Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript

Luke Lindemann and Claire Bowern

May 20, 2021

Note: This is an updated version of the article that we uploaded on October 27, 2020. For this update we developed an improved method for extracting language text from Wikipedia, removing metadata and wikicode, and we have rebuilt our corpus based on current wikipedia dumps. The new methodology is described in Section 3.2, and we have updated the statistics and graphs in Section 4 and Appendices B and D. None of our results have changed substantially, suggesting that character-level entropy was not greatly affected by the inclusion of stray metadata. However, our forthcoming work will use these corpora for word-level statistics including measures of repetition and type token ratio, for which metadata would have a much greater effect on results. We also made a minor alteration to the Maximal Voynich transcription system as described in Section 3.1.1. This also has a negligible effect on character entropy. The statistics in Section 4 and Appendix A have been updated accordingly.

Abstract

This paper outlines the creation of three corpora for multilingual comparison and analysis of the Voynich manuscript: a corpus of Voynich texts partitioned by Currier language, scribal hand, and transcription system, a corpus of 311 language samples compiled from Wikipedia, and a corpus of eighteen transcribed historical texts in eight languages. These corpora will be utilized in subsequent work by the Voynich Working Group at Yale University.

We demonstrate the utility of these corpora for studying characteristics of the Voynich script and language, with an analysis of conditional character entropy in Voynichese. We discuss the interaction between character entropy and language, script size and type, glyph compositionality, scribal conventions and abbreviations, positional character variants, and bigram frequency.

This analysis characterizes the interaction between script compositionality, character size, and predictability. We show that substantial manipulations of glyph composition are not sufficient to align conditional entropy levels with natural languages. The unusually predictable nature of the Voynichese script is not attributable to a particular script or transcription system, underlying language, or substitution cipher. Voynichese is distinct from every comparison text in our corpora because character placement is highly constrained within the word, and this may indicate the loss of phonemic distinctions from the underlying language.

Corpus materials and code are available from github.com/chirila/Voynich-public.
## Contents

1 Introduction 3

2 Structure and Content of the Voynich Manuscript 5

3 Description of the Corpora 6

### 3.1 The Voynich Corpus

- 3.1.1 Voynich Transcription Systems 8
- 3.1.2 Voynich Document Preparation 14

### 3.2 The Wikipedia Corpus

- 3.2.1 Wikipedia Document Preparation 15
- 3.2.2 Wikipedia Languages by Family 16

### 3.3 The Historical Corpus

- 3.3.1 Description of the Corpus 20
- 3.3.2 Historical Document Preparation 21

4 Conditional Character Entropy in Voynichese 22

### 4.1 Character Frequency Distribution 23

### 4.2 Entropy Variance and Document Length 23

### 4.3 Entropy in Voynichese 25

#### 4.3.1 Character Set Size 25
#### 4.3.2 Languages A and B 27

### 4.4 Comparison to the Wikipedia Corpus 27

### 4.5 Comparison to the Historical Corpus 29

#### 4.5.1 Parallel Diplomatic and Normalized texts 29

### 4.6 What makes Voynichese unique? 31

#### 4.6.1 Entropy and Bigram Frequency 31
#### 4.6.2 Compositionality, Word-finals, and Syllable Structure 34
#### 4.6.3 Bigrams with High Conditional Probability 36

### 4.7 Summary 37

A Voynich Corpus Statistics 38

B Wikipedia Corpus Statistics 39

C Historical Corpus Statistics 48

D The Sukhotin Algorithm for Vowel Detection 49
1 Introduction

The Voynich Manuscript (e.g. Figure 1) is an early 15th Century illustrated manuscript written by multiple unknown scribes (Davis 2020) in an unknown cipher or language. It contains about 38,000 words of text. The Voynich alphabet, which is not found in any other known work, has resisted nearly 110 years of modern attempts at decipherment (see Bowern and Lindemann 2020 and http://www.voynich.nu for overviews). This is despite the fact that there is clear evidence of language-like structure in the text, at least at the paragraph level (Reddy and Knight 2011; Amancio et al. 2013; Landini 2001).  

![Figure 1: A paragraph of text and labelled figures from the “Recipes” section of the Voynich Manuscript (f100r).](image)

The major unsolved question of the Voynich text is whether it represents meaningful language. It could be a medieval hoax that is designed to look like an esoteric alchemical text, in which no sequences of letters correspond to any meaningful words or concepts (see Rugg 2004; Timm and Schinner 2020, amongst others). If so, the creators did an incredible job of imitating the patterns of an authentic language text considering what was known about the structure of language at the time. In so doing, they must have modeled their fake language after a real language that they were familiar with, imbuing it with familiar, language-like patterns. If this is the case, it may still be possible to take clues from the structure of Voynichese to pinpoint a language or region of origin.

We find it more likely that Voynichese does represent meaningful language, and this opens the possibility that Voynichese may ultimately be deciphered. It is possible that the text was created

---

1 Thanks to members of the “Mystery of the Voynich Manuscript” class at Yale for discussion of some of these points. Division of labor: LL and CB planned the analyses; LL compiled the corpora and wrote the scripts; LL and CB analyzed the data; LL wrote the paper with input from CB.

2 In previous decades, there was the separate question of whether the manuscript is a modern hoax, i.e. a 20th century forgery of a medieval manuscript. This has been fairly decisively disproven by chemical analysis, though see Barlow (1986) for some earlier discussion that predates the results from carbon dating.
to encode meaning, but the nature of the encipherment obscured it in such a way that the original meaning is permanently irrecoverable. Even if this is the case, we may be able to glean information about language and content with a careful analysis of Voynichese structure. Note that others, such as Timm and Schinner (2020), argue strongly that the Voynich manuscript does not encode meaningful text, and so cannot be compared with natural language. The goal of this project is to analyze the structure and patterning of the Voynich manuscript, and to compare it to known texts. By comparing Voynichese to known languages and texts, we reduce the set of possible hypotheses about language, origin, and the question of meaningfulness. This paper describes the creation of text corpora for conducting experiments on the Voynich text. Note that our aim is not to advance a strong claim about exclusive identification (of the form “Voynichese is Hebrew” or “Voynichese is Occitan”) but rather to explore the relationships between the morphological and phonological profiles of Voynichese with a typologically broad range of natural and constructed languages.

We require as many examples of languages and scripts as possible in order to understand the range of possibilities in the structure of texts. We would also like to be able to classify these texts by language and language family in order to see whether closely related languages share affinities of structure, allowing us to narrow down the range of possible languages (or at least better understand possible encryption processes). For this reason, our first comparison corpus consists of Wikipedia articles written in 311 languages, representing thirty-nine language families.

The Voynich manuscript is also the product of a particular historical context. This includes the medieval scribal traditions which produced it, as well as the herbalogical, alchemical, and astrological knowledge which informed its content. To take a particular relevant example, medieval scribes made much more frequent use of abbreviations than we find in modern writing. We therefore include a second comparison corpus of historical manuscripts in English, Georgian, Hebrew, Icelandic, Italian, Latin, Persian, and Spanish.

This paper consists of three sections. In Section 2, we give an overview of the structure and content of the Voynich manuscript, focusing on what is known about the text. We give a brief description of each section and discuss the evidence for multiple Voynich languages and scribal hands.

In Section 3, we outline in detail the process of creating the three corpora used for this project: the Voynich Corpus, the Wikipedia Corpus, and the Historical Corpus. We discuss the issue of transcription in the Voynich text, and define the three transcriptions we use: Maximal, Maximal Simplified, and Minimal.

In Section 4, we demonstrate the utility of the corpora by comparing character-level properties

---

3Timm and Schinner 2021 make this point especially forcefully, even going as far as accusing us of a lack of scientific rigor for not finding their arguments convincing. While the aim of this paper is to make language comparisons across natural and constructed languages, rather than to make arguments about any particular theory, we reiterate a point we have made several times elsewhere (including in the review article that they criticize): that Voynichese appears unnatural only below the word level. At the level of page and paragraph, Voynichese is comparable to natural language and structured text. This finding must be taken into account when proposing that Voynichese does not encode meaningful text, as most methods of creating meaningless text will not exhibit this property. See Zandbergen (2021) for discussion of the Cardan Grille cypher method, which could be used to create meaningless text or to encode natural language.

4This is why we include samples of languages which are extremely unlikely to underlie the Voynich manuscript, including Indigenous languages of the Americas and modern constructed languages. We include them not because we think they are likely Voynichese candidates (far from it), but rather so that we can better study the interaction of morphological and lexical typology and the statistical profiles of different languages.
of the Voynich manuscript with the Wikipedia and historical text samples. We discuss in detail the unusually low conditional character entropy of Voynichese, and compare the effect that language, script, transcription system, usage of abbreviations, and typographical convention has on this value.

2 Structure and Content of the Voynich Manuscript

The Voynich manuscript contains 102 folios in its current form. There is evidence that some of the pages have been removed and rearranged from their original ordering. There are Arabic numerals 1-116 in the top right corner of each recto folio. These numerals were probably not written by the original authors, but were added at a later date. Ten of the folios fold out to reveal additional diagrams and text, the largest of which is the “Rose,” a complex six-page foldout. See Davis (2020) for a discussion of manuscript hands and foliation.

While the Voynich document does not appear to have section or chapter titles, it can be divided into five sections based upon the drawings and figures in each section:

1. The Herbal section is the first and largest section, taking up approximately half of the entire manuscript. Each folio contains an illustration of an herb or flower. One or more paragraphs of text are written around the illustration. There are no labels on the illustrations themselves in this section.

2. The Cosmological section consists of circular diagrams and charts that appear to be astrological in nature. Most of them include drawings of stars and stylized suns, with text written in spirals and copious labels. A few of the characters are recognizable medieval astrological symbols. There is also a twelve-page sequence of Zodiac illustrations within concentric circles of text and pictures of women with labels. In most cases, there are exactly thirty women per illustration. The Zodiac signs are correctly ordered and have been labelled at the center with corresponding month names in the Occitan dialect of French (which is probably a later addition and should not be mistaken for Voynichese). There is less running paragraph text in this section, but there are many labels and text written in a circular pattern.

3. The Balneological section contains pictures of what appear to be stylized women bathing in large basins and interconnected ornamental tubes. Each folio contains multiple paragraphs of running text, and many of the women are labelled. This is followed by the six-page Rose foldout, which on one side depicts nine interconnected circular diagrams, many centered with stars or tubes. The other side contains text and more circular diagrams.

---

5The fourteen missing folios are f12, f59, f60, f61, f62, f63, f64, f74, f91, f92, f97, f98, f109, and f110.

6For ease of reference, we have labelled sequentially coherent Section boundaries based on the current order of the manuscript. Therefore, those pages which consist of only text (and are therefore not obviously classifiable) are classified as part of the section in which they appear. Furthermore, the isolated Herbal pages that are found in the Cosmology and Recipe sections are classified as part of the Cosmology and Recipe sections respectively. This differs from the section coding schema employed by the interlinear gloss file available at http://www.voynich.nu. It is very likely that some pages in the manuscript are now in a different order than the order in which they were first composed.

7The number of divisions or points on many of these charts also suggest astrological concepts: twelve representing the houses of the Zodiac, seven representing the planets, four representing the elements (also humors, directions, qualities, or triplicities), eight representing the monastic hours of the day.

8Though it takes up twelve folios, the Zodiac is incomplete because it only depicts ten out of the twelve signs. Capricorn and Aquarius, the first and last signs, are missing. Ares and Taurus are each depicted twice on separate folios. Their charts depict only fifteen women, which suggests that each folio represents a half-month. The only chart with neither thirty nor fifteen women is Gemini, which has twenty-nine.
4. The Recipes section is distinguished by pages with paragraphs of text separated by assort-
ments of labelled herbs, leaves, or roots. To the left of the paragraphs there are what appear
to be ornate jars. In between these pages of “recipes” there is a central section of herbals in
the same style as the Herbal section.

5. The Stars section contains no illustrations and consists of densely packed short paragraphs
of text. Each page contains ten to twenty paragraphs which are marked on the lefthand side
by a seven-pointed star symbol.

There is evidence that more than one scribe produced the text. Currier (1976) noted the
existence of multiple scribal hands, and he also classified pages of the text into two different “lan-
guages” (Currier Language A and B) based upon consistent and marked differences in the frequency
of certain words and glyph combinations. The usage of the term “language” is misleading, because
Language A and Language B do not necessarily represent different natural languages. There are
substantial similarities of structure and vocabulary. They may represent different dialects of the
same language, or they may represent the same dialect but use a slightly different encoding scheme.
With a small number exceptions, every folio is written in only one Language and Scribal Hand,
and each Scribal Hand employs only one language. This implies to us that the scribes who made
the text were also its authors. If they were copying a previous work, we should not expect to find
such a close correlation between the language and scribal hand.

The first half of the Herbal Section is written in Language A, and the second half alternates
between Languages A and B. The Balneological and Stars sections are written entirely in B. The
“recipes” of the Recipes section are all written in A, while the “herbals” in the Recipes section are
written in A or B (suggesting that it was originally part of the Herbals section). The Cosmological
section, which contains mostly labelled diagrams rather than running text, was left unclassified by
Currier, although it most closely resembles Language B.

The recent analysis of Davis (2020) demonstrates evidence for five different scribal hands based
on variations in the formation of several glyphs. She finds that Language A is written entirely by
Hand 1 (with the exception of 58r), while the other hands write in Language B. Hand 2 is found
in the second half of the Herbal section, the entire Balneological section, and thirty-three lines on
a folio in the Stars section (115r) which is shared with Hand 3. Hand 3 is found at the end of the
Herbal and Cosmological sections, the “herbal” portion of the Recipes, and every folio of the Stars
section. The Cosmological section is written almost entirely in Hand 4. Hand 5 is found only in
the second half of the Herbal section.

The amount of text written by Hands 1, 2, and 3 is approximately equal: 10-12 thousand words
each. Hands 4 and 5 are found mostly on diagram labels rather than running text, and account
for less than four thousand words between them. Overall, 87% of the Voynich text is written in
paragraphs and 13% consists of labels on diagrams or drawings.

3 Description of the Corpora

The following sections describe the corpus materials used in the current study.

3.1 The Voynich Corpus

The Voynich corpus consists of digitally transcribed copies of the manuscript itself (see Section 3.1.2
below for more details on the transcription used). We also created separate documents for each
Voynich Language, Scribal Hand, and separated running text from the text found in labels and diagrams. The Voynich corpus used for this project consists of the following documents:

1. **Full Voynich**: The entire Voynich text, including running text, labels, and diagrams.
2. **Full Voynich Text**: The Voynich text written in paragraphs, without text in labels and diagrams.
3. Voynich A: Voynich written in Currier Language A, including running text, labels, and diagrams.

4. Voynich A Text: Voynich written in Currier Language A, without text in labels and diagrams.

5. Voynich B: Voynich written in Currier Language B, including running text, labels, and diagrams.

6. Voynich B Text: Voynich written in Currier Language B, without text in labels and diagrams.

7. Voynich 1: Voynich written in Hand 1, including running text, labels, and diagrams.

8. Voynich 1 Text: Voynich written in Hand 1, without text in labels and diagrams.

9. Voynich 2: Voynich written in Hand 2, including running text, labels, and diagrams.

10. Voynich 2 Text: Voynich written in Hand 2, without text in labels and diagrams.

11. Voynich 3: Voynich written in Hand 3, including running text, labels, and diagrams.

12. Voynich 3 Text: Voynich written in Hand 3, without text in labels and diagrams.

13. Voynich 4: Voynich written in Hand 4, including running text, labels, and diagrams.

14. Voynich 4 Text: Voynich written in Hand 4, without text in labels and diagrams.

15. Voynich 5: Voynich written in Hand 5, including running text, labels, and diagrams.

16. Voynich 5 Text: Voynich written in Hand 5, without text in labels and diagrams.

The particular transcription system used to write Voynichese can have a measurable effect on the statistical properties of the text itself. All of the documents above have been converted into three different transcription systems: Simplified Maximal, Full Maximal, and Minimal. The important issue of transcription is discussed in Section 3.1.1.

3.1.1 Voynich Transcription Systems

Scholars of the Voynich manuscript have proposed several transcription systems to assign characters to particular Voynich glyphs. These transcription systems include FSG, Bennett, Currier, Frogguy, EVA (Extensible Voynich Alphabet), and V101 (see Zandbergen 2010 for an overview). The most commonly used of these systems is EVA. The systems differ in the assumptions they make about what constitutes a single character or character variant, and these assumptions can have an effect on the statistical properties of the text. Research on the most plausible character set is ongoing, and therefore any analysis of Voynichese should take into account the particular assumptions of the given transcription system.

The most significant assumption is whether or not certain sequences of Voynich glyphs constitute a single character or a sequence of multiple characters. Many of these common glyph sequences occur either word-finally or word-initially. They are written as sequences of multiple characters in the EVA transcription system, while the earlier Currier transcription considers them to be single characters. Transcription systems can be ranked according to whether common Voynich glyph sequences are minimally or maximally decomposed into individual characters. The EVA
Figure 4: Glyph combinations and their Minimal and Maximal transliterations. Maximal Voynich is equivalent to EVA, except that plumes are represented by apostrophes. For Minimal Voynich, we have made the substitutions given above, following Currier’s schema. The last two were suggested by Zandbergen (2010). The ee combination is very common in the middle of words. The qo combination is almost always an initial sequence, with q being followed by o 98% of the time.
transcription was designed to be convertible to other major transcription systems. The characters of EVA are therefore intended to represent a lower bound on the length of characters to allow for the conversion to all possible compositions of characters in other systems. Whether or not EVA makes correct assumptions, it is the most convenient transcription system for analyzing Voynich in many cases because it allows for easy conversion into other systems. By contrast, Currier is the most minimally decomposed transcription system of the major systems. Common glyph common combinations tend to be represented as a single character rather than multiple characters.

We take EVA as the basis for our most decomposed system, i.e. the Maximal transcription. The only difference between our Maximal transcription and EVA is in the way that plumes are transcribed. Plumes are looping strokes which are found almost exclusively above the \textit{ch} bench characters (see the example in the Initials section of Figure 3.1.1). There are seven plumes found elsewhere in the document. In EVA, the \textit{c} with a plume above it is written as \textit{s}, and it is considered to be the same character as the glyph \textit{s} which is written with a single connected stroke and is found without an \textit{h} in words like \textit{sol}. For our Maximal transcription, we want the glyphs to be as decomposed as possible, and so we use the apostrophe character for all plumes. Thus for example the word which is transcribed as \textit{shedy} in EVA is transcribed as \textit{c’hedy} in the Maximal transcription. We consider the \textit{s} written with a single unconnected stroke to be a separate character.

For a Minimal transcription, we take EVA and substitute all of the glyph combinations in Currier’s system. We add two additional glyph combinations based on suggestions from Zandbergen (2010). The differences between the Minimal and Maximal transcriptions are outlined in Figure 4. Minimal Voynich represents our effort to create a transcription in which common glyph sequences are minimally decomposed into multiple characters. With future research, handwriting and script analysis will hopefully determine the plausibility of particular glyph decompositions with a higher degree of certainty. But for the present it is useful to compare two transcription systems which represent upper and lower bounds of compositionality.

Transcription systems also differ in the extent to which they represent ligatures and infrequent characters. Ligatures are horizontal lines that are sometimes employed to connect two characters. They may simply be the result of a fluent writing style. These are distinguished in the Full Maximal Transcription (and EVA) by capitalizing the first of the letters (see Figure 3.1.1).

There are sixteen characters which appear less than fifty times each in the entire manuscript. Some of the infrequent characters are recognizable astrological symbols or are found only on cosmological diagrams, while others appear to be variants of other characters or even typographical mistakes. There are also about 250 unreadable glyphs, which are represented in EVA by an asterisk symbol. Altogether, these infrequent characters account for less than 0.15% of the text.

Figure 3.1.1 demonstrates the differences between the Full Maximal and Simplified Maximal transcriptions. Full Maximal is EVA with ligatures and rare characters included. Simplified Maximal removes the ligatures and uses an asterisk to designate all rare characters (with the exception of the \textit{x}, which is the most frequent of the rare characters and appears in both diagrams and text).

\footnote{The exception is V101, which makes different assumptions about many character variants. In particular, V101 assumes that many similar-looking glyphs, which in other systems are considered to be variants of the same character, are different characters. It is not easy to convert between V101 and other systems, and V101 has a larger character inventory. We are utilizing EVA because it is the system used for the interlinear files and because we believe the character inventory size is more plausible.}

\footnote{In a previous version of this paper, the Maximal transcription was identical to EVA. Here we introduce a single difference involving the treatment of plume strokes. The effect on character entropy is minimal: conditional character entropy falls 0.0-3.2% depending on the sample text.}
| Voynich glyphs | Maximal (Full) | Maximal (Simplified) |
|---------------|---------------|---------------------|
|               |               |                     |
| **Ligatures** |               |                     |
| αα           | cth           | cth                 |
| αα            | cTh           | cth                 |
| ιι            | sol           | sol                 |
| ιι            | Sol           | sol                 |
| αα            | ar            | ar                  |
| αα            | Ar            | ar                  |
| **Rare**      |               |                     |
| x             | x             | x                   |
| π             | z             | *                   |
| α             | v             | *                   |
| ιι            | u             | *                   |
| ιι            | Ń             | *                   |

Figure 5: The Simplified Maximal transcription: ligatures are ignored, and rare characters (including unusual plumes) are represented by an asterisk.
The Simplified Maximal alphabet has about half as many characters (23 rather than 45), but this has a minimal effect on character statistics because of the infrequency of rare characters. Figure 6 shows the three transcriptions on a portion of folio 49 recto.
Figure 6: Full Maximal, Simplified Maximal, and Minimal transcriptions of folio 49 recto, paragraph 2, lines 1-5

Maximal Full (EVA):

podaiin cheo kcho daiin chcthy
Sod chol y tcheol daiin cthodd
q'o c'hoqoky c'hor c'heor otol daiin
ochol chol chody dchodaiin dain
q'oho cheey dhcey qotchody

Minimal:

pодаM Seo kSo daM SQy
sod Sol y tSeol daM Qodd
q'o Zоqky Zor Zeor otol daM
oSol Sol Sody dSodaM daM
q'oSo SEy dSey qtSody

Maximal Simplified:

podaiin cheo kcho daiin chcthy
sod chol y tcheol daiin cthodd
q'o c'hoqoky c'hor c'heor otol dain
ochol chol chody dchodaiin dain
q'oho cheey dhcey qotchody
3.1.2 Voynich Document Preparation

The Voynich texts created for this analysis were derived from the Landini-Stolfi Interlinear Gloss File (LSI), which contains multiple transcriptions of the Voynich manuscript in EVA. We used Takeshi Takahashi’s transcription for our corpus because it is the most complete.11 Voynichese in the LSI is written out line-by-line and accompanied by notes and metadata. We used R to parse this code into a long table in which each word of Voynichese is associated with its precise position in the text. This consisted of deleting the notes, copying the page-level and line-level metadata, and separating the words by word breaks.12 We then calculated the word’s position from the beginning and end of the line and from the beginning and end of the paragraph. In our long table, each Voynichese word is listed sequentially along with the following metadata:

1. Full Maximal (EVA) transcription of the word
2. Simplified Maximal transcription of the word
3. Distance from the Beginning of the Line (1, 2, 3, etc.)
4. Distance from the End of the Line
5. Distance from the Beginning of the Paragraph
6. Distance from the End of the Paragraph
7. Paragraph/Diagram designation
8. Line Number on the page
9. Folio Number
10. Quire Number
11. Section of the Manuscript
12. Language (Currier’s designation)
13. Hand (Davis’ designation)
14. Transcriber

The long table can then be consulted to create Voynich documents that focus on particular Voynich Languages, Hands, types of text, or positions within the folio, paragraph, or line. We used it to create the sixteen Voynich documents listed at the beginning of this section: Full Voynich, Full Voynich Text, Voynich A, Voynich A Text, Voynich B, Voynich B Text, Voynich 1, Voynich 1 Text, Voynich 2, Voynich 2 Text, Voynich 3, Voynich 3 Text, Voynich 4, Voynich 4 Text, Voynich 5, and Voynich 5 Text.

---

11 A full set of transliterations can be found at [http://www.voy nich.nu/transcr.html#links](http://www.voy nich.nu/transcr.html#links). There are a small number of gaps in Takahashi’s transcription, including labels on the Rose pages, partial text on the foldout pages for f101, and the small amount of (non-Voynich) text on f116v, the last page.

12 Credible word breaks in the LSI are represented by periods, and possible word breaks by commas. We followed the credible word breaks.
3.2 The Wikipedia Corpus

To date, the online encyclopedia Wikipedia has versions in approximately 319 different languages. These versions are separate collaboratively edited editions which range widely in size. The English edition boasts over six million articles, although only about half of the language editions contain more than 1,000 entries.\textsuperscript{13} While it varies from language to language, a single Wikipedia entry contains on average about 500 words, which means that the Voynich manuscript contains roughly the same amount of text as 75 Wikipedia articles.\textsuperscript{14} Our Wikipedia corpus consists of a sample of every language that has more than 100 articles.

The primary advantage of the Wikipedia Corpus is that we can compare Voynich text with that of many different languages, language families, and scripts, and see whether Voynich falls within the range of plausible languages or language families, and if so, which languages or families it most closely resembles statistically. The conventions, motivations, and contents of modern online encyclopedias are obviously very different from that of a medieval herbal and astrological manuscript. However, both consist of discrete collections of informative text on specialized topics. The language of Voynichese should in many structural aspects be more akin to modern Wikipedia entries than, for example, medieval diary entries or historical narratives. In contrast to encyclopedia entries, we would expect narratives to follow a temporal sequence throughout, for verbs to be predominantly in the past tense, and for certain names and pronouns to recur predictably. The Wikipedia corpus is thus particularly well-suited for comparison with the Voynich texts. It is also superior to more formal genres of corpora like newspaper corpora, which are written for a different purpose and contain far fewer languages.

The Wikipedia Corpus consists of 311 language samples written in thirty-four different scripts, categorized into thirty-nine major language families and seventy-three subfamilies. In most cases, the samples consist of the first 200,000 words from Wikipedia entries for that language edition, listed alphabetically by headword. The corpus includes samples of many languages which are plausible candidates for Voynichese, e.g., Romance dialects like Corsican and Lombard and Germanic dialects like Bavarian and Low Saxon. There are also samples of extinct languages like Gothic, Anglo-Saxon, and Pali, as well as nine modern artificially-constructed languages including Esperanto and Lojban.

Some language families are particularly well-represented in the Corpus, with ten or more language samples for each family. These are the Bantu, Germanic, Indic, Iranian, Malayo-Polynesian, Romance, Slavic, Tibeto-Burman, Turkic, and Uralic families.

3.2.1 Wikipedia Document Preparation

The documents used for this analysis were obtained February 2021 from Wikimedia dump files.\textsuperscript{15} These files are continually updated and available at \url{http://dumps.wikimedia.org}. We down-

\textsuperscript{13}See \url{http://meta.wikimedia.org/wiki/List_of_Wikipedias} for a current list of Wikipedia versions by number of entries.

\textsuperscript{14}The entry word count for English Wikipedia is much larger than most other languages because it tends to have longer entries. The average word count per entry in our English sample is over 3,700.

\textsuperscript{15}In a previous version of this paper, we described an earlier process for obtaining Wikipedia sample texts. In 2019, we processed Wikidump files using the Python \texttt{Gensim} module and limited our samples to the first 500 articles in each language. Our new methodology creates dramatically cleaner texts, and the text sizes are more even because we limit the text by number of words rather than number of articles. Between 2019 and 2021, Wikipedia added the following languages: Awadhi, Balinese, Guianan Creole, Kotava, Ladin, Madurese, Mon, Moroccan Arabic, Nias, N’Ko, Saraiki, Inari Sami, and Sakizaya. The Northern Luri Wikipedia was deleted. These changes have been made to the lists of languages and families given below.
loaded the BZIP2 compressed files titled *Articles, templates, media/file descriptions, and primary meta-pages*. For most of the dump files, we ran a python script to process them into raw text documents containing the first 200,000 words from wikipedia entries for that language, including only articles that consist of 100 or more words.\(^{16}\)

We deleted any remaining metadata and tables by hand. We then further processed the texts by removing punctuation and capitalization, and deleted any characters with less than a .01% occurrence in Latin, Cyrillic, Greek, Arabic, and Hebrew. In English, this filters everything but the lowercase letters \((a-z)\). We filtered the text by the unicode range of the particular script in order to delete irrelevant characters.

Wikipedia versions differ widely by average article length. A small number of languages, including Cree, Cheyenne, and Inupiak, contain mostly short single-line articles with a repetitive structure. For these samples we included articles of all sizes and hand-deleted repetitive text, but this means that the resulting sample texts are quite short. Some other wikipedia versions, most notably Cebuano and Waray-Waray, have an inflated number of articles because a high percentage of entries were created by bots. Bot-created entries also tend to be short and formulaic, although this should be a minimal issue for our corpus because we deleted articles of less than 100 words.\(^{17}\)

### 3.2.2 Wikipedia Languages by Family

This is a full list of the language samples in the Wikipedia Corpus categorized by language family and sub-family. An ideal corpus would contain a large number of languages from each language family, and the sub-family categories would represent languages at an approximately equal time depth of divergence. However, the list of Wikipedia languages, while representing an impressive diversity of language families, is nevertheless skewed heavily towards European languages.

The categorization below is an attempt to group together languages which are genetically similar while keeping the size of the categories approximately equal. Sub-families were chosen in language families with a large representation in the Corpus. If there are single languages from distinct sub-families, an “Other” category is used. For simplicity, extinct languages and proto-language progenitors of a family (e.g. Latin, Sanskrit, and Gothic) are also grouped into the “Other” category.

There are two categories in this list that are not based upon genetic relatedness. All artificially-constructed languages are grouped under a single category. The Constructed languages that have Wikipedia versions are all international auxiliary languages meant to facilitate communication (as opposed to artistic constructed language like Klingon or Quenya). They are all heavily based on the vocabulary and grammar of European languages.\(^{18}\) The second category is that of Creoles, which are not the product of language divergence in a single family but rather have a complex genetic relationship with two or more language families. Here they have been subcategorized by their lexifier language, which is the language from which most of their vocabulary is drawn.

\(^{16}\)The python script, along with the text files and a more in-depth description of the extraction process, is available at [http://www.lukelindemann.com/wiki_corpus.html](http://www.lukelindemann.com/wiki_corpus.html). For languages with scripts that do not use spaces for words, e.g. Thai and Chinese scripts, we instead extracted the first 100 articles. Some versions utilize two different scripts in different entries, e.g. Kashmiri has entries in either Devanagari or Arabic scripts, and we created two separate texts in this case.

\(^{17}\)Note that we have not filtered articles for whether they were created by bots, such as the Lsjbot (cf. [https://en.wikipedia.org/wiki/Lsjbot](https://en.wikipedia.org/wiki/Lsjbot)).

\(^{18}\)The exception is Lojban, which is a constructed logical language designed as an experiment in eliminating syntactic ambiguity.
1. **Afro-Asiatic:**
   (a) **Semitic:** Amharic, Arabic, Aramaic, Egyptian Arabic, Hebrew, Maltese, Moroccan Arabic, Tigrinya
   (b) **Other families:** Hausa, Kabyle, Oromo, Somali
2. **Albanian:** Albanian
3. **Algonquian:** Atikamekw, Cheyenne, Cree (Canadian Syllabics), Cree (Latin)
4. **Armenian:** Armenian, Western Armenian
5. **Athabaskan:** Navajo
6. **Austroasiatic:** Khmer, Mon, Santali, Vietnamese, Banjar
7. **Aymara:** Aymara
8. **Baltic:** Latgalian, Latvian, Lithuanian, Samogitian
9. **Caucasian:** Abkhazian, Adyghe, Avar, Chechen, Ingush, Kabardian Circassian, Lak, Lezgian
10. **Celtic:** Breton, Welsh, Irish, Scottish Gaelic, Manx, Cornish
11. **Constructed:** Esperanto, Interlingua, Interlingue, Ido, Kotava, Lojban, Lingua Franca Nova, Novial, Volapük
12. **Creoles:**
   (a) **English:** Bislama, Jamaican Patois, Norfolk, Sranan, Tok Pisin
   (b) **French:** Haitian
   (c) **Portuguese:** Guianan Creole, Papiamentu
   (d) **Spanish:** Zamboanga Chavacano
13. **Dravidian:** Kannada, Malayalam, Tamil, Tulu, Telugu
14. **Germanic:**
   (a) **Anglic:** English, Scots, Simple English
   (b) **Dutch:** Afrikaans, Dutch, West Flemish, Zeelandic
   (c) **Frisian:** North Frisian, West Frisian, Saterland Frisian
   (d) **High German:** Alemannic, Bavarian, German, Ripuarian, Luxembourghisch, Palatinate German, Yiddish
   (e) **Low German:** Low Saxon, Dutch Low Saxon
   (f) **North Germanic:** Danish, Faroese, Icelandic, Norwegian (Nynorsk), Norwegian (Bokmål), Swedish
   (g) **Other families/proto-languages:** Anglo-Saxon, Gothic, Limburgish
15. **Hellenic:** Greek, Pontic
16. **Indic:**
   (a) **Central:** Awadhi, Hindi, Urdu, Fiji Hindi,
   (b) **Eastern:** Assamese, Bengali, Maithili, Odia, Bihari, Bishnupriya Manipuri,
(c) Northern: Doteli, Nepali  
(d) Northwestern: Sindhi, Punjabi, Saraiki, Western Punjabi  
(e) Southern: Sinhalese, Divehi, Marathi, Goan Konkani  
(f) Western: Romani, Gujarati  
(g) Other families/proto-languages: Kashmiri (Arabic), Kashmiri (Devanagari), Sanskrit, Pali (Devanagari), Pali (Latin)  

17. Inuit: Greenlandic, Inuktutit (Canadian Syllabics), Inuktutit (Latin), Inupiaq  
18. Iranian: Sorani, Zazaki, Persian, Gilaki, Kurdish, Mazandarani, Ossetian, Pashto, Tajik  
19. Iroquoian: Cherokee  
20. Japonic: Japanese  
21. Kartvelian: Georgian, Mingrelian  
22. Koreanic: Korean  
23. Malayo-Polynesian:  
   (a) Javanesic: Banyumasan, Javanese  
   (b) Malayic: Indonesian, Malay, Minangkabau  
   (c) Polynesian: Tongan, Hawaiian, Maori, Samoan, Tahitian  
   (d) Phillipine: Central Bicolano, Cebuano, Gorontalo, Ilokano, Pangasinan, Kapampangan, Tagalog, Waray-Waray  
   (e) Other families: Acehnese, Balinese, Buginese (Buginese), Buginese (Latin), Chamorro, Fijian, Madurese, Malagasy, Nauruan, Nias, Sundanese, Tetum  
24. Mande: Bambara, N’Ko  
25. Moksha: Moksha  
26. Mongolic: Buryat, Mongolian, Kalmyk  
27. Niger-Congo:  
   (a) Bantu: Kongo, Lingala, Luganda, Northern Sotho, Chichewa, Kikuyu, Kinyarwanda, Kirundi, Shona, Swati, Sesotho, Swahili, Tswana, Tswa, Tumbuka, Twi, Venda, Xhosa, Zulu  
   (b) Other families: Akan, Ewe, Fula, Igbo, Kabiye, Sango, Wolof, Yoruba  
28. Nilotic: Dinka  
29. Quechua: Quechua  
30. Romance:  
   (a) Italo-Dalmatian: Italian, Corsican, Sicilian, Neapolitan, Venetian, Tarantino  
   (b) Gallo-Romance: Catalan, French, Franco-Provençal, Ladin, Occitan, Picard, Norman, Walloon  
   (c) Gallo-Italic: Piedmontese, Ligurian, Lombard, Emilian-Romagnol  
   (d) Iberian: Aragonese, Asturian, Extremaduran, Galician, Ladino, Mirandese, Portuguese, Spanish
(e) Other families/proto-languages: Aromanian, Friulian, Romanian, Romansh, Sardinian, Latin

31. Slavic

(a) East Slavic: Belarusian, Belarusian Taraskevica, Russian, Ukrainian
(b) West Slavic: Czech, Kashubian, Lower Sorbian, Upper Sorbian, Polish, Rusyn, Slovak, Silesian
(c) South Slavic: Bulgarian, Bosnian, Croatian, Macedonian, Serbo-Croatian, Serbian, Slovenian, Old Church Slavonic

32. Tai: Lao, Shan, Thai, Zhuang

33. Tibeto-Burman: Tibetan, Min Dong, Dzongkha, Gan, Hakka, Burmese, Newar, Wu, Chinese, Classical Chinese, Min Nan, Cantonese

34. Tupian: Guarani

35. Turkic:

(a) Oghuz: Azerbaijani, Chuvash, Gagauz, South Azerbaijani, Turkmen, Turkish
(b) Karlik: Uyghur, Uzbek
(c) Kipchak: Bashkir, Crimean Tatar, Karakalpak, Kazakh, Karachay-Balkar, Kirghiz, Tatar (Cyrillic), Tatar (Latin)
(d) Siberian: Sakha, Tuvin

36. Uralic:

(a) Finnic: Estonian, Finnish, Vepsian, Võro
(b) Permic: Komi-Permyak, Komi, Udmurt
(c) Mari: Meadow Mari, Hill Mari
(d) Sami: Inari Sami, Northern Sami
(e) Other families: Erzya, Hungarian, Livvi-Karelian, Northern Sami

37. Uto-Aztecan: Nahuatl

38. Vasconic: Basque

3.3 The Historical Corpus

The Wikipedia Corpus contains a large number of languages and language families, but it consists entirely of modern texts. It is therefore necessary to compare Voynichese to contemporaneous historical manuscripts as well, because there are important differences between modern and historical texts which are not typically addressed in statistical analyses of Voynichese.

One important point of difference is spelling standardization, which is much higher in most modern languages than it is in medieval manuscripts. This is less of an issue for medieval Latin texts, as Latin has been standard since the Classical Period, but it is an important consideration for the many written languages which had yet to standardize by the 15th century. Spelling variation will have an effect on statistics like type-token ratio because a single word will be represented by multiple types.
A second important difference concerns the typographical conventions of scribes. Because all literature was written and copied by hand during this period, scribes developed hundreds of abbreviations and symbols to represent frequently occurring phrases, words, and grammatical functions. This was especially prevalent in Latin texts. It introduces variability of a different type, and has an effect on statistics like the information entropy of the text. However, most modern transcriptions of historical manuscripts omit these abbreviations and conventions for readability.

3.3.1 Description of the Corpus

The Historical Corpus consists of transcriptions of manuscripts written between 400 and 1600 AD. The corpus is continuously updated as we discover new sources of digitally transcribed historical manuscripts. The languages represented in the corpus currently include English, Georgian, Hebrew, Icelandic, Italian, Latin, Persian, and Spanish. The majority of the texts are in Latin and English.

In order to match the presumed contents of the Voynich manuscript, we have made an effort to include texts on magic, astrology, and alchemy. Many of the important texts in this genre, including the highly influential *Secretum Secretorum*, were originally written in Arabic or Persian and were being translated into Latin and vernacular European languages during the time that the Voynich manuscript was created. We have included a Latin and English translation of the *Secretum Secretorum*, an English translation of the *Alphabet of Tales*, Agrippa’s *Three Books of Occult Philosophy*, a Spanish translation of *Picatrix*, and Bruno’s Latin *De Magia*. We have also included Trithemius’ *Steganographia*, which is ostensibly about magic and spirit communication but is in fact an enciphered treatise on cryptography. In the historically related topic of Medicine, we have included the *Science of Cirurgie* and the archives of Richard Napier’s medical records collected by the Casebooks Project at the University of Cambridge.

A secondary goal in the creation of the Historical Corpus is to collect manuscripts in parallel diplomatic and normalized versions. The diplomatic version of a manuscript uses special characters to faithfully replicate the original abbreviations and typographical conventions, while the normalized version does not use abbreviations and the orthography is typically modernized. This allows us to directly compare the effect of typographical conventions on the same text, which may provide insights into the peculiar properties of Voynichese. We have included parallel diplomatic and normalized versions of three texts: the Icelandic *Codex Wormianus*, the English *Medical Casebooks*, and the Latin *Necrologium Lundense*. For the other Latin texts we also created our own abbreviated forms of the texts based on widespread orthographic conventions; this material will be discussed in forthcoming work.

A similar issue is found with abjad scripts like Arabic and Hebrew, which are typically written without vowels. The exclusion of vowels has an effect on the entropy statistics of a text. We have included two versions from the Tanakh: one with and one without the *niqqud* diacritics which are used primarily to mark vowels.

Table 1 lists the historical manuscripts in the corpus, along with their language, script, approximate date of composition, and author (or translator or scribe).

---

19 A diplomatic transcription is technically distinct from a type facsimile, which uses digital fonts to replicate the exact appearance of the text. For our purposes, we consider the most faithful available reproduction of a text to be the diplomatic transcription.
### Table 1: Details of historical manuscripts

| Name                                      | Language | Script | Author                        | Date   |
|-------------------------------------------|----------|--------|-------------------------------|--------|
| Medical Casebooks                         | English  | Latin  | Richard Napier               | 1597   |
| Three Books of Occult Philosophy          | English  | Latin  | Heinrich Cornelius Agrippa   | 1509   |
| Science of Cirurgie                       | English  | Latin  | Lanfranc of Milan             | 1306   |
| Secretum Secretorum                      | English  | Latin  | Robert Copland (translator)  | 1528   |
| Alphabet of Tales                         | English  | Latin  | Etienne de Besançon           | 1400   |
| Amiran-Darejaniani                        | Georgian | Georgian | Mose Xoneli                 | 1150   |
| Mishneh Torah                             | Hebrew   | Hebrew | Maimonides                   | 1170   |
| Masoretic Tanakh                          | Hebrew   | Hebrew | Aaron ben Moses ben Asher (scribe) | 1008   |
| Codex Wormianus                           | Icelandic| Latin  | Unknown                      | 1350   |
| La Rettorica                              | Italian  | Latin  | Brunetto Latini              | 1261   |
| Necrologium Lundense                      | Latin    | Latin  | Unknown                      | 1123   |
| De Ortu et Tempo Antichristi             | Latin    | Latin  | Adso Deruensis               | 900    |
| Historia Hierosylmitanae Expeditionis    | Latin    | Latin  | Albert of Aix                | 1125   |
| De Magia                                  | Latin    | Latin  | Giordano Bruno               | 1590   |
| Secretum Secretorum                      | Latin    | Latin  | Philip of Tripoli (translator) | 1270   |
| Steganographia                           | Latin    | Latin  | Johannes Trithemius          | 1499   |
| Sindbad-Name                              | Persian  | Arabic | Zahiri Samarqandi            | 1362   |
| Picatrix                                  | Spanish  | Latin  | pseudo-Majriti              | 1256   |

#### 3.3.2 Historical Document Preparation

The transcribed texts were obtained from multiple sources. The Georgian, Italian, and Persian texts come from the TITUS Project at the University of Frankfurt.\(^{20}\) The Icelandic text comes from the Medieval Nordic Text Archive.\(^{21}\) The Hebrew Masoretic Tanakh comes from Sacred Texts\(^{22}\) and the Christian Classics Ethereal Library\(^{23}\), while the *Mishneh of Maimonides* was obtained here.\(^{24}\)

Of the Latin texts, the *Secretum Secretorum* comes from the Corpus Corporum of the University of Zurich\(^{25}\), the *Necrologium Lundense* comes from the Necrologium Lundense Online\(^{26}\), *De Ortu et Tempo Antichristi* and *Historia Hierosylmitanae Expeditionis* come from the Latin Library\(^{27}\), and *De Magia* and *Steganographia* come from the Twilit Grotto.\(^{28}\)

\(^{20}\) [http://titus.uni-frankfurt.de](http://titus.uni-frankfurt.de)

\(^{21}\) [http://clarino.uib.no/menota/page](http://clarino.uib.no/menota/page)

\(^{22}\) [http://www.sacred-texts.com/bib/tan/index.htm](http://www.sacred-texts.com/bib/tan/index.htm)

\(^{23}\) [http://www.ccel.org/a/anonymous/hebrewot/home.html](http://www.ccel.org/a/anonymous/hebrewot/home.html)

\(^{24}\) [http://kodesh.snunit.k12.il/i/0.htm](http://kodesh.snunit.k12.il/i/0.htm)

\(^{25}\) [http://mlat.uzh.ch/MLS/](http://mlat.uzh.ch/MLS/)

\(^{26}\) [http://notendur.hi.is/mjm7/](http://notendur.hi.is/mjm7/)

\(^{27}\) [http://www.thelatinlibrary.com](http://www.thelatinlibrary.com)

\(^{28}\) [http://esotericarchives.com](http://esotericarchives.com)
The English Alphabet of Tales and Science of Cirurgie come the Corpus of Middle English Prose and Verse at the University of Michigan.\textsuperscript{29} The Medical Casebooks come from the Casebooks Project of the University of Cambridge.\textsuperscript{30} The English translation of the Secretum Secretorum comes from Colour Country,\textsuperscript{31} and Agrippa’s Three Books of Occult Philosophy are from the Twilit Grotto.\textsuperscript{32}

Some of these texts are much longer than the Voynich manuscript, and so we have included only a portion of the entire text. We restricted the Masoretic Tanakh to the Bereshit, i.e. the Book of Genesis. We included the introduction and first two books of the Mishneh, the first sixty pages of the Codex Wormianus, and the first three books of the Science of Cirurgie. The Necrologium Lundense currently has normalized and diplomatic transcriptions of three folios (f124v, f125r, f125v), but they are substantive enough that we included them here. For the Medical Casebooks, we copied the first fifty chronologically sorted consultations taken by Richard Napier and written in his hand.

As with the Wikipedia Corpus, the historical documents were cleaned by removing capitalization and punctuation, as well as notes made by transcribers. For texts in Latin scripts, symbols with a less than .01% occurrence were removed. Texts written in non-Latin scripts were filtered by unicode range. For the diplomatic texts, special characters – including character variants and astrological symbols – were left intact.

4 Conditional Character Entropy in Voynichese

In this section, we demonstrate the usage of the Historical and Wikipedia corpora by examining the character-level properties of Voynichese. As discussed in Bowern and Lindemann (2020), we are particularly interested in the metric of conditional character entropy, or second-order character entropy ($h^2$). Conditional character entropy is a measure of the overall predictability of characters in a text. In his 1976 book on computational applications to scientific and engineering problems, Yale physicist William Bennett Jr. used a transcription of Voynichese to illustrate the concept of information entropy in language and its application to cryptography. He found the conditional character entropy of Voynichese to be surprisingly low compared to a sample of European plain texts and ciphers. This means that Voynichese characters are unusually predictable compared to most European languages. We discuss the definition of conditional character entropy and the history of its application to the Voynich manuscript more thoroughly in Bowern and Lindemann (2020).

This conditional character entropy value of a text, $h^2$, is dependent upon the conventions of the script in which it is written. For example, Bennett (1976) found that the $h^2$ of Voynichese was roughly equivalent to that of a Hawaiian text. Stallings (1998) pointed out that Bennett’s Hawaiian sample used a simplified orthography that did not contain glottal stops or distinguish between long and short vowels, and this has the effect of making the Hawaiian text look more predictable (and more like Voynichese). The following factors potentially have an effect on the $h^2$ value of a text:

1. Document Length
2. Character set size (total number of characters in the alphabet)

\textsuperscript{29}http://quod.lib.umich.edu
\textsuperscript{30}http://casebooks.lib.cam.ac.uk
\textsuperscript{31}http://www.colourcountry.net/secretum/
\textsuperscript{32}http://esotericarchives.com
3. Type of script (alphabet, syllabary, abugida, abjad, etc.)
4. Abbreviations and other typographical conventions
5. Encoding process (if the text is a cipher)

We discuss the first four of these factors in detail below (the fifth will be discussed in a forthcoming paper). We find that Voynich A and B are of a sufficient length that the \( h_2 \) values are reliable, and that the exclusion of rare characters (Maximal Simplified as opposed to the Full Maximal transcription) has a negligible effect on entropy.

The type of script (Maximal as opposed to Minimal) has a more appreciable effect on entropy, but all transcriptions of Voynich are significantly lower than any other text in the corpora. An analysis of script types in the Wikipedia and Historical corpora shows that Voynich most closely resembles an alphabetic script rather than an abjad, abugida, or syllabary.

We compare the parallel diplomatic and normalized versions of historical texts, as well as the forms of Hebrew with and without vowels, and conclude that the unusual character entropy of Voynich is not attributable to conventional scholarly abbreviations or the absence of characters that represent vowels.

At the character level, Voynichese most closely resembles tonal languages written in the Latin script and other languages in which there is a restricted set of word-final characters. This is likely the result of an encoding process, and may suggest that Voynichese simplifies the phonemic distinctions of the language it represents.

4.1 Character Frequency Distribution

In its simplest characteristics, Voynichese does not appear very different from other texts in the Historical and Wikipedia corpora. The character set size for both the Maximal transcription (42 characters) and Minimal transcription (45 characters) is well within the general range for alphabets: 25-92 characters.\(^{33}\) As discussed in Section 4.4, the Voynichese character set size is small compared to non-alphabetic script types like abugidas and syllabaries, but it is the right size for an alphabet.

Secondly, the character frequency distribution of Voynichese is fairly typical. This is demonstrated in Figure 7, which displays the ranked proportional frequencies of the thirty most frequent characters in Voynichese compared with those of texts in the Historical and Wikipedia corpora.

Character frequency distribution is related to and reflected in the metric of unigram character entropy \((H1)\), which measures character-level predictability irrespective of position within the text. Here again, Voynichese is not unusual: \( H1 \) is 3.94 for Minimal Voynich and 3.91 for Maximal Voynich, while the overall range for alphabets in the corpora is from 3.57-4.82 bits.

The Voynich text only begins to look unusual when we factor in the position of a character within the text. Conditional character entropy \((h2)\) measures the predictability of a character given the character that precedes it. This is the metric that Bennett (1976) found to produce unusually low values in Voynichese, and on which we focus this analysis.

4.2 Entropy Variance and Document Length

Two texts written in the same language and script may have slightly different \( h2 \) values due to differences in content and stylistic variation. With a large enough text sample, this variation is

\(^{33}\) The Simplified Maximal transcription has only 23 characters, but it excludes rare characters and therefore represents an absolute lowest estimate.
Figure 7: Proportional Frequency of the thirty most frequent characters in the Historical Corpus texts, the (alphabetic) Wikipedia Corpus texts, and Voynichese (Full Maximal and Minimal transcriptions). Simplified Maximal is identical to Full Maximal with regards to frequent characters. The first ranked character for each language is a space, and its frequency indicates average word length.

minimal. If a text sample is very short, there will not be enough data to obtain a reliable $h2$ result, and there will be more variation.

The Voynich manuscript contains roughly 38 thousand words, of which 11 thousand are in Voynich A and 23 thousand are in Voynich B. We need to know whether these lengths are sufficient for obtaining reasonably certain $h2$ values, and what sort of variance can be expected.

We tested $h2$ variance using the English wikipedia sample, which consists of 199,564 words. We calculated the $h2$ values samples of randomly selected sequences of text at various word lengths. The results are in Figure 8.

With a window of only 50 words, the average $h2$ is 2.62, there is a wide range of 2.0-2.9, and the standard deviation is 0.12. The average is much lower than the text’s overall $h2$ value of 3.40. As the window size increases, the variance tightens and the averages converge on the overall value. With a window of 10,000 words, the average is 3.35, the range is 3.22-3.42, and the standard deviation is 0.033. This means that 95% of the samples are within 0.066 bits of the average, and the average is within 0.05 bits of the text’s overall $h2$ value. For documents of around 10,000 words we should therefore reasonably expect $h2$ to be accurate to about one-tenth of a bit.

When running the same procedure on Voynich A and Voynich B, the $h2$ variance is comparable to that of English. The standard deviation at 50 words is 0.13 for A and 0.15 for B (compared to 0.13 in English), and at 5,000 words it is 0.049 for A and 0.072 for B (compared to 0.048 in English).

This means that the Voynich A and Voynich B sample are large enough to obtain reasonable entropy calculations. However, an analysis at the level of sections, scribal hands, or folios will be
somewhat less reliable.

4.3 Entropy in Voynichese

For Voynichese, Language A and Language B pattern differently. Conditional character entropy in Language B is lower regardless of the transcription system, while in Language A it is only slightly higher than in the combined Full text (see Figure 9). The compositionality of the transcription system has an effect on $h^2$. Maximal Voynich has a lower $h^2$ than Minimal Voynich, because glyph compositions are based upon common glyph sequences (making the text appear more predictable).

The $h^2$ values for running paragraph text (excluding labels on diagrams) is slightly lower than text with labels included. Voynichese in labels and diagrams has a higher $h^2$. Scribe Hand 1, which is used to write Language A, has a nearly identical $h^2$. Scribe Hands 2 and 3, which are used to write Language B, have values quite similar to B. Scribe Hands 4 and 5 have values slightly higher than Language B, as they are mostly employed in writing labels and diagrams. These values are listed in Appendix A.

In all cases, the character set size for Voynichese is between 21 and 45 characters, and the conditional character entropy ranges from 1.91 to 2.56 bits.

4.3.1 Character Set Size

The Full Maximal and Maximal Simplified transcriptions have nearly identical $h^2$ values despite a significant difference in the size of the character sets. With the Full Voynich text, the Simplified Maximal transcription has twenty-two fewer characters, but there is only a .08% difference in conditional character entropy (2.114 to 2.112).
Figure 9: Character set size is plotted against conditional character entropy ($h_2$). Lower $h_2$ values indicate more predictability in character bigrams.

This is significant, because conditional character entropy can be affected by character set size (as noted in Stallings 1998). It is potentially important because the upper bound for conditional character entropy is determined by the value of $H_0$, calculated as the logarithm of character set size. For example, a text that uses an alphabet of 16 characters will have a maximum conditional character entropy of 4 bits, while an alphabet of 49 characters will have a maximum $h_2$ of 7 bits.

However, if the additional characters are rare, the overall effect on $h_2$ is slight. The process of cleaning a text by removing highly infrequent characters does not have an appreciable effect on entropy. After removing the capitalization, numerals, and punctuation from the raw English sample, there are 75 remaining characters. These includes characters which are very rare in English or exist only in foreign words, such as ū and ç. Processing the sample involves removing characters which appear with a frequency less than 0.01%, after which only 27 characters remain. However, the difference in $h_2$ between the filtered and unfiltered English sample is only 0.08% (3.406 to 3.403).

On the other hand, there is an appreciable difference in entropy between the Minimal and Maximal transcriptions, despite the fact that the character set size is roughly similar. This demonstrates that decisions about the composition of high-frequency glyph sequences have a greater effect on entropy than decisions about the inclusion of low-frequency characters.
4.3.2 Languages A and B

Voynich Language A and Language B are similar at the character level. Despite the fact that they have different distributions at the word level, the most frequent character sequences are roughly the same in both languages. There are two exceptions, which illustrate the difference in entropy between the two texts.

The -edy glyph sequence found at the end of words is eighty-six times more common in Voynich B (one out of five words in Voynich B end with this sequence). Secondly, the qo- sequence at the beginning of a word is about twice as common in Voynich B (also found in one in five words). The frequency of these two sequences alone substantially increase the predictability of the Voynich B text, and this is the main source of the differences in conditional character entropy between A and B. If the two sequences are removed from both texts, then the \( h^2 \) value for Language A and Language B come within about 1% of each other.\(^{34}\)

4.4 Comparison to the Wikipedia Corpus

Of the 311 wikipedia language samples represented in the Wikipedia Corpus, none of them have an \( h^2 \) comparable to Voynichese. Voynichese has lower values, meaning that its text is more predictable. While the Minimal Voynich transcription is slightly higher (with an average \( h^2 \) of 2.48 rather than 2.11), this is still lower than the \( h^2 \) range in the Wikipedia Corpus, from 2.77-6.14.

Figure 10 depicts the character set size and conditional character entropy for texts that use between 20 and 55 characters and have an \( h^2 \) range from 2.5 to 4. This is the range of most of the alphabets in the Corpus. The majority of Wikipedia versions (202 languages) are written in the Latin script, although the Corpus also includes samples of Cyrillic, Georgian, Gothic, Greek, Ol Chiki and N’Ko. Languages written in Cyrillic, which are mostly Slavic and Turkic, tend to have a somewhat higher \( h^2 \).

Most of the abjads are also in this range. Abjads are writing systems in which consonants are written and vowels are (mostly) not represented. The abjads in the Corpus include Arabic and Hebrew. They are used to write Afro-Asiatic (specifically Semitic), Indic, Iranian, and Turkic languages, as well as the Germanic language Yiddish. They also have a somewhat higher \( h^2 \) on average.

The conditional character entropy of texts written in Latin scripts ranges from 2.8-3.8, and includes all of the languages with the lowest \( h^2 \) values. The languages with \( h^2 \) values closest to Voynichese are Hawaiian, Venda, Tsrama, and Min Dong. Hakka and Min Dong are Tibeto-Burman languages, while Venda and Tsrama are Bantu (Niger-Congo). It is noteworthy that all except Hawaiian are tonal languages that use a Latin script for their orthography. That is, they systematically collapse suprasegmental distinctions.

Expanding outward, Figure 11 includes languages that contain up to 130 characters and have a conditional entropy between 2.5 and 4.5. These primarily consist of the abugidas. Abugidas are writing systems in which consonant-vowel sequences are written as a unit, with consonants as the primary symbol and vowels added to it. The abugidas in the corpus include Bengali, Buginese, Devanagari, Gujarati, Gurmukhi, Kannada, Khmer, Lao, Malayalam, Myanmar, Odia, Sinhala, Tamil, Telugu, Thaana, Thai, and Tibetan. They are all derived from the Brahmi script, and are

\(^{34}\)The frequency of this common glyph sequence is partially attributable to a single word chedy, which is the most common word in B and almost entirely absent from A. But even when this word is disregarded, -edy is significantly more common in B.
used to write Austroasiatic, Dravidian, Indic, Tai, and Tibeto-Burman languages. The abugidas usually have many more characters but tend to have only a slightly higher entropy.

Expanding even further, Figure 12 includes languages that contain up to 7000 characters and have a conditional entropy up to 6.5. The languages that are written with logograms have the highest character set sizes and highest entropy values. For these languages, each character represents a morpheme or word, and thus character entropy is approximately equal to word entropy. Chinese logograms are used to write varieties of Chinese and other Tibeto-Burman languages: Cantonese, Chinese, Classical Chinese, Gan, and Wu. Also included under this category is Japanese, which uses a mixed writing system with logograms and two syllabaries, and ranges between syllabaries and logograms in character set size and $h^2$.

Between this extreme and the abugidas are syllabaries, in which a single character denotes a syllable. The syllabaries include scripts designed for the Iroquoian language Cherokee and the Inuit language Inuktitut. They also include the Hangul script used for Korean and the Ethiopian (Ge’ez) script used to write Amharic and Tigrinya. Technically, Ethiopian is an abugida and Hangul is an alphabet, but both are represented in unicode by separate codes for each full syllable rather than with combining characters. Thus they have the character set size and conditional entropy in the syllabary range. The full table of values for each language may be found in Appendix B.

The Voynichese script patterns most closely with the alphabets in the Wikipedia Corpus. The
character set size of Full Maximal and Minimal Voynichese are quite similar to most of the alphabets in the sample, but the $h^2$ values are lower than we see with any of the languages in the Wikipedia Corpus. The abjads are also similar, with an equivalent character set size but a slightly higher $h^2$. Voynichese clearly falls outside of the range of most abugidas, syllabaries, and logograms.

### 4.5 Comparison to the Historical Corpus

Figure 13 shows the range for the texts in the Historical Corpus. The texts written with alphabets (Latin and Georgian) have an $h^2$ range between 3 and 3.5, while the abjads (Hebrew and Arabic scripts) range from 3.5 to 4. The five English and five Latin texts demonstrate the variability in $h^2$ and character set size within the same language. Some of these differences are attributable to script variation. The English Medical Casebooks, like the Icelandic Codex Wormianus and Necrologium Lundense, has a larger alphabet because it contains somewhat more characters in the normalized versions as well as the diplomatic versions.

#### 4.5.1 Parallel Diplomatic and Normalized texts

Figure 14 shows how conditional entropy varies between parallel versions of the same text when different forms of transliteration are employed. For the English *Casebooks*, Latin *Necrologium*
Figure 12: Conditional Character Entropy and Character Size for Wikipedia Languages from 20-7000 characters. The grey box indicates the range of the previous graph.

*Lundense,* and Icelandic *Codex Wormianus,* this consists of the normalized and diplomatic versions. For the Hebrew Bereshit, this consists of the text with or without *niqqud* vowel-marking diacritics. We also compare the Full Maximal and Minimal transcriptions of the Voynich texts.

The graph illustrates the variation in the conditional entropy among the versions of historical texts, which relates primarily to character set size rather than to the conditional entropy values. The biggest difference in \( h_2 \) is from the Hebrew Bereshit text, where including *niqqud* lowers the values of \( h_2 \) by 0.25 bits. Note that the variation between Voynich hands and characters is of a similar order, but all the Voynich measurements are substantially lower than the historical samples.

The usage of abbreviations and special characters has the effect of raising the conditional character entropy of the English, Icelandic, and Latin texts and taking them further from the values we find for Voynichese. The Minimal transcription of Voynich has a slightly higher conditional character entropy, but it is clear that the extremely low conditional entropy of Voynichese is not simply attributable to a particular Voynich transcription system or the kinds of abbreviations and typographical conventions that were common in European manuscripts.

Reddy and Knight (2011) argue from the statistical distribution of letters and words that Voynichese most closely resembles an abjad. Many Voynichese characters are only found at the beginning or end of a word, which resembles the positional variants of letters in the Arabic script. Comparison with the Wikipedia and Historical corpora demonstrates that the abjad hypothesis is not, however, an explanation for the low conditional entropy in Voynichese. The abjads in our corpora have a higher conditional character entropy than the alphabets, and adding the vowels
back in with *niqqud* (essentially turning an abjad into an alphabet) lessens conditional character entropy substantially. If Voynichese is an abjad, it is a highly unusual one.\textsuperscript{35}

### 4.6 What makes Voynichese unique?

#### 4.6.1 Entropy and Bigram Frequency

Conditional character entropy tells us about the predictability of a text at the character level. One way of thinking about it is this: if you look at any character in a text, how certain can you be, on average, that you will be able to guess the character that follows it? In the English sample from the Wikipedia corpus, for example, the letter *q* is followed by the letter *u* 96% of the time.\textsuperscript{36} In other words, the conditional frequency of the bigram *qu* is 96%. So if we see a *q* in an English text, we can be reasonably certain that we know what the next letter is.

\textsuperscript{35}Reddy and Knight (2011) also argue that the Voynich Manuscript might be written in an abjad because of the results from their two-state HMM investigations. The HMM deduces a word formula of A*B rather than picking out a class of consonants and one of vowels; this pattern was also found with their Arabic tests. However, this does not necessarily mean that there are no vowels represented in the text, rather that the regularity and singularity of word-final items is swamping other possible groupings.

\textsuperscript{36}Some of the exceptions include the words *FAQs*, *Iraqi*, and *qi*. 

---

\textsuperscript{31}
Figure 14: Differences between Parallel Comparison Texts: abbreviated Latin has higher conditional character entropy than unabbreviated Latin, and the diplomatic transcriptions of Icelandic and medical English have higher conditional character entropy than the normalized transcriptions. The Minimal transcription of Voynichese has higher conditional character entropy than the Full Maximal transcription.

However, all the other letters in English are much less predictable. For example, the most common letter to follow p is e, but this only happens 17% of the time. Conditional character entropy is the average for all letters weighted by their frequency, and this gives us a measure of the overall level of disorder in the text. A text consisting of randomly generated characters has a higher conditional character entropy than a text that contains meaningful and ordered natural language. Voynichese, however, has starkly lower values than we see with any natural language, meaning many of its letters are like the English q rather than p.

Figures (15) and (16) consist of two heatmaps of the English bigram space. The top map is simply the conditional frequency of each bigram. Bright spots indicate bigrams with particularly high conditional frequencies: qu, y# (the # symbol indicates a space, i.e., y at the end of a word), ve and d#. In the bottom map, each of these values have been weighted by the overall frequency of the bigram itself. This weighting gives a much better picture of which bigrams contribute to the conditional character entropy of the text as a whole.

So while qu is a highly predictable pairing, the letter q itself is fairly rare, and therefore it has a negligible effect on the predictability of the text as a whole. Note from the right side of this map that certain letters at the end of the word (d#, s#, y#) make a contribution as a result of the frequent English morphological suffixes (-ed, -s, and -y). Other bigrams like th, he, #a, and in in
the most frequently used words in English (the, a, in). The bigram th is a digraph, meaning that two characters are used to represent a single phoneme. Common digraphs can contribute to low conditional character entropy values because the two letters together share the information load of a single phoneme.
4.6.2 Compositionality, Word-finals, and Syllable Structure

This is related to the issue of compositionality in the transcription of unknown texts like Voynichese. If certain glyphs occur primarily in a particular sequence, this may be evidence that the sequence of glyphs represents a single character. Thus Full Maximal Voynich (EVA), which is maximally decomposed, has a lower conditional entropy than Minimal Voynich, for which common glyph sequences are taken to be single characters. But even Minimal Voynich is much more predictable than any of the European languages.

Compare the weighted heatmap for English at the bottom of Figure 15 with the weighted heatmap for Venda (a Southern Bantu language of South Africa) in Figure 17. The Venda language has the second-lowest conditional character entropy of any non-Voynichese text in the corpora, and this added predictability is visible in the overall reddening of the heatmap. There are more bigrams in Venda which are both highly predictable and extremely frequent in the text. The vh bigram is a digraph found in some of the most common words in the language. Most notably, the letter a is very common at the end of a word, as are the other vowels. In fact, 97% of all words in Venda end with a vowel, and nearly half (48%) of all words end with an a.

This dramatic restriction of possible letters at the end of the word is common for languages with low conditional character entropy. For example, 98% of words in the Hawaiian sample end with a vowel, and in Min Dong 95% of words end with either a vowel, k, or g. This is due to restrictions on the possible syllable structures in these languages. While many Indo-European languages have fairly complex syllables with consonant clusters that can occur both at the beginning or end of the syllable, many languages of the Tibeto-Burman, Malayo-Polynesian, and Niger-Congo families do not allow syllables to end with a consonant, and they disallow many of the consonant clusters found in Indo-European languages. More complex syllables increase conditional character entropy because each consonant can be followed by a much larger number of possible consonants and vowels. Even among European languages, those that have more complex syllables, such as Slavic languages, tend to have slightly higher conditional character entropy. Abjads have even higher conditional character entropy because they lack written vowels altogether.

Secondly, Venda, like most of the other lowest conditional-entropy languages, is tonal. Words in tonal languages that differ only by tone may have distinct meanings, and so most orthographies of these languages have a means of indicating tone. In tonal languages written with the Latin script (like Vietnamese), diacritics over the vowels are often employed to indicate tone. The orthography employed for the Venda sample, however, does not distinguish tone at all. This has the effect of collapsing distinctions that are present in the spoken language: two words which are pronounced differently may be spelled the same, and this makes the text more predictable.

Voynichese has a lower conditional entropy than these other languages because it has even more frequent, highly predictable bigrams. Figure 18 maps the weighted conditional bigrams for (Simple) Maximal Voynich. As with Venda, certain characters are usually found at the end of words: 41% of words end with y, and 93% of words end with either y, n, l, r, m or s. Many other bigrams are prominent in other parts of the word: the ch bench characters are usually found together at the beginning of the word (but sometimes have an intervening gallows character), qo is found at the beginning of words, dy is a very common sequence at the end of words, o is almost always followed by l or a gallows symbol (usually t or k), and i is usually part of a word final sequence of in or iin. All characters are heavily restricted in whether they can appear at the beginning, middle, or end of the word, and which characters can come before or after.

The unusual predictability of Voynichese cannot be entirely attributed to the compositionality
of the transcription system. The Minimal transcription of Voynichese lacks many of these highly predictable bigrams, because common sequences like ch, iin, and go are represented as single characters (cf. Figure 4). But most characters are still restricted to certain positions in the word: S, Z, Q, W, X and Y at the beginning, a, E, e, i, t, k, p and f in the middle, and N, M, 3, K, L,
Figure 19: Prevalence of Bigrams with High Conditional Probability (>50%) for Various Languages. Red indicates the percentage of word-final bigrams (e.g. \( y\# \)).

5, \( T, U, 0, G, H \) and 1 at the end. Thus one cannot simply assume that the low character entropy is due to our over-splitting of characters; even when they are grouped together, Voynichese is still unusual compared to other language samples.

### 4.6.3 Bigrams with High Conditional Probability

Another way to investigate the relationship between script properties and entropy is to measure the percentage of a text that contains bigrams with high conditional probability. Most texts have relatively few bigrams with a conditional probability greater than 50%. In the English Wikipedia text, there are only four: \( qu \) has 96% conditional probability, \( y\# \) has 75%, \( ve \) has 59% conditional probability, and \( d\# \) has 54% conditional probability. Because most of these bigrams are relatively infrequent, they make up only 3.3% of the text as a whole. By contrast, Voynich in the Simple Maximum transcription contains 12 bigrams with high conditional probability, and they are much more frequent, making up 29.3% of the text. With the Minimal transcription, there are 23 bigrams with high conditional probability, and they make up 23.9% of the text.

Figure 19 shows for various languages the proportion of the text which consists of bigrams with high conditional probability. The only other texts that have a comparable amount of high conditional bigrams are those of Min Dong, Hakka, Venda, and Hawaiian, the four languages in the Wikipedia corpus with the lowest conditional character entropy. Min Dong is between the two Voynich transcriptions.

The red coloring denotes word-finals, i.e., characters which are followed by a space. In most languages of the sample, the majority of predictable bigrams are at the ends of words, whether or not their script contains characters which are found exclusively at the end of words (as with the positional character variants of Arabic, Hebrew, and Greek). This is true for English, because \( y\# \)
and $d#$ are much more common than $qu$ and $ve$. Voynichese, like Hakka and Min Dong, has a great deal of word-final bigrams in addition to other types of bigrams.

### 4.7 Summary

At the character level, Voynichese is strikingly different from any other text in the Wikipedia and Historical corpora. The character set size and frequency of characters is conventional, but the characters are combined in an extremely predictable way, as indicated by an unusual conditional character entropy that is distinctly lower than any of the 316 comparison texts.

This discrepancy is not attributable to the transcription system used to encode Voynich, although decisions about the compositionality of glyph sequences can have a significant effect on entropy. Nor is it the result of conventional scholarly abbreviations of the historical period or the absence of written vowels.

Rather, it is largely the result of common characters which are heavily restricted to certain positions within the word. Voynichese most closely resembles tonal languages written in the Latin script and languages with relatively limited syllabic inventories.

We do not take this as evidence that the language underlying Voynich is likely to be from the Niger-Congo, Malayo-Polynesian, or Tibeto-Burman families. A more reasonable scenario is that the script ignores certain sound distinctions that are made in the underlying language, as Bennett’s sample of Hawaiian did not include vowel distinctions and the Venda wikipedia text does not mark tone. Whatever method was used to generate the Voynichese script, it created written words that are highly constrained in form.

We have described the creation of a corpus of Voynich texts and two comparison corpora for analyzing Voynichese from a broad typological perspective. In addition to the character-level analysis given here, we have used these corpora for our overview in Bowern and Lindemann (2020) and will continue to use them in subsequent analyses.

One line of inquiry that will be addressed in forthcoming work is the effect of historical encipherment methods on character properties. Here we will make a few brief observations. If Voynichese is encoded text, it must be more complex than a simple substitution cipher. A simple monoalphabetic substitution cipher will have absolutely no effect on conditional character entropy, because the same characters simply shift places with one another. The addition of null (meaningless) characters and multiple variants of high-frequency characters were used to make it harder to identify uniquely frequent characters. This can have an effect on character entropy (as can more complex polyalphabetic ciphers), but we would then expect the character frequency distribution of Voynichese to be atypical. They also affect the distribution of minimal pairs (if null-insertion operates on an otherwise monoalphabetic cipher). We can also rule out most types of polyalphabetic cipher, since such ciphers would disrupt the regular encoding of identical words across pages, as well as increasing (rather than decreasing) $h2$.

Voynichese is a clear outlier at the character level, and we might be tempted to conclude from this that the text is meaningless. However, the Voynich text conceals sophisticated layers of structure. Subsequent work will examine the Voynich text at the word level, for which Voynichese is much more typical.
The following appendices provide some general information about the datasets used in this paper.

## A Voynich Corpus Statistics

This table displays basic statistics for each of the Voynich sample texts in the Maximal transcription, broken down by running paragraph text and labels.

|                  | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|------------------|-----------------|--------------------|------------|---------------|-------|
| **Full Voynich Text** |                 |                    |            |               |       |
| Paragraphs       | 205,014         | 36                 | 33,111     | 6,936         | 2.117 |
| Labels           | 29,389          | 36                 | 4,829      | 2,283         | 2.309 |
| **Voynich A**    |                 |                    |            |               |       |
| Paragraphs       | 66,477          | 33                 | 11,081     | 3,281         | 2.101 |
| Labels           | 2,134           | 25                 | 334        | 289           | 2.425 |
| **Voynich B**    |                 |                    |            |               |       |
| Paragraphs       | 136,046         | 30                 | 21,632     | 4,661         | 1.964 |
| Labels           | 9,698           | 26                 | 1,594      | 778           | 2.044 |
| **Hand 1**       |                 |                    |            |               |       |
| Paragraphs       | 61,963          | 33                 | 10,352     | 3,032         | 2.083 |
| Labels           | 2,783           | 31                 | 525        | 365           | 2.572 |
| **Hand 2**       |                 |                    |            |               |       |
| Paragraphs       | 61,698          | 27                 | 10,054     | 2,367         | 1.910 |
| Labels           | 6,230           | 21                 | 1,016      | 531           | 1.975 |
| **Hand 3**       |                 |                    |            |               |       |
| Paragraphs       | 72,550          | 30                 | 11,755     | 3,419         | 1.999 |
| Labels           | 2,631           | 22                 | 427        | 294           | 2.086 |
| **Hand 4**       |                 |                    |            |               |       |
| Paragraphs       | 17,850          | 25                 | 2,864      | 1,548         | 2.279 |
| Labels           | 15,630          | 25                 | 2,511      | 1,399         | 2.284 |
| **Hand 5**       |                 |                    |            |               |       |
| Paragraphs       | 3,662           | 22                 | 580        | 387           | 2.079 |
| Labels           | 2,111           | 26                 | 350        | 255           | 2.055 |

The following table compares the character set size and conditional character entropy ($h^2$) for each of the Voynich texts in each of the three transcription systems: Maximal, Maximal Simplified, and Minimal.
### Full Voynich

|                | Maximal | Maximal Simplified | Minimal |
|----------------|---------|--------------------|---------|
| **Full Voynich** | 45 / 2.114 | 23 / 2.112 | 41 / 2.475 |
| **Language A**  | 36 / 2.122 | 22 / 2.119 | 39 / 2.504 |
| **Language B**  | 32 / 1.973 | 23 / 1.973 | 40 / 2.304 |
| **Hand 1**      | 42 / 2.122 | 23 / 2.117 | 40 / 2.506 |
| **Hand 2**      | 27 / 1.921 | 23 / 1.921 | 39 / 2.219 |
| **Hand 3**      | 30 / 1.999 | 23 / 1.999 | 39 / 2.338 |
| **Hand 4**      | 25 / 2.279 | 22 / 2.279 | 36 / 2.558 |
| **Hand 5**      | 26 / 2.111 | 23 / 2.112 | 31 / 2.319 |

### B Wikipedia Corpus Statistics

The following table gives basic statistics for each of the sample languages in the Wikipedia Corpus:

| Language        | Wikicode | Family      | Script     | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$  |
|-----------------|----------|-------------|------------|-----------------|--------------------|------------|---------------|-------|
| Abkhazian       | ab       | Caucasian   | Cyrillic   | 474,421         | 44                 | 58,100     | 19,109        | 3.571 |
| Acehnese        | ace      | Malayo-Polynesian | Latin | 566,068      | 33                 | 90,025     | 11,400        | 3.208 |
| Adyghe          | ady      | Caucasian   | Cyrillic   | 86,867          | 36                 | 10,791     | 5,212         | 3.403 |
| Afrikaans       | af       | Germanic    | Latin      | 1,228,386       | 31                 | 198,349    | 22,349        | 3.375 |
| Akan            | ak       | Niger-Congo | Latin      | 409,746         | 31                 | 82,731     | 9,578         | 3.321 |
| Albanian        | sq       | Albanian    | Latin      | 1,162,617       | 30                 | 193,060    | 29,512        | 3.376 |
| Alemannic       | als      | Germanic    | Latin      | 1,197,089       | 38                 | 197,012    | 39,254        | 3.542 |
| Amharic         | am       | Afro-Asiatic| Ethiopic   | 940,578         | 293                | 193,388    | 49,646        | 4.503 |
| Anglo Saxon     | ang      | Germanic    | Latin      | 679,186         | 41                 | 109,182    | 26,219        | 3.605 |
| Arabic          | ar       | Afro-Asiatic| Arabic     | 1,164,523       | 44                 | 195,629    | 38,372        | 3.702 |
| Aragonese       | an       | Romance     | Latin      | 1,170,029       | 34                 | 210,697    | 23,412        | 3.293 |
| Aramaic         | arc      | Afro-Asiatic| Syriac     | 52,659          | 40                 | 9,236      | 3,853         | 3.483 |
| Armenian        | hy       | Armenian    | Armenian   | 1,521,970       | 40                 | 194,344    | 36,578        | 3.376 |
| Aromanian       | roa,rup  | Romance     | Latin      | 89,929          | 42                 | 15,126     | 5,248         | 3.440 |
| Assamese        | as       | Indic       | Bengali    | 1,203,527       | 63                 | 189,182    | 32,598        | 3.760 |
| Asturian        | ast      | Romance     | Latin      | 1,235,937       | 36                 | 206,485    | 24,907        | 3.308 |
| Atikamekw       | atj      | Algonquian  | Latin      | 245,152         | 28                 | 34,454     | 7,904         | 2.953 |
| Avar            | av       | Caucasian   | Cyrillic   | 705,663         | 37                 | 88,140     | 25,138        | 3.355 |
| Awadhi          | awa      | Indic       | Devanagari | 438,119         | 69                 | 84,403     | 14,615        | 3.746 |
| Aymara          | ay       | Aymara      | Devanagari | 202,890         | 39                 | 23,822     | 11,645        | 3.185 |
| Azerbaijani     | az       | Turkic      | Latin      | 1,492,347       | 34                 | 199,917    | 35,295        | 3.644 |
| Balinese         | ban      | Malayo-Polynesian | Latin | 1,317,387      | 27                 | 199,768    | 25,568        | 3.180 |
| Bambara         | bm       | Mande       | Latin      | 122,358         | 38                 | 25,163     | 4,806         | 3.191 |
| Banjar          | bja      | Austronesian| Latin      | 1,371,576       | 28                 | 199,761    | 27,426        | 3.175 |
| Banyumasan      | map,bms  | Malayo-Polynesian | Latin | 1,383,372      | 28                 | 198,810    | 23,499        | 3.257 |
| Language             | Wikicode | Family     | Script   | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|----------------------|----------|------------|----------|-----------------|-------------------|------------|---------------|------|
| Bashkir              | ba       | Turkic     | Cyrillic | 1,412,766       | 42                | 194,138    | 37,466        | 3.691|
| Basque               | eu       | Vasconic   | Latin    | 1,540,698       | 27                | 199,658    | 37,568        | 3.250|
| Bavarian             | bar      | Germanic   | Latin    | 1,178,419       | 34                | 197,795    | 41,284        | 3.550|
| Belarusian           | be       | Slavic     | Cyrillic | 1,383,017       | 33                | 192,440    | 39,472        | 3.566|
| Belarusian (Taraškievica) | be_x_old | Slavic     | Cyrillic | 1,379,366       | 35                | 189,806    | 39,306        | 3.567|
| Bengali              | bn       | Indic      | Bengali  | 1,301,543       | 62                | 191,876    | 28,481        | 3.665|
| Bihari               | bh       | Indic      | Devanagari | 940,050       | 68                | 189,977    | 18,150        | 3.618|
| Bishnupriya Manipuri | bpy      | Indic      | Bengali  | 517,177         | 63                | 79,490     | 16,408        | 3.545|
| Bislama              | bi       | Creole     | Latin    | 16,041          | 29                | 3,145      | 798           | 3.126|
| Bosnian              | bs       | Slavic     | Latin    | 1,305,485       | 33                | 195,997    | 37,659        | 3.468|
| Breton               | br       | Celtic     | Latin    | 1,095,236       | 35                | 206,119    | 20,998        | 3.322|
| Buginese (Buginese)  | bug      | Malayo-Polynesian | Buginese | 11,978         | 29                | 2,158      | 1,153         | 3.365|
| Buginese (Latin)     | bug      | Malayo-Polynesian | Latin    | 49,167         | 28                | 7,126      | 2,956         | 3.278|
| Bulgarian            | bg       | Slavic     | Cyrillic | 1,251,467       | 31                | 194,249    | 29,953        | 3.476|
| Burmese              | my       | Tibeto-Burman | Myanmar  | 1,002,038       | 62                | 75,527     | 44,643        | 3.324|
| Buryat               | bxr      | Mongolic   | Cyrillic | 1,418,044       | 35                | 192,042    | 38,617        | 3.503|
| Cantonese            | zh_yue   | Tibeto-Burman | Chinese  | 322,674         | 3,608             | 31,888     | 28,439        | 5.409|
| Catalan              | ca       | Romance    | Latin    | 1,189,313       | 41                | 209,944    | 23,100        | 3.370|
| Cebuano              | ceb      | Malayo-Polynesian | Latin    | 1,198,584       | 32                | 200,597    | 23,993        | 3.186|
| Central Bicolano     | bcl      | Malayo-Polynesian | Latin    | 1,197,161       | 35                | 197,619    | 23,546        | 3.212|
| Chamorro             | ch       | Malayo-Polynesian | Latin    | 37,146         | 34                | 6,547      | 1,903         | 3.291|
| Chechen              | ce       | Caucasian  | Cyrillic | 1,344,890       | 35                | 195,158    | 43,064        | 3.672|
| Cherokee             | chr      | Iroquoian  | Cherokee | 57,838         | 84                | 11,835     | 3,777         | 3.829|
| Cheyenne             | chy      | Algonquian | Latin    | 2,812          | 39                | 318        | 225           | 3.194|
| Chichewa             | ny       | Niger-Congo | Latin    | 689,446        | 28                | 99,634     | 17,337        | 3.246|
| Chinese              | zh       | Tibeto-Burman | Chinese  | 1,234,309       | 6,222             | 124,650    | 109,341       | 6.142|
| Chuvash              | cv       | Turkic     | Cyrillic | 1,184,848       | 38                | 188,153    | 35,546        | 3.620|
| Classical Chinese    | zh_classical | Tibeto-Burman | Chinese  | 80,593         | 3,292             | 12,778     | 11,912        | 5.176|
| Cornish              | kw       | Celtic     | Latin    | 749,995        | 30                | 139,379    | 18,315        | 3.385|
| Corsican             | co       | Romance    | Latin    | 1,134,720       | 33                | 207,593    | 28,600        | 3.232|
| Cree                 | cr       | Algonquian | Latin    | 3,127          | 29                | 325        | 229           | 2.992|
| Crimean Tatar        | crh      | Turkic     | Latin    | 397,729        | 35                | 54,339     | 16,048        | 3.589|
| Croatian             | hr       | Slavic     | Latin    | 1,293,468       | 32                | 194,591    | 39,602        | 3.464|
| Czech                | cs       | Slavic     | Latin    | 1,330,126       | 41                | 196,446    | 40,391        | 3.745|
| Danish               | da       | Germanic   | Latin    | 1,258,211       | 30                | 196,664    | 28,753        | 3.476|
| Dinka                | din      | Nilotic    | Latin    | 331,605        | 35                | 64,389     | 12,093        | 3.469|
| Divehi               | dv       | Indic      | Thaana   | 1,732,631       | 51                | 186,755    | 41,357        | 3.138|
| Doteli               | dty      | Indic      | Devanagari | 1,170,968   | 71                | 187,576    | 38,443        | 3.824|
| Dutch                | nl       | Germanic   | Latin    | 1,263,624       | 30                | 197,198    | 24,939        | 3.332|
| Dutch Low Saxon      | nds_nl   | Germanic   | Latin    | 1,191,824       | 35                | 199,215    | 31,151        | 3.393|
| Language          | Wikicode | Family       | Script   | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|-------------------|----------|--------------|----------|-----------------|--------------------|------------|---------------|-------|
| Dzongkha          | dz       | Tibeto-Burman| Tibetan  | 93,231          | 63                 | 24,147     | 1,735         | 3.076 |
| Egyptian Arabic   | arz      | Afro-Asiatic | Arabic   | 1,078,766       | 47                 | 195,437    | 36,930        | 3.663 |
| Emilian-Romagnol  | eml      | Romance      | Latin    | 784,411         | 53                 | 162,961    | 23,869        | 3.525 |
| English           | en       | Germanic     | Latin    | 1,216,023       | 27                 | 199,564    | 19,843        | 3.403 |
| Erzya             | myv      | Uralic       | Cyrillic | 1,353,230       | 34                 | 180,837    | 44,893        | 3.518 |
| Esperanto         | eo       | Constructed  | Latin    | 1,219,760       | 34                 | 197,285    | 30,975        | 3.305 |
| Estonian          | et       | Uralic       | Latin    | 1,487,423       | 34                 | 192,800    | 48,756        | 3.536 |
| Ewe               | ee       | Niger-Congo  | Latin    | 94,902          | 43                 | 18,632     | 4,058         | 3.188 |
| Extremaduran      | ext      | Romance      | Latin    | 1,175,298       | 35                 | 200,577    | 31,832        | 3.354 |
| Faroese           | fo       | Germanic     | Latin    | 1,181,388       | 34                 | 191,824    | 29,506        | 3.523 |
| Fiji Hindi        | hif      | Indic        | Latin    | 706,696         | 34                 | 128,450    | 15,253        | 3.411 |
| Fijian            | fj       | Malayo-Polynesian | Latin | 226,281         | 27                 | 42,360     | 4,345         | 2.888 |
| Finnish           | fi       | Uralic       | Latin    | 1,728,271       | 28                 | 196,573    | 55,335        | 3.475 |
| Franco-Provençal  | frp      | Romance      | Latin    | 314,875         | 41                 | 58,163     | 12,054        | 3.461 |
| French            | fr       | Romance      | Latin    | 1,371,315       | 39                 | 233,025    | 23,791        | 3.354 |
| Friulian          | fur      | Romance      | Latin    | 1,052,533       | 38                 | 198,007    | 21,733        | 3.284 |
| Fula              | ff       | Nigerian-Congo | Latin | 238,729        | 37                 | 40,601     | 10,899        | 3.364 |
| Gagauz            | gag      | Turkic       | Latin    | 484,481         | 38                 | 67,724     | 19,070        | 3.570 |
| Galician          | gl       | Romance      | Latin    | 1,225,239       | 34                 | 197,864    | 23,310        | 3.314 |
| Gan               | gan      | Tibeto-Burman| Chinese  | 48,003          | 2,938              | 6,514      | 5,837         | 4.597 |
| Georgian          | ka       | Kartvelian   | Georgian | 1,590,023       | 43                 | 190,379    | 42,981        | 3.566 |
| German            | de       | Germanic     | Latin    | 1,411,555       | 31                 | 197,197    | 32,028        | 3.394 |
| Gilaki            | glk      | Iranian      | Arabic   | 916,037         | 49                 | 180,675    | 31,070        | 3.854 |
| Goan Konkani      | gom      | Indic        | Devanagari | 1,317,991      | 76                 | 197,428    | 48,971        | 3.659 |
| Gorontalo         | gor      | Malayo-Polynesian | Latin | 359,794        | 32                 | 53,579     | 10,964        | 3.258 |
| Gothic            | got      | Germanic     | Gothic   | 63,136          | 28                 | 9,352      | 3,598         | 3.321 |
| Greek             | el       | Hellenic     | Greek    | 1,270,133       | 34                 | 192,047    | 26,770        | 3.519 |
| Greenlandic       | kl       | Inuit        | Latin    | 213,944         | 29                 | 18,203     | 9,980         | 3.055 |
| Guarani           | gn       | Tupian       | Latin    | 1,388,881       | 48                 | 198,012    | 32,287        | 3.435 |
| Guianan Creole    | gcr      | Creole       | Latin    | 610,232         | 31                 | 115,926    | 12,096        | 3.224 |
| Gujarati          | gu       | Indic        | Gujarati | 1,173,177       | 69                 | 195,421    | 36,078        | 3.732 |
| Haitian           | ht       | Creole       | Latin    | 955,666         | 35                 | 188,176    | 20,106        | 3.385 |
| Hakka             | hak      | Tibeto-Burman| Latin    | 369,395         | 44                 | 84,588     | 3,607         | 2.857 |
| Hausa             | ha       | Afro-Asiatic | Latin    | 1,116,212       | 31                 | 200,918    | 20,315        | 3.105 |
| Hawaiian          | haw      | Malayo-Polynesian | Latin | 561,237         | 31                 | 112,870    | 5,728         | 2.765 |
| Hebrew            | he       | Afro-Asiatic | Hebrew   | 1,123,178       | 28                 | 194,947    | 38,482        | 3.664 |
| Hill Mari         | mrj      | Uralic       | Cyrillic | 630,450         | 37                 | 91,189     | 22,380        | 3.701 |
| Hindi             | hi       | Indic        | Devanagari | 1,035,585      | 71                 | 195,809    | 20,001        | 3.623 |
| Hungarian         | hu       | Uralic       | Latin    | 1,437,916       | 35                 | 198,457    | 47,571        | 3.742 |
| Icelandic         | is       | Germanic     | Latin    | 1,237,880       | 36                 | 194,966    | 33,368        | 3.596 |
| Language                      | Wikicode | Family         | Script     | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|-------------------------------|----------|----------------|------------|-----------------|-------------------|------------|---------------|-------|
| Ido                           | io       | Constructed    | Latin      | 1,178,602       | 33                | 197,718    | 20,822        | 3.329 |
| Igbo                          | ig       | Niger-Congo    | Latin      | 984,726         | 44                | 188,440    | 20,250        | 3.380 |
| Ilokano                       | ilo      | Malayo-Polynesian | Latin    | 1,172,160       | 28                | 198,139    | 20,266        | 3.163 |
| Inari Sami                    | smn      | Uralic         | Latin      | 260,928         | 36                | 34,021     | 10,928        | 3.769 |
| Indonesian                    | id       | Malayo-Polynesian | Latin    | 1,446,218       | 26                | 201,308    | 17,949        | 3.222 |
| Ingush                        | inh      | Caucasian      | Cyrillic   | 191,553         | 37                | 27,865     | 10,432        | 3.649 |
| Interlingua                   | ia       | Constructed    | Latin      | 1,195,958       | 28                | 195,759    | 19,769        | 3.244 |
| Interlingue                   | ie       | Constructed    | Latin      | 1,076,710       | 38                | 187,169    | 18,211        | 3.380 |
| Inuktitut (Canadian Syllabics)| iu       | Inuit          | Canadian Syllabics | 13,614   | 117             | 1,923      | 1,193        | 3.732 |
| Inuktitut (Latin)             | iu       | Inuit          | Latin      | 17,119          | 25                | 1,415      | 1,097         | 3.129 |
| Impiak                        | ik       | Inuit          | Latin      | 6,536           | 39                | 692        | 500           | 3.350 |
| Irish                         | ga       | Celtic         | Latin      | 1,130,156       | 31                | 199,466    | 20,054        | 3.338 |
| Italian                       | it       | Romance        | Latin      | 1,268,739       | 33                | 204,004    | 26,338        | 3.272 |
| Jamaican Patois               | jam      | Creole         | Latin      | 535,308         | 27                | 95,784     | 13,421        | 3.387 |
| Japanese                      | ja       | Japonic        | Japanese   | 1,015,563       | 2,887             | 74,348     | 60,520        | 5.027 |
| Javanese                      | jv       | Malayo-Polynesian | Latin    | 1,351,976       | 29                | 199,101    | 27,774        | 3.238 |
| Kabardian Circassian          | kbd      | Caucasian      | Cyrillic   | 693,947         | 35                | 91,720     | 30,563        | 3.367 |
| Kabiyé                        | kbp      | Niger-Congo    | Latin      | 1,104,123       | 40                | 206,498    | 22,038        | 3.302 |
| Kabyle                        | kab      | Afro-Asiatic   | Latin      | 1,130,980       | 39                | 210,877    | 30,148        | 3.458 |
| Kalmyk                        | xal      | Mongolic       | Cyrillic   | 149,170         | 41                | 23,437     | 9,615         | 3.863 |
| Kannada                       | kn       | Dravidian      | Kannada    | 1,633,412       | 66                | 198,531    | 57,666        | 3.773 |
| Kapampangan                   | pam      | Malayo-Polynesian | Latin    | 1,207,315       | 31                | 193,277    | 23,199        | 3.063 |
| Karachay-Balkar               | krc      | Turkic         | Cyrillic   | 1,138,595       | 34                | 145,979    | 31,559        | 3.476 |
| Karakalpak                    | kaa      | Turkic         | Latin      | 958,773         | 37                | 121,182    | 32,916        | 3.509 |
| Kashmiri (Arabic)             | ks       | Indic          | Arabic     | 34,438          | 79                | 6,201      | 2,922         | 3.872 |
| Kashmiri (Devanagari)         | ks       | Indic          | Devanagari | 22,493          | 69                | 3,641      | 2,093         | 3.864 |
| Kashubian                     | csb      | Slavic         | Latin      | 792,280         | 40                | 120,564    | 37,608        | 3.750 |
| Kazakh                        | kk       | Turkic         | Cyrillic   | 1,478,373       | 40                | 195,972    | 38,033        | 3.496 |
| Khmer                         | km       | Austroasiatic  | Khmer      | 654,815         | 79                | 28,173     | 21,384        | 4.138 |
| Kikuyu                        | ki       | Niger-Congo    | Latin      | 62,285          | 32                | 10,016     | 3,532         | 3.152 |
| Kinyarwanda                   | rw       | Niger-Congo    | Latin      | 840,081         | 29                | 126,721    | 20,888        | 3.136 |
| Kirghiz                       | ky       | Turkic         | Cyrillic   | 1,477,252       | 37                | 194,186    | 37,921        | 3.605 |
| Kirundi                       | rn       | Niger-Congo    | Latin      | 218,081         | 28                | 32,282     | 8,974         | 3.074 |
| Komi                          | kv       | Uralic         | Cyrillic   | 952,576         | 36                | 139,004    | 30,271        | 3.706 |
| Komi-Permyak                  | koi      | Uralic         | Cyrillic   | 547,852         | 36                | 82,138     | 18,841        | 3.743 |
| Kongo                         | kg       | Niger-Congo    | Latin      | 126,751         | 38                | 21,763     | 4,324         | 3.141 |
| Korean                        | ko       | Koreanic       | Hangul     | 817,046         | 1,379             | 203,285    | 65,252        | 4.685 |
| Kotava                        | avk      | Constructed    | Latin      | 1,160,948       | 30                | 190,260    | 33,067        | 4.354 |
| Kurdish                       | ku       | Iranian        | Latin      | 1,097,852       | 35                | 201,217    | 29,586        | 3.386 |
| Ladin                         | lld      | Romance        | Latin      | 597,615         | 46                | 114,922    | 18,314        | 3.394 |
| Language         | Wikicode | Family      | Script | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|------------------|----------|-------------|--------|-----------------|-------------------|------------|---------------|-------|
| Ladino           | lad      | Romance     | Latin  | 1,115,534       | 32                | 193,345    | 26,514        | 3.305 |
| Lak              | lbe      | Caucasian   | Cyrillic | 170,141         | 36                | 22,564     | 5,981         | 3.454 |
| Lao              | l0       | Tai         | Lao    | 343,956         | 55                | 21,063     | 13,620        | 4.030 |
| Latgalian        | ltg      | Baltic      | Latin  | 299,177         | 36                | 41,482     | 13,350        | 3.606 |
| Latin            | la       | Romance     | Latin  | 1,430,965       | 28                | 193,514    | 42,260        | 3.412 |
| Latvian          | lv       | Baltic      | Latin  | 1,408,947       | 36                | 193,369    | 36,694        | 3.634 |
| Lezgian          | lez      | Caucasian   | Cyrillic | 1,426,123       | 33                | 194,005    | 32,609        | 3.411 |
| Ligurian         | lij      | Romance     | Latin  | 1,108,522       | 51                | 214,470    | 37,687        | 3.475 |
| Limburgish       | li       | Germanic    | Latin  | 1,193,438       | 34                | 201,569    | 27,862        | 3.436 |
| Lingala          | ln       | Niger-Congo | Latin  | 383,704         | 46                | 64,511     | 11,314        | 3.411 |
| Lingua Franca Nova | lfn      | Constructed | Latin  | 1,022,406       | 28                | 196,299    | 15,063        | 3.111 |
| Lithuanian       | lt       | Baltic      | Latin  | 1,477,695       | 36                | 194,671    | 43,340        | 3.573 |
| Livvi-Karelian   | olo      | Uralic      | Latin  | 987,848         | 31                | 126,766    | 33,422        | 3.571 |
| Lojban           | jbo      | Constructed | Latin  | 288,970         | 28                | 59,390     | 6,453         | 3.140 |
| Lombard          | lmo      | Romance     | Latin  | 1,036,268       | 41                | 208,682    | 29,793        | 3.444 |
| Low Saxon        | nds      | Germanic    | Latin  | 1,177,767       | 30                | 199,049    | 27,399        | 3.390 |
| Lower Sorbian    | dsb      | Slavic      | Latin  | 863,845         | 39                | 131,759    | 30,568        | 3.626 |
| Luganda          | lg       | Niger-Congo | Latin  | 1,427,630       | 27                | 204,897    | 31,783        | 3.158 |
| Luxembourgisch   | lb       | Germanic    | Latin  | 1,254,155       | 33                | 200,807    | 28,191        | 3.423 |
| Macedonian       | mk       | Slavic      | Cyrillic | 1,302,905       | 31                | 199,719    | 30,368        | 3.407 |
| Madurese         | mad      | Malayo-Polynesian | Latin  | 312,666       | 34                | 46,468     | 8,560         | 3.394 |
| Maithili         | mai      | Indic       | Devanagari | 1,105,898     | 73                | 186,573    | 34,716        | 3.831 |
| Malagasy         | mg       | Malayo-Polynesian | Latin  | 1,245,462     | 34                | 186,592    | 21,030        | 2.944 |
| Malay            | ms       | Malayo-Polynesian | Latin  | 1,406,260     | 27                | 198,741    | 18,313        | 3.186 |
| Malayalam        | ml       | Dravidian   | Malayalam | 797,436       | 77                | 78,073     | 35,542        | 3.702 |
| Maltese          | mt       | Afro-Asiatic | Latin  | 1,418,795     | 32                | 241,726    | 24,892        | 3.515 |
| Manx             | gv       | Celtic      | Latin  | 1,166,926      | 33                | 208,179    | 18,495        | 3.373 |
| Maori            | mi       | Malayo-Polynesian | Latin  | 417,506       | 33                | 88,263     | 6,548         | 2.875 |
| Marathi          | mr       | Indic       | Devanagari | 1,297,866     | 78                | 188,409    | 42,144        | 3.689 |
| Mazandarani      | mzm      | Iranian     | Arabic  | 993,445        | 46                | 193,871    | 24,773        | 3.660 |
| Meadow Mari      | mhr      | Uralic      | Cyrillic | 1,371,590     | 38                | 199,570    | 34,921        | 3.549 |
| Min Dong         | cdo      | Tibeto-Burman | Latin  | 430,592       | 48                | 101,336    | 2,796         | 2.841 |
| Min Nan          | zh_min_nan | Tibeto-Burman | Latin  | 1,267,107     | 49                | 309,129    | 8,318         | 3.066 |
| Minangkabau      | min      | Malayo-Polynesian | Latin  | 1,083,848     | 27                | 158,742    | 19,545        | 3.181 |
| Mingrelian       | xmf      | Kartvelian  | Georgian | 1,452,488     | 40                | 187,555    | 49,075        | 3.580 |
| Miranda          | mw      | Romance     | Latin  | 1,210,157      | 33                | 203,845    | 22,829        | 3.309 |
| Moksha           | mdf      | Moksha      | Cyrillic | 254,066       | 34                | 34,186     | 12,397        | 3.474 |
| Mon              | maw      | Austroasiatic | Myanmar | 510,675       | 74                | 42,496     | 24,496        | 3.570 |
| Mongolian        | mn       | Mongolic    | Cyrillic | 1,321,073     | 35                | 193,476    | 29,282        | 3.559 |
| Moroccan Arabic  | ary      | Afro-Asiatic | Arabic  | 792,354       | 46                | 143,786    | 29,576        | 3.725 |

43
| Language               | Wikicode | Family             | Script    | Character Count | Character Set Size | Word Count    | Word Set Size | h2 |
|------------------------|----------|--------------------|-----------|-----------------|--------------------|---------------|---------------|----|
| N’Ko                   | nqo      | Mande              | N’Ko      | 1,122,166       | 48                 | 196,488       | 22,518        | 2.970 |
| NahuaTL                | nah      | Uto-Aztecan        | Latin     | 508,194         | 36                 | 62,805        | 15,605        | 3.227 |
| Nauruan                | na       | Malayo-Polynesian  | Latin     | 59,243          | 51                 | 9,774         | 2,819         | 3.405 |
| Navajo                 | nv       | Athabaskan         | Latin     | 202,428         | 37                 | 28,125        | 7,125         | 3.366 |
| Neapolitan             | nap      | Romance            | Latin     | 1,031,893       | 40                 | 185,831       | 32,261        | 3.291 |
| Nepali                 | ne       | Indic              | Devanagari| 1,170,229       | 68                 | 179,096       | 36,637        | 3.770 |
| Newar                  | new      | Tibeto-Burman      | Devanagari| 770,077         | 76                 | 122,917       | 27,822        | 3.750 |
| Nias                   | nia      | Malayo-Polynesian  | Latin     | 293,580         | 29                 | 52,144        | 7,274         | 3.077 |
| Norfolk                | pih      | Creole             | Latin     | 114,354         | 34                 | 21,202        | 5,008         | 3.497 |
| Norman                 | nrm      | Romance            | Latin     | 877,907         | 41                 | 165,642       | 20,754        | 3.364 |
| North Frisian          | frr      | Germanic           | Latin     | 1,073,028       | 36                 | 191,321       | 26,674        | 3.499 |
| Northern Sami          | se       | Uralic             | Latin     | 793,557         | 41                 | 100,856       | 26,797        | 3.607 |
| Northern Sotho         | nso      | Niger-Congo        | Latin     | 137,660         | 30                 | 25,626        | 4,583         | 2.950 |
| Norwegian (Bokmaal)    | no       | Germanic           | Latin     | 1,247,199       | 32                 | 196,357       | 29,556        | 3.475 |
| Norwegian (Nynorsk)    | nn       | Germanic           | Latin     | 1,185,458       | 32                 | 193,537       | 27,119        | 3.466 |
| Novial                 | nov      | Constructed        | Latin     | 220,177         | 33                 | 38,318        | 8,423         | 3.325 |
| Occitan                | oc       | Romance            | Latin     | 1,238,392       | 38                 | 208,924       | 20,667        | 3.358 |
| Old Church Slavonic    | cu       | Slavic             | Cyrillic  | 68,106          | 50                 | 10,108        | 3,258         | 3.398 |
| Oriya                  | or       | Indic              | Odia      | 1,249,524       | 66                 | 184,928       | 32,844        | 3.742 |
| Oromo                  | om       | Afro-Asiatic       | Latin     | 1,025,327       | 29                 | 233,367       | 40,522        | 3.193 |
| Ossetian               | os       | Iranian            | Cyrillic  | 1,198,686       | 33                 | 194,122       | 34,753        | 3.651 |
| Palatinate German      | pfl      | Germanic           | Latin     | 1,250,014       | 38                 | 197,124       | 42,618        | 3.440 |
| Pali (Devanagari)      | pi       | Indic              | Devanagari| 7,512           | 62                 | 1,157         | 686           | 3.694 |
| Pali (Latin)           | pi       | Indic              | Latin     | 56,922          | 32                 | 5,971         | 1,953         | 2.905 |
| Pangasinan             | pags     | Malayo-Polynesian  | Latin     | 176,803         | 27                 | 31,387        | 6,904         | 3.166 |
| Papiamentu             | pap      | Creole             | Latin     | 1,049,584       | 38                 | 195,235       | 20,194        | 3.344 |
| Pashto                 | ps       | Iranian            | Arabic    | 892,610         | 52                 | 196,694       | 25,578        | 3.651 |
| Pennsylvania German    | pdc      | Germanic           | Latin     | 226,954         | 31                 | 37,682        | 8,436         | 3.381 |
| Persian                | fa       | Iranian            | Arabic    | 1,024,245       | 42                 | 197,880       | 19,807        | 3.651 |
| Picard                 | pcd      | Romance            | Latin     | 543,922         | 39                 | 102,478       | 17,963        | 3.397 |
| Piedmontese            | pms      | Romance            | Latin     | 1,062,707       | 35                 | 218,355       | 19,308        | 3.318 |
| Polish                 | pl       | Slavic             | Latin     | 1,410,675       | 35                 | 195,119       | 44,911        | 3.638 |
| Pontic                 | pmt      | Hellenic           | Greek     | 127,501         | 38                 | 20,894        | 6,214         | 3.534 |
| Portuguese             | pt       | Romance            | Latin     | 1,235,214       | 39                 | 198,672       | 22,429        | 3.316 |
| Punjabi                | pa       | Indic              | Gurmukhi  | 970,590         | 64                 | 159,713       | 22,171        | 3.577 |
| Quechua                | qu       | Quechua            | Latin     | 335,139         | 35                 | 40,329        | 14,685        | 3.197 |
| Ripuarian              | ksh      | Germanic           | Latin     | 1,110,517       | 35                 | 198,243       | 34,554        | 3.556 |
| Romani                 | rmy      | Indic              | Latin     | 91,925          | 38                 | 15,487        | 4,116         | 3.294 |
| Romanian               | ro       | Romance            | Latin     | 1,348,034       | 31                 | 210,279       | 28,217        | 3.416 |
| Romansh               | rm       | Romance            | Latin     | 1,203,902       | 34                 | 208,626       | 18,734        | 3.306 |

44
| Language          | Wicone | Family    | Script | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|------------------|--------|-----------|--------|-----------------|-------------------|------------|---------------|-------|
| Russian          | ru     | Slavic    | Cyrillic | 1,500,801       | 34                | 200,538    | 40,697        | 3.619 |
| Rusyn            | rue    | Slavic    | Cyrillic | 1,244,127       | 38                | 187,752    | 48,078        | 3.740 |
| Saaraike         | skr    | Indic     | Arabic  | 927,476         | 57                | 197,187    | 23,313        | 3.679 |
| Sakha            | sah    | Turkic    | Cyrillic | 1,488,350       | 39                | 191,813    | 44,495        | 3.617 |
| Sakizaya         | szy    | Austronesian | Latin  | 1,067,950       | 27                | 194,641    | 15,539        | 2.948 |
| Samoan           | sm     | Malayo-Polynesian | Latin  | 399,484         | 32                | 83,345     | 6,083         | 2.992 |
| Samogitian       | bsi_bmg | Baltic   | Latin   | 719,074         | 36                | 105,660    | 34,440        | 3.599 |
| Sango            | sg     | Niger-Congo | Latin  | 17,054          | 38                | 3,746      | 640           | 2.862 |
| Sanskrit         | sa     | Indic     | Devanagari | 1,507,088     | 76                | 177,498    | 60,631        | 3.728 |
| Santalí          | sat    | Austroasiatic | Ol Chiki | 1,013,781     | 37                | 187,788    | 19,498        | 3.362 |
| Sardinian        | sc     | Romance   | Latin   | 1,157,097       | 33                | 204,144    | 32,653        | 3.230 |
| Saterland Frisian| stq    | Germanic  | Latin   | 1,146,938       | 30                | 186,897    | 25,801        | 3.384 |
| Scots            | sco    | Germanic  | Latin   | 1,122,338       | 27                | 199,515    | 22,474        | 3.417 |
| Scottish Gaelic  | gd     | Celtic    | Latin   | 1,137,203       | 33                | 206,651    | 18,580        | 3.145 |
| Serbian          | sr     | Slavic    | Cyrillic | 1,241,960       | 31                | 190,415    | 38,120        | 3.458 |
| Serbo-Croatian   | sh     | Slavic    | Latin   | 1,271,863       | 31                | 192,522    | 40,828        | 3.472 |
| Sesotho          | st     | Niger-Congo | Latin  | 303,739         | 28                | 55,756     | 7,713         | 3.004 |
| Shan             | shn    | Tai       | Myanmar | 494,159         | 105               | 31,048     | 22,064        | 3.154 |
| Shona            | sn     | Niger-Congo | Latin  | 626,613         | 27                | 76,706     | 22,057        | 3.169 |
| Sicilian         | scn    | Romance   | Latin   | 1,198,370       | 38                | 204,500    | 29,579        | 3.214 |
| Silesian         | szl    | Slavic    | Latin   | 891,117         | 46                | 132,556    | 42,331        | 3.767 |
| Simple English   | simple | Germanic  | Latin   | 1,128,909       | 27                | 198,481    | 14,313        | 3.362 |
| Sindhi           | sd     | Indic     | Arabic  | 875,905         | 62                | 184,506    | 20,142        | 3.694 |
| Sinhalese        | si     | Indic     | Sinhala | 1,182,228       | 76                | 188,511    | 34,716        | 3.799 |
| Slovak           | sk     | Slavic    | Latin   | 1,343,586       | 42                | 195,932    | 42,027        | 3.737 |
| Slovenian        | sl     | Slavic    | Latin   | 1,039,854       | 31                | 161,193    | 29,227        | 3.516 |
| Somali           | so     | Afro-Asiatic | Latin  | 1,245,451       | 28                | 194,843    | 30,491        | 3.268 |
| Sorani           | ckb    | Iranian   | Arabic  | 1,220,535       | 39                | 194,694    | 38,589        | 3.452 |
| South Azerbaijani| azb    | Turkic    | Arabic  | 1,325,401       | 49                | 194,657    | 50,268        | 3.717 |
| Spanish          | es     | Romance   | Latin   | 1,251,204       | 33                | 202,983    | 23,801        | 3.291 |
| Sranan           | srn    | Creole    | Latin   | 61,506          | 35                | 12,032     | 2,520         | 3.185 |
| Sundanese        | su     | Malayo-Polynesian | Latin  | 703,485         | 28                | 103,181    | 14,482        | 3.368 |
| Swahili          | sw     | Niger-Congo | Latin  | 1,233,964       | 26                | 195,587    | 21,695        | 3.084 |
| Swati            | ss     | Niger-Congo | Latin  | 201,303         | 29                | 23,888     | 10,554        | 3.195 |
| Swedish          | sv     | Germanic  | Latin   | 1,301,628       | 31                | 198,498    | 32,104        | 3.539 |
| Tagalog          | tl     | Malayo-Polynesian | Latin  | 1,213,045       | 27                | 196,218    | 19,488        | 3.059 |
| Tahitian         | ty     | Malayo-Polynesian | Latin  | 16,927          | 43                | 3,858      | 694           | 2.846 |
| Tajik            | tg     | Iranian   | Cyrillic | 1,248,236       | 40                | 190,743    | 34,147        | 3.614 |
| Tamil            | ta     | Dravidian | Tamil   | 1,783,419       | 49                | 192,511    | 58,671        | 3.327 |
| Tarantino        | roa_roa | Romance  | Latin   | 1,148,683       | 35                | 205,377    | 21,844        | 3.109 |
| Language              | Wikicode | Family      | Script       | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|-----------------------|----------|-------------|--------------|-----------------|-------------------|------------|---------------|-------|
| Tatar (Cyrillic)      | tt       | Turkic      | Cyrillic     | 940,273         | 38                | 128,935   | 27,004        | 3.670 |
| Tatar (Latin)         | tt       | Turkic      | Latin        | 429,020         | 41                | 60,991    | 18,904        | 3.749 |
| Telugu                | te       | Dravidian   | Telugu       | 1,551,065       | 67                | 193,017   | 56,921        | 3.783 |
| Tetum                 | tet      | Malayo-Polynesian | Latin   | 660,727         | 37                | 108,003   | 14,469        | 3.446 |
| Thai                  | th       | Tai         | Thai         | 953,229         | 68                | 38,480    | 27,031        | 4.247 |
| Tibetan               | bo       | Tibeto-Burman | Tibetan     | 1,655,843       | 83                | 428,495   | 6,127         | 3.143 |
| Tigrinya              | ti       | Afro-Asiatic | Ethnic       | 96,980          | 230               | 21,081    | 7,467         | 4.401 |
| Tok Pisin             | tpi      | Creole      | Latin        | 87,118          | 27                | 16,046    | 2,136         | 3.135 |
| Tongan                | to       | Malayo-Polynesian | Latin   | 167,431         | 34                | 36,884    | 3,972         | 2.945 |
| Tsonga                | ts       | Niger-Congo | Latin        | 522,298         | 28                | 88,346    | 11,540        | 2.970 |
| Tswana                | tn       | Niger-Congo | Latin        | 1,068,169       | 28                | 203,475   | 16,097        | 2.816 |
| Tulu                  | tcy      | Dravidian   | Kannada      | 1,243,356       | 68                | 168,376   | 51,885        | 3.695 |
| Tumbuka               | tum      | Niger-Congo | Latin        | 62,879          | 29                | 9,480     | 2,803         | 3.446 |
| Turkish               | tr       | Turkic      | Latin        | 1,503,312       | 34                | 202,601   | 39,777        | 3.549 |
| Turkmen               | tk       | Turkic      | Latin        | 1,510,318       | 37                | 202,877   | 43,688        | 3.603 |
| Tuvan                 | tyy      | Turkic      | Cyrillic     | 1,451,621       | 40                | 208,827   | 34,787        | 3.494 |
| Twi                   | tw       | Niger-Congo | Latin        | 173,715         | 30                | 35,061    | 5,612         | 3.289 |
| Udmurt                | udm      | Uralic      | Cyrillic     | 539,349         | 40                | 75,154    | 19,440        | 3.669 |
| Ukrainian             | uk       | Slavic      | Cyrillic     | 1,401,545       | 34                | 193,133   | 40,936        | 3.653 |
| Upper Sorbian         | hsb      | Slavic      | Latin        | 1,272,611       | 38                | 190,832   | 32,500        | 3.559 |
| Urdu                  | ur       | Indic       | Arabic       | 912,459         | 52                | 195,713   | 15,390        | 2.970 |
| Uyghur                | ug       | Turkic      | Arabic       | 1,257,845       | 39                | 164,208   | 31,147        | 3.479 |
| Uzbek                 | uz       | Turkic      | Latin        | 1,548,096       | 29                | 197,038   | 37,078        | 3.436 |
| Venda                 | ve       | Niger-Congo | Latin        | 99,874          | 30                | 17,566    | 3,452         | 2.792 |
| Venetian              | vec      | Romance     | Latin        | 1,046,029       | 37                | 200,507   | 30,156        | 3.265 |
| Vepsian               | vep      | Uralic      | Latin        | 1,366,394       | 30                | 179,658   | 23,467        | 3.501 |
| Vietnamese            | vi       | Austroasiatic | Latin   | 901,129         | 92                | 200,641   | 6,639         | 3.168 |
| Volapüik              | vo       | Constructed | Latin        | 1,001,117       | 30                | 163,400   | 14,480        | 3.237 |
| Võro                  | fiu_vro  | Uralic      | Latin        | 829,251         | 34                | 122,791   | 31,833        | 3.635 |
| Walloon               | wa       | Romance     | Latin        | 1,006,614       | 37                | 197,108   | 20,966        | 3.379 |
| Waray-Waray           | war      | Malayo-Polynesian | Latin   | 1,170,812       | 32                | 200,277   | 19,247        | 3.133 |
| Welsh                 | cy       | Celtic      | Latin        | 1,125,686       | 32                | 204,242   | 19,707        | 3.465 |
| West Flemish          | vls      | Germanic    | Latin        | 1,123,656       | 32                | 200,520   | 30,198        | 3.434 |
| West Frisian          | fy       | Germanic    | Latin        | 1,127,112       | 33                | 194,755   | 21,690        | 3.374 |
| Western Armenian      | hyw      | Armenian    | Armenian     | 1,406,809       | 41                | 196,090   | 42,235        | 3.479 |
| Western Punjabi       | pnb      | Indic       | Arabic       | 922,341         | 50                | 196,296   | 18,042        | 3.595 |
| Wolof                 | wo       | Niger-Congo | Latin        | 983,997         | 34                | 203,888   | 11,051        | 3.292 |
| Wu                    | wuu      | Tibeto-Burman | Chinese     | 54,208          | 3,187             | 6,303     | 5,755         | 4.388 |
| Xhosa                 | xh       | Niger-Congo | Latin        | 1,064,361       | 29                | 130,727   | 41,754        | 3.267 |
| Yiddish               | yi       | Germanic    | Hebrew       | 1,133,439       | 30                | 199,605   | 18,652        | 3.311 |

46
| Language         | Wikicode | Family     | Script | Character Count | Character Set Size | Word Count | Word Set Size | $h_2$ |
|------------------|----------|------------|--------|-----------------|--------------------|------------|---------------|-------|
| Yoruba           | yo       | Niger-Congo| Latin  | 683,230         | 43                 | 153,063    | 18,010        | 3.591 |
| Zamboanga Chavacano | cbk_zam  | Creole     | Latin  | 1,147,970       | 33                 | 195,097    | 24,341        | 3.310 |
| Zazaki           | diq      | Iranian    | Latin  | 1,093,565       | 38                 | 198,835    | 35,611        | 3.528 |
| Zeelandic        | zea      | Germanic   | Latin  | 1,135,510       | 36                 | 204,440    | 29,827        | 3.388 |
| Zhuang           | za       | Tai        | Latin  | 28,961          | 29                 | 4,509      | 1,599         | 3.069 |
| Zulu             | zu       | Niger-Congo| Latin  | 866,234         | 28                 | 102,758    | 33,905        | 3.262 |
C Historical Corpus Statistics

The following table gives basic statistics for each of the Historical Texts, along with separate statistics for parallel texts with normalized and diplomatic versions.

| Text                              | Language | Script | Character Count | Character Set Size | Word Count | Word Set Size | $h^2$ |
|-----------------------------------|----------|--------|----------------|--------------------|------------|---------------|-------|
| Medical Casebooks                 | English  | Latin  | 3,057          | 884                | 15,195     | 41            | 3.418 |
| Normalized                        |          |        | 3,069          | 901                | 14,646     | 54            | 3.435 |
| Diplomatic                        |          |        |                |                    |            |               |       |
| Three Books of Occult Philosophy  | English  | Latin  | 56,914         | 6,675              | 314,259    | 28            | 3.241 |
| Science of Cirurgie               | English  | Latin  | 97,949         | 7,443              | 483,823    | 30            | 3.240 |
| Secretum Secretorum              | English  | Latin  | 18,350         | 2,754              | 96,373     | 27            | 3.111 |
| Alphabet of Tales                 | English  | Latin  | 177,763        | 14,083             | 891,000    | 33            | 3.254 |
| Amiran-Darejaniani                | Georgian | Georgian| 45,169         | 12,888             | 336,149    | 37            | 3.420 |
| Mishneh Torah                     | Hebrew   | Hebrew | 27,261         | 7,857              | 143,516    | 29            | 3.637 |
| Masoretic Tanakh (Bereshit)       | Hebrew   | Hebrew | 17,802         | 6,327              | 99,024     | 29            | 3.526 |
| Without niqqud                    |          |        | 17,802         | 7,091              | 164,569    | 43            | 3.256 |
| With niqqud                       |          |        |                |                    |            |               |       |
| Codex Wormianus                   | Icelandic| Latin  | 31,592         | 8,393              | 162,890    | 48            | 3.390 |
| Normalized                        |          |        | 31,442         | 8,637              | 150,374    | 60            | 3.490 |
| Diplomatic                        |          |        |                |                    |            |               |       |
| La Rettorica                      | Italian  | Latin  | 32,230         | 6,789              | 198,138    | 29            | 3.141 |
| Necrologium Lundense              | Latin    | Latin  | 309            | 222                | 1,723      | 72            | 3.348 |
| Normalized                        |          |        | 314            | 194                | 2,046      | 41            | 3.204 |
| Diplomatic                        |          |        |                |                    |            |               |       |
| De Ortu Et Tempo Antichristi     | Latin    | Latin  | 1,939          | 975                | 13,078     | 24            | 3.252 |
| Historia Hierosylmitanae Expeditionis | Latin  | Latin  | 125,987        | 20,082             | 900,781    | 27            | 3.368 |
| De Magia                          | Latin    | Latin  | 11,790         | 4,067              | 81,028     | 24            | 3.315 |
| Secretum Secretorum              | Latin    | Latin  | 39,349         | 9,294              | 262,206    | 25            | 3.277 |
| Steganographia                   | Latin    | Latin  | 21,529         | 7,559              | 154,739    | 33            | 3.424 |
| Sindbad-Name                      | Persian  | Arabic | 19,751         | 8,672              | 103,002    | 68            | 3.871 |
| Picatrix                          | Spanish  | Latin  | 110,684        | 19,420             | 642,787    | 32            | 3.244 |
D The Sukhotin Algorithm for Vowel Detection

The Sukhotin Algorithm is a procedure for determining which characters of an encoded text are vowels (Sukhotin 1962, Guy 1991b). Guy (1991a) applied the algorithm to two folios of the Voynich text. For further discussion of the Sukhotin algorithm and its usage in Voynich analysis, see Bowern and Lindemann (2020).

The vowel determination for Voynichese is slightly different between Voynich languages, and is heavily dependent upon the transcription system. Here we present the vowel determination results of the Sukhotin Algorithm for multiple languages and scripts in comparison to Voynichese. We include two results: one in which spaces between words are included as a separate character (designated by the # symbol), and one in which they are ignored. If spaces are included, they are almost always identified first by the algorithm, and the overall results are better. In most cases, the exclusion of spaces produces a similar result, but may include consonants that are ‘vowel-like’ (e.g. y) or tend to be found at the beginning or end of words.

| Language | With Spaces | Without Spaces |
|----------|-------------|----------------|
| **English** | #, a, e, i, o, u | a, e, i, o, t, u, y |
| **French** | #, a, â, e, ê, é, ê, i, i, o, ô, u, ü, y | a, â, e, ê, é, ê, i, i, o, ô, u, ü, y |
| **German** | #, a, â, c, e, i, o, ö, u, ü, y | a, â, e, h, i, o, ö, p, u, ü, y |
| **Greek** | #, α, ά, ε, έ, η, ή, ι, ι, o, ô, ω, ώ | α, ά, ε, έ, η, ή, ι, ι, o, ô, ω, ώ |
| **Russian** | #, a, e, ê, i, o, y, b, m, b,  ámb | a, ê, i, o, y, b, m, b,  ámb |
| **Georgian** | #, ღ, ღ, o, ღ, ღ, ღ | ღ, ღ, ღ, o, ღ, ღ, ღ, ღ, ღ |

Figure 20: Results of the Sukhotin Algorithm for Wikipedia Samples in the Latin alphabet (English, French, German), Greek alphabet (Greek), and Georgian alphabet (Georgian).

Figure 20 shows the results for the Wikipedia samples of English, French, German, Greek, Russian, and Georgian, which are written in the Latin, Greek, Cyrillic, and Georgian alphabets. Running the algorithm with spaces does a fairly good job of picking out the vowels exhaustively and exclusively. Without spaces, the algorithm picks out t in English, which is the first letter of the most common word. German also includes the characters c and h, which are commonly used in digraphs. For Georgian, the results are mixed: including spaces, the algorithm correctly identifies the vowels but includes some consonants as well, and it does worse if spaces are excluded.
Figure 21 shows the results of the Sukhotin algorithm for Wikipedia samples of languages in three non-alphabetic scripts: Hebrew, Arabic, and Devanagari (run on the Hindi sample). Hebrew and Arabic are abjads which do not include vowels. The algorithm identifies some consonants which double as vowels, but otherwise has a tendency to identify the final or initial forms of consonants, particularly if spaces are excluded. In Arabic the harakat vowel marking is identified first. The algorithm is somewhat successful for Hindi, written in an abugida (alphasyllabary). Without spaces, it initially identifies the freestanding forms of the vowels, and then the separate combining forms of the vowels. With spaces, it only includes the combining forms.

Figure 22 shows the results of the Sukhotin algorithm for Full Voynich, A, and B in each of the three transcriptions. In the Maximal Simplified transcription, the algorithm identifies common characters which are found primarily in the middle of words and which bear a resemblance to Latin script vowels (a, c, h, i, and o). The only difference between A and B is whether c and h are included. Without spaces, the algorithm additionally identifies word-finals: n, y, and g. For Voynich B it also identifies the rare character symbol, and the f gallows character which marks the beginning of a paragraph.

In the full Maximal transcription, the algorithm identifies the same characters as vowels, but the results are swamped by the addition of the ligature forms of consonants and vowels, and, when spaces are excluded, other extremely rare characters. Similarly, in the Minimal transcription the same characters are identified as vowels, with the exception of i and the inclusion of the digraph ee. Characters which are only found at the beginning of the words are also included: q, X, Q, and S and C.

The Sukhotin algorithm is based upon the observation that vowels are more likely to be adjacent to consonants than to other vowels (Guy 1991b). This generalization is ultimately a reflection of universal properties of sonority in the world’s languages, but its validity for any one language is subject to phonotactics (language-particular rules about possible phoneme sequences) and the peculiarities of the script. In addition to identifying vowels, it tends to identify rare characters and word-initial/final characters.

---

37 For Hebrew and Arabic the ordering of results is given from right to left to reflect the order in which these characters are read. Devanagari is ordered from left to right.
### Voynich Maximal Simplified

| Voynich Maximal Simplified | Full Voynich | With Spaces | Without Spaces | #, α (a), ε (e), γ (h), ñ (i), σ (o) |
|----------------------------|--------------|-------------|----------------|---------------------------------------|
|                            | Vojnich A    | With Spaces | Without Spaces | #, α (a), γ (h), ñ (i), σ (o)          |
|                            | Vojnich B    | With Spaces | Without Spaces | #, α (a), ε (e), γ (h), ñ (i), σ (o)   |

### Voynich Maximal

| Voynich Maximal | Full Voynich | With Spaces | Without Spaces | #, α (a), ε (e), γ (F), γ (h), γ (i), γ (I), γ (K), σ (o), γ (O), γ (Y) |
|-----------------|--------------|-------------|----------------|--------------------------------------------------------------------------------|
| Vojnich A       | With Spaces | Without Spaces | #, α (a), ε (e), γ (F), γ (h), γ (i), γ (I), γ (K), σ (o), γ (Y) |
| Vojnich B       | With Spaces | Without Spaces | #, α (a), ε (e), γ (F), γ (h), γ (i), γ (I), γ (K), σ (o), γ (Y) |

### Voynich Minimal

| Voynich Minimal | Full Voynich | With Spaces | Without Spaces | #, α (a), ε (e), γ (E), γ (h), σ (o), γ (X) |
|-----------------|--------------|-------------|----------------|------------------------------------------|
| Vojnich A       | With Spaces | Without Spaces | #, α (a), ε (e), γ (E), γ (h), σ (o), γ (X) |
| Vojnich B       | With Spaces | Without Spaces | #, α (a), ε (e), γ (E), γ (h), σ (o), γ (X) |

Figure 22: Results of the Sukhotin Algorithm for Voynichese in the Maximal Simplified, Full Maximal, and Minimal transcriptions.
References

Amancio, D. R., Altmann, E. G., Rybski, D., Oliveira Jr, O. N., and Costa, L. d. F. (2013). Probing the statistical properties of unknown texts: application to the voynich manuscript. *PLoS One*, 8(7).

Barlow, M. (1986). The Voynich Manuscript - By Voynich? *Cryptologia*, 10(4):210–216. Citation Key Alias: barlow1986a.

Bennett, W. R. (1976). *Scientific and engineering problem-solving with the computer*, chapter 4. Language, pages 103–198. Prentice Hall series in automatic computation. Prentice Hall, Englewood Cliffs, N.J.

Bowern, C. and Lindemann, L. (2020). The Linguistics of the Voynich Manuscript. *Annual Review of Linguistics*.

Currier, P. (1976). Papers on the Voynich Manuscript. In D’Imperio, M., editor, *New Research on the Voynich Manuscript*, Washington, DC.

Davis, L. F. (2020). How many glyphs and how many scribes: Digital paleography and the voynich manuscript. *Manuscript Studies*, V(1):162–178.

Guy, J. B. (1991a). Statistical properties of two folios of the Voynich Manuscript. *Cryptologia*, 15(3):207–218.

Guy, J. B. (1991b). Vowel identification: an old (but good) algorithm. *Cryptologia*, 15(3):258–262.

Landini, G. (2001). Evidence of linguistic structure in the Voynich manuscript using spectral analysis. *Cryptologia*, 25(4):275–295.

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. Association for Computational Linguistics.

Rugg, G. (2004). An elegant hoax? a possible solution to the voynich manuscript. *Cryptologia*, 28(1):31–46.

Stallings, D. (1998). Understanding the second-order entropies of Voynich text. [http://ixoloxi.com/voynich/mbpaper.htm](http://ixoloxi.com/voynich/mbpaper.htm).

Sukhotin, B. (1962). Eksperimental’noe vydelenie klassov bukv s pomoseju evm. *Problemny strukturoj lingvistiki*, 234:189–206.

Timm, T. and Schinner, A. (2020). A possible generating algorithm of the Voynich manuscript. *Cryptologia*, 44(1):1–19.

Timm, T. and Schinner, A. (2021). Review of the linguistics of the voynich manuscript by claire bowern and luke lindemann. *Cryptologia*, 0(0):1–5.

Zandbergen, R. (2010). Voynich ms. [http://www.voynich.nu/index.html](http://www.voynich.nu/index.html). [Online; accessed 24-April-2018].
Zandbergen, R. (2021). The cardan grille approach to the voynich ms taken to the next level. arXiv:2104.12548 [cs]. arXiv: 2104.12548.