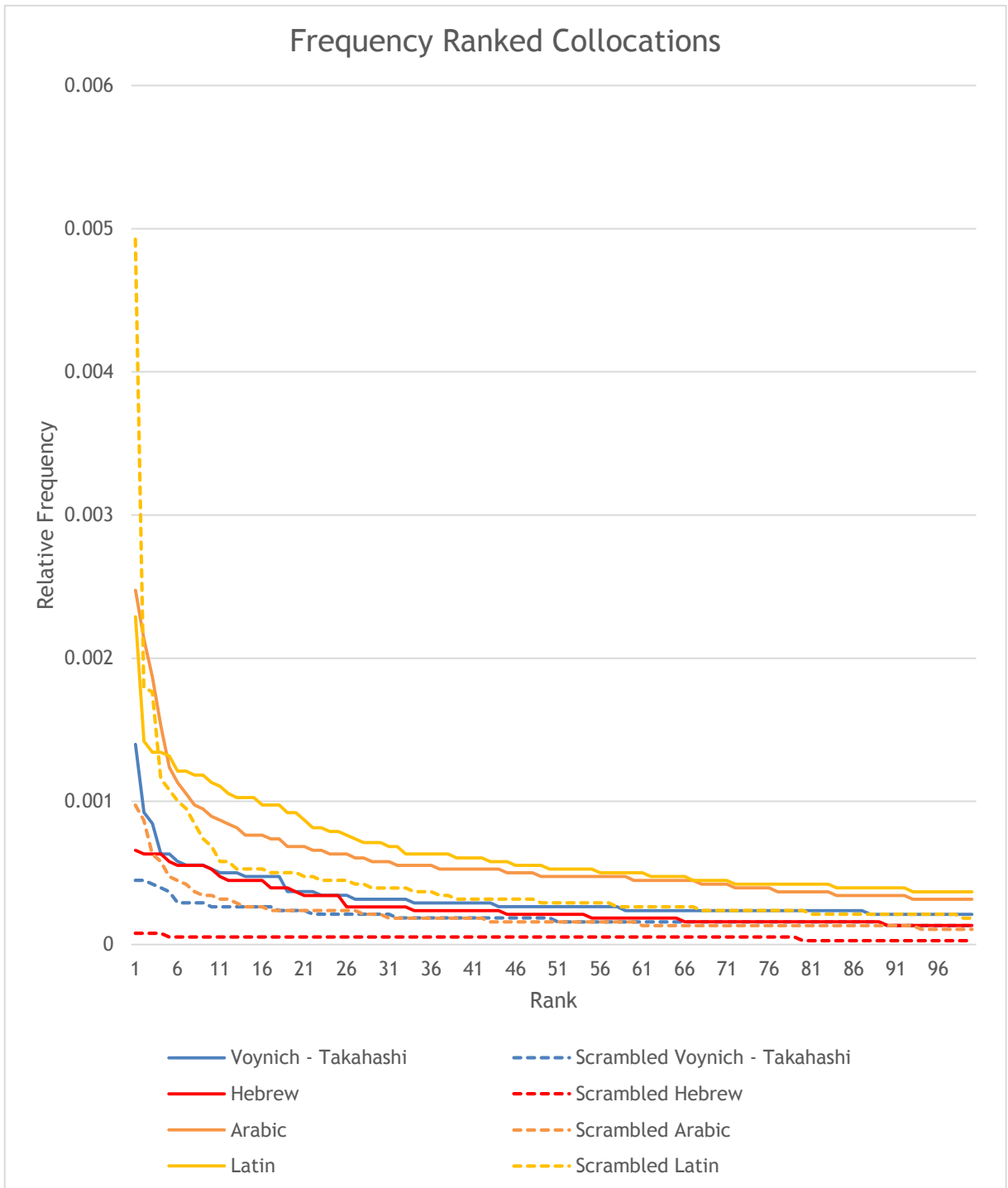


Figure 5-2: Frequency Ranked (1-100) Collocations of Scrambled Texts closest to the Voynich

This final graph in Figure 5-3 shows the results obtained with using the PMI metric to rank each of the collocations.

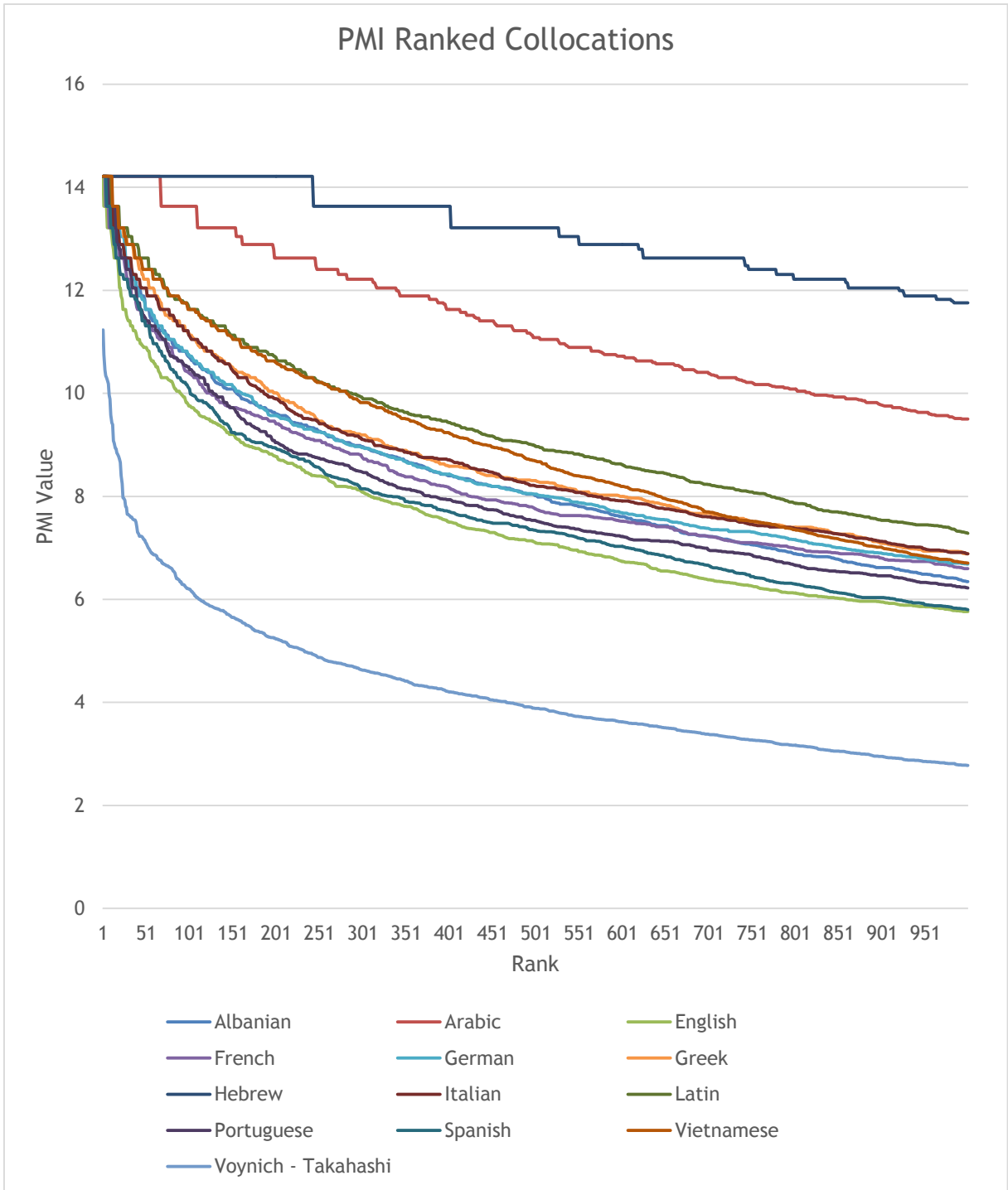


Figure 5-3: PMI Ranked (1-1000) Collocations

5.5 Discussion

From the frequency ranking graph in Figure 5-1 it can be clearly seen that the Voynich has a very low measure of word association when compared to most other languages. However, it can also be seen that initially, the Voynich has a higher measure of word association than that of Hebrew before becoming very similar as the ranks increase. Note that while this measure does show an interesting result, and can be considered reasonably viable using simple frequency as the metric [28], it can only show a very limited number of rankings meaning it only shows a small representation of the entire data.

The second frequency graph shows particular languages and the Voynich Manuscript that displayed a possible relationship and their corresponding scrambling of the same text. As shown in research, weak word association may be the result of a hoax from random placement of words (gibberish) [11] or the scrambling of a text [2]. If the Voynich Manuscript is either of these the scrambling of the text may not have any significant effect on ranking curve. However the scrambling of the text does show a minor drop in the curve albeit with a generally smaller difference than shown by the other languages suggesting that there is some level of association between words within the Voynich.

A particular issue with using frequency as a ranking method are function words within languages. Examining the collocations of English and the other languages show that many collocations involve function words. These words, such as ‘the’ or ‘and’, appear very frequently biasing the results. The choice to use pointwise mutual information as another metric for word association was to attempt to relieve the data of this bias and to give more viable results [28]. The drawback to this is that PMI gives a bias to less frequent words [7] [11] hence any words or collocations that appeared only once were ignored.

The PMI ranked graph in Figure 5-3 again shows that the Voynich Manuscript has a very weak measure of word association, far below that of any of the other languages. By accounting for the function word bias in the frequency ranking, the graph also shows that Hebrew has a high measure of word association unlike in the frequency ranking graph. This result for Hebrew is significantly different than that obtained by Shi and Roush [11]. A multitude of possible differences may have accounted for this difference including different exclusion parameters or even a different section of Hebrew. However, due to the much more flat curve and the representation of the words within the text file, there may be errors in the ranking of Hebrew and Arabic.

As the actual contents of the Voynich Manuscript are unknown, the results shown may be due to differing text domains [27] [7]. As the corpus is compiled from various different translations of the Old Testament, the results of the other languages are biased to that of biblical texts. This may allow for better comparisons of the known languages but may also have significantly different collocation statistics than those of other texts in different domains of the same languages.

5.6 Conclusion

Based on these results we can conclude that the Voynich Manuscript generally has a weaker measure of word association than that of the other tested languages. If only comparing using a simple frequency metric, the Voynich does however show a possible relationship to Hebrew or scrambled Arabic.

From a more general perspective the weak measure of word association may also be related to a hoax or a type of code that hides the word order [2] [11]. It may also be due to lax spelling due to less standardized written language. English itself went through many linguistic changes throughout the 11th century to the 15th century where Old English may have been almost incomprehensible [30]. With the Voynich being carbon dated back to the 15th century it is possible that a non-standardized language or part thereof was used throughout the writing of the Voynich.

6 Discussion

The results of the experiments above show some interesting results, albeit very basic in the world of linguistics. From the basic characterisation of the text in Section 2, the Voynich Manuscript does appear to follow Zipf's Law giving plausibility to it being a natural language. However the data also shows a binomial distribution of word lengths which can contradict Zipf's Law as this may suggest the Voynich is a type of cipher or code and not a natural language. As shown in other research, it can also be related to a type of abjad.

The results obtained from the addition of character bigram data showed that single characters can be categorised into alphabet and non-alphabet characters through the use of statistics. It also highlighted that the writing-style can affect the overall data obtained from a text, even if in the same language, and required careful thought to keep all the texts within the same style when making comparisons. While the final extraction and categorisation of the basic Voynich characters may not have returned meaningful results, it does suggest that single characters may not distinctly represent punctuation, in a traditional sense, or numerical tokens. A different approach may lead to different results.

The affix extraction showed that a naïve approach to affixes, particularly suffixes, allowed for a simple comparison between the Voynich and other languages. Possible relationships were found between the Voynich and French, Greek and Latin but does not give any definitive conclusions. Previous research seemed to suggest that morphology needed significantly larger sample sizes than what was available from the Voynich and would require a deeper knowledge of the language. Unfortunately the significant complexity of the morphological structure of words and how it could vastly differ between languages at a much lower level was beyond the scope of this paper.

Collocations seemed to give simple comparative measures on the surface, showing that Hebrew and scrambled Arabic may have a word association relationship to that of the Voynich. However, when utilizing a different association metric it was shown that the Voynich had no relationship with any of the tested languages. The language relationship results therefore depended on the choice of word association metric. The results did show that, regardless of the word association metric, the Voynich continued to have a weak measure of word association suggesting that it may be in some form of code or cipher that hides the word association or a form of hoax.

It can be clearly seen that no one linguistic property, on a basic level, can conclusively point to a definitive relationship or hypothesis in regards to the Voynich Manuscript. Nevertheless these basic results can help point towards certain areas and languages for future research. Obviously more in-depth research is required.

7 Conclusion

The paper shows that by using simple statistical measures found within written texts, it is possible to indicate possible linguistic relationships between the Voynich and different linguistic properties of other languages.

It is seen that while the Voynich appears to follow Zipf's Law, suggesting it is a natural language, the binomial distribution of the word lengths also suggest it may be a type of code, cipher or abjad.

Basic character and bigram frequencies can be used to identify possible alphabet and non-alphabet characters but can be influenced by the sample size and the writing-style of a given text.

Using a simple affix definitions it is determined that the Voynich may have weak relationships with French, Greek and Latin. This however relies on the simple, restrictive definition for affixes which, in terms of morphological structure, isn't necessarily simple nor as restrictive.

Similarly, a simple definition of collocations shows possible relationships between the Voynich and Hebrew. As with the affixes, this also relies on a simple definition. This also showed that, even with the same definition, the metric used to rank the collocations could greatly vary the relationships between languages but did keep a very low word association measure for the Voynich. This again suggests that the Voynich may be a type of code, cipher or even a hoax.

Without much more in-depth research and testing, the relationships found using simple statistical measures lack conclusive evidence. The results found for each different linguistic property tested showed features that could be related to multiple languages or hypotheses, narrowing down possible options for future research.

8 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," in *LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2011.
- [3] G. Landini, "Evidence of Linguistic Structure in the Voynich Manuscript using Spectral Analysis," in *Cryptologia*, 2001, pp. 275-295.
- [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*, vol. 8, no. 7, pp. 1-10, 2013.
- [6] Unite for Sight, "Importance of Quality Sample Size," Unite for Sight, [Online]. Available: <http://www.uniteforsight.org/global-health-university/importance-of-quality-sample-size>. [Accessed 10 October 2015].
- [7] A. Thanopoulos, N. Fakotakis and G. Kokkinakis, "Comparative Evaluation of Collocation Extraction Metrics," *LREC*, vol. 2, pp. 620-625, 2002.
- [8] K. W. Church and W. A. Gale, "A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams," *Computer Speech & Languages*, vol. 5, no. 1, pp. 19-54, 1991.
- [9] J. Lyons, "Practical Cryptography - Text Characterisation," [Online]. Available: <http://practicalcryptography.com/cryptanalysis/text-characterisation/>. [Accessed 8 October 2015].
- [10] 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.
- [11] B. Shi and P. Roush, "Semester B Final Report 2014 - Cracking the Voynich Code," University of Adelaide, Adelaide, 2014.
- [12] S. T. Piantadosi, "Zipf's word frequency law in natural language: a critical review and future directions," 2015.
- [13] R. L. Solso and C. L. Juel, "Positional frequency and versatility of bigrams for two-through nine-letter English words," *Behaviour Research Methods & Instrumentation*,

vol. 12, no. 3, pp. 297-343, 1980.

- [14] M. N. Jones and D. J. K. Mewhort, "Case-sensitive letter and bigram frequency counts from large-scale English corpora," *Behaviour Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 388-396, 2004.
- [15] G. Booij, *The Grammar of Words: An Introduction to Linguistic Morphology*, Oxford: Oxford University Press, 2012.
- [16] S. R. Anderson, "Encyclopedia of Cognitive Science - Morphology," Macmillan Reference, Ltd., [Online]. Available: http://cowgill.ling.yale.edu/sra/morphology_ecs.htm. [Accessed 8 October 2015].
- [17] R. H. Baayen, "Corpus linguistics in morphology: morphological productivity," in *Corpus Linguistics. An International Handbook*, 2009, pp. 900-919.
- [18] M. G. Snover and M. R. Brent, "A Probabilistic Model for Learning Concatenative Morphology," *Advances in Neural Information Processing Systems*, pp. 1513-1520, 2002.
- [19] O. Kohonen, S. Virpioja and K. Lagus, "Semi-supervised learning of concaenative morphology," in *11th Meeting of the ACL-SIGMORPHON, ACL 2010*, Uppsala, 2010.
- [20] H. Hammarström, "A Naive Theory of Affixation and an Algorithm for Extraction," in *SIGPHON '06 Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology*, 2006.
- [21] G. Eryiğit and E. Adalı, "An Affix Stripping Morphological Analyzer for Turkish," in *IASTED International Conference, Artificial Intelligence and Applications*, Innsbruck, 2004.
- [22] G. Minnen, J. Carroll and D. Pearce, "Applied morphological processing of English," *Natural Language Engineering*, vol. 7, no. 3, pp. 207-223, 2001.
- [23] J. Stolfi, "Voynich Manuscript stuff," 23 May 2005. [Online]. Available: <http://www.ic.unicamp.br/~stolfi/EXPORT/projects/voynich/Welcome.html>.
- [24] 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/>.
- [25] 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/>.
- [26] I. A. Mel'čuk, "Collocations and Lexical Functions," in *Phraseology. Theory, Analysis, and Applications.*, Oxford, Clarendon Press, 1998, pp. 23-53.
- [27] F. Smadja, "Retrieving Collocations from Text: Xtract," *Computational Linguistics*, vol. 19, no. 1, pp. 143-177, 1993.

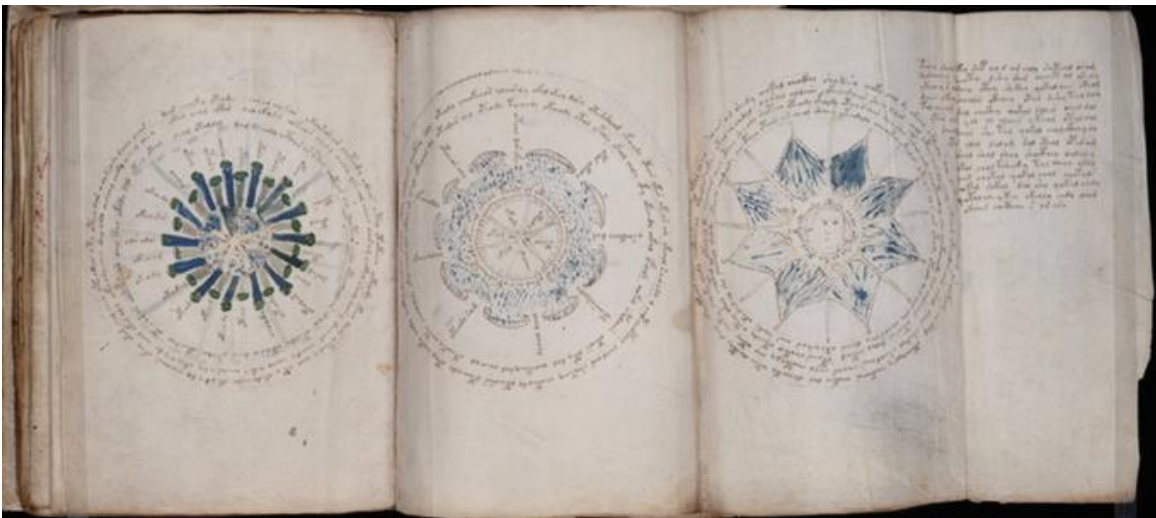
- [28] J. Wermter and U. Hahn, "You Can't Beat Frequency (Unless You Use Linguistic Knowledge) - A Qualitative Evaluation of Association Measures for Collocation and Term Extraction," in *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Stroudsburg, 2006.
- [29] D. Pearce, "A Comparative Evaluation of Collocation Extraction Techniques," in *LREC*, 2002.
- [30] L. Mastin, "The History of English - Middle English," 2011. [Online]. Available: http://www.thehistoryofenglish.com/history_middle.html. [Accessed 14 October 2015].
- [31] "The Voynich Manuscript," 22 March 2015. [Online]. Available: <https://archive.org/details/TheVoynichManuscript>.
- [32] R. Zandbergen, "Analysis of the text," 13 April 2015. [Online]. Available: <http://www.voynich.nu/analysis.html>.

Appendix A: The Voynich Manuscript Examples

The following images are of the Voynich Manuscript. These images have been reproduced from the Internet Archive [31]. 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



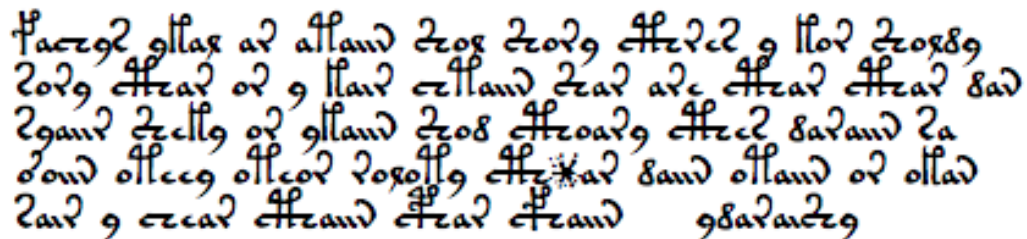
The pharmaceutical section, folios 87r - 102v.



The recipes section, folios 103r - 116v.

Appendix B: The Interlinear Archive

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 [32].



Faczge gllax ar alland zoz zozg cthreel g llor zozdg
zozg cthreel or g llard cthreel zrad are cthreel cthreel sad
zganr zrellg or glland zoz cthreel cthreel sadand za
dond ollecg ollecg rozollg cthreel sad and olland or ollad
zard g cthreel cthreel cthreel cthreel ydaraiShy

fachys ykal ar atain Shol Shory cThres y kor Sholdy
sory cThar or y kair chtain Shar are cThar cThar dan
syair Sheky or ykain Shod cThoary cThes darain sa
o'oiin oteey oteor rototy cTh*ar daiin otain or okan
sair y chear cThain cPhar cFhain ydaraiShy

Appendix C: The European Voynich Alphabet

The EVA as shown on René Zandbergen's website [32].

BASIC EVA CHARACTERS		
	EVA	Capitalised EVA
'	ʹ	
a	ᵃ	ᵃ
b	ᵇ	
c	ᶜ	
d	ᵈ	
e	ᵉ	ᵉ
f	ᶠ	ᶠ
g	ᵍ	
h	ᵂ	ᵂ
i	ᶦ	ᶦ
j	ᵋ	
k	ᵏ	ᵏ
l	ᶫ	
m	ᵐ	
n	ᵎ	
o	ᵒ	ᵒ
p	ᵖ	ᵖ
q	ᵑ	
r	ᵣ	
s	ᵈ	ᵈ
t	ᵗ	ᵗ
u	ᵘ	
v	ᵛ	
x	ˣ	
y	ʸ	ʸ
z	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"		
"	ᶘ	plume on top of connector
+	ᶙ	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	,	𐌶	-	-	-	190	¼	𐌹	
131	f	𐌶	161	i	𐌶	191	ı	𐌺	
132	„	𐌶	162	¢	𐌶	192	À	𐌺	
133	...	𐌶	163	£	𐌶	193	Á	𐌺	
134	†	𐌶	164	¤	𐌶	194	Â	𐌶	
135	‡	𐌶	165	¥	𐌶	195	Ã	𐌶	
136	^	𐌶	166	¦	𐌶	196	Ä	𐌶	
137	‰	𐌶	167	§	𐌶	197	Å	𐌶	
138	Š	𐌶	168	¨	𐌶	198	Æ	𐌶	
139	‹	𐌶	169	©	𐌶	199	Ç	𐌶	
140	œ	𐌶	170	ª	𐌶	200	È	𐌶	
141	□	𐌶	171	«	𐌶	201	É	𐌶	
142	□	𐌶	172	¬	𐌶	202	Ê	𐌶	
143	□	𐌶	173	-	𐌶	203	Ë	𐌶	
144	□	𐌶	174	®	𐌶	204	Ì	𐌶	
145	‘	𐌶	175	¯	𐌶	205	Í	𐌶	
146	’	𐌶	176	°	𐌶	206	Î	𐌶	
147	“	𐌶	177	±	𐌶	207	Ï	𐌶	
148	”	𐌶	178	²	𐌶	208	Ð	𐌶	
149	•	𐌶	179	³	𐌶	209	Ñ	𐌶	
150	-	𐌶	180	´	𐌶	210	Ò	𐌶	
151	—	𐌶	181	µ	𐌶	211	Ó	𐌶	
152	~	𐌶	182	¶	𐌶	212	Ô	𐌶	
153	™	𐌶	183	·	𐌶	213	Ö	𐌶	
154	š	𐌶	184	,	𐌶	214	Ï	𐌶	
155	›	𐌶	185	¹	𐌶	215	×	𐌶	
156	œ	𐌶	186	º	𐌶	216	Ø	𐌶	
157	□	𐌶	187	»	𐌶				
158	□	𐌶	188	¼	𐌶				
159	ÿ	𐌶	189	½	𐌶				

Appendix D: Interlinear Archive Processing Output Example

The following gives an example of the pre-processing that was completed on the Interlinear Archive.

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.okey!y-{plant}  
<f17v.P.4;F>    kchor.fchol.cphol.olcheol.okeeey-{plant}
```



Processed File for H

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

Appendix E: Corpora

The Table E.1 below shows the texts used within the English corpus.

Text Name	Author	Obtained From
The Merry Adventures of Robin Hood	Howard Pyle	Project Gutenberg https://www.gutenberg.org
Macbeth	William Shakespeare	Project Gutenberg https://www.gutenberg.org
The New Testament (King James)	Various	Project Gutenberg https://www.gutenberg.org
An Account of the Foxglove and its Medical Uses	William Withering	Project Gutenberg https://www.gutenberg.org
The Story of the Heavens	Robert S. Ball	Project Gutenberg https://www.gutenberg.org

Table E.1: English Corpus

The Table E.2 below shows the various different languages used within the Old Testament corpus. All utilised the first sections of the Old Testament until a total word count of 38000 was reached.

Language	Obtained From
Albanian	The Unbound Bible Project https://unbound.biola.edu
Arabic	The Unbound Bible Project https://unbound.biola.edu
English	The Unbound Bible Project https://unbound.biola.edu
French	The Unbound Bible Project https://unbound.biola.edu
German	The Unbound Bible Project https://unbound.biola.edu
Greek	The Unbound Bible Project https://unbound.biola.edu
Hebrew	The Unbound Bible Project https://unbound.biola.edu
Italian	The Unbound Bible Project https://unbound.biola.edu
Latin	The Unbound Bible Project https://unbound.biola.edu
Spanish	The Unbound Bible Project https://unbound.biola.edu
Vietnamese	The Unbound Bible Project https://unbound.biola.edu

Table E.2: Language Comparison Corpus

Appendix F: Example Bigram Table

The table below shows a small subsection of a bigram table used in the English Investigation of the New Testament.

<i>Bigram</i>	<i>Frequency</i>	<i>Relative Frequency</i>
---------------	------------------	---------------------------

a,	1153	0.04%	b,	466	0.02%	c,	47	0.00%
a-	0	0.00%	b-	0	0.00%	c-	0	0.00%
a.	212	0.01%	b.	125	0.00%	c.	13	0.00%
a0	0	0.00%	b0	0	0.00%	c0	0	0.00%
a1	0	0.00%	b1	0	0.00%	c1	0	0.00%
a2	0	0.00%	b2	0	0.00%	c2	0	0.00%
a3	0	0.00%	b3	0	0.00%	c3	0	0.00%
a4	0	0.00%	b4	0	0.00%	c4	0	0.00%
a5	0	0.00%	b5	0	0.00%	c5	0	0.00%
a6	0	0.00%	b6	0	0.00%	c6	0	0.00%
a7	0	0.00%	b7	0	0.00%	c7	0	0.00%
a8	0	0.00%	b8	0	0.00%	c8	0	0.00%
a9	0	0.00%	b9	0	0.00%	c9	0	0.00%
a:	111	0.00%	b:	79	0.00%	c:	5	0.00%
a;	99	0.00%	b;	47	0.00%	c;	9	0.00%
a?	11	0.00%	b?	15	0.00%	c?	0	0.00%
aA	0	0.00%	bA	0	0.00%	cA	0	0.00%
aB	0	0.00%	bB	0	0.00%	cB	0	0.00%
aC	0	0.00%	bC	0	0.00%	cC	0	0.00%
aD	0	0.00%	bD	0	0.00%	cD	0	0.00%
aE	0	0.00%	bE	0	0.00%	cE	0	0.00%
aF	0	0.00%	bF	0	0.00%	cF	0	0.00%
aG	0	0.00%	bG	0	0.00%	cG	0	0.00%
aH	0	0.00%	bH	0	0.00%	cH	0	0.00%
aI	0	0.00%	bI	0	0.00%	cI	0	0.00%
aJ	0	0.00%	bJ	0	0.00%	cJ	0	0.00%
aK	0	0.00%	bK	0	0.00%	cK	0	0.00%
aL	0	0.00%	bL	0	0.00%	cL	0	0.00%
aM	0	0.00%	bM	0	0.00%	cM	0	0.00%
aN	0	0.00%	bN	0	0.00%	cN	0	0.00%
aO	0	0.00%	bO	0	0.00%	cO	0	0.00%
aP	0	0.00%	bP	0	0.00%	cP	0	0.00%
aQ	0	0.00%	bQ	0	0.00%	cQ	0	0.00%
aR	0	0.00%	bR	0	0.00%	cR	0	0.00%
aS	0	0.00%	bS	0	0.00%	cS	0	0.00%
aT	0	0.00%	bT	0	0.00%	cT	0	0.00%
aU	0	0.00%	bU	0	0.00%	cU	0	0.00%
aV	0	0.00%	bV	0	0.00%	cV	0	0.00%
aW	0	0.00%	bW	0	0.00%	cW	0	0.00%
aY	0	0.00%	bY	0	0.00%	cY	0	0.00%

Appendix G: Example Character Frequency Table

The table below shows a subsection of a character frequency table used in the English Investigation of the New Testament.

6	6465	0.19%
7	5911	0.17%
8	5769	0.17%
9	5632	0.16%
A	17842	0.51%
B	4670	0.13%
C	1661	0.05%
D	8740	0.25%
E	2585	0.07%
F	2326	0.07%
G	6093	0.17%
H	3181	0.09%
I	13202	0.38%
J	6364	0.18%
K	535	0.02%
L	9168	0.26%
M	3032	0.09%
N	1835	0.05%
O	8840	0.25%
P	1770	0.05%
Q	5	0.00%
R	7500	0.22%
S	4837	0.14%
T	7659	0.22%
U	290	0.01%
V	99	0.00%
W	2394	0.07%
Y	541	0.02%
Z	904	0.03%
a	256886	7.37%
b	43921	1.26%
c	52813	1.52%
d	148834	4.27%
e	407830	11.70%
f	80834	2.32%
g	48788	1.40%
h	278976	8.01%

<i>Character</i>	<i>Frequency</i>	<i>Relative Frequency</i>
------------------	------------------	---------------------------

Appendix H: Basic Character Frequencies Comparison

The following figure gives a basic frequency comparison of the lower-case alphabet and numerical tokens found in the New Testament and Robin Hood texts.

