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

Luke Lindemann and Claire Bowern

October 27, 2020

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 294 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

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 (Reddy and Knight 2011; Amancio et al. 2013; Landini 2001).[1]

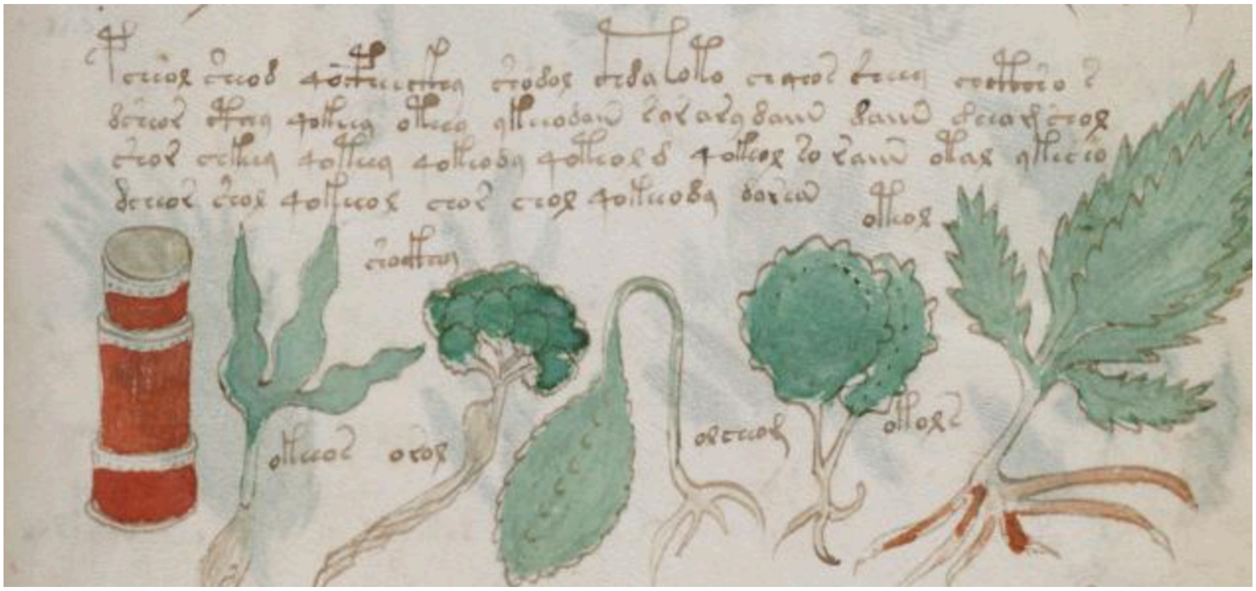


Figure 1: A paragraph of text and labelled figures from the "Recipes" section of the Voynich Manuscript (f100r).

The major unsolved question of the Voynich text is whether it represents meaningful language.[2] 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 to encode meaning, but the nature of the encipherment obscured it in such a way that the original

[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.

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.

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 languages.[3]

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 294 languages, representing thirty-eight 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 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.[4] These numerals were probably not written by

[3] This 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.

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

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:[5]

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.[6] 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).[7] 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 suns and containing stars or tubes. The other side contains text and more circular diagrams.
4. The Recipes section is distinguished by pages with paragraphs of text separated by assortments 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.

[5] For 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.

[6] The 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.

[7] Though 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.

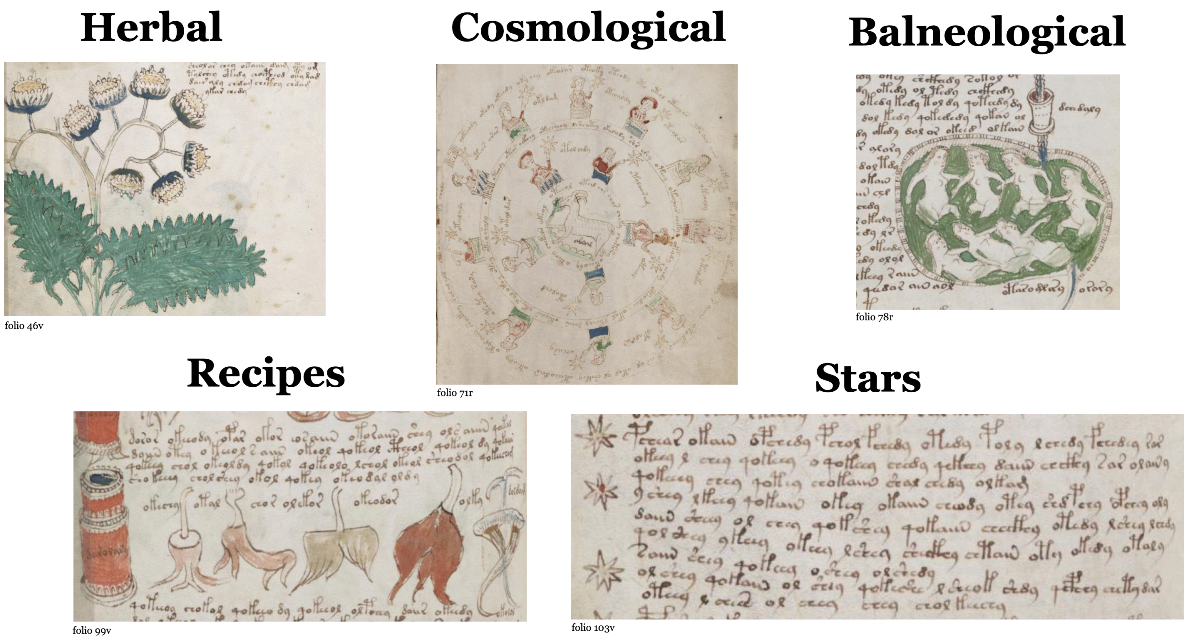


Figure 2: Selected examples from each section of the Voynich Manuscript

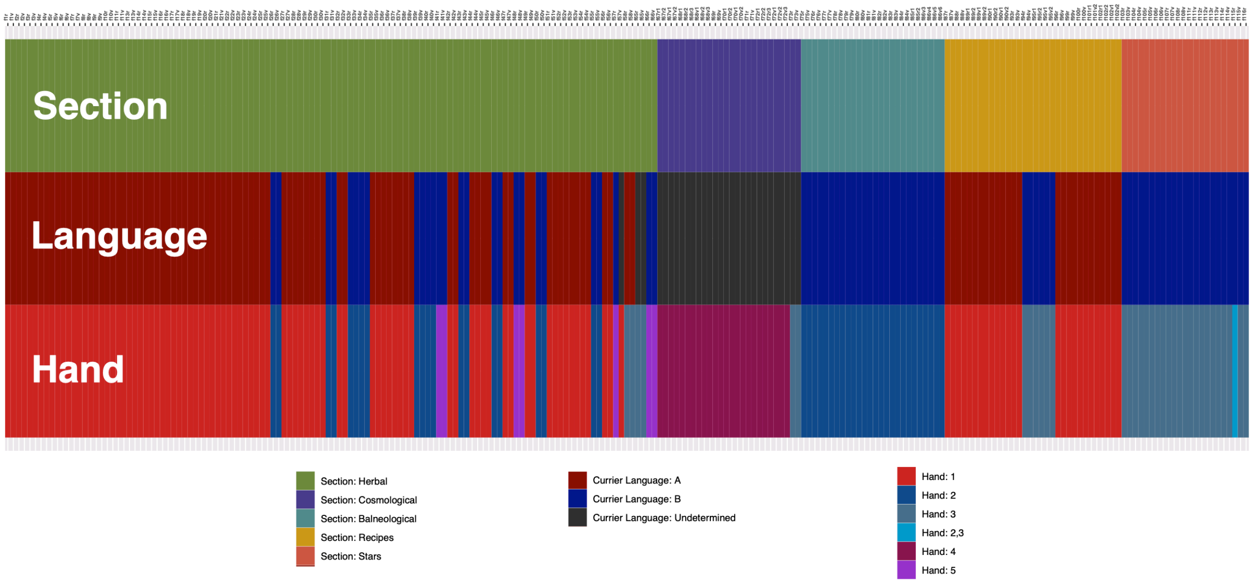


Figure 3: Map of Sections, Languages (Currier 1976) and Scribal Hands (Davis 2020)

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 "languages" (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 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.[8]

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.

[8] 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.

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.

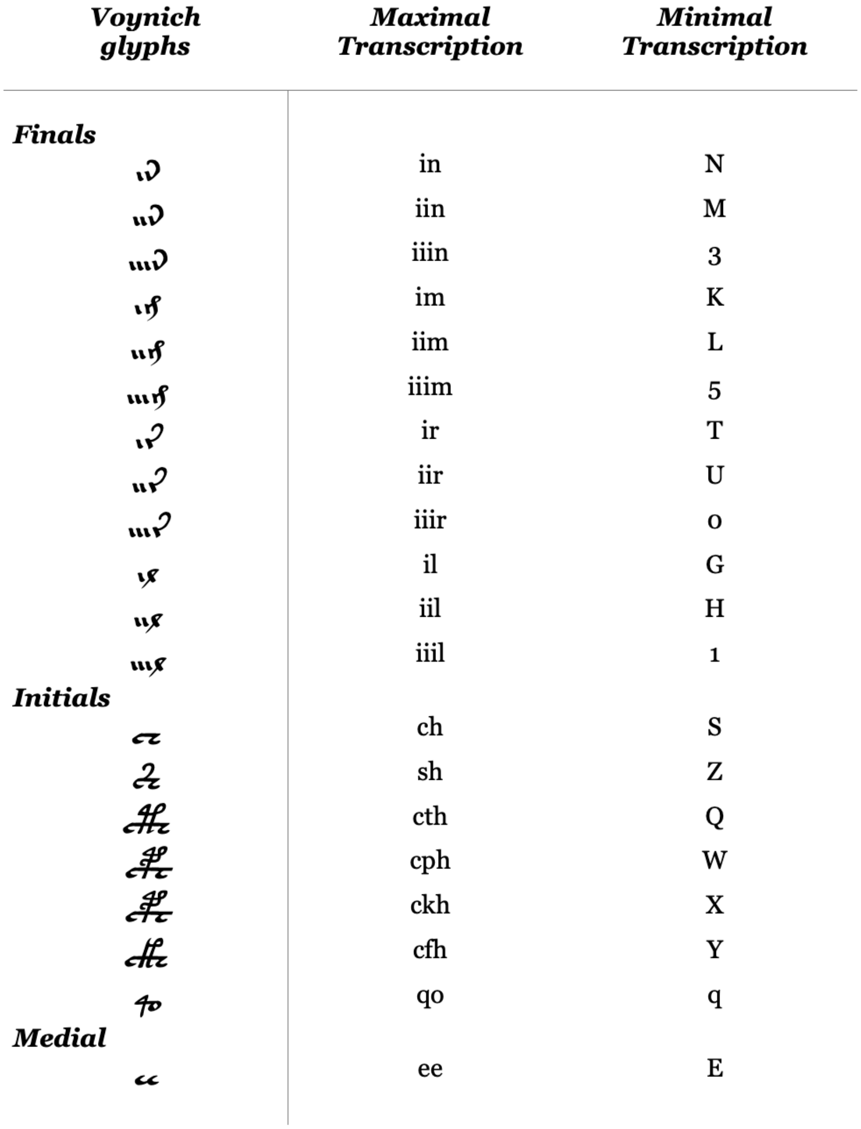
The particular transcription system used to write Voynichese can have a measurable effect on the statistical properties of the text itself. All of the documents above have been converted into three different transcription systems: Simplified Maximal, Full Maximal, and Minimal. The important issue of 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 transcription was designed to be convertible to other major transcription systems.[9] The characters of EVA therefore represent a lower bound on the length of characters to allow for the conversion

[9] 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.



Voynich glyphs	Maximal Transcription	Minimal Transcription
Finals		
	in	N
	iin	M
	iiin	3
	im	K
	iim	L
	iiim	5
	ir	T
	iir	U
	iiir	0
	il	G
	iil	H
	iiil	1
Initials		
	ch	S
	sh	Z
	cth	Q
	cph	W
	ckh	X
	cfh	Y
	qo	q
Medial		
	ee	E

Figure 4: Glyph combinations and their Minimal and Maximal transliterations. Maximal Voynich is equivalent to EVA. 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.

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.

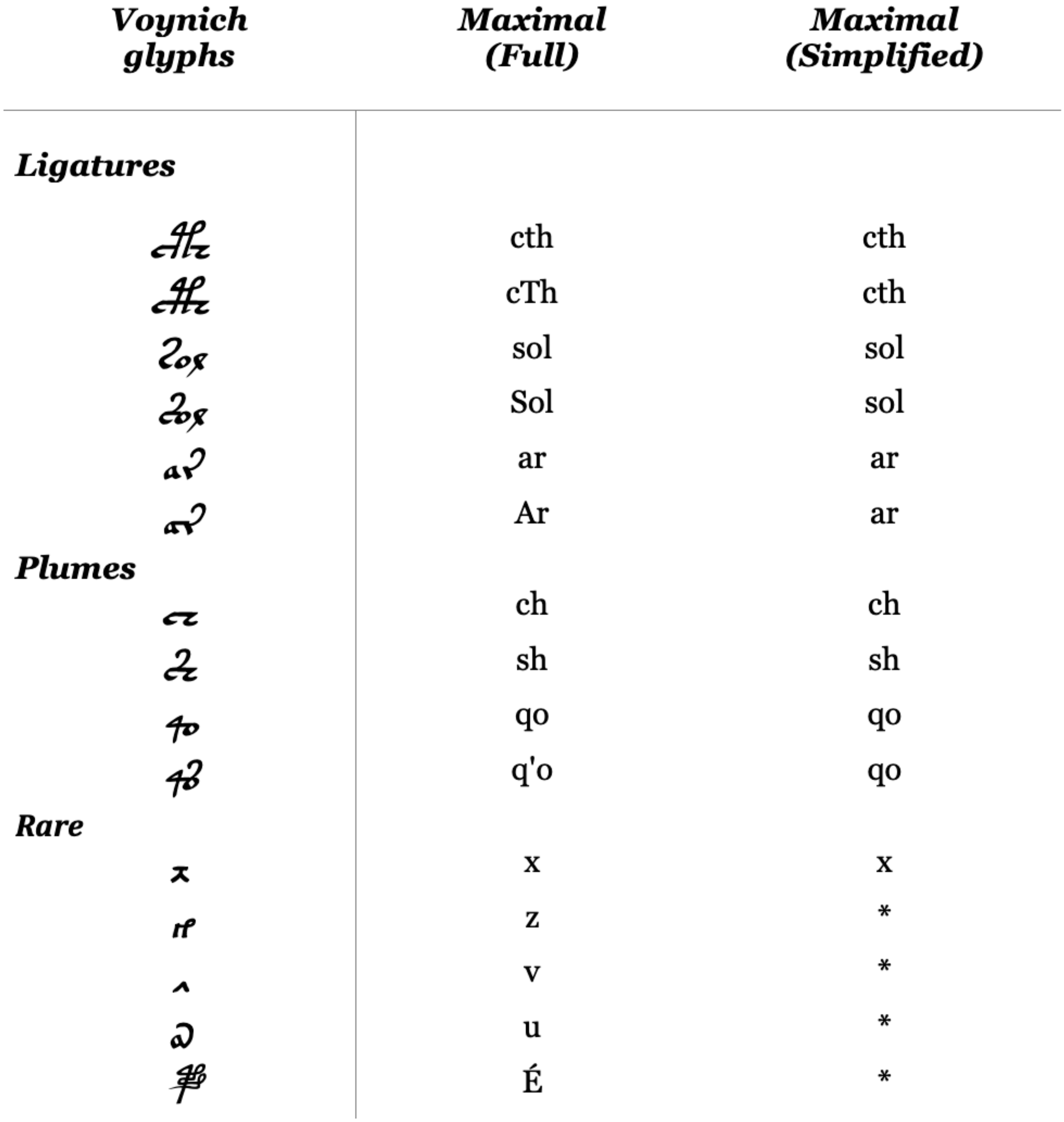
We take EVA to be the most decomposed system, i.e. the Maximal transcription. For a Minimal transcription, we take EVA and substitute all of the glyph combinations in Currier's system. We add two additional glyph combinations based on suggestions from Zandbergen (2010). The differences between the Minimal 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, plumes, 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. EVA distinguishes ligatures by capitalizing the first of the letters (see Figure 3.1.1).

Plumes are loops which are almost exclusively found above the *c* character. In EVA, the *c* with a loop above it is written as *s*. There are seven plumes found elsewhere in the document, and they are written in EVA with an apostrophe (*'*).

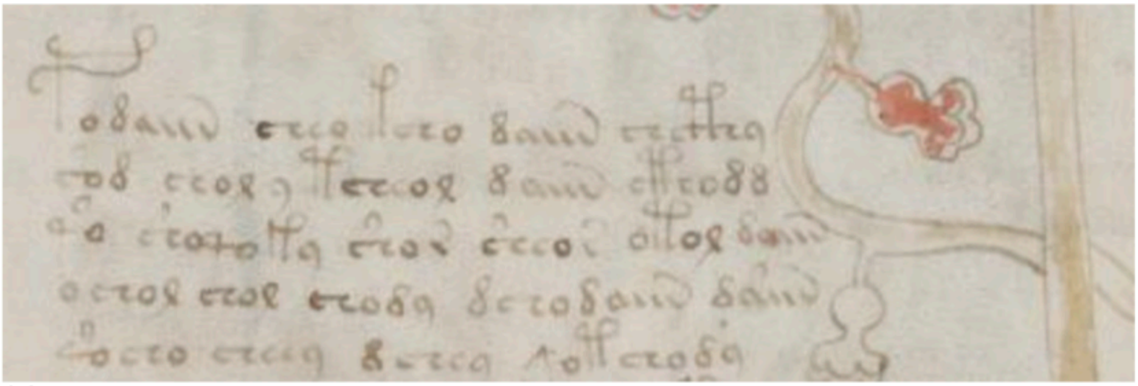
There are sixteen characters which appear less than fifty times each in the entire manuscript. Some of the infrequent characters are recognizable astrological symbols or are found only on cosmological diagrams, while others appear to be variants of other characters or even typographical mistakes. There 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, plumes, and rare characters included. Simplified Maximal removes the ligatures, keeps the *c/s* distinction but considers other plumes to be rare characters, and uses an asterisk to designate all rare characters (with the exception of the *x*, which is the most frequent of the rare characters and appears in both diagrams and text). The Simplified Maximal alphabet has about half as many characters (22 rather than 38), 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.

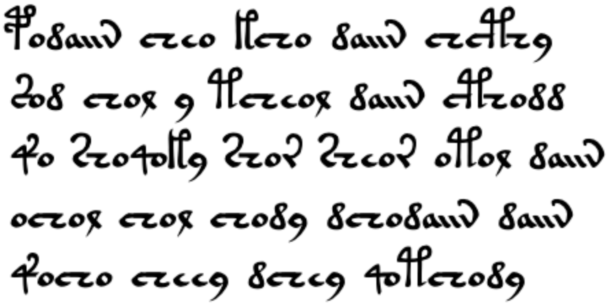


Voynich glyphs	Maximal (Full)	Maximal (Simplified)
Ligatures		
	cth	cth
	cTh	cth
	sol	sol
	Sol	sol
	ar	ar
	Ar	ar
Plumes		
	ch	ch
	sh	sh
	qo	qo
	q'o	qo
Rare		
	x	x
	z	*
	v	*
	u	*
	É	*

Figure 5: The Simplified Maximal transcription: ligatures are ignored, and rare characters (including unusual plumes) are represented by an asterisk.



folio 49r



Minimal:

podaM Seo kSo daM SQy
sod Sol y tSeol daM Qodd
q*o Zoqky Zor Zeor otol daM
oSol Sol Sody dSodaM daM
q*oSo SEy dSey qtSody

Maximal Full (EVA):

podaiin cheo kcho daiin chcthy
Sod chol y tcheol daiin cthodd
q'o shoqoky shor sheor otol daiin
ochol chol chody dchodaiin daiin
q'ocho cheey dchey qotchody

Maximal Simplified:

podaiin cheo kcho daiin chcthy
sod chol y tcheol daiin cthodd
q*o shoqoky shor sheor otol daiin
ochol chol chody dchodaiin daiin
q*ocho cheey dchey qotchody

Figure 6: Full Maximal, Simplified Maximal, and Minimal transcriptions of folio 49 recto, paragraph 2, lines 1-5

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.[10] 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.[11] 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.*

[10] A full set of transliterations can be found at `http://www.voynich.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.

[11] 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 300 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.[12] 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.[13] 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 294 language samples written in thirty-two different scripts, categorized into thirty-eight major language families and seventy-one subfamilies. In most cases, the samples consist of the first 500 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 eight 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 August 2019 from wikimedia dump files. These files are continually updated and available at `http://dumps.wikimedia.org`. We downloaded 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 500 wikipedia entries for that language.[14] The script also removes capitalization

[12] See `http://meta.wikimedia.org/wiki/List_of_Wikipedias` for a current list of wikipedia versions by number of entries.

[13] 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.

[14] This is a modified version of the extremely useful script written by Matthew Mayo at `http://www.kdnuggets.com/2017/11/building-wikipedia-text-corpus-nlp.html`. However, this method fails on most abugida (alphasyllabary) scripts, because it deletes vowel-representing combining characters. For the Bengali, Devanagari, Gujarati,

and punctuation.

We additionally 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*) and (*é*). We filtered the text by the unicode range of the particular language's script in order to delete irrelevant characters as well as metadata and tags (which are mostly written in the Latin script), and filtered out particular English metadata tags. However, it is not always possible to completely remove all English metadata from language written in the Latin script. This may have an effect on languages with very little text and short entries. A small number of wikipedia versions (e.g. Piedmontese and Cree) contain relatively little running text. Rather, they have short pages with extremely formulaic entries and/or vocabulary lists with glosses. They tend to be clear outliers in measurements of natural language word statistics like type-token ratio.

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.[15] 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.

1. **Afro-Asiatic**:
   (a) **Semitic:** Amharic, Arabic, Aramaic, Egyptian Arabic, Hebrew, Maltese, Tigrinya
   (b) **Other families:** Hausa, Kabyle, Oromo, Somali
2. **Albanian**: Albanian
3. **Algonquian**: Atikamekw, Cheyenne, Cree

Gurmukhi, Kannada, Khmer, Lao, Malayalam, Myanmar, Odia, Sinhala, Tamil, Telugu, Thaana, Thai, and Tibetan scripts, we simply filtered the Latin text out of the first 100,000 lines of the dump file. The result is primarily running text because most of the metadata in the code is written in the Latin script. For Pali and Cree, entries were simply copied by hand because of the prohibitively large ratio of text to metadata.

[15] The exception is Lojban, which is a constructed logical language designed as an experiment in eliminating syntactic ambiguity.

4. **Armenian**: Armenian, Western Armenian
5. **Athabaskan**: Navajo
6. **Austroasiatic**: Khmer, 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, Lojban, Lingua Franca Nova, Novial, Volapük
12. **Creoles**:
    (a) **English:** Bislama, Jamaican Patois, Norfolk, Sranan, Tok Pisin
    (b) **French:** Haitian
    (c) **Portuguese:** 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, Luxembourgisch, 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:** Hindi, Urdu, Fiji Hindi,
    (b) **Eastern:** Assamese, Bengali, Maithili, Odia, Bihari, Bishnupriya Manipuri,
    (c) **Northern:** Doteli, Nepali
    (d) **Northwestern:** Sindhi, Punjabi, Western Punjabi
    (e) **Southern:** Sinhalese, Divehi, Marathi, Goan Konkani
    (f) **Western:** Romani, Gujarati
    (g) **Other families/proto-languages:** Kashmiri, Sanskrit, Pali
17. **Iranian**: Sorani, Zazaki, Persian, Gilaki, Kurdish, Northern Luri, Mazandarani, Ossetian, Pashto, Tajik

18. **Iroquoian**: Cherokee
19. **Japonic**: Japanese
20. **Kartvelian**: Georgian, Mingrelian
21. **Koreanic**: Korean
22. **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, Buginese, Chamorro, Fijian, Malagasy, Nauruan, Sundanese, Tetum
23. **Mande**: Bambara
24. **Moksha**: Moksha
25. **Mongolic**: Buryat, Mongolian, Kalmyk
26. **Niger-Congo**:
    (a) **Bantu:** Kongo, Lingala, Luganda, Northern Sotho, Chichewa, Kikuyu, Kinyarwanda, Kirundi, Shona, Swati, Sesotho, Swahili, Tswana, Tsonga, Tumbuka, Twi, Venda, Xhosa, Zulu
    (b) **Other families:** Akan, Ewe, Fula, Igbo, Kabiye, Sango, Wolof, Yoruba
27. **Nilotic**: Dinka
28. **Quechua**: Quechua
29. **Romance**:
    (a) **Italo-Dalmatian:** Italian, Corsican, Sicilian, Neapolitan, Venetian, Tarantino
    (b) **Gallo-Romance:** Catalan, French, Franco-Provençal, 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
30. **Slavic**
    (a) **East Slavic:** Belarusian, Belarusian Taraškievica, 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

31. **Tai**: Lao, Shan, Thai, Zhuang
32. **Tibeto-Burman**: Tibetan, Min Dong, Dzongkha, Gan, Hakka, Burmese, Newar, Wu, Chinese, Classical Chinese, Min Nan, Cantonese
33. **Tupian**: Guarani
34. **Turkic**:
    (a) **Oghuz:** Azerbaijani, Chuvash, Gagauz, South Azerbaijani, Turkmen, Turkish
    (b) **Karluk:** Uyghur, Uzbek
    (c) **Kipchak:** Bashkir, Crimean Tatar, Karakalpak, Kazakh, Karachay-Balkar, Kirghiz, Tatar
    (d) **Siberian:** Sakha, Tuvan
35. **Uralic**:
    (a) **Finnic:** Estonian, Finnish, Vepsian, Võro
    (b) **Permic:** Komi-Permyak, Komi, Udmurt
    (c) **Mari:** Meadow Mari, Hill Mari
    (d) **Other families:** Erzya, Hungarian, Livvi-Karelian, Northern Sami
36. **Uto-Aztecan**: Nahuatl
37. **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.[16] 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).

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.[17] The Icelandic text comes from the Medieval Nordic Text Archive.[18] The Hebrew Masoretic Tanakh comes from Sacred Texts[19] and the Christian Classics Ethereal Library[20], while the *Mishneh of Maimonidies* was obtained here.[21]

Of the Latin texts, the *Secretum Secretorum* comes from the Corpus Corporum of the University

[16] 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.

[17] `http://titus.uni-frankfurt.de`

[18] `http://clarino.uib.no/menota/page`

[19] `http://www.sacred-texts.com/bib/tan/index.htm`

[20] `http://www.ccel.org/a/anonymous/hebrewot/home.html`

[21] `http://kodesh.snunit.k12.il/i/0.htm`

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

Table 1: Details of historical manuscripts

of Zurich[22], the *Necrologium Lundense* comes from the Necrologium Lundense Online[23], *De Ortu et Tempo Antichristi* and *Historia Hierosylmitanae Expeditionis* come from the Latin Library[24], and *De Magia* and *Steganographia* come from the Twilit Grotto.[25]

The English *Alphabet of Tales* and *Science of Cirurgie* come the Corpus of Middle English Prose and Verse at the University of Michigan.[26] The *Medical Casebooks* come from the Casebooks Project of the University of Cambridge.[27] The English translation of the *Secretum Secretorum* comes from Colour Country,[28] and Agrippa's *Three Books of Occult Philosophy* are from the Twilit Grotto.[29]

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

[22] `http://mlat.uzh.ch/MLS/`

[23] `http://notendur.hi.is/mjm7/`

[24] `http://www.thelatinlibrary.com`

[25] `http://esotericarchives.com`

[26] `http://quod.lib.umich.edu`

[27] `http://casebooks.lib.cam.ac.uk`

[28] `http://www.colourcountry.net/secretum/`

[29] `http://esotericarchives.com`

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 (*h2*). 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, *h2*, is dependent upon the conventions of the script in which it is written. For example, Bennett (1976) found that the *h2* 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 *h2* value of a text:

1. Document Length
2. Character set size (total number of characters in the alphabet)
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 *h2* 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 Full Maximal transcription (41 characters) and Minimal transcription (45 characters) is well within the general range for alphabets: 24-92 characters.[30] 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.

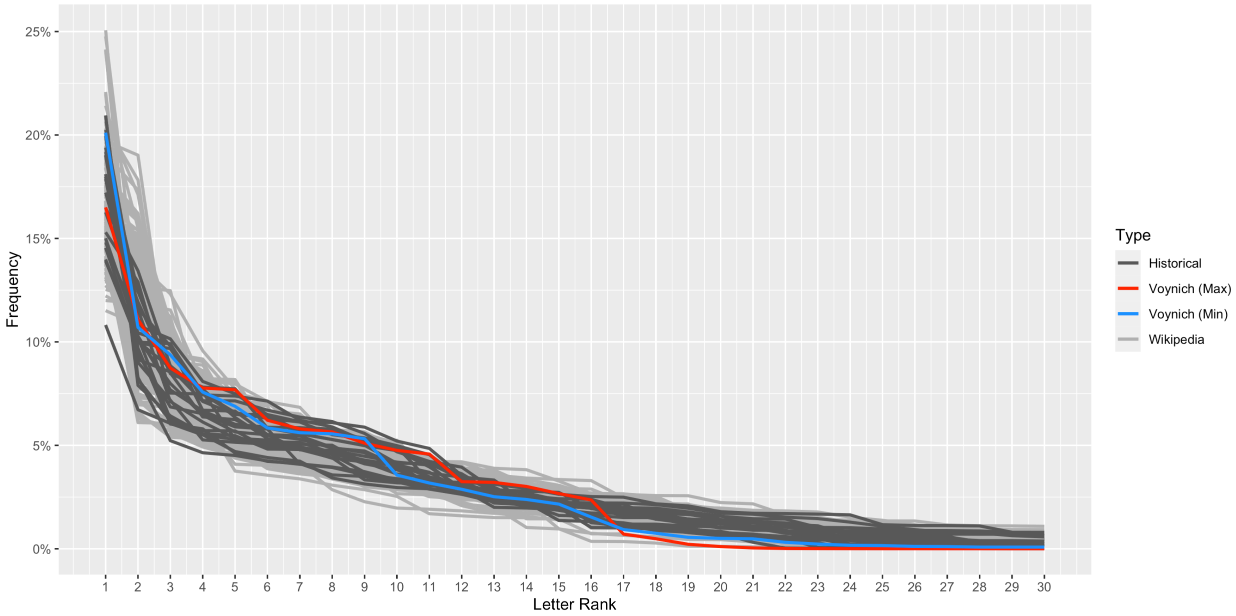


Figure 7: Proportional Frequency of the thirty most frequent characters in the Historical Corpus texts, the (alphabetic) Wikipedia Corpus texts, and Voynichese (Full Maximal and Minimal transcriptions). Simplified Maximal is identical to Full Maximal with regards to frequent characters. The first ranked character for each language is a space, and its frequency indicates average word length.

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

[30] The Simplified Maximal transcription has only 21 characters, but it excludes rare characters and therefore represents an absolute lowest estimate.

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 3.88 for Minimal Voynich and 3.94 for Maximal Voynich, while the overall range in the corpora is from 3.72-4.82 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 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 largest text in our corpora, the English wikipedia sample, which consists of 1,881,447 words. We calculated the *h2* values samples of randomly selected sequences of text at various word lengths. The results are in Figure 8.

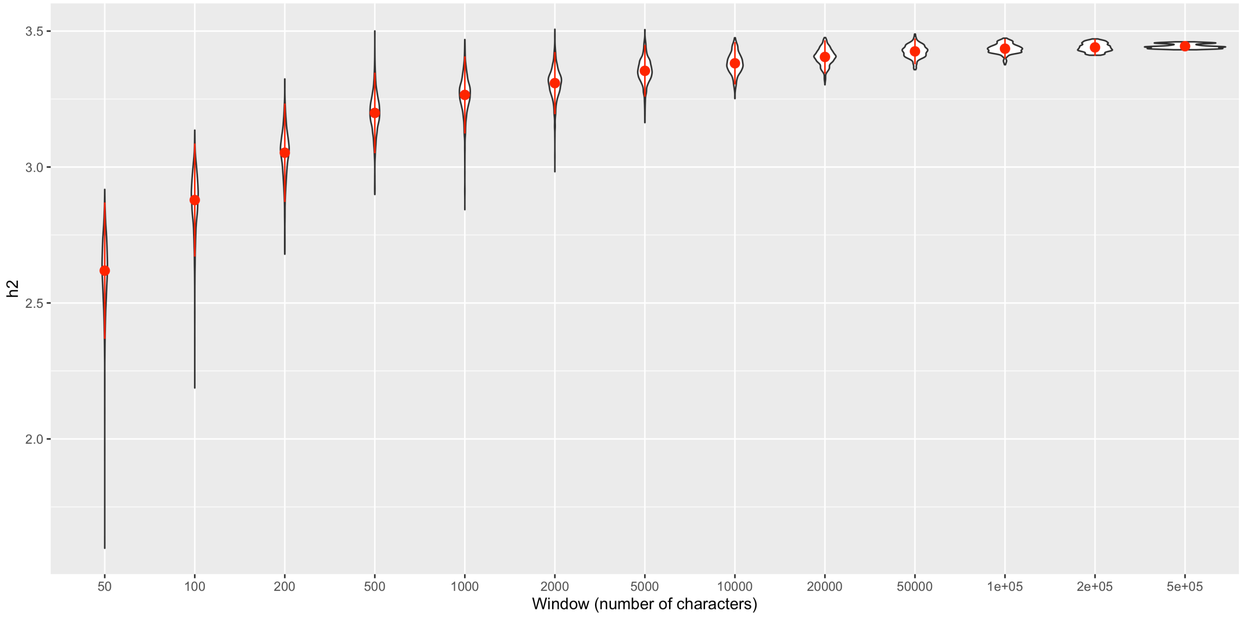


Figure 8: Density plot of *h2* values for random samples of English at multiple word lengths from 50-500,000. Each window size is sampled 1,000 times. The red dot is the average and the line indicates one standard deviation from the mean.

With a window of only 50 words, the average *h2* is 2.62, there is a wide range of 1.6-3.9, and the standard deviation is 0.13. The average is much lower than the text's overall *h2* value of 3.45. 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 3.38, the range is 3.25-3.48, and the standard deviation is 0.039. This means that 95% of the samples are within 0.078 bits of the average, and the average is within 0.07 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 0.14 for A and 0.15 for B (compared to 0.13 in English), and at 5,000 words it is 0.047 for A and 0.067 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 *h2*. Maximal Voynich has a lower *h2* than Minimal Voynich, because glyph compositions are based upon common glyph sequences (making the text appear more predictable).

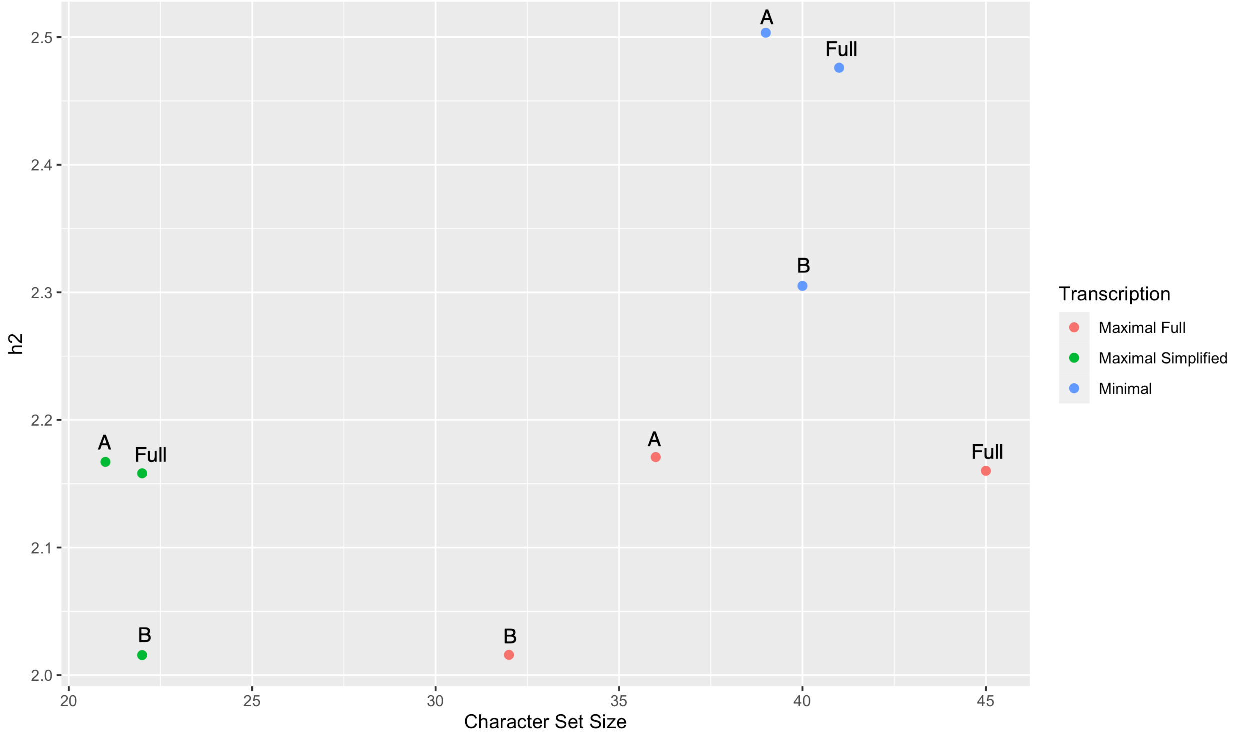


Figure 9: Character set size is plotted against conditional character entropy (*h2*). Lower *h2* values indicate more predictability in character bigrams.

The *h2* 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 *h2*. Scribe Hand 1, which is used to write Language A, has a nearly identical *h2*. 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 20 and 45 characters, and the conditional character entropy ranges from 1.95 to 2.50 bits.

4.3.1 Character Set Size

The Full Maximal and Maximal Simplified transcriptions have nearly identical *h2* values despite a significant difference in the size of the character sets. With the Full Voynich text, the Simplified Maximal transcription has twenty fewer characters, but there is only a .09% difference in conditional character entropy (2.159 to 2.157).

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 *H0*, 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 *h2* of 7 bits.

However, if the additional characters are rare, the overall effect on *h2* is slight. The process of cleaning a text by removing highly infrequent characters does not have an appreciable effect on entropy. For example, the unprocessed English wikipedia sample has an alphabet of 198 characters, which includes characters like *ü* and *ç*. Processing the sample involves removing characters which appear with a frequency of less than 0.01%, after which only 28 letters remain (including the space and *é*). However, the difference in *h2* between the unprocessed and processed English sample is only 0.15% (3.453 to 3.448).

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 *h2* value for Language A and Language B come within 3% of each other.[31]

[31] The frequency of this common glyph sequence is partially attributable to a single word *chedy*, which is the most

4.4 Comparison to the Wikipedia Corpus

Of the 294 wikipedia language samples represented in the Wikipedia Corpus, none of them have an *h2* comparable to Voynichese. Voynichese has lower values, meaning that its text is more predictable. While the Minimal Voynich transcription is slightly closer to the lowest values we find in the Wikipedia Corpus (with an average *h2* of 2.43 rather than 2.16), this is still lower than the *h2* range in the Wikipedia Corpus, from 2.85-6.25.

Figure 10 depicts the character set size and conditional character entropy for texts that use between 20 and 55 characters and have an *h2* range from 2 to 4. This is the range of most of the alphabets in the Corpus. The majority of Wikipedia versions (193 languages) are written in the Latin script, although the Corpus also includes samples of Cyrillic, Georgian, Gothic, Greek, and Ol Chiki. Languages written in Cyrillic, which are mostly Slavic and Turkic, tend to have a somewhat higher *h2*.

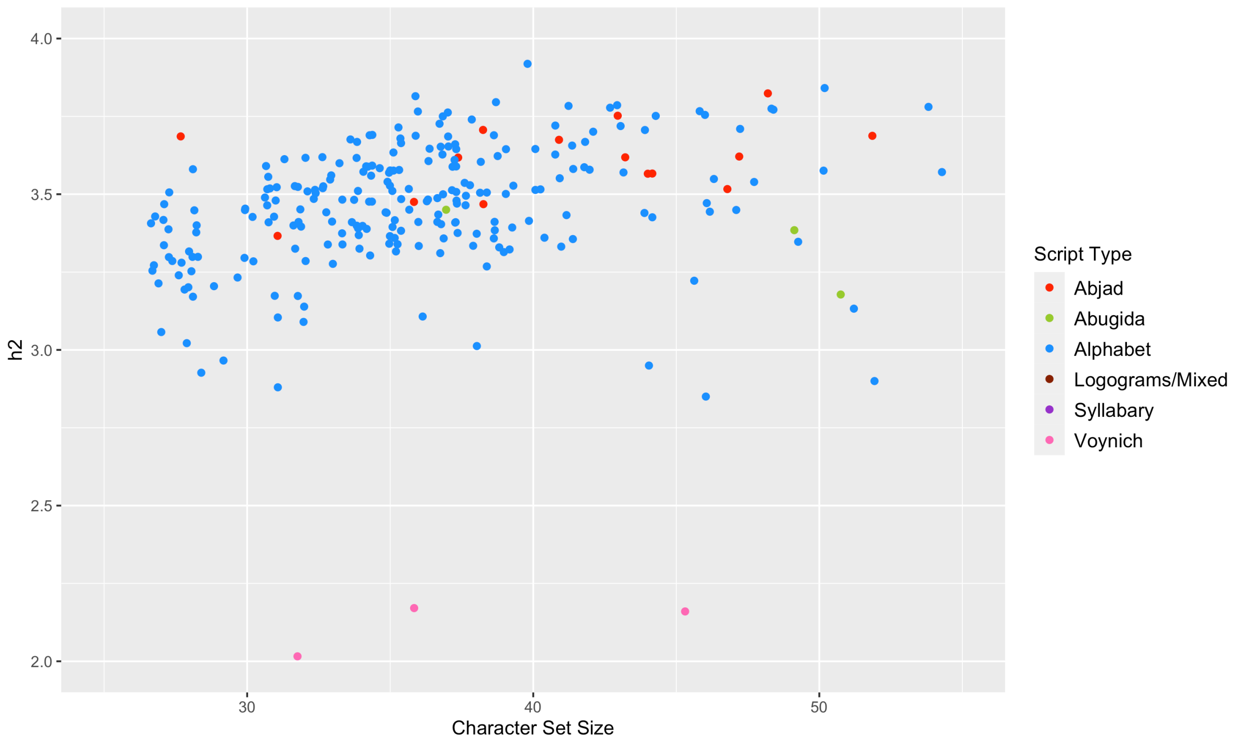


Figure 10: Conditional Character Entropy and Character Size for Wikipedia Languages from 20-130 characters. Full Maximal Voynich (A, B, and Full) is shown at the bottom. The languages with the closest *h2* values are Hakka (2.85), Venda (2.88), Min Dong (2.90), Tswana (2.93), Pali (2.95), Atikamekw (2.97), Hawaiian (3.01), and Lojban (3.02).

All but two 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

common word in B and almost entirely absent from A. But even when this word is disregarded, -*edy* is significantly more common in B.

languages, as well as the Germanic language Yiddish. They also have a somewhat higher *h2* 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 *h2* values. The languages with *h2* values closest to Voynichese are Hakka, Venda, Min Dong, and Tswana. Hakka and Min Dong are Tibeto-Burman languages, while Venda and Tswana are Bantu (Niger-Congo). It is noteworthy that all four 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, 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 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.

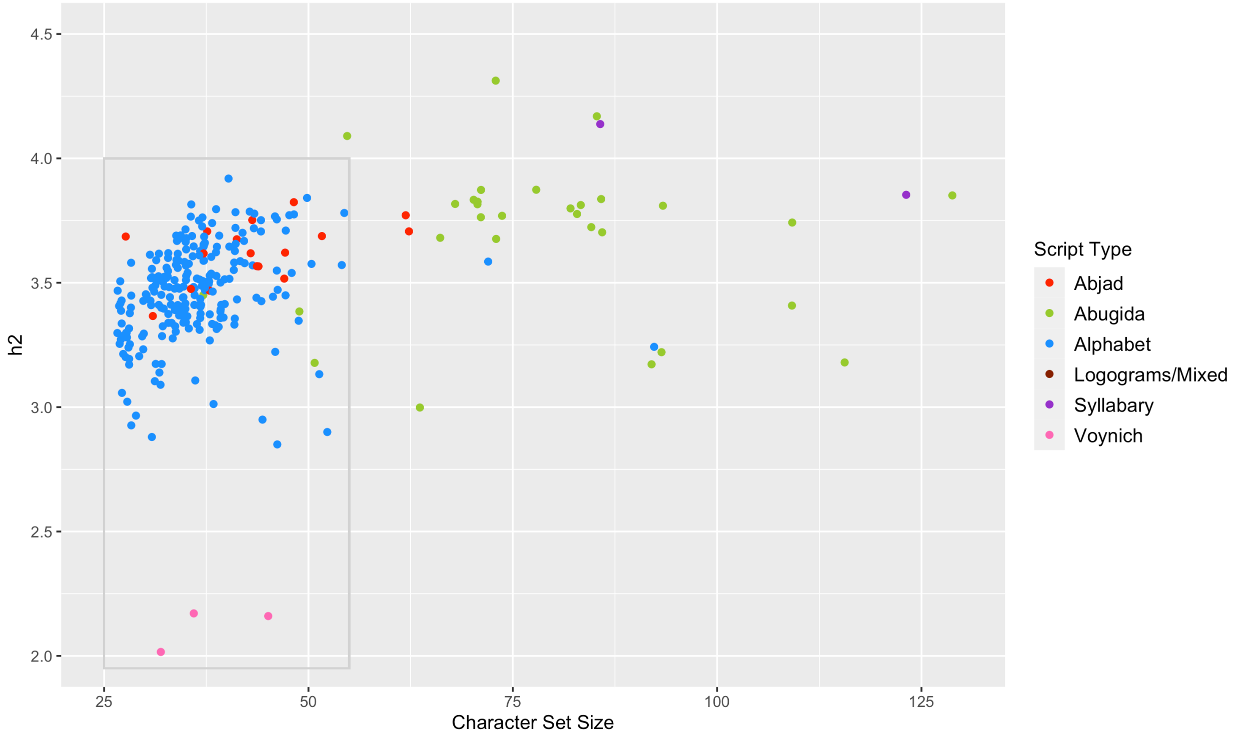


Figure 11: Conditional Character Entropy and Character Size for Wikipedia Languages from 20-130 characters. Most of the Abugidas are in this range, of which the Tibeto-Burman and Indic languages tend to have lower *h2* values and the Austroasiatic and Tai languages have higher *h2* values. The grey box indicates the range of the previous graph.

Expanding even further, Figure 12 includes languages that contain up to 7600 characters and have a conditional entropy up to 6.5. The languages that are written with logograms have the high-

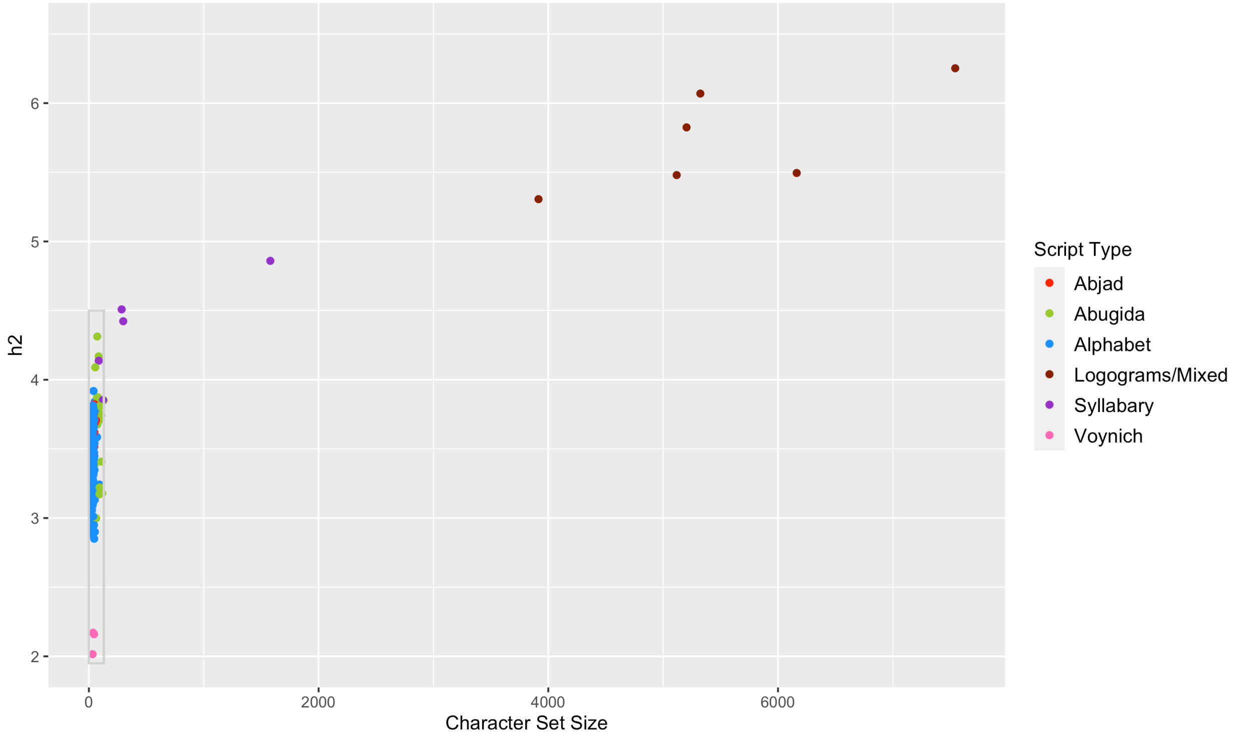


Figure 12: Conditional Character Entropy and Character Size for Wikipedia Languages from 20-7600 characters. The grey box indicates the range of the previous graph.

est 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 *h2*.

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 Ethiopic (Ge'ez) script used to write Amharic and Tigrinya. Technically, Ethiopic 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 *h2* 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 *h2*. Voynichese clearly falls outside of the range of most abugidas, syllabaries, and logograms.

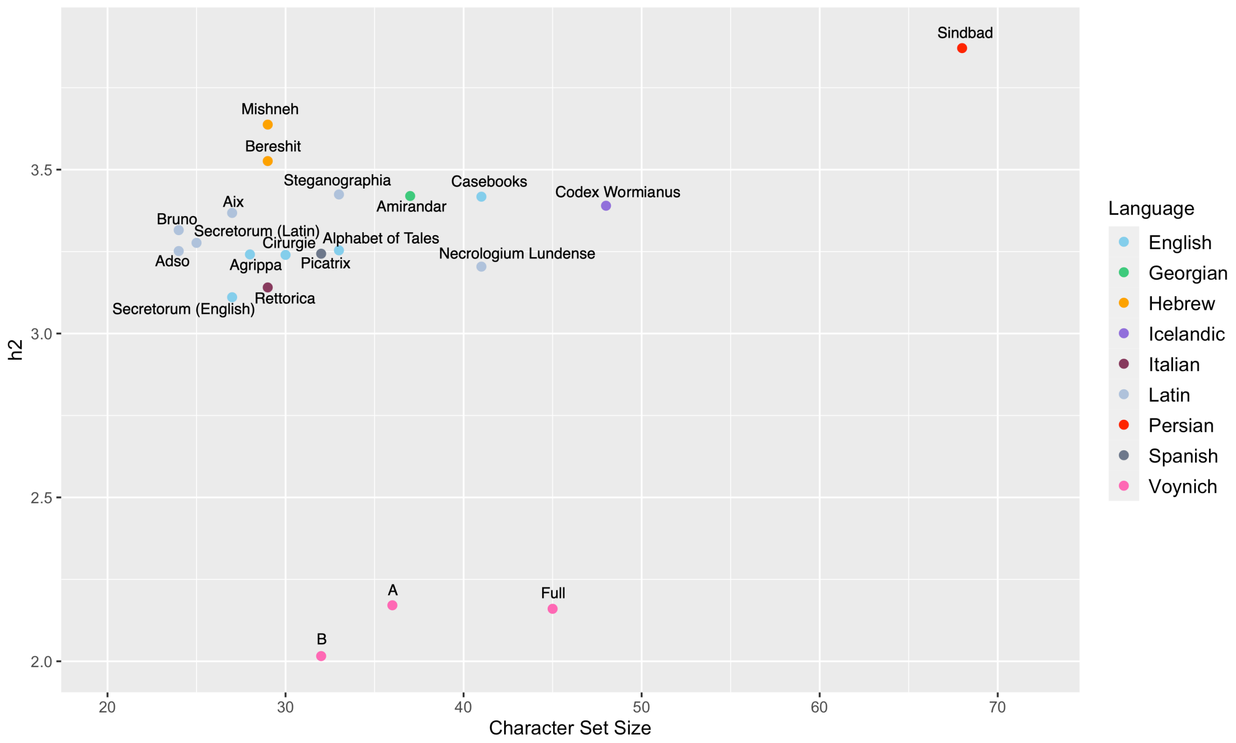


Figure 13: Conditional Character Entropy and Character Size for Historical Texts. The Voynich transcription is Full Maximal, and the historical scripts are normalized and unabbreviated. Hebrew is calculated without vowel markings. The alphabets range from 3-3.5 and the abjads range from 3.5-4.

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 *h2* 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 *h2* 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 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.

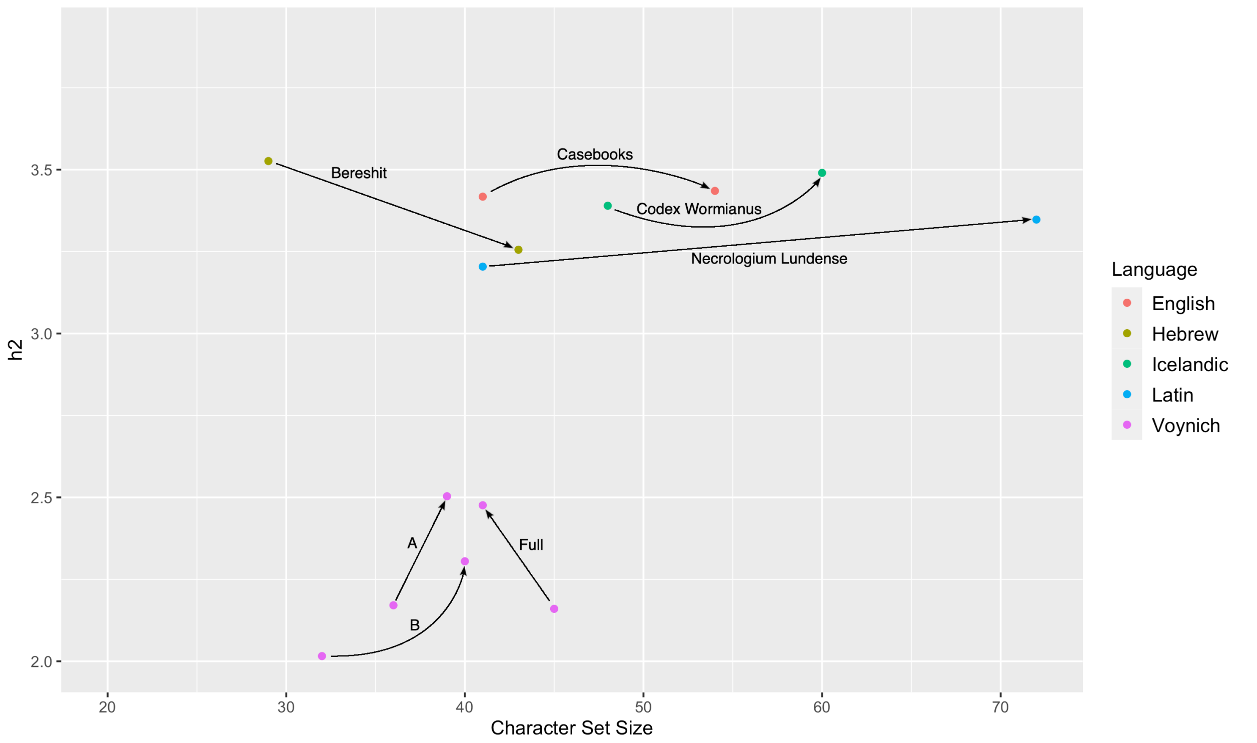


Figure 14: Differences between Parallel Comparison Texts: abbreviated Latin has higher conditional character entropy than unabbreviated Latin, and the diplomatic transcriptions of Icelandic and medical English have higher conditional character entropy than the normalized transcriptions. The Minimal transcription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

The biggest difference in *h2* is from the Hebrew Bereshit text, where including *niqqud* lowers the values of *h2* 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.[32]

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* 93% of the time.[33] In other words, the conditional frequency of the bigram *qu* is 93%. So if we see a *q* in an English text, we can be reasonably certain that we know what the next letter is.

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 16% 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.

[32] 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. 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.

[33] Some of the exceptions include the words *FAQs*, *Iraqi*, and *qi*.

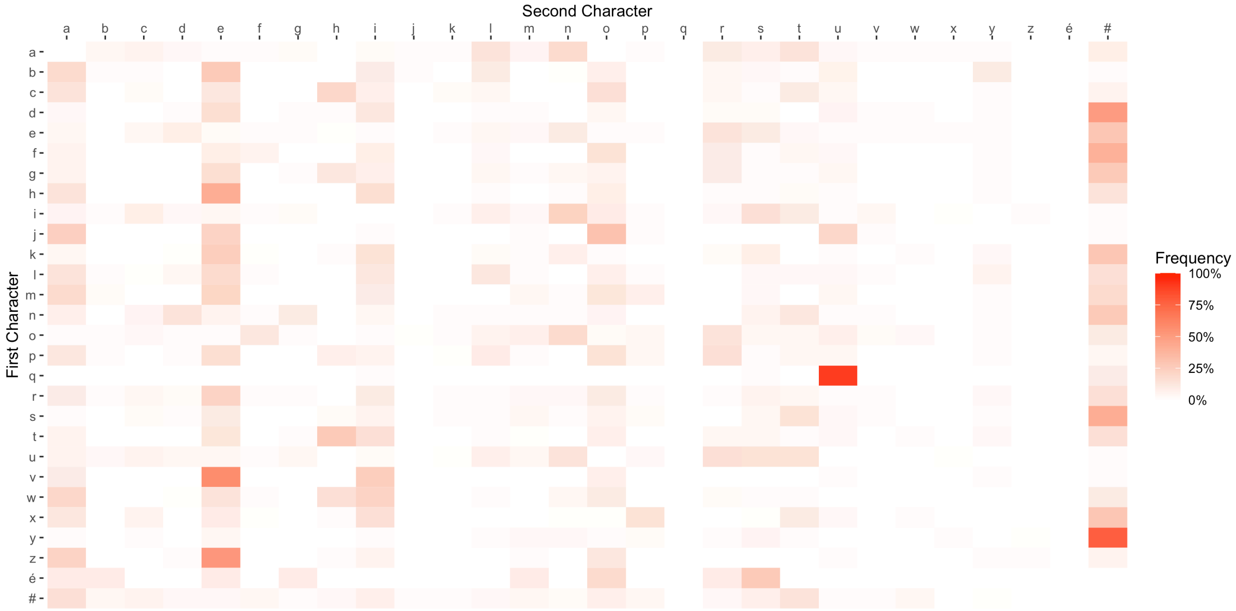


Figure 15: Conditional Frequency of English Bigrams

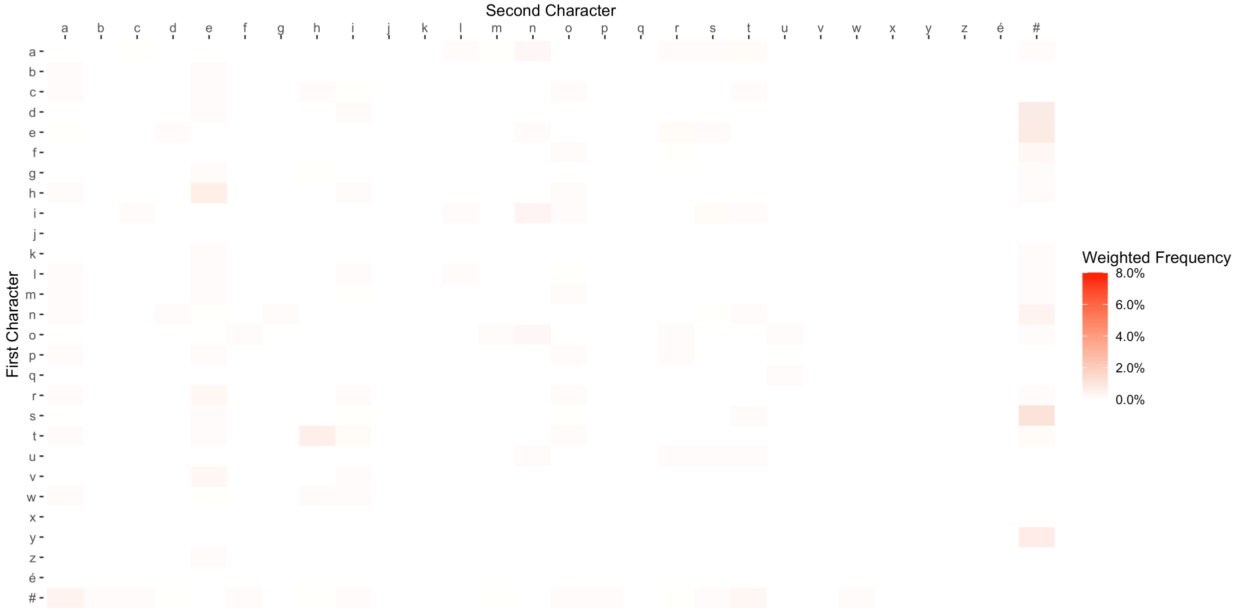


Figure 16: Conditional Frequency of English Bigrams Weighted by Overall Bigram Frequency.

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

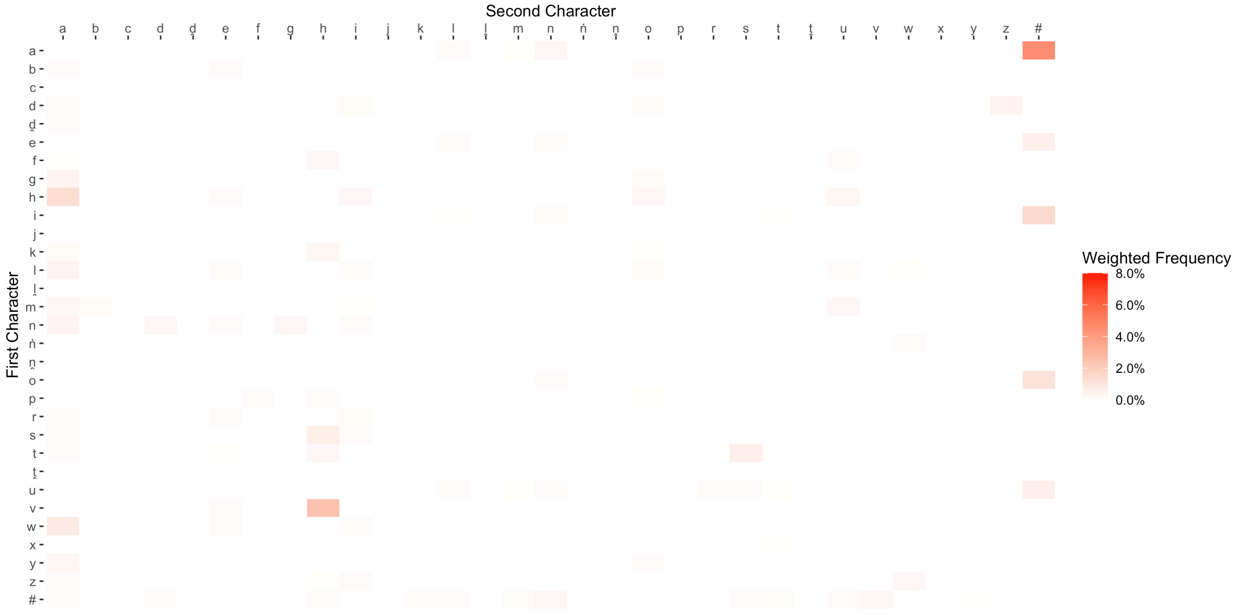


Figure 17: Conditional Frequency of Venda Bigrams Weighted by Overall Bigram Frequency

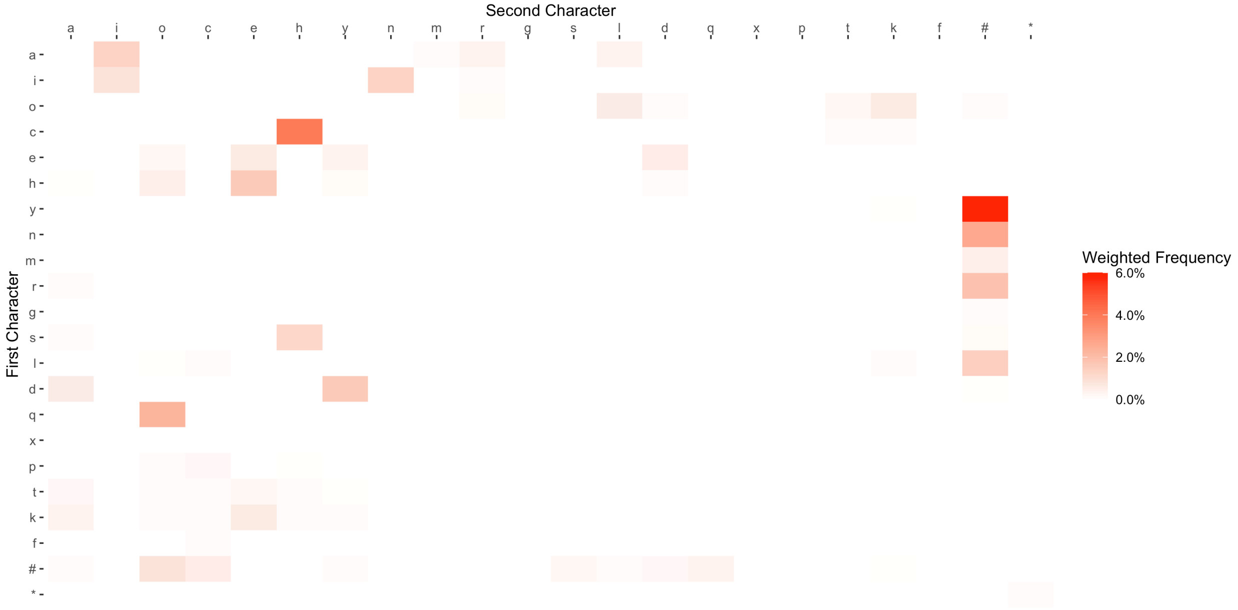


Figure 18: Conditional Frequency of (Simple Maximal) Voynich Bigrams Weighted by Overall Bigram Frequency

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, 95% of all words in Venda end with a vowel, and nearly half (46%) 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, 94% of words in the Hawaiian sample end with a vowel, and in Min Dong 84% 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 a tonal. Words in tonal languages that differ only by tone can be distinct from one another, 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 *qo* 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, 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 93% conditional probability, *y#* has 74%, *ve* has 56% conditional probability, and *d#* has 52% conditional probability. Because most of these bigrams are relatively infrequent, they make up only 3.1% of the text as a whole. By contrast, Voynich in the Simple Maximum transcription contains 11 bigrams with high conditional probability, and they are much

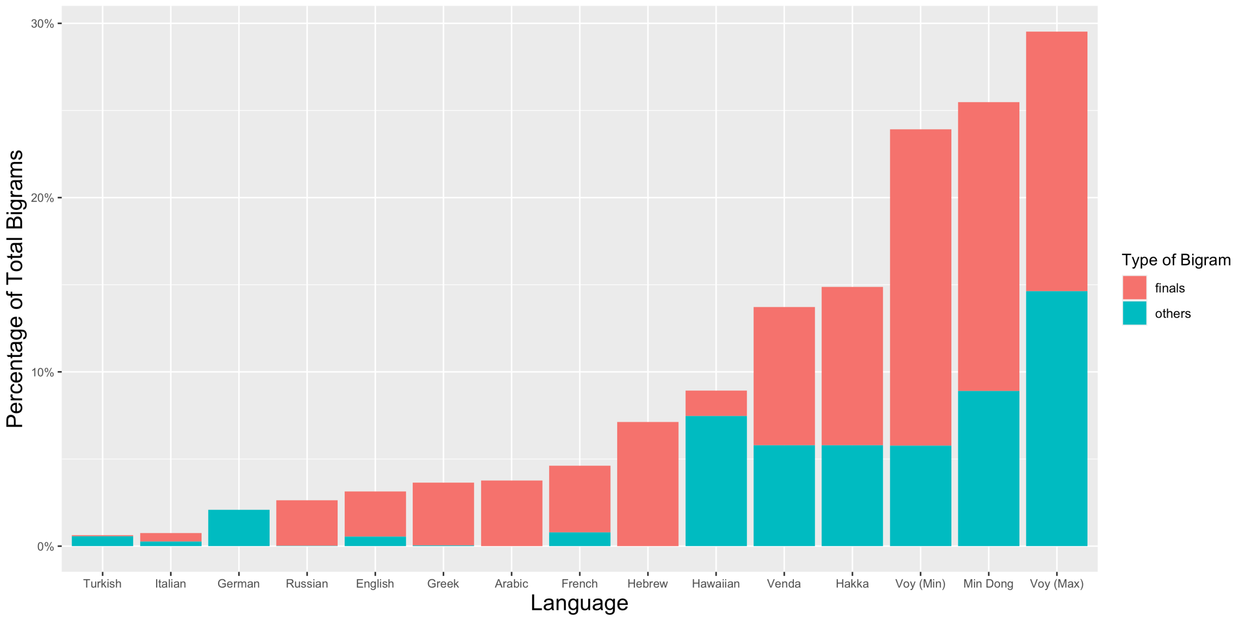


Figure 19: Prevalence of Bigrams with High Conditional Probability (>50%) for Various Languages. Red indicates the percentage of word-final bigrams (e.g. *y#*)

more frequent, making up 29.5% 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, and Venda, the three 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 (EVA) transcription, broken down by running paragraph text and labels.

	Character	Character	Word	Word	*h2*
	Count	Set Size	Count	Set Size	
Full Voynich Text	229,905	45	37,940	8,172	2.159
Paragraphs	200,988	36	33,111	6,936	2.117
Labels	28,916	36	4,829	2,283	2.354
Voynich A	67,244	36	11,415	3,460	2.171
Paragraphs	65,125	33	11,081	3,281	2.149
Labels	2,118	24	334	289	2.449
Voynich B	142,885	32	23,226	4,947	2.015
Paragraphs	133,442	30	21,632	4,661	2.005
Labels	9,442	26	1,594	778	2.099
Hand 1	63,455	42	10,877	3,260	2.171
Paragraphs	60,691	33	10,352	3,032	2.131
Labels	2,763	30	525	365	2.596
Hand 2	66,381	26	11,070	2,590	1.967
Paragraphs	60,296	26	10,054	2,367	1.955
Labels	60,84	20	1,016	531	2.023
Hand 3	73,939	30	11,755	3,419	2.036
Paragraphs	71,361	30	11,328	3,302	2.027
Labels	2,577	21	427	294	2.140
Hand 4	17,603	24	2,864	1,548	2.320
Paragraphs	2,158	20	353	268	2.150
Labels	15,444	24	2,511	1,399	2.320
Hand 5	5,639	26	930	563	2.165
Paragraphs	3,594	21	580	387	2.123
Labels	2,044	26	350	255	2.122

The following table compares the character set size and conditional character entropy (*h2*) for each of the Voynich texts in each of the three transcription systems: Maximal (EVA), Maximal Simplified, and Minimal.

	Maximal (EVA)	**Maximal Simplified**	**Minimal**
Full Voynich	45 / 2.159	22 / 2.157	41 / 2.475
Language A	36 / 2.171	21 / 2.167	39 / 2.445
Language B	32 / 2.015	22 / 2.014	40 / 2.304
Hand 1	42 / 2.170	22 / 2.165	40 / 2.506
Hand 2	26 / 1.967	22 / 1.967	39 / 2.219
Hand 3	30 / 2.036	22 / 2.036	39 / 2.338
Hand 4	24 / 2.320	21 / 2.319	36 / 2.558
Hand 5	26 / 2.165	22 / 2.167	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**	***h2***
Abkhazian	ab	Caucasian	Cyrillic	357,933	50	47,109	17,744	3.841
Acehnese	ace	Malayo-Polynesian	Latin	615,968	36	99,016	14,150	3.334
Adyghe	ady	Caucasian	Cyrillic	209,352	39	28,294	10,043	3.527
Afrikaans	af	Germanic	Latin	2,797,272	32	437,036	41,438	3.451
Akan	ak	Niger-Congo	Latin	112,256	31	22,277	4,626	3.428
Albanian	sq	Albanian	Latin	2,695,315	30	439,910	55,551	3.454
Alemannic	als	Germanic	Latin	2,193,017	38	349,723	65,263	3.604
Amharic	am	Afro-Asiatic	Ethiopic	523,265	285	111,388	31,280	4.509
Anglo Saxon	ang	Germanic	Latin	851,117	44	132,845	32,440	3.751
Arabic	ar	Afro-Asiatic	Arabic	6,726,649	38	1,173,265	110,553	3.706
Aragonese	an	Romance	Latin	1,588,470	34	262,896	29,252	3.398
Aramaic	arc	Afro-Asiatic	Syriac	500,856	47	79,259	14,018	3.517
Armenian	hy	Armenian	Armenian	6,740,053	40	858,736	96,478	3.414
Aromanian	roa_rup	Romance	Latin	885,806	42	132,053	46,509	3.668
Assamese	as	Indic	Bengali	2,065,918	68	307,160	49,631	3.817
Asturian	ast	Romance	Latin	1,795,258	36	289,119	35,805	3.411
Atikamekw	atj	Algonquian	Latin	332,978	29	45,951	9,288	2.966
Avar	av	Caucasian	Cyrillic	948,289	37	120,967	33,304	3.487
Aymara	ay	Aymara	Latin	647,865	38	100,137	18,898	3.505
Azerbaijani	az	Turkic	Latin	3,849,157	34	513,172	67,464	3.689
Bambara	bm	Mande	Latin	221,729	39	43,959	8,273	3.323
Banjar	bjn	Austronesian	Latin	759,796	33	110,217	27,830	3.412
Banyumasan	map_bms	Malayo-Polynesian	Latin	1,029,369	28	148,876	22,542	3.316
Bashkir	ba	Turkic	Cyrillic	1,871,550	43	256,982	47,889	3.778
Basque	eu	Vasconic	Latin	1,831,871	34	242,647	37,628	3.390

Language	**Wikicode**	**Family**	**Script**	**Character**	**Character**	**Word**	**Word**	*h2*
				Count	**Set Size**	**Count**	**Set Size**	
Bavarian	bar	Germanic	Latin	1,738,331	35	276,978	58,108	3.634
Belarusian	be	Slavic	Cyrillic	1,252,216	34	168,920	32,840	3.588
Belarusian (Taraškievica)	be_x_old	Slavic	Cyrillic	2,019,448	34	273,172	49,479	3.590
Bengali	bn	Indic	Bengali	4,415,799	74	603,086	77,530	3.769
Bihari	bh	Indic	Devanagari	1,050,483	109	189,042	22,326	3.742
Bishnupriya Manipuri	bpy	Indic	Bengali	666,464	64	86,631	4,179	2.998
Bislama	bi	Creole	Latin	93,438	37	16,804	3,625	3.435
Bosnian	bs	Slavic	Latin	3,143,762	32	469,831	69,807	3.514
Breton	br	Celtic	Latin	1,208,324	34	225,449	26,570	3.410
Buginese	bug	Malayo-Polynesian	Latin	405,038	35	58,945	22,591	3.578
Bulgarian	bg	Slavic	Cyrillic	2,720,484	31	419,128	38,615	3.489
Burmese	my	Tibeto-Burman	Myanmar	4,189,301	109	321,177	158,297	3.408
Buryat	bxr	Mongolic	Cyrillic	1,307,192	37	182,529	41,117	3.610
Cantonese	zh_yue	Tibeto-Burman	Chinese	920,409	5,204	125,737	87,423	5.825
Catalan	ca	Romance	Latin	5,813,470	37	1,017,022	65,428	3.404
Cebuano	ceb	Malayo-Polynesian	Latin	405,486	28	64,820	8,555	3.194
Central Bicolano	bcl	Malayo-Polynesian	Latin	787,663	34	126,432	21,075	3.303
Chamorro	ch	Malayo-Polynesian	Latin	82,626	35	14,630	3,302	3.442
Chechen	ce	Caucasian	Cyrillic	1,210,953	35	178,648	34,371	3.527
Cherokee	chr	Iroquoian	Cherokee	119,988	86	24,864	7,573	4.138
Cheyenne	chy	Algonquian	Latin	32,252	42	5,338	2,275	3.579
Chichewa	ny	Niger-Congo	Latin	594,079	28	85,216	15,980	3.240
Chinese	zh	Tibeto-Burman	Chinese	3,360,334	7,543	419,350	292,259	6.252
Chuvash	cv	Turkic	Cyrillic	580,445	42	90,034	19,715	3.701
Classical Chinese	zh_classical	Tibeto-Burman	Chinese	435,761	5,324	76,688	63,910	6.069
Cornish	kw	Celtic	Latin	480,704	33	84,259	13,076	3.483
Corsican	co	Romance	Latin	554,905	33	99,028	16,657	3.276
Cree	cr	Algonquian	Latin	21,289	46	3,001	1,620	3.755
Crimean Tatar	crh	Turkic	Latin	767,603	36	105,364	25,704	3.688
Croatian	hr	Slavic	Latin	2,220,315	33	322,131	57,765	3.547
Czech	cs	Slavic	Latin	3,238,277	41	477,234	73,833	3.784
Danish	da	Germanic	Latin	3,078,289	33	470,640	61,138	3.560
Dinka	din	Nilotic	Latin	358,610	34	74,081	12,305	3.476
Divehi	dv	Indic	Thaana	2,367,003	51	256,168	50,848	3.178
Doteli	dty	Indic	Devanagari	1,429,359	78	211,783	44,452	3.874
Dutch	nl	Germanic	Latin	5,837,024	32	898,128	79,343	3.411
Dutch Low Saxon	nds_nl	Germanic	Latin	1,158,671	35	186,629	31,809	3.440
Dzongkha	dz	Tibeto-Burman	Tibetan	279,374	93	71,786	3,746	3.221
Egyptian Arabic	arz	Afro-Asiatic	Arabic	1,314,446	41	237,389	42,398	3.675
Emilian-Romagnol	eml	Romance	Latin	1,047,147	54	212,759	30,288	3.571

Language	**Wikicode**	**Family**	**Script**	**Character**	**Character**	**Word**	**Word**	*h2*
				Count	**Set Size**	**Count**	**Set Size**	
English	en	Germanic	Latin	11,468,763	28	1,881,447	82,745	3.448
Erzya	myv	Uralic	Cyrillic	870,363	33	121,004	30,529	3.519
Esperanto	eo	Constructed	Latin	2,914,899	39	465,324	65,700	3.411
Estonian	et	Uralic	Latin	3,721,966	34	487,713	99,545	3.572
Ewe	ee	Niger-Congo	Latin	126,332	41	23,546	5,137	3.356
Extremaduran	ext	Romance	Latin	1,257,010	37	210,516	33,578	3.410
Faroese	fo	Germanic	Latin	1,093,895	36	167,401	29,402	3.606
Fiji Hindi	hif	Indic	Latin	506,617	27	89,631	8,539	3.429
Fijian	fj	Malayo-Polynesian	Latin	54,525	32	10,407	2,052	3.139
Finnish	fi	Uralic	Latin	5,688,095	31	654,746	136,198	3.516
Franco-Provençal	frp	Romance	Latin	460,017	38	76,538	15,203	3.536
French	fr	Romance	Latin	10,038,500	39	1,682,160	86,596	3.393
Friulian	fur	Romance	Latin	608,989	39	106,957	16,833	3.384
Fula	ff	Niger-Congo	Latin	201,549	33	33,667	9,054	3.442
Gagauz	gag	Turkic	Latin	858,522	46	121,746	28,436	3.767
Galician	gl	Romance	Latin	4,280,927	34	685,505	54,915	3.369
Gan	gan	Tibeto-Burman	Chinese	365,679	6,163	64,364	41,872	5.495
Georgian	ka	Kartvelian	Georgian	2,190,953	72	262,878	55,644	3.585
German	de	Germanic	Latin	10,668,182	32	1,461,075	160,396	3.485
Gilaki	glk	Iranian	Arabic	477,352	43	93,379	17,314	3.752
Goan Konkani	gom	Indic	Devanagari	4,247,253	86	622,853	116,391	3.703
Gorontalo	gor	Malayo-Polynesian	Latin	767,622	28	110,937	19,173	3.4
Gothic	got	Germanic	Gothic	315,083	32	40,358	11,109	3.396
Greek	el	Hellenic	Greek	4,735,773	35	669,498	64,534	3.576
Greenlandic	kl	Inuit	Latin	410,032	37	43,396	16,584	3.411
Guarani	gn	Tupian	Latin	2,050,689	46	321,172	39,838	3.444
Gujarati	gu	Indic	Gujarati	3,317,693	71	526,046	82,096	3.815
Haitian	ht	Creole	Latin	475,296	34	85,827	12,262	3.389
Hakka	hak	Tibeto-Burman	Latin	315,853	46	76,235	3,385	2.850
Hausa	ha	Afro-Asiatic	Latin	798,797	31	142,386	17,455	3.174
Hawaiian	haw	Malayo-Polynesian	Latin	721,609	38	137,288	11,408	3.012
Hebrew	he	Afro-Asiatic	Hebrew	5,544,646	28	970,710	100,103	3.686
Hill Mari	mrj	Uralic	Cyrillic	832,549	37	119,850	28,884	3.763
Hindi	hi	Indic	Devanagari	4,699,542	85	853,127	68,494	3.724
Hungarian	hu	Uralic	Latin	5,267,175	36	725,339	119,290	3.815
Icelandic	is	Germanic	Latin	2,065,885	37	298,641	49,323	3.726
Ido	io	Constructed	Latin	1,746,660	35	282,584	30,402	3.417
Igbo	ig	Niger-Congo	Latin	289,010	46	55,972	8,385	3.549
Ilokano	ilo	Malayo-Polynesian	Latin	860,601	28	141,462	20,954	3.378
Indonesian	id	Malayo-Polynesian	Latin	5,277,656	28	745,241	50,804	3.299

Language	**Wikicode**	**Family**	**Script**	**Character Count**	**Character Set Size**	**Word Count**	**Word Set Size**	***h2***
Ingush	inh	Caucasian	Cyrillic	365,498	37	55,207	15,818	3.685
Interlingua	ia	Constructed	Latin	972,274	30	156,253	19,893	3.296
Interlingue	ie	Constructed	Latin	492,025	38	81,461	15,440	3.495
Inuktitut	iu	Inuit	Inuktitut	13,898	123	2,108	1,373	3.853
Inupiak	ik	Inuit	Latin	26,593	39	3,753	2,102	3.623
Irish	ga	Celtic	Latin	1,630,276	32	280,774	29,495	3.4
Italian	it	Romance	Latin	5,978,519	34	944,446	71,366	3.325
Jamaican Patois	jam	Creole	Latin	530,035	27	94,546	14,103	3.417
Japanese	ja	Japonic	Japanese	3,171,925	3,915	388,362	203,650	5.306
Javanese	jv	Malayo-Polynesian	Latin	1,044,195	32	153,282	26,542	3.325
Kabardian Circassian	kbd	Caucasian	Cyrillic	883,684	34	116,025	33,991	3.399
Kabiye	kbp	Niger-Congo	Latin	772,117	40	144,624	18,177	3.360
Kabyle	kab	Afro-Asiatic	Latin	1,145,063	43	207,438	33,745	3.570
Kalmyk	xal	Mongolic	Cyrillic	341,107	40	53,160	18,227	3.919
Kannada	kn	Dravidian	Kannada	4,528,033	70	543,980	130,703	3.833
Kapampangan	pam	Malayo-Polynesian	Latin	999,691	29	157,002	23,045	3.205
Karachay-Balkar	krc	Turkic	Cyrillic	1,314,614	33	170,483	33,970	3.525
Karakalpak	kaa	Turkic	Latin	1,059,888	31	168,812	31,276	3.480
Kashmiri	ks	Indic	Arabic	19,794	62	4,998	1,833	3.771
Kashubian	csb	Slavic	Latin	520,258	41	74,950	19,624	3.720
Kazakh	kk	Turkic	Cyrillic	3,975,954	40	530,836	74,330	3.516
Khmer	km	Austroasiatic	Khmer	1,816,386	85	96,956	57,964	4.169
Kikuyu	ki	Niger-Congo	Latin	115,272	35	18,014	6,052	3.383
Kinyarwanda	rw	Niger-Congo	Latin	639,458	44	99,497	22,482	3.440
Kirghiz	ky	Turkic	Cyrillic	2,721,106	37	362,583	56,958	3.653
Kirundi	rn	Niger-Congo	Latin	249,286	36	38,728	9,118	3.107
Komi	kv	Uralic	Cyrillic	766,908	37	115,163	23,876	3.750
Komi-Permyak	koi	Uralic	Cyrillic	812,340	36	118,919	22,495	3.766
Kongo	kg	Niger-Congo	Latin	98,925	46	16,820	4,551	3.222
Korean	ko	Koreanic	Hangul	1,901,932	1,580	483,787	121,286	4.860
Kurdish	ku	Iranian	Latin	1,211,195	36	212,059	33,845	3.483
Ladino	lad	Romance	Latin	1,026,930	35	169,905	27,377	3.365
Lak	lbe	Caucasian	Cyrillic	145,125	35	19,692	7,705	3.569
Lao	lo	Tai	Lao	1,705,222	55	103,164	54,498	4.090
Latgalian	ltg	Baltic	Latin	625,295	37	85,764	23,121	3.653
Latin	la	Romance	Latin	1,972,637	32	270,487	58,497	3.503
Latvian	lv	Baltic	Latin	3,117,365	37	423,546	63,378	3.660
Lezgian	lez	Caucasian	Cyrillic	1,341,518	35	181,758	34,927	3.510
Ligurian	lij	Romance	Latin	550,591	48	101,330	22,329	3.539
Limburgish	li	Germanic	Latin	1,985,566	36	322,800	44,635	3.517

Language	**Wikicode**	**Family**	**Script**	**Character Count**	**Character Set Size**	**Word Count**	**Word Set Size**	***h2***
Lingala	ln	Niger-Congo	Latin	575,428	46	98,558	17,861	3.471
Lingua Franca Nova	lfn	Constructed	Latin	1,173,626	30	219,852	24,557	3.232
Lithuanian	lt	Baltic	Latin	3,230,055	35	426,482	73,475	3.584
Livvi-Karelian	olo	Uralic	Latin	565,364	34	73,593	22,907	3.617
Lojban	jbo	Constructed	Latin	896,007	28	215,747	14,426	3.022
Lombard	lmo	Romance	Latin	728,275	41	140,038	31,197	3.581
Low Saxon	nds	Germanic	Latin	1,552,696	31	249,094	37,198	3.464
Lower Sorbian	dsb	Slavic	Latin	782,094	48	114,672	29,540	3.772
Luganda	lg	Niger-Congo	Latin	1,564,739	27	223,643	35,508	3.214
Luxembourgisch	lb	Germanic	Latin	759,413	35	116,671	22,570	3.484
Macedonian	mk	Slavic	Cyrillic	6,099,603	31	936,128	78,754	3.410
Maithili	mai	Indic	Devanagari	763,408	71	114,509	24,456	3.873
Malagasy	mg	Malayo-Polynesian	Latin	1,283,206	38	192,913	32,230	3.268
Malay	ms	Malayo-Polynesian	Latin	2,631,843	27	377,157	30,784	3.286
Malayalam	ml	Dravidian	Malayalam	4,684,305	82	460,277	142,026	3.799
Maltese	mt	Afro-Asiatic	Latin	2,113,008	34	351,823	36,292	3.592
Manx	gv	Celtic	Latin	891,933	36	149,874	18,678	3.450
Maori	mi	Malayo-Polynesian	Latin	503,904	32	98,030	9,870	3.173
Marathi	mr	Indic	Devanagari	3,277,623	86	446,153	81,575	3.836
Mazandarani	mzn	Iranian	Arabic	660,593	44	136,473	15,278	3.566
Meadow Mari	mhr	Uralic	Cyrillic	858,781	37	120,131	15,090	3.479
Min Dong	cdo	Tibeto-Burman	Latin	565,516	52	140,080	3,762	2.9
Min Nan	zh_min_nan	Tibeto-Burman	Latin	786,929	51	197,226	8,416	3.132
Minangkabau	min	Malayo-Polynesian	Latin	804,495	28	118,452	17,397	3.280
Mingrelian	xmf	Kartvelian	Georgian	950,448	42	122,736	32,603	3.587
Mirandese	mwl	Romance	Latin	5,466,247	33	913,310	61,920	3.339
Moksha	mdf	Moksha	Cyrillic	493,259	34	64,800	17,500	3.476
Mongolian	mn	Mongolic	Cyrillic	2,278,033	35	333,393	41,507	3.576
Nahuatl	nah	Uto-Aztecan	Latin	1,122,934	37	149,448	34,405	3.513
Nauruan	na	Malayo-Polynesian	Latin	121,321	54	19,150	7,118	3.780
Navajo	nv	Athabaskan	Latin	482,150	37	72,601	10,752	3.409
Neapolitan	nap	Romance	Latin	815,987	41	143,266	23,372	3.332
Nepali	ne	Indic	Devanagari	1,940,280	83	282,248	51,693	3.812
Newar	new	Tibeto-Burman	Devanagari	1,016,021	93	145,909	34,502	3.810
Norfolk	pih	Creole	Latin	84,456	31	14,722	4,347	3.613
Norman	nrm	Romance	Latin	566,223	41	102,057	15,301	3.433
North Frisian	frr	Germanic	Latin	687,133	37	111,620	22,152	3.628
Northern Luri	lrc	Iranian	Arabic	697,527	48	137,583	26,788	3.824
Northern Sami	se	Uralic	Latin	476,737	44	59,754	20,366	3.706
Northern Sotho	nso	Niger-Congo	Latin	322,281	30	54,844	9,734	3.284

Language	**Wikicode**	**Family**	**Script**	**Character Count**	**Character Set Size**	**Word Count**	**Word Set Size**	***h2***
Norwegian (Bokmål)	no	Germanic	Latin	5,269,784	31	816,781	82,476	3.522
Norwegian (Nynorsk)	nn	Germanic	Latin	1,099,070	37	165,788	27,543	3.588
Novial	nov	Constructed	Latin	616,905	27	99,204	8,091	3.272
Occitan	oc	Romance	Latin	1,457,361	38	236,727	28,236	3.465
Old Church Slavonic	cu	Slavic	Cyrillic	213,007	47	32,824	7,802	3.449
Oriya	or	Indic	Odia	649,200	66	88,813	17,196	3.681
Oromo	om	Afro-Asiatic	Latin	1,470,805	27	206,815	39,546	3.298
Ossetian	os	Iranian	Cyrillic	672,527	34	106,260	21,657	3.676
Palatinate German	pfl	Germanic	Latin	1,194,952	39	182,273	41,183	3.501
Pali	pi	Indic	Latin	58,673	44	6,206	2,225	2.950
Pangasinan	pag	Malayo-Polynesian	Latin	522,565	28	84,721	18,104	3.299
Papiamentu	pap	Creole	Latin	1,239,993	37	234,470	23,078	3.376
Pashto	ps	Iranian	Arabic	2,152,502	52	466,643	48,180	3.688
Pennsylvania German	pdc	Germanic	Latin	536,973	32	84,493	15,500	3.526
Persian	fa	Iranian	Arabic	4,044,947	37	824,252	37,447	3.618
Picard	pcd	Romance	Latin	987,378	41	166,221	30,636	3.551
Piedmontese	pms	Romance	Latin	381,855	35	71,197	7,706	3.341
Polish	pl	Slavic	Latin	3,803,094	35	524,423	88,992	3.664
Pontic	pnt	Hellenic	Greek	309,509	35	49,571	10,693	3.540
Portuguese	pt	Romance	Latin	9,201,460	39	1,486,578	78,037	3.358
Punjabi	pa	Indic	Gurmukhi	1,942,611	73	365,155	36,770	3.676
Quechua	qu	Quechua	Latin	789,125	37	110,389	27,236	3.5
Ripuarian	ksh	Germanic	Latin	665,795	37	111,183	26,111	3.589
Romani	rmy	Indic	Latin	327,885	50	51,074	13,441	3.576
Romanian	ro	Romance	Latin	6,254,820	34	968,215	85,544	3.511
Romansh	rm	Romance	Latin	2,319,022	35	396,148	33,637	3.359
Russian	ru	Slavic	Cyrillic	8,586,678	34	1,160,076	139,132	3.668
Rusyn	rue	Slavic	Cyrillic	405,846	38	58,884	19,697	3.740
Sakha	sah	Turkic	Cyrillic	1,131,761	40	147,492	34,576	3.645
Samoan	sm	Malayo-Polynesian	Latin	550,217	32	114,252	8,659	3.090
Samogitian	bat_smg	Baltic	Latin	562,800	37	78,069	24,399	3.645
Sango	sg	Niger-Congo	Latin	41,494	44	8,227	1,905	3.426
Sanskrit	sa	Indic	Devanagari	2,975,900	83	328,303	99,635	3.776
Santali	sat	Austroasiatic	Ol Chiki	1,248,005	37	220,338	26,027	3.450
Sardinian	sc	Romance	Latin	1,056,680	35	183,292	33,421	3.316
Saterland Frisian	stq	Germanic	Latin	1,121,438	32	187,942	30,541	3.523
Scots	sco	Germanic	Latin	1,158,338	27	194,690	24,219	3.506
Scottish Gaelic	gd	Celtic	Latin	602,579	38	98,551	17,046	3.373
Serbian	sr	Slavic	Cyrillic	4,686,143	31	701,133	86,366	3.519
Serbo-Croatian	sh	Slavic	Latin	4,695,790	31	710,129	105,774	3.556

Language	**Wikicode**	**Family**	**Script**	**Character Count**	**Character Set Size**	**Word Count**	**Word Set Size**	***h2***
Sesotho	st	Niger-Congo	Latin	159,438	32	28,054	6,185	3.285
Shan	shn	Tai	Myanmar	2,647,884	116	168,495	95,352	3.180
Shona	sn	Niger-Congo	Latin	618,453	27	82,143	23,910	3.254
Sicilian	scn	Romance	Latin	777,746	38	129,487	22,966	3.334
Silesian	szl	Slavic	Latin	534,353	48	78,291	27,610	3.775
Simple English	simple	Germanic	Latin	1,919,962	27	332,485	21,352	3.407
Sindhi	sd	Indic	Arabic	2,018,087	62	429,726	31,874	3.707
Sinhalese	si	Indic	Sinhala	1,826,685	129	279,796	50,072	3.851
Slovak	sk	Slavic	Latin	2,233,010	43	323,153	61,471	3.786
Slovenian	sl	Slavic	Latin	2,474,592	33	364,934	57,774	3.619
Somali	so	Afro-Asiatic	Latin	1,557,643	27	241,690	38,419	3.336
Sorani	ckb	Iranian	Arabic	1,857,930	38	301,550	51,320	3.468
South Azerbaijani	azb	Turkic	Arabic	668,986	43	111,001	23,388	3.619
Spanish	es	Romance	Latin	8,009,861	33	1,294,134	81,703	3.338
Sranan	srn	Creole	Latin	364,486	31	73,118	5,166	3.104
Sundanese	su	Malayo-Polynesian	Latin	1,758,880	28	269,727	29,644	3.580
Swahili	sw	Niger-Congo	Latin	1,897,209	28	300,271	33,393	3.253
Swati	ss	Niger-Congo	Latin	295,650	37	36,144	14,493	3.311
Swedish	sv	Germanic	Latin	3,454,252	31	515,025	66,553	3.591
Tagalog	tl	Malayo-Polynesian	Latin	2,298,511	28	368,300	35,454	3.171
Tahitian	ty	Malayo-Polynesian	Latin	84,760	49	15,559	3,878	3.347
Tajik	tg	Iranian	Cyrillic	1,672,125	41	253,532	44,166	3.656
Tamil	ta	Dravidian	Tamil	4,485,864	49	475,986	118,916	3.385
Tarantino	roa_tara	Romance	Latin	2,222,268	39	386,118	56,571	3.329
Tatar	tt	Turkic	Latin	724,029	39	104,638	29,022	3.796
Telugu	te	Dravidian	Telugu	4,050,780	71	499,679	112,362	3.826
Tetum	tet	Malayo-Polynesian	Latin	614,296	38	100,065	13,348	3.505
Thai	th	Tai	Thai	4,086,477	73	195,870	113,268	4.312
Tibetan	bo	Tibeto-Burman	Tibetan	1,655,561	92	425,748	7,457	3.172
Tigrinya	ti	Afro-Asiatic	Ethiopic	91,327	299	20,262	7,231	4.423
Tok Pisin	tpi	Creole	Latin	222,269	33	38,863	6,676	3.374
Tongan	to	Malayo-Polynesian	Latin	409,219	37	79,799	9,249	3.358
Tsonga	ts	Niger-Congo	Latin	591,967	27	99,748	13,664	3.057
Tswana	tn	Niger-Congo	Latin	1,365,773	28	253,949	21,931	2.927
Tulu	tcy	Dravidian	Kannada	1,885,690	71	249,534	69,766	3.763
Tumbuka	tum	Niger-Congo	Latin	111,104	30	18,114	4,634	3.449
Turkish	tr	Turkic	Latin	4,499,751	36	621,683	86,321	3.646
Turkmen	tk	Turkic	Latin	1,257,976	39	170,555	39,106	3.645
Tuvan	tyv	Turkic	Cyrillic	1,250,039	38	179,530	31,400	3.529
Twi	tw	Niger-Congo	Latin	134,699	36	25,603	6,224	3.477

Language	**Wikicode**	**Family**	**Script**	**Character**	**Character**	**Word**	**Word**	***h2***
				Count	**Set Size**	**Count**	**Set Size**	
Udmurt	udm	Uralic	Cyrillic	598,527	39	84,075	17,671	3.689
Ukrainian	uk	Slavic	Cyrillic	3,776,955	34	506,113	79,490	3.690
Upper Sorbian	hsb	Slavic	Latin	865,119	47	125,004	30,040	3.710
Urdu	ur	Indic	Arabic	2,570,761	47	553,505	30,420	3.621
Uyghur	ug	Turkic	Arabic	1,405,930	36	183,913	29,279	3.475
Uzbek	uz	Turkic	Latin	4,571,576	27	594,169	77,020	3.468
Võro	fiu_vro	Uralic	Latin	722,187	35	102,235	24,671	3.679
Venda	ve	Niger-Congo	Latin	130,928	31	22,724	4,474	2.880
Venetian	vec	Romance	Latin	622,277	39	114,784	20,886	3.314
Vepsian	vep	Uralic	Latin	889,501	34	116,516	21,667	3.560
Vietnamese	vi	Austroasiatic	Latin	8,140,355	92	1,797,350	31,255	3.242
Volapük	vo	Constructed	Latin	661,589	32	100,332	18,416	3.617
Walloon	wa	Romance	Latin	706,736	37	133,188	23,738	3.507
Waray-Waray	war	Malayo-Polynesian	Latin	595,631	28	95,382	11,175	3.201
Welsh	cy	Celtic	Latin	1,075,706	33	170,854	24,785	3.6
West Flemish	vls	Germanic	Latin	1,059,540	32	175,526	31,286	3.509
West Frisian	fy	Germanic	Latin	1,292,482	34	199,809	26,634	3.482
Western Armenian	hyw	Armenian	Armenian	563,204	40	81,645	18,882	3.514
Western Punjabi	pnb	Indic	Arabic	1,025,781	44	217,707	20,569	3.566
Wolof	wo	Niger-Congo	Latin	2,404,439	35	495,852	23,023	3.339
Wu	wuu	Tibeto-Burman	Chinese	380,052	5,118	56,866	39,886	5.480
Xhosa	xh	Niger-Congo	Latin	1,109,675	27	141,382	44,652	3.387
Yiddish	yi	Germanic	Hebrew	1,651,194	31	302,081	25,956	3.366
Yoruba	yo	Niger-Congo	Latin	1,496,023	43	328,852	34,576	3.719
Zamboanga Chavacano	cbk_zam	Creole	Latin	1,844,820	35	306,250	36,404	3.395
Zazaki	diq	Iranian	Latin	663,577	41	111,379	26,560	3.627
Zeelandic	zea	Germanic	Latin	916,386	37	152,466	26,800	3.470
Zhuang	za	Tai	Latin	246,715	35	37,719	10,857	3.715
Zulu	zu	Niger-Congo	Latin	886,083	30	112,619	35,454	3.427

C Historical Corpus Statistics

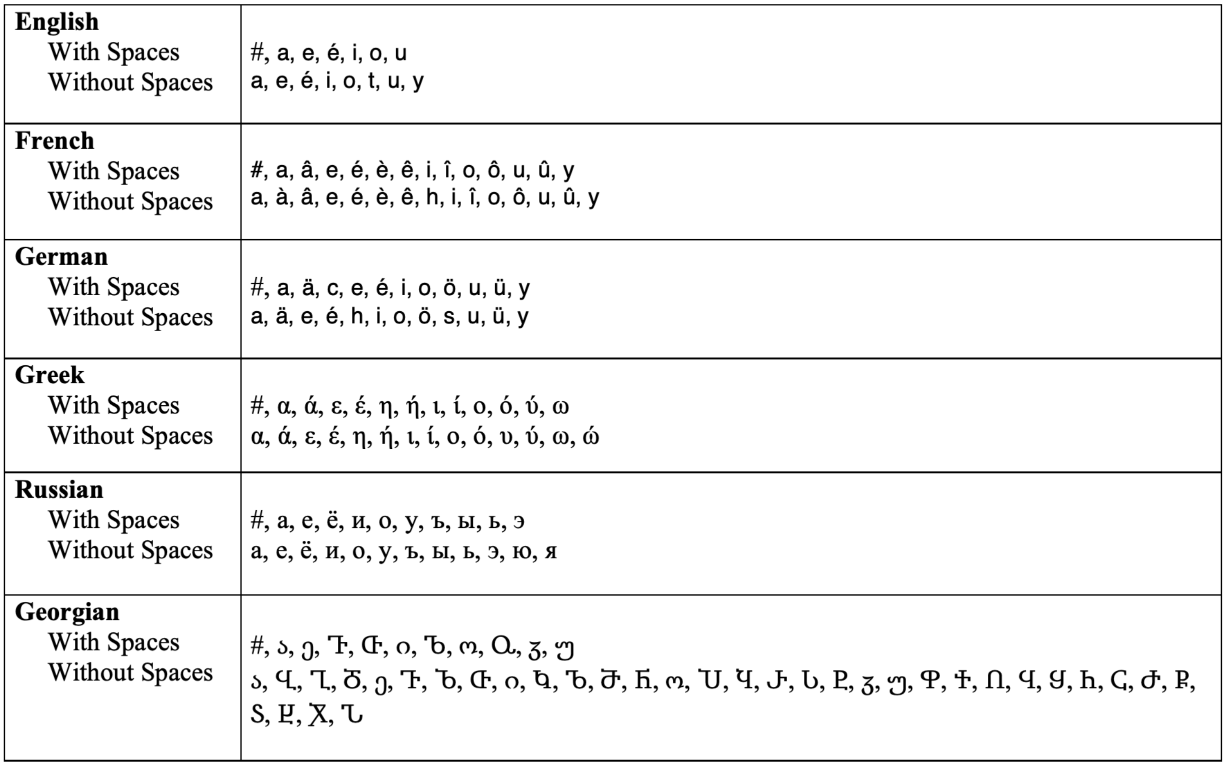
The following table gives basic statistics for each of the Historical Texts, along with separate statistics for parallel texts with normalized and diplomatic versions.

Text	Language	Script	Character Count	Character Set Size	Word Count	Word Set Size	*h2*
Medical Casebooks	English	Latin					
Normalized			3,057	884	15,195	41	3.418
Diplomatic			3,069	901	14,646	54	3.435
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					
Without *niqqud*			17,802	6,327	99,024	29	3.526
With *niqqud*			17,802	7,091	164,569	43	3.256
Codex Wormianus	Icelandic	Latin					
Normalized			31,592	8,393	162,890	48	3.390
Diplomatic			31,442	8,637	150,374	60	3.490
La Rettorica	Italian	Latin	32,230	6,789	198,138	29	3.141
Necrologium Lundense	Latin	Latin					
Normalized			309	222	1,723	72	3.348
Diplomatic			314	194	2,046	41	3.204
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.



English	
With Spaces	#, a, e, é, i, o, u
Without Spaces	a, e, é, i, o, t, u, y
French	
With Spaces	#, a, â, e, é, è, ê, i, î, o, ô, u, û, y
Without Spaces	a, à, â, e, é, è, ê, h, i, î, o, ô, u, û, y
German	
With Spaces	#, a, ä, c, e, é, i, o, ö, u, ü, y
Without Spaces	a, ä, e, é, h, i, o, ö, s, u, ü, y
Greek	
With Spaces	#, α, ά, ε, έ, η, ή, ι, ί, ο, ό, ύ, ω
Without Spaces	α, ά, ε, έ, η, ή, ι, ί, ο, ό, υ, ύ, ω, ώ
Russian	
With Spaces	#, а, е, ё, и, о, у, ъ, ы, ь, э
Without Spaces	а, е, ё, и, о, у, ъ, ы, ь, э, ю, я
Georgian	
With Spaces	#, ა, ე, Ⴈ, Ⴄ, ი, Ⴍ, ო, Ⴀ, ვ, უ
Without Spaces	ა, Ⴁ, Ⴂ, Ⴃ, ე, Ⴈ, Ⴍ, Ⴄ, ი, Ⴉ, Ⴌ, Ⴅ, Ⴆ, ო, Ⴓ, Ⴗ, Ⴇ, Ⴑ, Ⴐ, ვ, უ, Ⴔ, Ⴕ, Ⴖ, Ⴎ, Ⴘ, Ⴙ, Ⴚ, Ⴛ, Ⴜ, Ⴝ, Ⴞ, Ⴟ, Ⴠ

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. For English and German, the results include the low-frequency character *é*, which is only used in loanwords (as is the German *y*), but this may reflect a more general tendency for

Hebrew	
With Spaces	#, ו, י, מ, ר
Without Spaces	ה, ו, י, ל, מ, ף, ר
Arabic	
With Spaces	#, ء, ؤ, ى, ئ, ت, ر, ص, ض, ط, ظ, ل, و, ي
Without Spaces	ـ, آ, أ, إ, ا, ت, ذ, ص, ط, ظ, ع, ك, م, و, ى, ي
Hindi	
With Spaces	़, #, अ, इ, उ, ऍ, ऑ, औ, ा, ि, ी, ु, ू, ृ, ॄ, ॅ, ॆ, े, ै, ॉ, ो, ौ, ्
Without Spaces	॔, ़, ं, ः, ॐ, अ, आ, इ, ई, उ, ऊ, ऋ, ॠ, ऍ, ऎ, ए, ऐ, ऑ, औ, ड़, ऩ, ऽ, ा, ि, ी, ु, ू, ृ, ॄ, ॅ, ॆ, े, ै, ॉ, ॊ, ो, ौ, ्

Figure 21: Results of the Sukhotin Algorithm for Wikipedia Samples in the Hebrew and Arabic abjads and the Devanagari abugida (Hindi).

the algorithm to identify low-frequency characters as vowels. German also includes the characters *s*, *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 fails completely if spaces are excluded.

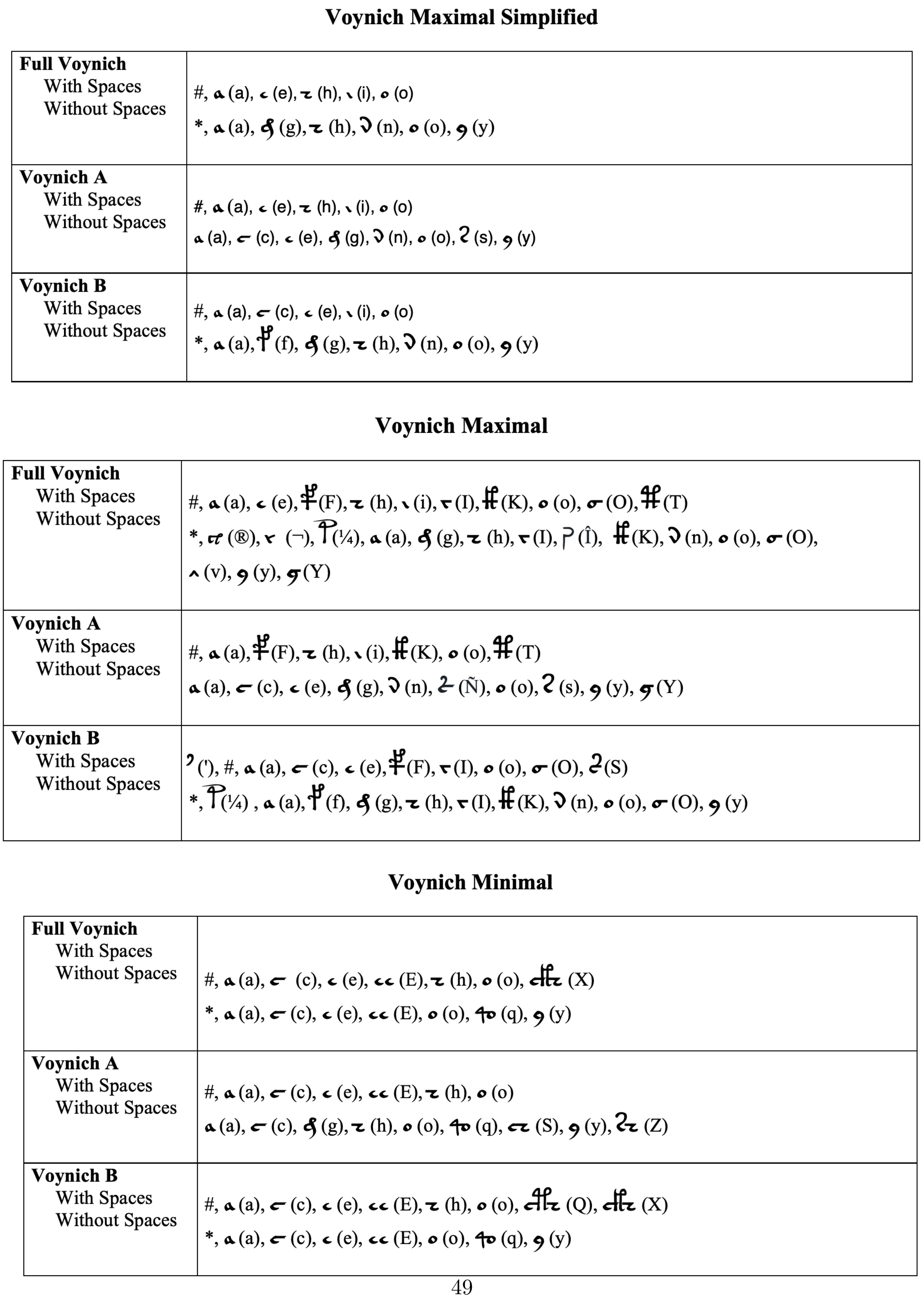
Figure 21 shows the results of the Sukhotin algorithm for Wikipedia samples of languages in three non-alphabetic scripts: Hebrew, Arabic, and Devanagari (run on the Hindi sample). Hebrew and Arabic are abjads which do not include vowels.[34] The algorithm identifies some consonants which double as vowels, but otherwise has a tendency to identify the final or initial forms of consonants, particularly if spaces are excluded. The algorithm is quite successful for Hindi, written in an abugida (alphasyllabary). It initially identifies the freestanding forms of the vowels, and then the separate combining forms of the vowels. When spaces are excluded, it also includes many of the diacritics and rare consonants.

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

[34] 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

Full Voynich With Spaces Without Spaces	#, (a), (e), (h), (i), (o) *, (a), (g), (h), (n), (o), (y)
Voynich A With Spaces Without Spaces	#, (a), (e), (h), (i), (o) (a), (c), (e), (g), (n), (o), (s), (y)
Voynich B With Spaces Without Spaces	#, (a), (c), (e), (i), (o) *, (a), (f), (g), (h), (n), (o), (y)

Voynich Maximal

Full Voynich With Spaces Without Spaces	#, (a), (e), (F), (h), (i), (I), (K), (o), (O), (T) *, (®), (¬), (¼), (a), (g), (h), (I), (Î), (K), (n), (o), (O), (v), (y), (Y)
Voynich A With Spaces Without Spaces	#, (a), (F), (h), (i), (K), (o), (T) (a), (c), (e), (g), (n), (Ñ), (o), (s), (y), (Y)
Voynich B With Spaces Without Spaces	('), #, (a), (c), (e), (F), (I), (o), (O), (S) *, (¼) , (a), (f), (g), (h), (I), (K), (n), (o), (O), (y)

Voynich Minimal

Full Voynich With Spaces Without Spaces	#, (a), (c), (e), (E), (h), (o), (X) *, (a), (c), (e), (E), (o), (q), (y)
Voynich A With Spaces Without Spaces	#, (a), (c), (e), (E), (h), (o) (a), (c), (g), (h), (o), (q), (S), (y), (Z)
Voynich B With Spaces Without Spaces	#, (a), (c), (e), (E), (h), (o), (Q), (X) *, (a), (c), (e), (E), (o), (q), (y)

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

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.

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`.

Sukhotin, B. (1962). Eksperimental'noe vydelenie klassov bukv s pomoscju evm. *Problemy strukturnoj lingvistiki*, 234:189–206.

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

Zandbergen, R. (2010). Voynich ms. `http://www.voynich.nu/index.html`. [Online; accessed 24-April-2018].