Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 74–78
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Word Probability Findings in the Voynich Manuscript

Colin Layfield[*], Lonneke van der Plas[†], Michael Rosner[‡], John Abela[*]
University of Malta
[*] Department of Computer Information Systems
[†] Institute of Linguistics and Language Technology
[‡] Department of Artificial Intelligence
Msida MSD2080, Malta
{colin.layfield, lonneke.vanderplas, mike.rosner, john.abela}@um.edu.mt

Abstract

The Voynich Manuscript has baffled scholars for centuries. Some believe the elaborate 15th century codex to be a hoax whilst others believe it is a real medieval manuscript whose contents are as yet unknown. In this paper, we provide additional evidence that the text of the manuscript displays the hallmarks of a proper natural language with respect to the relationship between word probabilities and (i) average information per subword segment and (ii) the relative positioning of consecutive subword segments necessary to uniquely identify words of different probabilities.

Keywords: Voynich Manuscript, Word Probabilities, Segment Information, Uniqueness Point

1. Introduction

The Voynich Manuscript (VM) is a codex or bound manuscript whose name derives from Wilfrid Michael Voynich, an antiquarian book dealer who purchased it in 1912 from the Jesuit Villa Mondragone in Frascati, near Rome. Recent radiocarbon tests at the University of Arizona have reliably dated the vellum to 1404-1438. The ink and colours used, although difficult to date directly, are not inconsistent with the time period nor suspicious (Stolte, 2011). It currently resides in the Beinecke Rare Book and Manuscript Library at Yale University as 'MS408'.

The physical manuscript is fairly modest upon first inspection, measuring about 10 inches high, 7 inches wide and about 2 inches thick (slightly larger than a typical modern paperback book). There is no indication of a title or an author for the work. The manuscript itself is made up of 116 numbered folios mostly of 2 pages with the exception of 10 foldouts of up to 6 pages most of which include both illustrations and text. VM comprises a total of about 35,000 words 170,000 characters written using between 24 and 30 letters of the unique VM alphabet[1], so it is clearly a very small corpus by modern standards (Zyats et al., 2016; Prinke and Zandbergen, 2017). An example of a page from the Herbal section, showing both the unusual text as well as drawings, can be found in Figure 1. Apart from these relatively concrete facts, very little is known about VM. The combination of illustrations and careful penmanship have led some researchers to suggest that VM is divided into sections devoted to astrology, cosmology, biology, pharmacology, herbs, and recipes (consisting of mostly text with star like 'bullet point' illustrations). Others have suggested that its overall purpose is to convey secrets of magic and alchemy. In short, there is no shortage of research that attempts or purports to unlock the secrets of this manuscript, but this does not fall into any coherent pattern of enquiry and is often of a highly speculative and/or subjective nature.

The authors believe that in order to make progress it is necessary to adopt a clearly articulated scientific approach in which goals, methodology and evidence are all clearly delimited. The present paper is a first step in that direction which provides some further evidence against theories which claim that VM is a hoax.

Figure 1: Page 16v from the Manuscript - Herbal Section (from Beinecke Library, accessed from https://archive.org/details/voynich)

[1]There is some debate around the number of individual characters since there appears to be some ligatures.

2. Background and Other Works

Mary D'Imperio, in her opening remarks at an early seminar on VM (when interest in it was renewed in the 1970s (Prinke and Zandbergen, 2017)) made the important observation that there was little agreement on the real nature of the document. She noticed that presenters classified it in one of five ways (D'Imperio, 1976):

- a natural language - not enciphered or concealed in any way but written in an unfamiliar script.
- a form of natural language but enciphered in some way.
- not a natural language at all, but rather a code or a synthetic language, like Esperanto, using a made up alphabet for further concealment.
- an artificial fabrication containing randomly generated meaningless padding, i.e. a hoax.
- completely meaningless doodling, produced by either a disturbed or eccentric person(s).

Knowledge of these classes provides some perspective for positioning research that has been carried out since. Thus the first 3 categories imply that the text has meaning and purpose, motivating attempts to "crack the Voynich code", whilst the last 2 negate the rationale for such efforts. Research that has been carried out can be roughly characterised under one or more of the following themes:

1. Character-level mapping
2. Word-level mapping and sentence interpretation
3. Investigations on statistical characteristics
4. Hoax-related investigations

The first theme is covered by work which aims to establish character-level correspondences with known writing systems or sounds. For example Bax (2014) exploited the fact that VM contains several examples of plant names adjacent to associated images. Through detailed micro-analysis matching sounds to symbols he proposed mappings for fourteen of the Voynich symbols used in ten words. Cheshire's work (Cheshire, 2019) not only proposes mappings for a larger set (33) of Voynich symbols but ventures into theme 2 by suggesting word mappings for certain sentences which are used to offer an unparalleled level of interpretation. The main problems here are that the samples are highly selective and justification for many of the assertions made is partial at best.

Work covering the third theme is often used to provide evidence for or against the fourth theme which is itself connected to the 5-way classification of VM mentioned earlier (e.g. if it is a fabrication it is also a hoax).

Experts are unsure whether the Voynich manuscript is written in some unknown language or is a hoax. Rugg (2004) claimed that the manuscript could have been written by constructing words from a grid of word prefixes, stems, and suffixes by means of a simple device known as a *Cardan grille* - an encryption tool used in the 16th century. Other researchers have proposed other hoax hypotheses.

Schinner (2007) attempted to show that the text was, statistically, consistent with stochastic text generation techniques similar to those proposed by Rugg. Not everyone agrees with Rugg and Schinner. Montemurro and Zanette (2013) conducted a study that shows that the text in the Voynich manuscript has similar word frequency distributions to text in natural languages. The authors claim that *"Here we analyse the long-range structure of the manuscript using methods from information theory. We show that the Voynich manuscript presents a complex organization in the distribution of words that is compatible with those found in real language sequences. These results together with some previously known statistical features of the Voynich manuscript, give support to the presence of a genuine message inside the book."*

Rugg and Taylor (2016) countered by stating than an "elaborate language" such as that in the Voynich manuscript can easily be created by using simple coding methods. At the moment there is disagreement on whether the Voynich manuscript is an elaborate hoax or whether it is a meaningful text in some code. This remains a hotly-debated topic amongst the experts.

Over the past 100 years or so, various researchers have applied a gamut of statistical analysis techniques. Many of these were used to find evidence that either supported or rejected the hoax hypothesis. Apart from Rugg, Montemurro, and Schinner, other researchers have used computational techniques to analyse, decipher, interpret, and to try to ultimately understand the manuscript.

In Mary D'Imperio's highly-cited book (D'Imperio, 1978), *The Voynich Manuscript: An Elegant Enigma*, she collected, analysed, and curated most of the research available up to that time.

Reddy and Knight (2011) investigated the VM's linguistic characteristics using a combination of statistical techniques and probabilistic models at page, paragraph, word and character levels. They found, *inter alia*, that VM characters within words were relatively more predictable than for English, Arabic, and Pinyin. Additional character-level analysis was performed by Landini (2001) and Zandbergen (2020) exploring topics such as entropy and spectral analysis of the text.

In 2015, McInnes and Wang (2015) published a comprehensive report on the application of statistical methods and data mining techniques that they used in order to discover linguistic features, relationship, and correlations in the Voynich text. The authors created an extensive, and comprehensive Wiki (Abbott, 2015) with all the results. A year later, Hauer and Kondrak (2016) proposed a suite of unsupervised techniques for determining the source language of text that has been enciphered with a monoalphabetic substitution cipher. The best method in the suite achieved an accuracy of 97% on the Universal Declaration of Human Rights in 380 languages. In the same paper the authors also present a novel approach to decoding anagrammed substitution ciphers that achieved an average decryption accuracy of 93% on a set of 50 ciphertexts. Where these methods were applied to the Voynich manuscript the results suggested Hebrew as the source language of the manuscript. This work has been criticised for not being scientifically

rigorous enough (Hauer and Kondrak, 2018).
As recently as June 2019, Zelinka et al. (2019) applied somewhat unorthodox, albeit very interesting, techniques to analyse the text in the manuscript. They concluded that their results indicated that the manuscript was likely written in a natural language since its fractal dimension was similar to that of Hemingway's novel, *The Old Man and the Sea*. The authors also reported that *complex network maps* (CNMs) generated from the Voynich manuscript were different from CNMs generated from random texts.

3. Motivation and Objectives

The main motivation for the programme of work we propose is to take stock of the diverse approaches towards the VM that have been taken so far and to investigate whether consistent application of solidly motivated computational techniques will advance our understanding in measurable ways.
The work reported in this paper focuses on theme 3, with implications for theme 4 as it shows further evidence for the claim that the VM has several characteristics of a natural language. The main novelty is the nature of the metric. King and Wedel (2020) have shown that there are certain patterns in the sequences of sounds and their position within word boundaries that are shared across a dataset of diverse languages. In particular, they demonstrate that less-probable words not only contain more sounds, they also contain sounds that convey more disambiguating information overall, and this pattern tends to be strongest at word-beginnings, where sounds can contribute the most information. We reproduced their experiments on the VM and found similar patterns.

4. Method

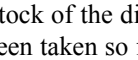
4.1. Data Used

The dataset used for the experiment is a transliteration file using the EVA (Extensible Voynich Alphabet) alphabet representation in the IVTFF (Intermediate Voynich Transliteration File Format). Version '1b' of the 'ZL' version of the file was used with version 1.5 of the IVTFF[2]. Only words that have been transcribed with a high degree of certainty were kept for our experiments (words with uncertain characters, character sequences or uncertain spaces were omitted). In total the transcription file contains 36,249 words of which 32,216 were retained for the work done here and, of those, 7,283 were unique (René Zandbergen, 2017).
It is noteworthy, at this point, to observe that the transliteration files available are evolving documents. These transliterations of the Voynich text are constantly being improved and modified to better reflect the content in the manuscript.

4.2. Approach

In order to investigate whether the relation between segment information and word probability follows a pattern similar to that found by King and Wedel (2020) across a large number of natural languages, we first computed the context-free word probabilities for all words retained from the transcription file, by dividing the counts for a given word by the total number of words as seen in Equation 1

$$p(word) = \frac{count(word)}{\sum_{word'} count(word')} \quad (1)$$

We also computed mean segment information for each word form up until the *uniqueness point* (Marslen-Wilson and Welsh, 1978) for that given word, that is, the point at which it is the only remaining word in the cohort starting with same sequence of segments. For example the Voychanese 'word'  has a uniqueness point of (5) as no other word in the Voychanese lexicon begins with those characters (in fact, the only other word, appearing once, that starts with the same 4 characters is).
The mean segment information calculation itself (token based) is calculated as seen in Equation 2:

$$h^*(seg_n) = -\log_2 \frac{count(seg_1 \dots seg_n) - count(word)}{count(seg_1 \dots seg_{n-1}) - count(word)} \quad (2)$$

It can be seen that the information for each segment of length n is the count of the first n segments (Voychanese characters) minus the total count of the word over the count of the segment that is one letter shorter minus the count of the word. The count of the word is removed to eliminate the correlation that the frequency of an entire word contributes to the calculation of the information of its segments.

5. Results

In Figure 2, we see the best-fit regression lines for mean token-based segment information by word probability, for word lengths four to eight[3], for corpora in five languages in addition to the VM[4]. The VM follows the same pattern as the other five natural languages in that it shows that less probable words contain more informative segments.
Figure 3, shows linear regression models predicting the relative position of the uniqueness-point for the words in the given corpora. Less probable words have significantly earlier uniqueness points for all four word lengths in VM. Also here, VM shares characteristics of the natural languages presented in the study by King and Wedel (2020).

6. Discussion

As explained in Section 2., previous work used statistical methods to research whether the Voynich manuscript behaves like a natural language. Some focus on word level

[2]A good reference site, as well as detailed information and download links for transliteration versions of the Voynich Manuscript, can be found on René Zandbergen's excellent website dedicated to the manuscript `http://www.voynich.nu/`.

[3]We follow King and Wedel (2020) in their selection of this range in word length, and note that 84% of the total word occurrences in the VM lie within the word length range from four to eight

[4]Due to space limitations we show the graphs for 5 languages, varied in terms of their language families and morphological complexity, focussing on the Indo-European language family because of their relevance for the VM in terms of the location in which they are spoken (for comparison with another 15 languages see King and Wedel (2020))

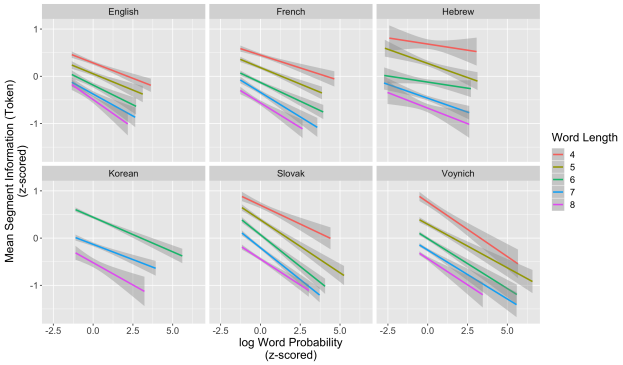


Figure 2: Relationship between log word probability and mean token-based segment information for words of length 4-8

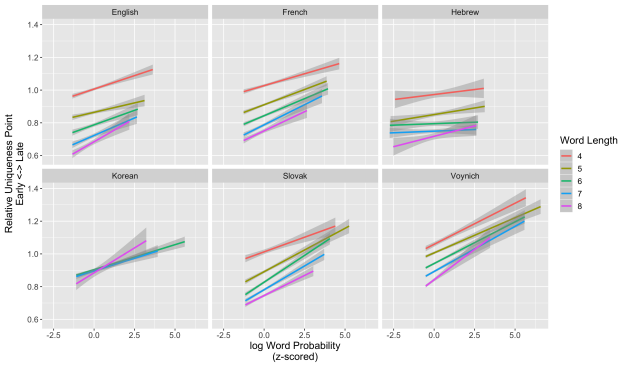


Figure 3: Relationship between log word probability and relative position of uniqueness-point for words of length 4-8

(Montemurro and Zanette, 2013; Zelinka et al., 2019) and find positive results.

The results above show several indications that not only at word level but also at the level of segments, VM shares characteristics with other natural languages. However, others, such as Zandbergen (2020) and Landini (2001) performed character-level analysis and show mixed results.

Landini's spectral analysis points in the same direction as our results, namely that the VM is a natural language, but it is hard to compare their results to ours, because of the different nature of their analysis.

Reddy and Knight (2011) compare the unigram and bigram predictability of VM characters with those of English, Arabic and Pinyin. Especially at bigram-level, VM is more predictable than English and Arabic, more closely resembling Pinyin.

This result is consistent with Zandbergen (2020) who shows that the entropy of characters in the VM is lower than for many other languages and in particular Indo-European languages. However, he also notes that the results differ depending on the position of the character. Characters at the 1st and 2nd position are more predictable than in Latin, but the 3rd and 4th characters are less predictable.

These works emphasise the difference between VM and other Indo-European languages, but also show the importance of character position. In contrast, our experiments show that when focusing on the relationship between word probability and character information, both on average and based on position (cf. Figure 2 and Figure 3), the same type of relation is found in the VM as in other text corpora.

A couple of caveats are needed: The comparisons in this paper are between VM and contemporary languages and larger corpora, in general. A better comparison would be between languages from roughly the same time period and corpora of the same size. Also, we do not have phonemic transcriptions of the VM and based these on the written characters.[5]

7. Conclusions and Future Work

In this paper, we showed more support for the claim that the VM is written in a natural language and therefore is not a hoax. Although several scholars have found statistical evidence pointing in the same direction, more evidence is needed, particularly to establish whether there is a known language family to which VM can plausibly be assigned. In future work, we would like to compare the results from VM with corpora from the same period that are also similar in size.

8. Acknowledgements

The authors extend their gratitude to Adam King who kindly assisted us and answered questions on the approach he used besides providing some code used to generate his results for comparison purposes. We would also like to thank the reviewers for their helpful comments and suggestions.

[5] Previous work (Mahowald et al., 2018) has used orthographics as a proxy for phonetics.

9. Bibliographical References

Abbott, D. (2015). Cracking the Voynich Code 2015 - Final Report. https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Cracking_the_Voynich_Code_2015_-_Final_Report. [Online; accessed 19-February-2020].

Bax, S. (2014). A Proposed Partial Decoding of the Voynich Script. http://stephenbax.net/wp-content/uploads/2014/01/Voynich-a-provisional-partial-decoding-BAX.pdf. [Online; accessed 19-February-2020].

Cheshire, G. (2019). The Language and Writing System of MS408 (Voynich) Explained. *Romance Studies*, 37(1):30–67.

M. D'Imperio, editor. (1976). *New Research on the Voynich Manuscript: Proceedings of a seminar*.

D'Imperio, M. (1978). *The Voynich Manuscript: An Elegant Enigma*. National Security Agency, US.

Hauer, B. and Kondrak, G. (2016). Decoding Anagrammed Texts Written in an Unknown Language and Script. *Transactions of the Association for Computational Linguistics*, 4:75–86.

Hauer, B. and Kondrak, G. (2018). AI didn't decode the cryptic Voynich manuscript — it just added to the mystery. *The Verge*, 1st February, 2018.

King, A. and Wedel, A. (2020). Greater Early Disambiguating Information for Less-Probable Words: The Lexicon Is Shaped by Incremental Processing: Early Information for Low-Probability Words. *Open Mind*, To appear.

Landini, G. (2001). Evidence of linguistic structure in the Voynich manuscript using spectral analysis. *Cryptologica*, 25(4).

Mahowald, K., Dautriche, I., Gibson, E., and Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science*, 42(8):3116–3134.

Marslen-Wilson, W. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1):29–63.

McInnes, A. and Wang, L. (2015). *Statistical Analysis of Unknown Written Language: The Voynich Manuscript - Project Group 31*. University of Adelaide, Australia.

Montemurro, M. A. and Zanette, D. H. (2013). Keywords and Co-occurence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis. *Plos One*, 8(6).

Prinke, R. T. and Zandbergen, R., (2017). *The Voynich Manuscript*, chapter The Unsolved Enigma of the Voynich Manuscript, pages 15–40. Watkins Publishing.

Reddy, S. and Knight, K. (2011). 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*, pages 78–86, Portland, OR, USA, June. Association for Computational Linguistics.

Rugg, G. and Taylor, G. (2016). Hoaxing statistical features of the Voynich Manuscript. *Cryptologia*, 41(3):247–268.

Rugg, G. (2004). The Mystery of the Voynich Manuscript. *Scientific American*, 291(1):104–109.

Schinner, A. (2007). The Voynich Manuscript: Evidence of the Hoax Hypothesis. *Cryptologia*, 31(2):95–107.

Stolte, D. (2011). UA Experts Determine Age of Book 'Nobody Can Read'. *UA News*. https://uanews.arizona.edu/story/ua-experts-determine-age-of-book-nobody-can-read.

Zandbergen, R. (2020). Voynich MS. http://www.voynich.nu/extra/sol_ent.html. Online; accessed 27 March, 2020.

Zelinka, I., Zmeskal, O., Windsor, L., and Cai, Z. (2019). Unconventional Methods in Voynich Manuscript Analysis. *MENDEL*, 25(1):1–14.

Zyats, P., Mysak, E., Stenger, J., Lemay, M.-F., Bezur, A., and Driscoll, D., (2016). *The Voynich Manuscript*, chapter Physical Findings, pages 23–37. Yale University Press.

10. Language Resource References

René Zandbergen. (2017). *ZL Transliteration of Voynich Manuscript*. http://www.voynich.nu/transcr.html, 1b, 1.5 IVTFF, 2017/09/24.