

An Analysis of the Relationship between Words within the Voynich Manuscript

Andrew Caruana¹, Colin Layfield¹ and John Abela¹

¹Department of Computer Information Systems, Faculty of ICT, University of Malta, Msida, Malta

Abstract

This paper investigates the presence of linguistic structure within the Voynich Manuscript by analysing various properties of word-pairs found in the manuscript as well as in other works written in natural languages. Apart from the Voynich, the other manuscripts we analysed are the Bible, Dante's *La Divina Commedia*, and Shakespeare's *Macbeth* and *Julius Caesar*. The order of words in the word-pairs in these manuscripts is analyzed and this analysis indicated that several pairs were more likely to appear in one order than the other. These are called 'skewed pairs'. The ratio of the number of skewed pairs in each work is to the number of all pairs in each work is plotted, along with this same ratio but for random shuffles of each work. The results indicated that there is a substantial difference in all natural language documents between their normal and shuffled counterparts. The difference is not as large within the Voynich Manuscript but the word-pair occurrence ratio of the original is still considerably higher than the ratio of the shuffled manuscript. This could indicate that the Voynich Manuscript is not random text but may be a language or a cipher.

Keywords

Voynich, word-pairs, natural language, linguistic structure, skewed pairs

1. Introduction

1.1. Background

The Voynich Manuscript (VM) is a medieval manuscript written in a language that, so far, has not been identified or successfully deciphered. It is named after Wilfrid Voynich, a Polish rare book dealer who bought the VM along with other rare manuscripts in 1912. Its origin can be traced as far back as the 15th century and eventually made its way into the hands of Voynich. The manuscript contains numerous illustrations of what appear to be alien plants, zodiac signs and astronomical depictions. What makes this manuscript perplexing however, is the fact that the language it is written in (known as Voynichese) is completely unique and has not been seen in any other work presently discovered. Thus, this makes the manuscript a topic of interest among many scholars throughout various fields of academia as well as independent Voynich Researchers [1, 2, 3].


International Conference on the Voynich Manuscript 2022, November 30–December 1, 2022, University of Malta.

✉ andrew.v.caruana.19@um.edu.mt (A. Caruana); colin.layfield@um.edu.mt (C. Layfield); john.abela@um.edu.mt (J. Abela)

ORCID 0000-0002-1868-4258 (C. Layfield)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

1.2. Aims and Motivation

The main objective of this work is to investigate, and analyse, word-pair occurrences in the Voynich and other manuscripts (some contemporary to the Voynich) in order to determine if the word-pair occurrence patterns in the other manuscript are also found in the Voynich. There are currently three main hypotheses that are considered by Voynich researchers - that the manuscript is a hoax, a cipher, or a natural language.

In this work, word-pairs within the manuscript and other medieval works are analysed in order to determine whether certain word-pairs appear more or less frequently than their inverses. If many pairs are found to satisfy this pattern in the known natural language manuscripts, the same analysis is then done on the Voynich Manuscript to see whether the same results can be observed. If there is a similarity between the natural language documents and the VM, it might indicate that the VM has some sort of linguistic structure to it.

1.3. Previous Work

Previous work has been done on the Voynich to try identify the nature of the text within the manuscript. Some scholars, such as Schinner, who published an article analysing the manuscript using random walk mapping and token/syllable repetition statistics, believe that the manuscript is a hoax [4]. That being said, many scholars support the natural language hypothesis, such as Landini, who performed spectral analysis on the manuscript and other natural language works and found similarities [5] and also Bax, who claimed to have deciphered some of the manuscript by comparing herbal illustrations to their believed real-life counterparts [6].

Furthermore, similar work to this paper involving co-occurrence graphs has been done on the manuscript before, namely by Montemurro who used them to determine a list of words with greater semantic meaning [7]. The results of this paper seem to indicate similarities to natural language documents and show some evidence of a linguistic structure present in the manuscript.

These works above served as a good foundation and guide for the work done in this paper both in terms of content and how it is written.

2. Methodology & Implementation

This section describes the pre-processing of all the data, the generation the word-pair co-occurrence graphs, and the extraction and formatting of the results.

2.1. Initial Setup & Data Preprocessing

Python was chosen as the ideal programming language for this task, due to its plentiful external libraries in addition to being very popular in the fields of Artificial Intelligence and Data Processing. Four main external libraries are used, the first of which being the natural language

toolkit (nltk) [8], a Python package which includes many powerful computational linguistics algorithms. In addition, the following libraries were also used: voynich [9], graphviz [10], matplotlib [11] and numpy [12].

The first step was to obtain natural language datasets of various manuscripts. The datasets chosen for this are books from the King James Bible (Genesis, Exodus, The Book of Job, The Gospel According to Saint Matthew and The Acts of the Apostles), Dante's La Divina Commedia and Shakespeare's Julius Caesar and Macbeth. It should be noted that Caesar, Macbeth and the books from the Bible are written in Old English and La Divina Commedia is in Old Italian. Furthermore, those particular books from the Bible were chosen to emulate the different sections of the manuscript and their unclear relation between said sections. The rest of the datasets were chosen to incorporate another language and due to ease of acquisition since they were present in *nltk*.

With regards to the VM, it is processed using the Voynich library previously mentioned. However, modifications were done on a local version of the library to modify the section labelling to better suit their illustration type. It should also be noted that the transliteration used is the ZL transliteration version 1r [1].

Following this, the text in the natural language documents is cleaned up by removing digits, punctuation and converting all the text to lowercase. Then, the stop words were removed for all documents via a frequency analysis process. The algorithm removes the top X percent of words ranked by their frequency. A suitable value for X is found through manual experimentation and trial and error for English and Italian, those being 0.001% and 0.002% respectively. The value of X for the Voynich manuscript is taken as the average of the two, since in this case, X could not be found in the same manner.

2.2. Graph Generation

After all of the pre-processing is completed, word-pair co-occurrence graphs were generated for each section of every document. This is done to serve as an underlying data structure and to allow the visualisation of the word pairs and their frequencies. This is done using the *graphviz* library.

The graphs were generated using a sliding window to find and associate words with each other and form pairs, meaning that a word-pair may not necessarily consist of adjacent words, but words found within a certain distance (in number of words) from each other. For this work, a window of size 3 was selected after experimenting with different values. A window of size 4 proved to be too large as a lot of unrelated words started being labelled as word-pairs, and the inverse was true with a size of 1 or 2. Furthermore, a window of size 3 means it looks 3 words before and 3 words after a given word. An example of a co-occurrence graph can be seen in Figure 1. Each node in the graph represents a word, and each edge between nodes indicates that those words form a pair. The number seen above each edge is the pair's frequency and the direction indicates the order of the word-pair. It should also be noted that the graph seen

in Figure 1 is not the full graph, but rather a snippet, as the full one is too large to show in its entirety.

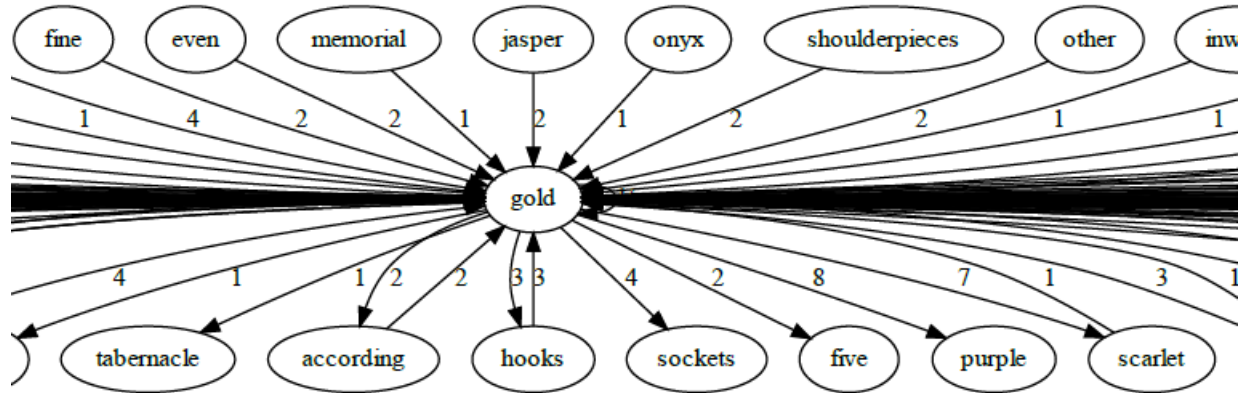


Figure 1: Co-Occurrence Graph for "gold" in the Book of Exodus in the Bible

2.3. Skewed Pairs Calculations

Let P be a word-pair (A, B) and let Q be its inverse (B, A) where A and B are both words. This can be denoted as:

$$P = [A, B]$$

$$Q = [B, A]$$

. The term *skewed pair* is used to denote that case where P appears much more frequently than its inverse Q . This would then be compared to a chosen frequency threshold T in order allow for the removal of lower scoring skewed pairs. This can be shown as:

$$F_P - F_Q > T$$

Where F_P is the frequency of P in the section and F_Q is the frequency of Q in this section. For a pair to be marked as a skewed pair, it would have to follow this condition. For this work, T was set to 5 after experimenting with higher and lower values as most documents other than the Bible did not have many skewed pair occurrences greater than 12, hence setting it to 5 would provide a good balance of quantity of pairs and relevance of pairs.

This concept however, is better illustrated with an example. Take the pair $[brown, dog]$, the word *brown* is much more likely to appear before the word *dog* than vice versa because that is how English is structured. Since English has this kind of pattern, it would be interesting to see if this applies to other languages and the Voynich Manuscript itself.

Since the skewed pairs could now be identified, the next step is to generate them for each document and plot the results.

3. Results & Evaluation

In this section we interpret the results obtained. This includes the skewed pairs ratio, the ratio of unique words and the connections per word.

3.1. Ratio of Skewed Pairs

Taking the skewed pairs calculations mentioned in the previous section, a graph is plotted with the ratio of the number of skewed pairs to the amount of pairs in each document. This is done for the normal document and for 3 random shuffles of each document in order to take an average. The initial hypothesis was that there would be a higher ratio in regular, ordered, natural languages rather than random shuffles, hence, this could provide some evidence in favour or against the hoax hypothesis of the Voynich Manuscript. This graph is plotted and can be seen in Figure 3.

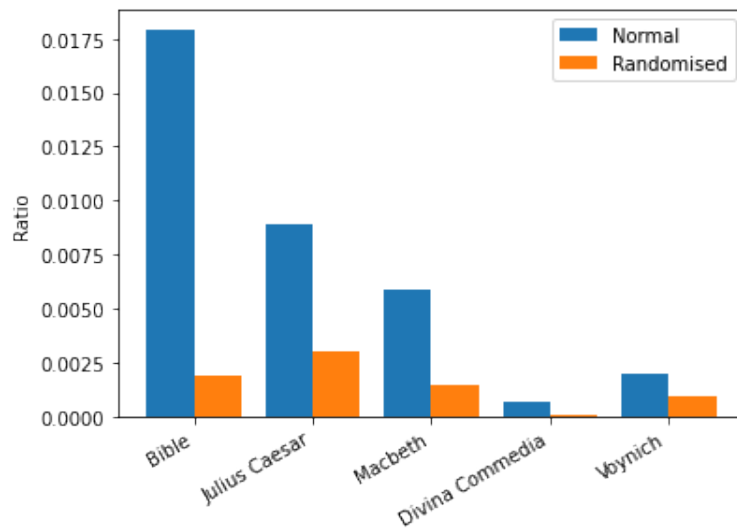


Figure 2: Ratio of Skewed Pairs among Documents

This graph proved to be quite interesting as there is, as was expected, a significant difference between ratios of normal texts and randomised texts in Old English and Old Italian as expected. However, with the Voynich, the ratio of skewed pairs for the normal document is only slightly more than double that of the random shuffled document. However, this gap for the Voynich text variants is still large enough to imply that there is some sort of linguistic structure within the manuscript.

3.2. Ratio of Unique Words

The results of skewed pair calculations led to investigation of the ratio of unique words. This can be defined as the ratio of the number of words used in a document without repetition is

to the total word count of said document. This utilised randomly selected text segments from each document (each of 10000 words in length to avoid any bias brought upon by Heaps' Law [13]) and aimed to find a reason why works like the Bible had so many skewed pairs whereas works like Divina Commedia had so few. Due to this random selection, the calculations were run three times using different randomly generated segments in order to minimise bias. The average of these three runs is then calculated and plotted.

What these results (see Figure 3) seem to indicate is that there is a roughly inverse relation between the ratio of unique words and the ratio of skewed pairs. The reason this may be the case is that if a language has fewer unique words in a given document, the more likely the document contains more repeated words, and hence the more likely for those repeated words to form skewed pairs.

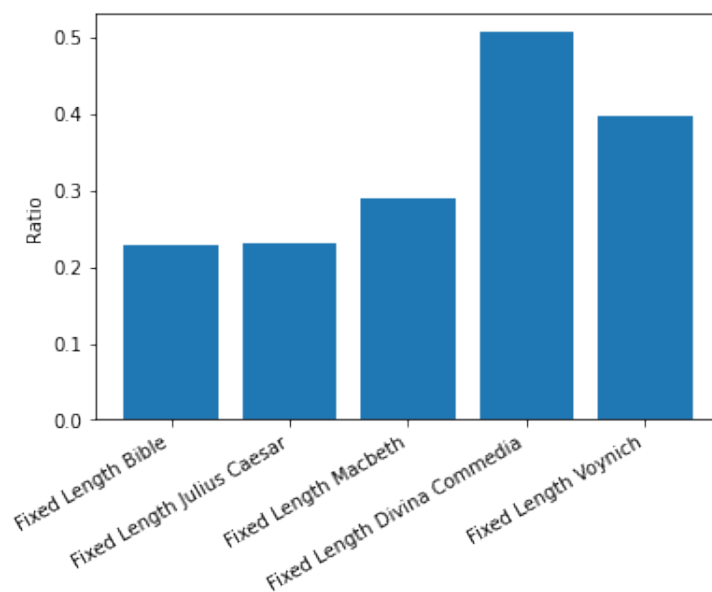


Figure 3: Ratio of Unique Words

3.3. Connections Per Word

As a final statistic, the number of connections each word has, meaning how many word-pairs each word is in, is investigated for each whole document.

These results, show an exponential decay in each instance, which is similar to Zipf's Law of Word Frequencies discussed previously. It should also be noted that each of these graphs was plotted with the most frequent words removed as an X percent of them were taken as stop words as described previously. Furthermore, the word connections in the Voynich Manuscript followed the same patterns as the natural languages, which may suggest that it is not a hoax.

Moreover, the number of connections for La Divina Commedia is much lower than the other documents and is much less smooth. This could tie into its low score in the skewed pairs ratio as if it does not have many connections per word, it is less likely that skewed pairs would appear. In Figure 4, one can see the results for La Divina Commedia and the Voynich Manuscript.

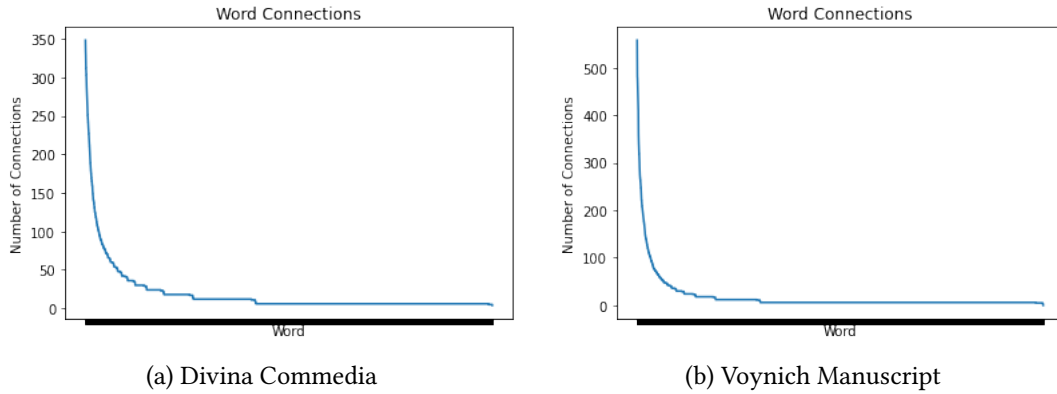


Figure 4: Connections per Word

4. Conclusions

This section concludes this paper and discusses the limitations of this work, along with potential future work.

4.1. Limitations

Whilst the aims for this work were successfully accomplished, there are a few limitations present. The first major one being the fact that the Voynich was only compared to two other natural languages, Old English and Old Italian. It is possible that the Voynich is a natural language but has nearly nothing in common with them, but rather has similarities to other languages. However, comparing the manuscript to many more languages would be rather complex, especially languages with very different structures such as Japanese or Chinese. Furthermore, it would be rather tedious to try this with every family of language due to the sheer quantity of them and the lack of available resources for low-resource languages such as Maltese.

In addition, another limitation is the fact that the stop word detection had to be done automatically. This is mentioned in previous sections, however since we do not know what the text in the Voynich means, it is impossible to compile a list of definite stop words based on meaning. Due to this, stop words were detected and removed based on frequency, which is undoubtedly less accurate than using a pre-determined list.

4.2. Future Work

This work could be expanded upon by taking the same skewed pair and unique word calculations but adding a third category: ciphers. To put it simply, given any document, one would perform the same calculations but on several variants, including some which have been encoded using certain cryptographic techniques known at the time. This would be done for several documents including the Voynich manuscript itself, and results of this experiment might be able to shed some light on whether the Voynich is a cipher or not. Moreover, a category of artificial languages, such as programming languages, could be added to observe what kind of score these languages would obtain and compare them to the scores of the other categories.

In addition, one could conduct further experimentation with utilising alternate methods of stop word detection rather than just frequency, such as methods described by Zou et al. at the City University of Hong Kong [14]. On top of that, methods such as TF-IDF could also be applied to these texts to filter out words depending on their relevance to the document.

4.3. Final Remarks

The results obtained from this work seem to indicate that the Voynich Manuscript is not a hoax, however, it must be reiterated that this does not prove that the Voynich is a natural language, but it can only lend evidence to one hypothesis or another.

In conclusion, whilst the Voynich manuscript still remains shrouded in mystery, this work has helped produce some evidence against the hoax hypothesis. One can only hope that somewhere down the line, this work is developed even further to attempt to uncover more of the Voynich Manuscript's hidden secrets.

References

- [1] R. Zandbergen, The Voynich Manuscript, 2021. URL: <http://voynich.nu/>, accessed on 2022-08-02.
- [2] S. Reddy, K. Knight, What We Know About The Voynich Manuscript, in: Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities, 2011, pp. 78–86.
- [3] M. E. d’Imperio, The Voynich Manuscript: An Elegant Enigma, National Security Agency/Central Security Service, 1978.
- [4] A. Schinner, The Voynich Manuscript: Evidence of the Hoax Hypothesis, *Cryptologia* 31 (2007) 95–107.
- [5] G. Landini, Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis, *Cryptologia* 25 (2001) 275–295.
- [6] S. Bax, A Proposed Partial Decoding of the Voynich Script, University of Bedfordshire. <http://stephenbax.net/wp-content/uploads/2014/01/Voynich-a-provisionalpartial-decoding-BAX.pdf> (2014).

- [7] M. A. Montemurro, D. H. Zanette, Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis, *PloS one* 8 (2013) e66344.
- [8] S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, " O'Reilly Media, Inc.", 2009.
- [9] J. Lee, voynich, 2022. URL: <https://github.com/jacoble628/voynich>, accessed on 2022-08-02.
- [10] S. Bank, voynich, 2013. URL: <https://graphviz.readthedocs.io/en/stable/>, accessed on 2022-08-02.
- [11] M. D. Team, matplotlib, 2003. URL: <https://matplotlib.org/>, accessed on 2022-08-02.
- [12] N. D. Team, numpy, 2005. URL: <https://numpy.org/>, accessed on 2022-08-02.
- [13] L. Egghe, Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments, *Journal of the American Society for Information Science and Technology* 58 (2007) 702–709.
- [14] F. Zou, F. L. Wang, X. Deng, S. Han, L. S. Wang, Automatic Construction of Chinese Stop Word List, in: *Proceedings of the 5th WSEAS international conference on Applied computer science*, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point ..., 2006, pp. 1010–1015.