

# The Voynich Manuscript

### Evelyn Keane

### C00239302

### FUNCTIONAL SPEC

Contents

[ABSTRACT 3](#_Toc102153882)

[INTRODUCTION 3](#_Toc102153883)

[ENTROPY 3](#_Toc102153884)

[ENTROPIES OF THE VOYNICH 4](#_Toc102153885)

[CALCULATING ENTROPIES 4](#_Toc102153886)

[CLASSICAL COMMON GAELIC TEXTS 6](#_Toc102153887)

[NORMALISATION PROCESS 6](#_Toc102153888)

[CALCULATING ENTROPIES 8](#_Toc102153889)

[GRAPHING 9](#_Toc102153890)

# 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.

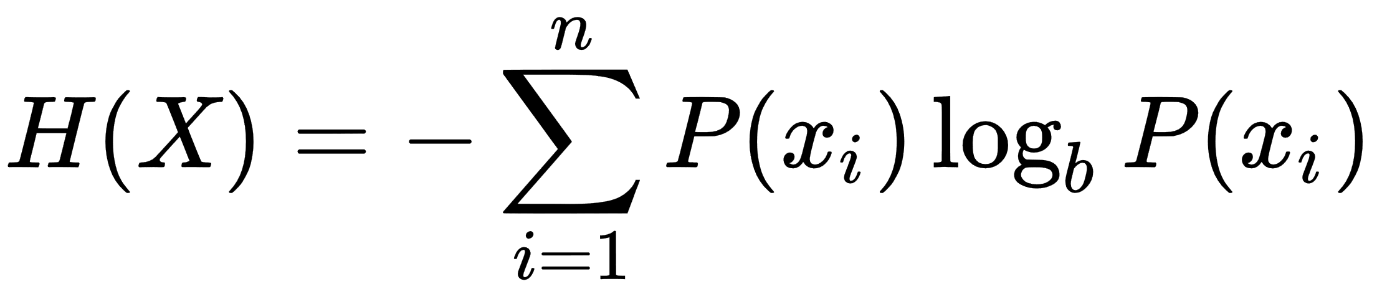
The Voynich manuscript has been studied by professional and amateur cryptographers alike, including American and British codebreakers from both World War I and World War II. The manuscript has never been demonstrably deciphered, and none of the many hypotheses proposed over the last hundred years have 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.[[2]](#endnote-2)

# ENTROPY

In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. This is defined as



The concept of information entropy was introduced by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication", and is also referred to as Shannon entropy. Shannon's theory defines a data communication system composed of three elements: a source of data, a communication channel, and a receiver. The "fundamental problem of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was generated by the source, based on the signal it receives through the channel. Shannon considered various ways to encode, compress, and transmit messages from a data source, and proved in his famous source coding theorem that the entropy represents an absolute mathematical limit on how well data from the source can be losslessly compressed onto a perfectly noiseless channel. Shannon strengthened this result considerably for noisy channels in his noisy-channel coding theorem.[[3]](#endnote-3)

The amount of information carried in the arrangement of words is the same across all languages, even languages that aren't related to each other. This consistency could hint at a single common ancestral language, or universal features of how human brains process speech.

Language carries meaning both in the words we choose, and the order we put them in. Some languages, like Finnish, carry most of their meaning in tags on the words themselves, and are fairly free-form in how words are arranged. Others, like English, are more strict -- "John loves Mary" means something different from "Mary loves John."[[4]](#endnote-4)

# ENTROPIES OF THE VOYNICH

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.

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)[[5]](#endnote-5). 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.[[6]](#endnote-6)

# CALCULATING ENTROPIES

The conditional entropies were used as follows. Say that H1 is the absolute first-order entropy and H2 is the absolute second-order entropy. Then h1 and h2 are the first- and second-order conditional entropies. h2 = H2-H1, since it is conditional on more than one character. h1 = H1, since it depends on only single characters; thus h1 is really not conditional.

The following measures were considered:

* 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, which equals the difference
* H1-h2: the difference between the first-order absolute entropy and the second-order conditional entropy.

As will be seen, there is a need here to compare systems with very different numbers of characters, to scale the statistics somehow to the size of the character set. 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. It is perhaps the best single, simple measure of what is being sought.

The % of the second-order maximum absolute entropy might have been used. One could calculate the % of H2 from the total H2 that could be delivered by each alphabet. Using digraphs with an alphabet of m characters, H2(max) is:

log2(m^2)

and the %H2(max) is:

(H2/log2(m^2))/100

However, the H2(max) depends tremendously on m, the size of the character set chosen. For Voynich text, Currier has 36 characters and Basic Frogguy has 23 characters. Characters that are hardly ever used have little effect on h1 and h2, but could make a tremendous difference in H2(max). Therefore, this measure was not used.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |

[[7]](#endnote-7)

Here are entropy results for Voynich texts, a sample of Herbal-A and Herbal-B. The Herbal-A sample's h1-h2 ranges 1.479-1.945, depending on which transcription alphabet is used. The Herbal-B sample's h1-h2 ranges 1.529-1.897. All these are far greater than the 0.93 for English and 0.70 for Latin.

The choice of transcription alphabet also makes an enormous difference. From Currier to Frogguy the range of h1-h2 is 1.5-1.9. The direction is what one would expect. Currier is the most synthetic, while Frogguy is the most analytical, decomposing single Currier characters into several Frogguy characters. Thus Currier Q = Frogguy cqpt.

# 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 |

I sourced over 20 manuscripts from celt.ucc.ie and their corpus of electronic texts[[8]](#endnote-8). 18 proved to be useable in the end. These texts were on topics similar to those seen in the Voynich – science and medicine. 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. Above is the finished entropy calculations for normalised contemporary texts written in Classical Common Gaelic.

# NORMALISATION PROCESS

These were preceded by carrying out the following normalisations:

• 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. 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 extent 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. The entirety of this process, as seen when performed above, takes a normal text and returns it in its most reduced form – just spaces and the lowercase alphabet. This is what entropy checks; if not normalised, the calculations would take into account all the numbers, symbols, other languages, and extraneous remarks, and thus the resultant entropy values would be rendered miles off correct.

# CALCULATING ENTROPIES

The calculations were found using the following short python program[[9]](#endnote-9):

Text

Description automatically generated

Text

Description automatically generated

Conditional (second order entropy) compares a value to the value preceding it, so I created text files for each CCG script but with a space added at the start of each START file and a space at the end of each END file. This creates two identical texts, one offset by one character as compared to the other. The code is simplistic and simply performs the relevant entropy formula to the data given.

# GRAPHING

MS Word was used to produce the following graphs from my results tables seen above:

Chart, scatter chart

Description automatically generated[[10]](#endnote-10)

REFERENCES

1. https://collections.library.yale.edu/catalog/2002046 [↑](#endnote-ref-1)
2. Landini, Gabriel: Evidence of linguistic structure in the Voynich MS using spectral analysis, in Cryptologia Vol. 25 Issue 4, 2001, pp. 275-295. [↑](#endnote-ref-2)
3. https://en.wikipedia.org/wiki/Entropy\_(information\_theory) [↑](#endnote-ref-3)
4. https://arxiv.org/ftp/arxiv/papers/1405/1405.2061.pdf [↑](#endnote-ref-4)
5. Bennett, William Ralph: Scientific and Engineering Problem Solving with the Computer. Englewood Cliffs: Prentice-Hall, 1976. [↑](#endnote-ref-5)
6. Lindemann, Luke and Claire Bowern: Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript. 27 October 2020. [↑](#endnote-ref-6)
7. Bennett, William Ralph: Scientific and Engineering Problem Solving with the Computer. Englewood Cliffs: Prentice-Hall, 1976. [↑](#endnote-ref-7)
8. https://celt.ucc.ie/publishd.html#ism [↑](#endnote-ref-8)
9. https://replit.com/@evlnkn/entropyCalculator [↑](#endnote-ref-9)
10. http://www.voynich.nu/extra/sol\_ent.html [↑](#endnote-ref-10)