

# The Voynich Manuscript

### Evelyn Keane

### C00239302

### fINAL REPORT

Contents

[ABSTRACT 3](#_Toc102159157)

[INTRODUCTION 4](#_Toc102159158)

[WHY ENTROPY 4](#_Toc102159159)

[PROCESS 4](#_Toc102159160)

[NORMALISATION 5](#_Toc102159161)

[CCG ENTROPY 6](#_Toc102159162)

[VMS ENTROPIES 7](#_Toc102159163)

[CCG VERSUS VMS/FINISHED PRODUCT 8](#_Toc102159164)

[CONCLUSION 12](#_Toc102159165)

[TECHNICAL AND PERSONAL LEARNING AND ACHIEVEMENTS 12](#_Toc102159166)

[PROJECT REVIEW 12](#_Toc102159167)

[ACKOWLEDGMENTS 13](#_Toc102159168)

[REFERENCES 13](#_Toc102159169)

# ABSTRACT

The Voynich manuscript is an illustrated codex hand-written in an otherwise unknown writing system, referred to as 'Voynichese'. The vellum on which it is written has been carbon-dated to the early 15th century (1404–1438), and stylistic analysis indicates it may have been composed in Italy during the Italian Renaissance. The origins, authorship and purpose of the manuscript are debated. Various hypotheses have been suggested, including that it is an otherwise unrecorded script for a natural language or constructed language; an unread code, cypher, or other form of cryptography; or simply a meaningless hoax.

The manuscript currently consists of around 240 pages, but there is evidence that additional pages are missing. Some pages are foldable sheets of varying size. Most of the pages have fantastical illustrations or diagrams, some crudely coloured, with sections of the manuscript showing people, fictitious plants, astrological symbols, etc. The text is written from left to right. The manuscript is named after Wilfrid Voynich, a Polish book dealer who purchased it in 1912. Since 1969, it has been held in Yale University's Beinecke Rare Book and Manuscript Library.

The Voynich manuscript has been studied by professional and amateur cryptographers alike, including American and British codebreakers from both World War I and World War II. The manuscript has never been demonstrably deciphered, and none of the many hypotheses proposed over the last hundred years have been independently verified. The mystery of its meaning and origin has excited the popular imagination, making it the subject of study and speculation.[[1]](#endnote-1)

# INTRODUCTION

Undertaking this as a research project will be a hefty challenge. I don’t presume to claim I’ll translate the script, but I will utilise cryptographical and mathematical analyses to find patterns that may aid in doing so. Anything surmised will be helpful in narrowing down the geographical origin of the script, whether it’s a real language, dead or alive, enciphered or cleartext, etc. I have resolved to approach this in an agnostic, linguistic manner. By that, I mean to ignore the issues of content and language. The question of which word means what, and in which language, are left to the very end and asked only once the underlying language is better understood with no assumptions being made.

# WHY ENTROPY

The anomalous second-order entropies of Voynich text are among its most puzzling features. h1-h2, the difference between conditional first- and second-order entropies, equals the difference H1-h2, the difference between the first-order absolute entropy and the second- order conditional entropy. h1-h2 or H1-h2 is a theoretically significant number; it denotes the average information carried by the first character in a digraph about the second one. Therefor it was chosen as a simple measure of what is being sought, although the whole entropy profile of text samples was considered.[[2]](#endnote-2)

Tests show that Voynich text does not have its low h2 measures solely because of a repetitious underlying text, that is, one that often repeats the same words and phrases. Tests also show that the low h2 measures are probably not due to an underlying low-entropy natural language. A verbose cipher, one which substitutes several ciphertext characters for one plaintext character, can produce the entropy profile of Voynich text.

William Ralph Bennett first applied the entropy concept to the study of the Voynich Manuscript in his Scientific and Engineering Problem Solving with the Computer (Englewood Cliffs: Prentice-Hall, 1976)[[3]](#endnote-3). His book has introduced many people to the VMs.

The repetitive nature of VMs text is obvious to casual examination. Entropy is one possible numerical measure of a text's repetitiousness. The higher the text's repetitiousness, the lower the second-order entropy (information carried in letter pairs). Bennett noted that only some Polynesian languages have second-order entropies as low as VMs text. Typical ciphers do not have a low second-order entropy either.

# PROCESS

I aimed to gather a statistically useful amount of corpora written in Early Modern Irish. Early Modern Irish represented a transition between Middle Irish and Modern Irish. Its literary form, Classical Gaelic, was used in Ireland and Scotland from the 13th to the 18th century; this ensures that a wide range of manuscripts can be examined that still cover the same dialect, and also encompasses the years in which the VMS was created, ensuring contemporaneity. I obtained these from celt.ucc.ie and their Corpus of Electronic Texts[[4]](#endnote-4). Some of the original manuscripts can be viewed on isos.dias.ie[[5]](#endnote-5). I collected over 20 and in the end 18 proved usable, whether due to insufficient word count or unwanted language age.

After amassing the collection, I set about normalising the contents by hand. With 18 total being possible to use in the end, this is enough to produce statistically significant results, confirmed by the ending results all being almost identical.

## NORMALISATION

The normalisation process consisted of:

• Remove editor’s marks, eg page numbers, manuscript numbers, line numbers, marks indicating page end, etc

• Labels will be left in since labels will be included in the VMS calulations

• Remove digits since there isn’t any in the VMS that we can see and numbers are therefore likely spelled out or letters are used to stand for numbers

• Remove symbols and punctuation since again there isn’t any in the VMS

• Change & to “agus”

• Change diacritics eg fadas to plain letters without the diacritics because I believe we’re working with something similar to an a priori syllabary which wouldn’t have diacritics

• Change capital letters to lowercase so they don’t get counted as new unique letters to match the VMS which we don’t see evidence of capital letters in

• Take out Latin (mainly Latin but other languages feature) sentences or chunks. Single words are ok

• Remove line breaks

A point to remember: symbols, punctuation, and grammar were inserted by the modern transcribers and translators, not the original historical writers, because grammar and punctuation weren’t much known nor cared for in the past except for copies of the bible and other books that would be distributed to scholars or used for teaching. This is documented in the preamble of each text and essentially I am undoing the normalisation that the transcribers/translators did to modernise the text and make it readable to our current high lexicographical standards.

A sample of text before and after normalising:

**{MS page/column 14/27} de cinamomo’ .i. inté caithius cainel161 go minic ni recha d'ég do truailledh na lendann oir toirmisgid e da mbia an follamhnughadh go maith o soin amach. Et dlighear a fis gon dligheand an t'uisgi beith glan ocus glantur an t'aer go h'ealadhanach le teine162 maith muna faghtur glan gu nadurdha e. Et is lór so ge do fédfuighi moran eili do radh ann163.**

Becomes

**de cinamomo i inte caithius cainel go minic ni recha deg do truailledh na lendann oir toirmisgid e da mbia an follamhnughadh go maith o soin amach et dlighear a fis gon dligheand an tuisgi beith glan ocus glantur an taer go healadhanach le teine maith muna faghtur glan gu nadurdha e et is lor so ge do fedfuighi moran eili do radh ann**

This process had to be done mostly by hand to ensure accuracy. The texts had to be read through to remove any Latin or other languages featured – this choice was much easier than automated language checking as it would be hard to make anything that remove all languages but the desired one, not to mention no known software can recognise Early Modern Irish as a language. Removing all numbers, symbols, and punctuation was easy enough with ctrl+F in MS Word. I manually removed parts containing legitimate letters, for example **{MS page/column 14/27}**, and remarks such as **[gap extant four words]**. Changing & to agus was done with ctrl+F again. Diacritics were removed the same way: ctrl+f to find á, é, í, ó, ú, and replace each with a, e, i, o, and u respectively. Capital letters were removed by selecting the entire text, then pressing shift+f3 twice to turn everything to uppercase and then back. Line breaks were removed and space was treated as a delimiting character.

Essentially I am reducing a corpus of text to nothing but lowercase alphabetical letters in one language as this is what entropy is to be calculated using. Any extraneous content will increase the resultant entropy value and render it inaccurate.

## CCG ENTROPY

The following measures were considered when calculating the entropy:

• h0: zero-order entropy (log2 of the number of different characters)

• h1: first-order conditional or absolute entropy

• h2: second-order conditional entropy

• h1-h2: difference between conditional first- and second order entropies

• H1-h2: the difference between the first-order absolute entropy and the second-order conditional entropy

The following table displays my results for the entropies of the Classical Common Gaelic texts.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MS code** | **Number of characters** | **Character count** | **h0** | **h1** | **h2** | **h1-h2** | **Translated in** |
| Advocates 72.1.2 (olim Gaelic II) | 22 | 101152 | 4.459 | 3.838 | 3.133 | 0.705 | 1500-1550 |
| Gaelic Ms. XLII | 24 | 111309 | 4.585 | 3.870 | 3.109 | 0.761 | 1694 |
| RIA Stowe, B II 1 | 21 | 101152 | 4.392 | 3.814 | 3.149 | 0.665 | 1300-1350 |
| RIA 23 P 10 | 23 | 68490 | 4.524 | 3.869 | 3.108 | 0.761 | 1500s |
| Harley MS 456 | 22 | 48527 | 4.459 | 3.880 | 3.177 | 0.703 | 1459 |
| MS 1299 (olim H 2 8) | 21 | 55291 | 4.392 | 3.862 | 3.135 | 0.727 | 1500-1600 |
| MS 23 N 16 (MS 443) | 22 | 23320 | 4.459 | 3.913 | 3.171 | 0.712 | 1606 |
| MS RIA 23 P 20 | 21 | 7060 | 4.392 | 3.861 | 3.057 | 0.804 | 1400–1520 |
| RIA MS 23 F 19 | 19 | 3043 | 4.248 | 3.856 | 3.077 | 0.782 | 1352 |
| MS TCD 1435 | 19 | 7693 | 4.248 | 3.862 | 3.127 | 0.735 | 1500s |
| RIA MS 23 F 19 | 21 | 8119 | 4.392 | 3.858 | 3.146 | 0.712 | 1352 |
| RIA MS 23 F 19 | 22 | 67442 | 4.459 | 3.882 | 3.182 | 0.701 | 1352 |
| MS NLI G 12 | 22 | 30739 | 4.459 | 3.893 | 3.169 | 0.725 | 1563 |
| MS RIA 23 P 20 | 24 | 184580 | 4.585 | 3.886 | 3.195 | 0.691 | 1400-1520 |
| TCL MS 1343 (H. 3. 22) | 23 | 230262 | 4.524 | 3.868 | 3.101 | 0.768 | 1415 |
| MS RIA 24 P 26 | 24 | 103440 | 4.585 | 3.881 | 3.147 | 0.734 | 1469 |
| MS RIA 24 P 26 | 21 | 137798 | 4.392 | 3.852 | 3.124 | 0.728 | 1469 |
| MS RIA 24 P 26 | 22 | 159040 | 4.459 | 3.869 | 3.153 | 0.716 | 1469 |

## VMS ENTROPIES

The Voynich Manuscript has a wildly different h1 AND h2 from normal modern languages, as seen below[[6]](#endnote-6).

Chart, scatter chart

Description automatically generated

The Voynich entropies values for four different transcription methods were used to show how transcriptions effect entropy values: Currier, FSG, EVA, and Frogguy. By using four, we can look at the average of them to lessen the impact this issue creates.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Type of Voynich Text** | **Transcription Alphabet** | **Number of Characters** | **Character count** | **h0** | **h1** | **h2** | **h1-h2** |
| Herbal-A | Currier | 33 | 9804 | 5.044 | 3.792 | 2.313 | 1.479 |
| Herbal-A | FSG | 24 | 10074 | 4.585 | 3.801 | 2.286 | 1.515 |
| Herbal-A | EVA | 21 | 12218 | 4.392 | 3.802 | 1.990 | 1.812 |
| Herbal-A | Frogguy | 21 | 13479 | 4.392 | 3.826 | 1.882 | 1.945 |
| Herbal-B | Currier | 34 | 13858 | 5.087 | 3.796 | 2.267 | 1.529 |
| Herbal-B | FSG | 24 | 14203 | 4.585 | 3.804 | 2.244 | 1.560 |
| Herbal-B | EVA | 21 | 16061 | 4.392 | 3.859 | 2.081 | 1.778 |
| Herbal-B | Frogguy | 21 | 17909 | 4.392 | 3.846 | 1.949 | 1.897 |

## CCG VERSUS VMS/FINISHED PRODUCT

I graphed h1 versus h2 for each of the Classical Common Gaelic texts and all the VMS transcriptions.

This graph was then repeated but starting at 0,0 to provide context for how close the numbers really are.

The resultant graphs show that the previously-thought low entropy values for the Voynich may not be so surprising. The graphs indicate that the h1 (first order entropy) of the Voynich is actually in line with a language being spoken when the Voynich was created – indicating a natural language that was spoken at the time.

For a more summarised graph, I plotted the average entropy values for all the Classical Common Gaelic texts against the average of each of the Voynich transcription, using the values in the table below.

|  |  |  |
| --- | --- | --- |
| **Translations/Transcriptions** | **Average H1** | **Average H2** |
| Classical Common Gaelic | 3.867 | 3.137 |
| Voynich: Currier | 3.794 | 2.29 |
| Voynich: FSG | 3.803 | 2.265 |
| Voynich: EVA | 3.831 | 2.036 |
| Voynich: Frogguy | 3.836 | 1.916 |

And graphed them as follows:

To finish, and conceptualise my results, I added the average h1 and h2 of the Irish scripts to Rene Zandbergen's graph[[7]](#endnote-7) displaying the entropies of the VMS versus European languages.

Chart, scatter chart

Description automatically generated

# CONCLUSION

My analysis for first order entropy makes it clear that the h1 measure is very close to the level of the Voynich text. That is useful to show the Voynich is not outside of the boundary for a natural language, and fits with natural languages spoken at the time.

However, the h2 measures still remain a problem. This measure of conditional entropy (or, even better, the ratio of h1 to h2) is what worries researchers most. It can be inferred from my analysis that the Voynich was written in a natural language with something having been done to it in the act of writing it down. Possibilities include enciphering, encoding, syllabication, and more.

This calls for further work on entropy, investigating ways in which a natural language could be altered when enscribing to cause the second order entropy to be lowered. Future comparisions of entropy and similar mathematical practices should be carried out on contemporary languages to the Voynich as, as proven in this investigation, the language of the Voynich is much closer to natural languages than previously thought.

## TECHNICAL AND PERSONAL LEARNING AND ACHIEVEMENTS

This project was moreso one of the mind rather than one of the coding prowess. Containing minimal coding, it challenged me to come up with a novel idea and prove it, which was tough work but worthwhile in the end. Less glamorous is the admission that most of the work was taken up by the normalisation process – doing it mostly by hand was laborious but I could be sure it was done correctly and thoroughly. MS Word made it easy to transform complicated tables to clear, informative graphs, so I could convey my point easily.

Personally I grew more confident during the project – lord knows I wasn’t lacking in it in the first place but I feel academically equal to much more accomplished people now. Having proved a new feature of the Voynich is something few people can say, and I have no reservations being proud of myself. My patience was tested particularly in the months spent reading Early Modern Irish to normalise it but I never gave up on the idea. Interpersonally I learned to fight for a hypothesis I thought was right and others didn’t. Doing a research project rather than a production project allowed me to break out from what I spent the past four years doing and let me realise a passion for linguistics.

## PROJECT REVIEW

I can’t say anything in particular went wrong – the normalisation process was tedious and I think I could have shortened it, but I’m glad I could be confident in the thoroughness with which I did it. What went right was – my idea. I might have been crushed if I couldn’t find out anything new. So many people were interested in my project and were happy to hear about it, particularly experts who offered feedback and encouraged me to continue my work and write it into a paper to be submitted to the University of Malta’s Voynich Manuscript Conference in late November 2022.

I can’t imagine someone picking a similar project in the future, so I’ll give advice to someone who wants to do a research project: fight for it, but pick your battles. If people say “well, you could, but here’s some legitimate reasons why you might regret it”, listen to them. Start your work early, document everything, and don’t be afraid to reach out for help and advice – lecturers, friends, experts in the field can all offer different positive aspects. Don’t give up if you get to the middle or end and realise your hypothesis or idea was wrong. Finish it, and say it was wrong. You learned from it, and proving something is wrong can be as helpful as proving something is right.

## ACKOWLEDGMENTS

celt.ucc.ie and isos.dias.ie proved invaluable resources for historical Irish corpa. Members of voynich.ninja for inspiring and encouraging. Experts who offered feedback, particularly Emma May Smith[[8]](#endnote-8). My supervisor Richie Butler who went to bat for me to work on such a wild idea in the first place and supported me all along. My friend Brendan Butler for making a documentary about my work. My friend Ben Guilfoyle and my partner Oisín Chelmiah for helping me with the smallest infuriating issues to the huge glaring ones.

# REFERENCES

1. https://collections.library.yale.edu/catalog/2002046 [↑](#endnote-ref-1)
2. Lindemann, Luke and Claire Bowern: Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript. 27 October 2020. [↑](#endnote-ref-2)
3. Bennett, William Ralph: Scientific and Engineering Problem Solving with the Computer. Englewood Cliffs: Prentice-Hall, 1976. [↑](#endnote-ref-3)
4. https://celt.ucc.ie [↑](#endnote-ref-4)
5. https://www.isos.dias.ie/ [↑](#endnote-ref-5)
6. http://www.voynich.nu/a2\_char.html#entropy [↑](#endnote-ref-6)
7. http://www.voynich.nu/a2\_char.html#entropy [↑](#endnote-ref-7)
8. Text, application

   Description automatically generated [↑](#endnote-ref-8)