ANNUAL REVIEWS

Annual Review of Linguistics

The Linguistics of the Voynich Manuscript

Claire L. Bowern[1] and Luke Lindemann[2]

[1]Department of Linguistics, Yale University, New Haven, Connecticut 06511, USA; email: claire.bowern@yale.edu

[2]Yale Center for Medical Informatics, Yale University, New Haven, Connecticut 06511, USA

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Annu. Rev. Linguist. 2021. 7:285–308

First published as a Review in Advance on November 11, 2020

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-011619-030613

Copyright © 2021 by Annual Reviews. All rights reserved

Keywords

ciphers, unknown languages, documentation, Voynich Manuscript, morphology, orthography

Abstract

The Voynich Manuscript is a fifteenth-century illustrated cipher manuscript. In this overview of recent approaches to the Voynich Manuscript, we summarize and evaluate current work on the language that underlies this document. We provide arguments for treating the document as natural language (rather than a medieval hoax) and show how statistical arguments can be made about the phonology, morphology, and structure of the document even though the contents remain undecipherable.

1. INTRODUCTION

Manuscript 408 in Yale University's Beinecke Rare Book & Manuscript Library[1]—otherwise known as the Voynich Manuscript—is a curious five-part codex written in an unknown language and unknown script. It has captured the imagination of cryptographers and linguists and is the subject of numerous claims about its content (Kennedy & Churchill 2006). While some argue that the work is a hoax (modern or medieval) or gibberish, others have advanced claims regarding what language the text is written in or what type of cipher may have been used to encode the text. To date, there is no consensus on the language that underlies the manuscript.

In this review, we survey linguistic aspects of the Voynich Manuscript and summarize issues around decipherment. We concentrate on features of the text—"phonology," morphology, and syntax—to shed light on the question of gibberish versus natural language on the one hand and encoded known language versus otherwise unattested language on the other.

The manuscript itself is bound in vellum (the binding is old but not original). The Voynich Manuscript has 116 folios (i.e., 232 pages) bound in 18 quires of varying size. Several pages are clearly missing. Thematically, the manuscript has five parts. The longest, which covers more than half the manuscript, contains drawings of plants (e.g., folio 19r, shown in **Figure 1**). The astrology section (e.g., folio 69v) contains astral charts, a zodiac, and other diagrams that may be related to astrology but that are not currently identifiable. The so-called balneological section is almost unparalleled elsewhere: It is text with illustrations of naked women in green water.[2] The fourth section shows illustrations of medical bottles and plant roots, whereas the fifth is unillustrated and

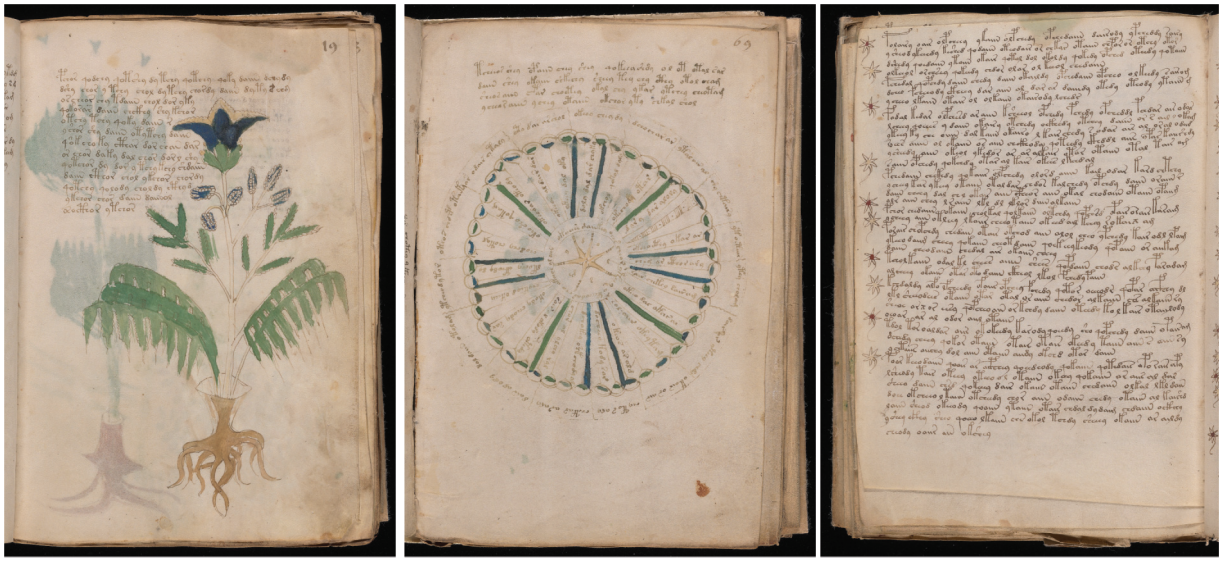


Figure 1

Folios 19r, 69v, and 105v of the manuscript, illustrating the subject themes of herbology, alchemy, and astrology. Images courtesy of the Beinecke Rare Book and Manuscript Library, Yale University.

[1] High-resolution images of all pages of the Voynich Manuscript are available at **https://brbl-dl.library.yale.edu/vufind/Record/3519597**.

[2] Some readers have seen parallels between the balneological section and illustrated medieval women's health manuals, such as the *Trotula*. It should be noted, however, that while the *Trotula*'s illustrations are readily

consists of paragraphs demarcated by stars (see folio 105v in **Figure 1**). Some pages are fold-outs. The text is written in iron-gall ink (typical of manuscripts of the time), and the illustrations are ink and pigment.

The manuscript is named after the book dealer Wilfrid Voynich, who acquired it from the Villa Mondragone (outside of Rome) in 1910 or 1911. The villa housed a former Jesuit library and had been sold in about 1874; presumably, the manuscript had been in the collection before that. The circumstances regarding the manuscript's acquisition are unclear. Voynich was very circumspect at the time, and the manuscript does not have the usual information about provenance that other manuscript purchases had—factors that add to the mystery surrounding the manuscript. After Voynich's death, the manuscript passed to his wife, the novelist Ethel Voynich. It was donated to Yale's Beinecke Library in 1969 by H.P. Kraus (for more information about these points, see Kennedy & Churchill 2006; see also **http://voynich.nu**).

The Voynich Manuscript may seem to be a curious topic for a linguistics article, given that the language of the manuscript is unknown. We argue that by applying linguistic methods—using insights from typology, language documentation, and statistical arguments related to a wide variety of languages (ancient, medieval, and modern)—we may shed light on the properties of the linguistic system underlying "Voynichese." Moreover, much of the work that attempts to decipher the manuscript claims to do so on linguistic principles, and that work should be evaluated on linguistic principles as well (see Section 5 below).

In this first section, we give some brief information about the manuscript and summarize arguments over whether the text is linguistic material—that is, whether the Voynich Manuscript is gibberish, a constructed language, or natural language. In Section 2, we discuss the manuscript's phonology (word-internal structure). In Section 3, we give an overview of the current state of knowledge with respect to morphology and word structure, and in Section 4 we discuss the syntax and discourse structure of the Voynich Manuscript. In Section 5, we discuss some of the current theories about the language underlying the manuscript.

In our investigations of the Voynich Manuscript, we do not take a traditional decipherment approach, such as those used in early cryptographic analyses (Currier 1976, D'Imperio 1978). Rather, we draw on our experience as documentary linguists to deduce aspects of linguistic structure, and we use our cross-linguistic experience as typologists to compare those structures and attempt to narrow down the possible target languages. Our aim is to shed light on the composition of the Voynich Manuscript by making two types of tests: comparing known languages and comparing enciphered tests of known languages that manipulate strings in various ways. While this approach is unlikely to yield a direct language match, it provides useful information about how sensitive tests are to both morphology and encipherment methods. This knowledge allows us to rule in or rule out various language families. Our comparisons are based on a corpus of historical materials and a cleaned sample of language data from Wikipedia. For more information on the corpus, we refer readers to Lindemann & Bowern (2020).

In this review, we concentrate on what is known about the language of the manuscript, not on its thematic content or the historical context of its composition, though both are important. We also do not address the history of the manuscript after its composition or construction.[3]

---

interpretable to anyone with a passing knowledge of childbirth, the same cannot be said of the Voynich Manuscript's illustrations.

[3]We recognize that the circumstances of composition are very relevant to the language. For example, if Manuscript 408 is a fifteenth-century copy of a possibly much earlier work, then different languages likely underlie the manuscript text.

We hope this review will ground some of the discussions about the manuscript and aid in determining the plausibility of theories. We also hope our review will serve as an example of how language documentary methods can be usefully applied to ancient languages and cryptographic problems.

1.1. Natural Language or Gibberish

The first question we need to address is whether the "language" of the Voynich Manuscript is natural language at all. That is, it could be gibberish of some type—a hoax made to look like a cipher. Proponents of the gibberish hypothesis point out the abnormal regularity and high degree of similarity in some adjacent words, which make it look unlike many other types of text (cf. Barlow 1986). The linguistic status of the Voynich Manuscript has been the subject of controversy. While authors such as Reddy & Knight (2011) and Hauer & Kondrak (2016) have worked from principles of decipherment, others have claimed to show that the language of the Voynich is not language at all. Rugg (2004) and Rugg & Taylor (2016) suggest that the language might have been generated by fifteenth-century cryptographic techniques, which could produce either enciphered text or gibberish. Rugg (2004) suggests that the manuscript may have been encoded with the so-called Cardan grille technique, while Rugg & Taylor (2016) suggest that the material is more compatible with a hoax than enciphered language. Daruka (2020) likewise concludes that the Voynich Manuscript is a hoax and contains gibberish, though created by different means than those suggested by Rugg. However, as shown by Landini (2001), Montemurro et al. (2013), Reddy & Knight (2011), and Sterneck & Bowern (2020), the gibberish account does not explain the higher-level document structure, which we discuss further in Section 4. That is, it is implausible for a fake, gibberish-based document to have internal structure of this type, and it is more likely that the Voynich Manuscript is a cipher or other encoding of a natural language.

At this stage, we are not persuaded by any of the arguments in favor of the Voynich text being gibberish. Because gibberish is by nature random, it should not display any of the higher-level organizational properties that the Voynich Manuscript displays (as summarized in Sections 3.3 and 4). The Voynich Manuscript is highly unusual and non-language-like at the character level. If we look beyond individual words to the line and paragraph levels, as well as the distribution of words across the manuscript, the text of the manuscript looks like encoded natural language rather than gibberish. Because the measures used to track the paragraph structure are unlikely to be directly manipulated, they are a good indicator of real structure. If we rule out gibberish, the question then becomes what type of encoding is represented by the Voynich writing system. While this is still unknown, some types of code can be ruled out. For example, it is probably not a simple substitution cipher because, if it were, it likely would have been deciphered by now. Conversely, polyalphabetic ciphers are unlikely both because they were probably not used in the early fifteenth century and because a polyalphabetic cipher would lead to identical words being encoded differently in different parts of the manuscript. We would not see the same (or similar) words being the most frequent on every page (see Section 3).

Finally, the manuscript could be in a constructed language (conlang). To our knowledge, the most extensive premodern conlang is the Turkish-, Persian-, and Arabic-based language Balaibalan (Häberl 2015, Koç 2005). Balaibalan, a Sufi esoteric language, is represented in three manuscripts that date from roughly 1580 but were likely collective efforts at composition over many years. There are two other well-attested ancient conlangs: the Latin- and German-based Lingua Ignota, created by Hildegard von Bingen (Higley 2007) in the twelfth century; and Enochian, an "angelic" English-like language invented in the sixteenth century by Edward Kelly and John Dee (Laycock 2001). All three conlangs are heavily based on natural languages and consist of made-up roots embedded in the morphology and syntax of natural languages. Thus, if Voynichese is a conlang, we

might expect it to pattern morphosyntactically with other natural languages but to be anomalous at the root level.

1.2. Background to the Manuscript

In the discussion that follows, we assume some points that are still debated (to our minds, unproductively) in Voynich studies. First, on the basis of the carbon dates of the vellum (as well as its overall appearance and the views of experts in medieval studies), we assume that the manuscript is a genuine medieval object (Clemens 2016, Clement 1997) and not a modern forgery. Those who argue that the manuscript is a modern hoax must assume that Voynich (or another person) obtained a large amount of untouched medieval parchment and made ink highly consistent with medieval practices, in an era before methods for accurate dating of parchment had been developed. That is, they must anachronistically assume that a modern hoaxer was trying to prevent detection by circumventing tests that had not, at that point, been invented.

We also consider it unlikely that the Voynich Manuscript is a medieval hoax. The cost associated with the production of such a manuscript and the number of people involved make it unlikely that it was created purely to deceive. A much smaller hoax would have served the same purpose at much less expense. Moreover, people who assume that the manuscript is a medieval hoax massively underestimate the amount of effort required to produce sustained language-like nonsense.[4]

The physical codex dates from the period 1404–1438 (Clemens 2016) on the basis of carbon dating, but we do not know whether the extant physical codex was copied from some earlier source. For this reason, we do not assume that the language must be medieval. Following Davis (2020), we assume multiple scribes. Davis provides evidence from the glyph shape that at least four (and more likely five) different hands were involved in the production of the manuscript.

Following the initial observation of Currier (1976) and subsequent work by Davis (2020) and Reddy & Knight (2011) (among others), we assume that there are two "languages" in the manuscript, labeled here for convenience as Voynich A and Voynich B. More precisely, there are two methods of encoding at least one natural language. While we use the term language here (following convention), it is not clear that the differences between Voynich A and Voynich B are due to different underlying languages or varieties (though this is possible). Voynich A is used in part of the herbal and pharmaceutical sections, while Voynich B is used in the balneological section, in some folios of the medicinal and herbal sections, and in the astrological section. The main differences between the two languages are in word frequency (which we discuss below in Section 3). Certain character sequences are very common in Voynich A (*ol* ol and *or* or) and relatively uncommon in Voynich B, and vice versa (*dy* dy). We do not address the connection between the two languages extensively here, but we do treat the two varieties separately for analytical purposes. Note that there is isomorphism between Davis's hands and Currier's two languages, with scribe 1 writing in Voynich A and scribes 2, 3, 4, and 5 in Voynich B.

2. PHONOLOGY AND GRAPHOLOGY

2.1. Scripts in the Voynich Manuscript

The Voynich Manuscript does not only contain Voynichese. While this article focuses on the material in the manuscript in the Voynich script, there are instances of other orthographies: the Occitan month names in the Zodiac pages (folios 70v–73v), the partially obscured phrase in

[4]We tested this point in an undergraduate class and found that beyond about 100 words, the task of writing language-like nonlanguage is very difficult. It is too easy to make local repetitions and words from other languages.

Table 1 List of most common Voynich characters

Transliteration	Glyph	Transliteration	Glyph
’	⸱	l	⸱
a	⸱	m	⸱
c	⸱	n	⸱
d	⸱	o	⸱
e	⸱	p	⸱
f	⸱	q	⸱
g	⸱	r	⸱
h	⸱	s	⸱
i	⸱	t	⸱
k	⸱	v	⸱
x	⸱	y	⸱
Combined characters			
ch	⸱	sh	⸱
cfh	⸱	cth	⸱
cph	⸱	ckh	⸱

Latin script at the end of the manuscript (folio 116v), and the now-invisible signature of Jacobus Horčický de Tepenec on the first page (see Skinner et al. 2017). The page numbers were added after the composition of the manuscript, but those numbers do not feature in our discussion.

For the purposes of examining properties of Voynichese, we use a conventionalized 1:1 mapping between Voynich characters and ASCII characters known as EVA (Extensible Voynich Alphabet, formerly European Voynich Alphabet), as in Takahashi's digital transcription (see **http://www.voynich.nu/transcr.html**) (**Table 1**). The purpose of the transliteration is to enable statistical-methods-based investigations of the text. We recognize that this mapping is a simplification and that we cannot properly draw conclusions about the phonological structure of Voynichese without understanding the structure of the orthography. To date, however, the Takahashi transcription offers the only possibility for computational-methods-based textual analysis. We recognize that this limits the conclusions we can draw.

2.2. The Script

The Voynich script includes Latin characters from several traditions (including Carolingian and Beneventan; see Clemens & Graham 2007), numerals, and characters that have no counterpart in other manuscripts except perhaps as ornamental flourishes (see Cappelli 1899, tavola IV, between pages LXVI and LXVII) (see also **Figure 2**). Some Voynich characters are clearly part of the Latin alphabet (⸱, ⸱, ⸱, ⸱). Others closely resemble numbers (⸱, ⸱, ⸱). Others appear to have no parallels in other scripts, including the four gallows characters ⸱, ⸱, ⸱, and ⸱. These are known as gallows characters in writings on the Voynich script because of their superficial appearance to gallows. The only known parallel of these characters is in the ornaments described by Cappelli (1899, tavola IV). While the ornamental characters in Cappelli's work are ligatures created by joining digraphs together, no such assumption can be made about the Voynich gallows characters. The other type of character in the Voynich script is known as a bench (e.g., ⸱, ⸱) because it looks somewhat like a bench.

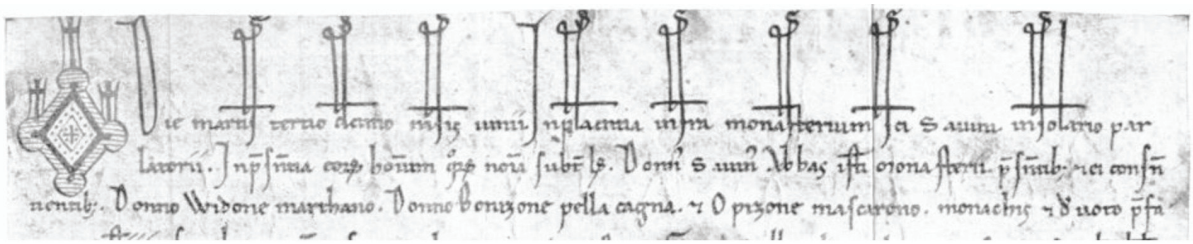


Figure 2

Example of the use of gallows-like characters in another manuscript of a similar epoch (reproduced from Cappelli 1899, tavola IV, between pages LXVI and LXVII). In this manuscript, the gallows characters are purely decorative and can be easily read as ornamental script.

In addition to the characters discussed here, there are other forms that appear only once, twice, or a few times in the manuscript (see Davis 2020, p. 170). The most common of these is the alchemical symbol ⸸.

There are 21 or 22 common glyphs, plus approximately that many rarer character forms. The total number of glyphs in the data set depends heavily on how one classifies the variable shapes of the graphemes. For example, there are two common transliterations in use, EVA and v101, which use different characters. We use EVA in this review because it is both the most widely used and the basis for the machine-readable transliterations on which our statistical work in the following sections depends.

EVA represents each Voynich letter with an ASCII character. It was designed to produce machine-readable versions of the text and is now widely used. The mappings of the Voynich letters to ASCII characters were based on the pseudonymous cryptographer Glen Claston's assumptions about what characters in the Latin alphabet the Voynich characters might be closest to (see **http://www.voynich.nu/transcr.html**). For example, ᦡ is transliterated in EVA as *d*. The transliteration equivalents are heavily influenced by European/Latinate considerations. For example, the sequence ᦠᦵ is represented as *qo* in large part because ᦠ never appears in the text except before ᦵ—a pattern reminiscent of the distribution of *q* with *u* in Latin. However, there is no other independent reason for assuming that Voynich ᦠ is Latin script *q*.

In our opinion, EVA probably underdifferentiates characters (grouping together two variants of ᦡ, for example). It also creates digraphs from characters that may be better understood as single glyphs. EVA ᦍᦌ, for example, comprises two distinct components: ᦍ and ᦌ (rendered in ASCII as *ch*), even though ᦌ is never found separately, and ᦍ is otherwise identical to ᦍ. EVA is, however, the transcription system that is used in the machine-readable version of the Voynich Manuscript, so we adopt it here pending a thorough review of the transcription system (for other issues regarding transcription, see the sidebar titled Some Difficulties in Creating a Transcription of Voynichese).

SOME DIFFICULTIES IN CREATING A TRANSCRIPTION OF VOYNICHESE

- The word breaks are possibly unreliable.
- The relationships between the characters are unknown.
- The relationship between the graphemes and the underlying phonology is unknown.

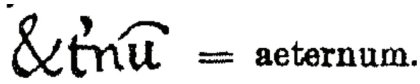


Figure 3

Apostrophe used as an abbreviation character (Cappelli 1982, p. 18); compare the Voynich combined character ⸗.

Some characters occur both in combination and separately in the manuscript. The bench character ⸗ combines with the gallows characters to form ⸗, ⸗, ⸗ and ⸗. The characters ⸗ and ⸗ appear to be distinguished only by the plume over the bench. This ligature is reminiscent of the character used for abbreviations in many medieval manuscripts, as illustrated in **Figure 3**. This same ligature may also distinguish ⸗ and ⸗. ⸗ and ⸗ appear to be distinguished by a downward stroke. However, it is not known whether these are chance similarities (compare the relationship between the Latin characters *o*, *b*, and *d*) or reflect some underlying principle of regularity in the script.

Finally, although the EVA transliteration makes Voynich material easy to pronounce, it has no basis beyond some similarity with some of the letters in the Voynich script. For example, transcribing the gallows characters as *f*, *k*, *t*, and *p* is purely a convention.

2.3. Phonology and Orthography

In this section, we consider the phonology of the linguistic system underlying the Voynich Manuscript. The examination conducted here is not phonology in the purely linguistic sense, of course, since we have no knowledge of the abstract sound organization represented by the orthography or of the relationship between the orthography and the phonology (proper). However, we can draw some conclusions about the representation of the linguistic system compared with orthographic representations of other languages.

2.3.1. Vowels versus consonants. Both Reddy & Knight (2011) and Guy (1991) discuss whether the Voynich writing system is an alphabet (i.e., containing both vowels and consonants) or an abjad (representing consonants only). One test for this classification uses the Sukhotin algorithm (Guy 1991), which works on the premise that in most (if not all) natural languages, vowels are more likely to be adjacent to consonants than to other vowels. That is, syllables of the shape CV or CVC are more common than those of the form V or VC. The algorithm computes an adjacency matrix for all characters. One successively sums the rows, assumes that the most frequently occurring segment is a vowel, and then removes twice the number of occurrences from the adjacency matrix. One continues identifying potential vowels until no positive sums remain.

For Voynichese overall, the Sukhotin algorithm identifies five vowels: *a*, *o*, *e*, *h*, and *i* (i.e., ⸗, ⸗, ⸗, ⸗, and ⸗). It is worth noting that three of these characters (⸗, ⸗, and ⸗) are similar to characters that represent vowels in the Latin alphabet. The letter *e* is slightly more common in Voynich B, and the letter *h* is slightly more common in Voynich A.

Some additional issues remain. First, the Sukhotin algorithm is sensitive to whether word breaks are included in the calculations. When word breaks are excluded, the characters identified as vowels are ⸗, ⸗, ⸗, ⸗, ⸗, and ⸗. That is, ⸗ is no longer identified as a vowel, and three characters that are almost exclusively word-final are identified (⸗, ⸗, and ⸗). Secondly, different characters are identified between the two Voynich languages, Voynich A and Voynich B.

Reddy & Knight (2011) argue that Voynichese shows more properties of an abjad than of an alphabetic script. Their argument draws in part on the induction of character classes from the clustering behavior of characters. They use a two-state bigram Hidden Markov Model over

letters in Voynichese and then induce two classes. For alphabets, these classes usually correspond to consonants and vowels. For Voynichese, however, the two classes correspond to the final character in a word versus the rest of the word. This result could be driven primarily by the characters that are found only word-finally (such as ſ, ৩, and ſ). That is, the result may indicate not that the Voynich script is an abjad but, rather, that there are positional variants for character forms. Furthermore, though the Sukhotin algorithm picks out both consonants and vowels in abjads, in abjad scripts not all words contain one of the vocalic characters (even though all words contain a vowel phonemically). In contrast, almost every word in the Voynich Manuscript has at least one of the characters that the Sukhotin algorithm picks out as vowels.[5]

2.3.2. Character entropy. The information entropy of a text can be thought of as the amount of unpredictability or disorder present in the text. Character-level entropy defines the average amount of information carried by a single character (usually measured in bits, which are also called shannons). The concept was introduced by Shannon (1949) and arises in the field of information theory, in which it is important for measuring the theoretical rate of information transmission. Character entropy is another metric by which we can compare languages with each other and with Voynichese.

Bennett (1976) notes the unusual nature of the Voynich script in his discussion of conditional character entropy (also known as second-order character entropy or h2). Conditional character entropy can be thought of as the overall predictability of a letter given the preceding letter. For example, in English texts, the letter *q* is almost invariably followed by *u*. The conditional probability of the bigram *qu* (the probability of *u* given that the previous letter is *q*) is close to 1. The overall conditional entropy is calculated from the conditional probabilities of each bigram, weighted by their overall occurrence in the text, as in the following equation:

$$H(X|Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(y_j)}{P(x_i, y_j)}.$$

Bennett (1976) compares the Voynich Manuscript with texts in four modern European languages and finds conditional entropy in Voynichese to be much smaller. That is, character sequences within the words in Voynich text are unusually predictable compared with European languages. Voynich characters appear in unusually predictable sequences, with certain characters found at the beginning or end of the word or only after certain characters. Bennett finds the Voynich character entropy to be comparable to those of Hawaiian and other Polynesian languages with small phoneme inventories and limited syllable shapes. However, Stallings (1998) notes that Bennett's Hawaiian text uses a simplified orthography that does not distinguish long vowels or glottal stops, which would have the effect of decreasing entropy.

In fact, the conditional character entropy in the EVA transcription of the Voynich text is significantly lower than in any other language text in our sample. **Figure 4** shows the conditional character entropy (h2) and character set size for 250 languages in the sample, coded for type of script. In **Figure 4**, the conditional character entropy of Voynichese (Voynich A, Voynich B, and the full sample with rare characters included) is much lower than that of any of the natural language samples.

The entropy of Voynichese is unlike that of any other language or script. Plausible manipulations of the script have been investigated, including various shorthand abbreviations and devoweling of the script. These manipulations affect the character entropy but not to the extent

[5]Furthermore, character entropy, as discussed in Section 2.3.2, is usually higher for abjads than for alphabetic scripts. For Voynichese, the entropy is lower.

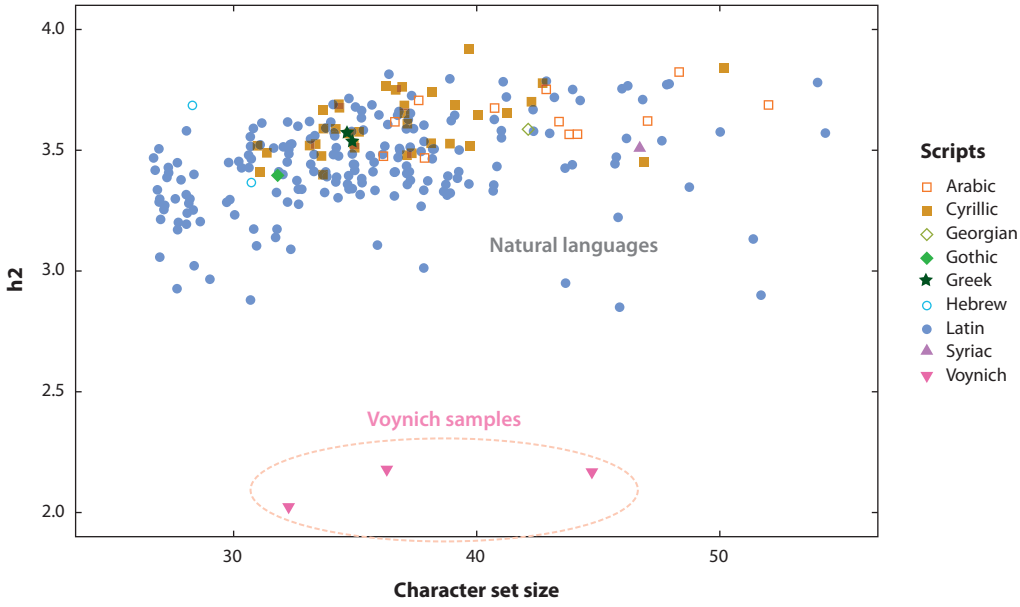


Figure 4

Conditional character entropy (h2) versus number of characters for Voynichese and other languages. This chart is restricted to languages written with alphabets (Latin, Cyrillic, Gothic, Georgian) or abjads (Hebrew, Arabic, Syriac). Alphasyllabaries like Devanagari have a greater character set size but a similar h2. Logograms like Chinese have a much greater character set size and a much higher h2. All of the lowest-h2 languages are written with alphabets.

that would be required to bring Voynichese close to the level of other languages. The only manipulation of this type that brings the conditional entropy to Voynich levels is systematic conflation of phonemic distinctions, such as conflating all vowels to a single character, recoding by dividing characters into whether they occur in the first or second half of the alphabet, or sorting all characters in a word into alphabetical order.

It is worth noting that values for conditional entropy are, to some extent, affected by how characters are divided. For example, if ϟ and ο are treated as separate characters, then entropy is not (fully) a function of misdividing characters. Changing the character divisions can lower the entropy, but not to the levels seen in Voynichese. Thus, the very low conditional entropy values are not simply a result of misparsing Voynich characters.

Entropy is also not a function of abbreviated coding, at least using the common abbreviations that scribes used. Conditional entropy of abbreviated texts is actually slightly higher (e.g., 3.4 for the abbreviated text of the Latin *Secreta Secretorum* versus 3.2 for the plain text version of the same work).

In summary, Voynichese has much lower conditional entropy than other texts and languages to which it has been compared. Its entropy, though not a function of script transliteration, may provide a clue to the type of encipherment used.

3. WORD-LEVEL MORPHOLOGY

3.1. Evidence for Words

Some textual traditions use spaces to visually separate words from each other, while others do not use spaces for this purpose. The Voynich Manuscript contains spaces that separate the text

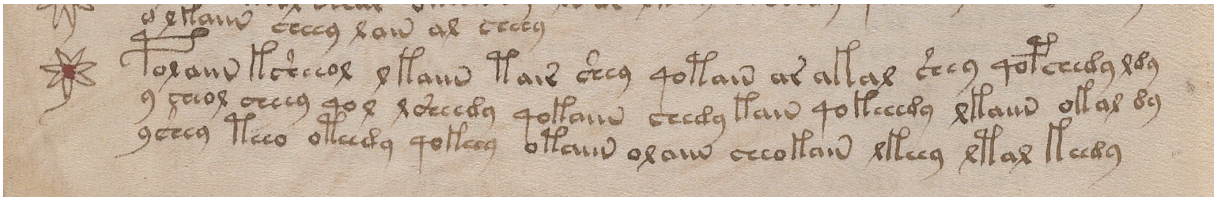


Figure 5

Example of a Voynich paragraph with word breaks. Currier (1976) considers the line to be a meaningful functional unit (i.e., similar to a clause or sentence), and some words are more frequent at line beginnings or ends. Image courtesy of the Beinecke Rare Book and Manuscript Library, Yale University.

into word-sized chunks. If we assume that these chunks do in fact represent words, then we can investigate the morphology of Voynichese. That is, we can look for evidence of an internal structure to Voynich words, such as the existence of prefixes and suffixes, which might represent grammatical properties like case or agreement. We can also look at the word as an atomic unit and examine word-level patterns that abstract away from character-level issues.

At some points in the manuscript, it is difficult to tell where the breaks are. For example, in **Figure 5**, the last two characters of the middle line, **ol**, are more set off from the characters that precede them compared with the word directly below, which ends in the same sequence. We could therefore either treat **ol** as a separate (short) word or as part of the preceding word with an unusual gap. However, despite such ambiguities, some consistent patterns about word structure can be deduced if it is assumed that visual spaces are word breaks.

3.2. Structure in Voynich Words

Many Voynich characters and character combinations are restricted to certain parts of words. Tiltman (1967) [quoted by D'Imperio (1978)] proposes that Voynich words consist of three separate "fields," with particular symbols occurring at the beginning, middle, or end of the word. The subsequent proposals and analyses of Stolfi (2000) and Reddy & Knight (2011) differ in complexity and coverage, but they maintain this basic notion that there are separate fields for particular characters and character combinations.

A few of the most common character combinations in each field are listed below:

1. Prefixes:[6] *qo-* **qo**, *o-* **o**, *y-* **y**, *ch-* **ch**, *sh-* **sh**, *d-* **d**
2. Roots/midfixes: *p* **p**, *t* **t**, *k* **k**, *f* **f**, *e* **e**, *ee* **ee**, *o* **o**, *a* **a**
3. Suffixes: *-y* **y**, *-dy* **dy**, *-l* **l**, *-r* **r**, *-m* **m**, *-iin* **iin**, *-in* **in**

This structure is similar in all words in all sections of the manuscript and in both Voynich A and Voynich B. There is some minor variation in the frequency of particular affixes between Voynich A and B. Most significantly, Voynich B has a higher frequency of both the *qo-* **qo** prefix and the *-dy* **dy** suffix (they are about two times and three times more common, respectively).[7]

Unlike other textual traditions, in Voynichese many characters and character combinations are found exclusively in one of the three character fields. This pattern is the ultimate source of

[6]The first three prefixes are usually followed by a gallows character (*p* **p**, *t* **t**, *k* **k**, *f* **f**), whereas the others are not. Some words instead begin with a bench-and-gallows combination (*cph* **cph**, *cth* **cth**, *ckh* **ckh**, *cfh* **cfh**).
[7]As opposed to running text, labels tend to lack the *qo-* **qo** prefix; this difference may be evidence that the prefix encodes a grammatical property like inflection or agreement.

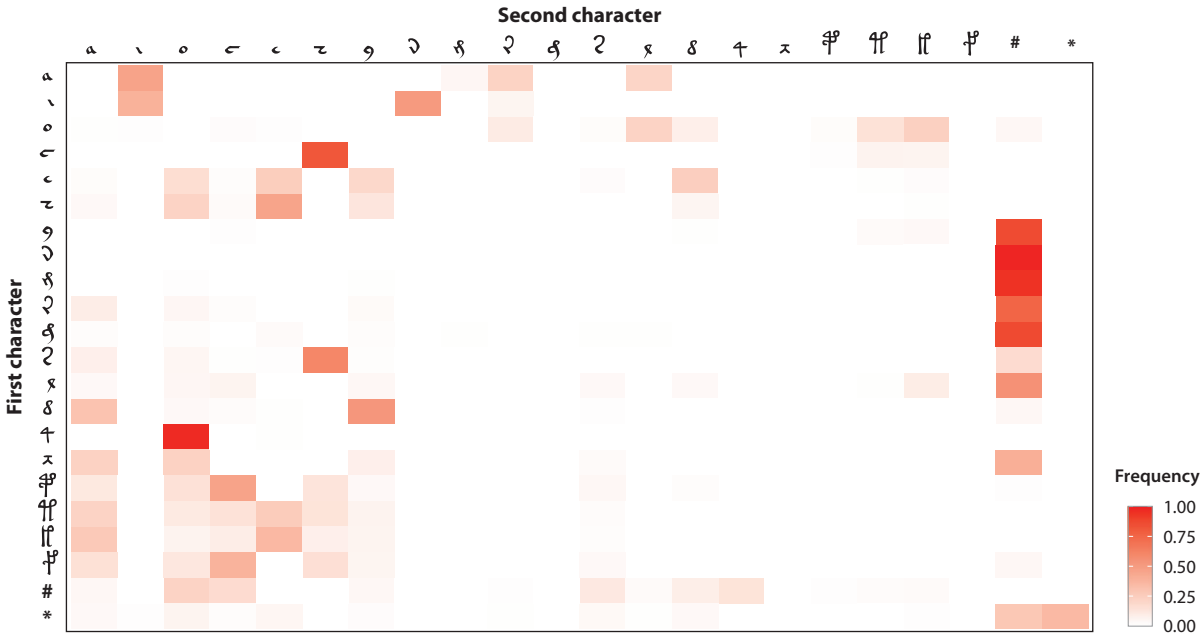


Figure 6

Frequency of each Voynich character given the previous character. Rows show the first characters in two-character sequences; columns show the second characters [pound signs (#) denote word boundaries, and asterisks (*) denote illegible characters]. Nonoccurring bigrams are unshaded; the darker the shading, the higher the percentage of bigrams where the two items co-occur in that order.

the text's unusually low conditional character entropy (h2). The text is highly predictable because certain letters occur only in certain parts of words and in a relatively small number of different combinations. For example, Voynich A has a somewhat higher h2 (2.17) than Voynich B (2.01). This difference is largely due to the increased frequency of *qo-* and *-dy* affixes mentioned above. If we delete these two affixes from both texts, then the h2 of Voynich A and Voynich B become nearly identical (2.23 and 2.24, respectively).

As discussed above, certain characters appear exclusively or almost exclusively at the beginnings (*q*) or ends (*m*, *g*, *r*, *n*) of words. The closest similarity to this phenomenon in other texts is in abjads such as Arabic and Hebrew, which contain a small number of glyph variants that occur only word-finally. The distributions of Voynich characters in bigrams are given in **Figure 6**.

In practice, it can be difficult to disentangle hypotheses about word structure from hypotheses about the script. For example, the very common word-final sequence *-iin* consists of three separate characters in the EVA transcription but only one character in the Currier transcription. If it is a sequence of multiple characters, then it most likely represents an entire suffix. If it is a single character, then it may not.

3.3. Distribution of Words in the Manuscript

Currier (1976) notes that certain frequent word forms recur throughout the Voynich Manuscript. However, the most common words differ between Voynich A and Voynich B. **Table 2** shows the 10 most common words in Voynich A and Voynich B and their proportional frequencies.

Although there is some overlap, the most common vocabulary items of Voynich A and Voynich B are substantially different. While the words in both languages are built from the same

Table 2 Most common words in Voynich A and B

Voynich A		Voynich B	
Transliteration (Voynich)	**Frequency (%)**	**Transliteration (Voynich)**	**Frequency (%)**
daiin (ꞃꞅꞅꞅ)	4.5	chedy (ꞅꞅꞅꞃꞅꞃ)	2.1
chol (ꞅꞅꞃꞅ)	2.5	ol (ꞅꞃ)	1.8
chor (ꞅꞅꞃꞅ)	1.6	shedy (ꞅꞅꞅꞃꞅꞃ)	1.8
s (ꞅ)	1.4	aiin (ꞅꞅꞅ)	1.5
dy (ꞃꞅ)	1.1	daiin (ꞃꞅꞅꞅ)	1.4
shol (ꞅꞅꞅꞃꞅ)	1.0	qokeedy (ꞅꞅꞅꞃꞅꞃ)	1.3
sho (ꞅꞅꞅ)	0.9	qokain (ꞅꞅꞅꞅꞅꞅ)	1.2
chy (ꞅꞅꞅ)	0.9	qokedy (ꞅꞅꞅꞃꞅꞃ)	1.2
cthy (ꞅꞅꞅꞅꞅꞅ)	0.9	qokeey (ꞅꞅꞅꞅꞅꞅ)	1.1
ol (ꞅꞃ)	0.9	chey (ꞅꞅꞅꞅ)	1.0
Total	15.7	**Total**	14.5

three-field structure, they do not clearly correspond to each other. They might be the result of different encoding processes, or they might represent different underlying natural languages.

Significantly, both Landini (2001) and Reddy & Knight (2011) note that the distribution of words in the Voynich Manuscript follows Zipf's law—a power law that relates the frequency of a word to its rank. Thus, if we rank each word by frequency count, we expect the second word to be roughly half as frequent as the first word, and the third word to be a third as frequent as the first word. A chart of frequency by word rank depicts a characteristic Zipf curve for the Voynich text. **Figure 7** compares the distribution curves of the first 100 words in Voynich and four other languages.

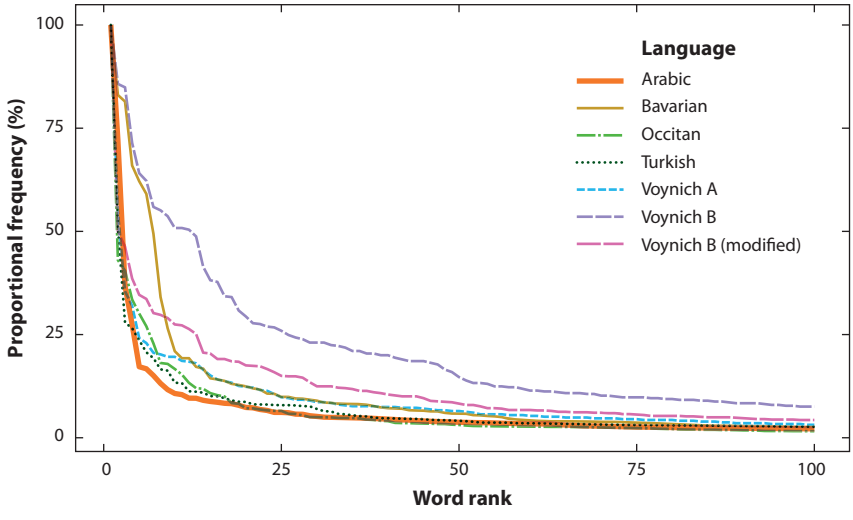


Figure 7

Word frequency distribution of Voynichese and selected languages. The x-axis shows the frequency rank of each word; the y-axis gives the proportional frequency. Frequency is given as a percentage of the most common word (e.g., word 2 in Occitan is 43% as frequent as word 1, word 3 is 40% as frequent as word 1, and so on).

Both Voynich languages follow a Zipfian distribution. Voynich B is a clear outlier in this sample, largely because its three most common words are of approximately equal frequency. It is possible that *chedy* ⟨ch.ed.y⟩ and *shedy* ⟨sh.ed.y⟩ represent the same word as they are distinguished only by whether there is a plume stroke over the bench character. If we make this assumption [represented in **Figure 7** as Voynich B (modified)], Voynich B is less of an outlier.

Zipf's law was originally formulated to describe word distributions in natural language corpora, although it has been found to apply to various other social phenomena. The fact that Voynich word frequencies follow a Zipfian distribution does not prove that the text is linguistically meaningful. However, the word distribution does not look highly unusual compared with natural languages, which we might expect if the text were naively created gibberish. As Reddy & Knight (2011, p. 80) comment, a Zipfian distribution "is a necessary (though not sufficient) test of linguistic plausibility."

Furthermore, if we believe that Voynich is an encoded form of natural language, any hypothesis about encoding must take into account the fact that the Zipfian distribution is preserved. Some forms of encoding will have the effect of diminishing or eliminating the distribution, while others will not. For example, encipherment methods that continuously rotate alphabets will flatten the frequency of the most common lexical items because those words will be enciphered differently on different pages. Regardless of the method used to encipher the Voynich Manuscript languages, the distribution is mostly consistent within each language.

The proportional frequencies of the most common words in linguistic texts are also useful for diagnosing linguistic structure. The most common word in Voynich A, *daiin* ⟨d.aii.n⟩, accounts for 4.5% of the words in that text, while the most common word in Voynich B, *chedy* ⟨ch.ed.y⟩, takes up 2.1%. These proportional frequencies are well within the expected range for most natural languages. The most common word in many natural languages is a definite article like 'the', a connective like 'and', or a preposition like 'in' or 'of'. In Voynich A, *daiin* ⟨d.aii.n⟩ is never found at the beginning of a paragraph; therefore, it may be a connective.

The proportional frequency of the top 10 most common words together is also within the typical range for natural languages. For Voynich A, this frequency is 15.7%, and for Voynich B it is 14.5% (see **Figure 8** for a comparison with other language families). The Voynich languages are within the range of each of these families and are closest to the averages for Semitic, Iranian, and Germanic. It should be noted that there is an inverse correlation between this statistic and morphological complexity. The percentage tends to be lower for languages that have many words with complex morphological structure, such as languages in the Turkic, Kartvelian, and Dravidian families. It tends to be higher in languages with less morphological complexity, such as those in the Romance family. While this statistic alone is not exact enough to match Voynich to a particular language family, it suggests that Voynichese has a medium level of morphological complexity.

3.4. Moving-Average Type–Token Ratio

Another statistic that is useful for determining the lexical diversity of a text is the type–token ratio. Languages with greater morphological complexity typically have higher type–token ratios as the number of distinct types approaches the number of overall tokens in a text. However, this statistic is heavily dependent on the length of the text. Gheuens (2019) introduces the Moving-Average Type–Token Ratio (MATTR) index, which takes the average type–token ratio over a set word window, as a way to measure lexical diversity irrespective of text length. Gheuens (2019) examines MATTR in Voynichese compared with a sample of language texts and concludes that 2,000 words is an ideal window. We have also found 2,000 words to be a good window for comparing MATTR in language families.

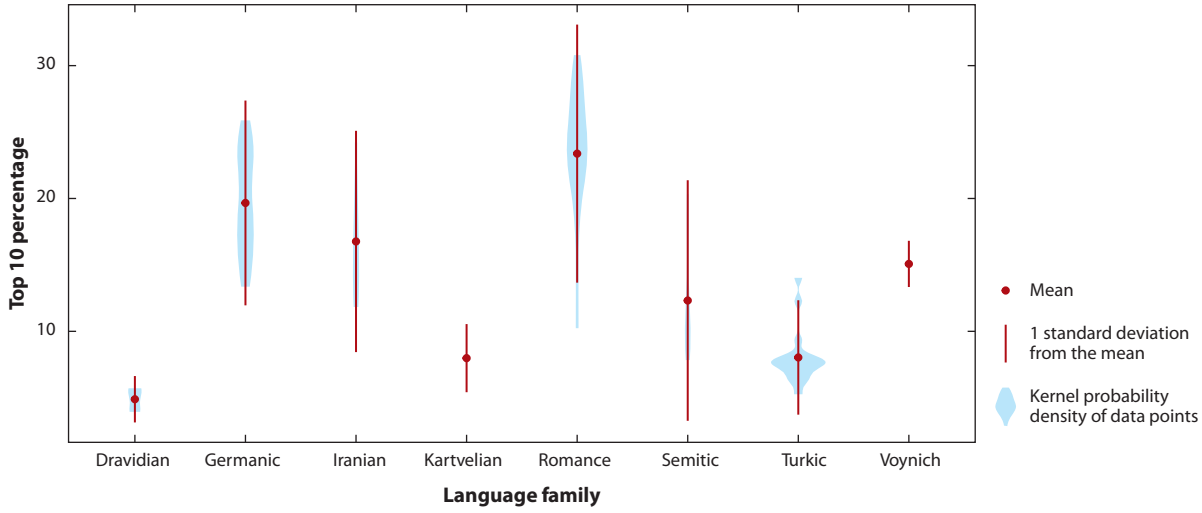


Figure 8

Proportional frequency (i.e., the combined frequencies of a language's 10 most frequent words) by language family. The seven language families shown here represent a total of 101 languages. Red lines indicate one standard deviation from the mean.

Figure 9 shows MATTR summaries across language families. The families with higher morphological complexity, like Dravidian and Kartvelian, have higher MATTR values. The Voynich texts are once again in the medium range; they are closest to Iranian, Germanic, and Romance.

The MATTR and proportional frequency measures provide distinct, largely complementary evidence that Voynichese represents a language of medium morphological complexity. The averages for Voynichese most closely resemble those for Germanic and Iranian and least resemble those for Turkic, Dravidian, and Kartvelian. As with the Zipfian word distribution, we find Voynichese to be well within the expected values for natural language texts and far from random

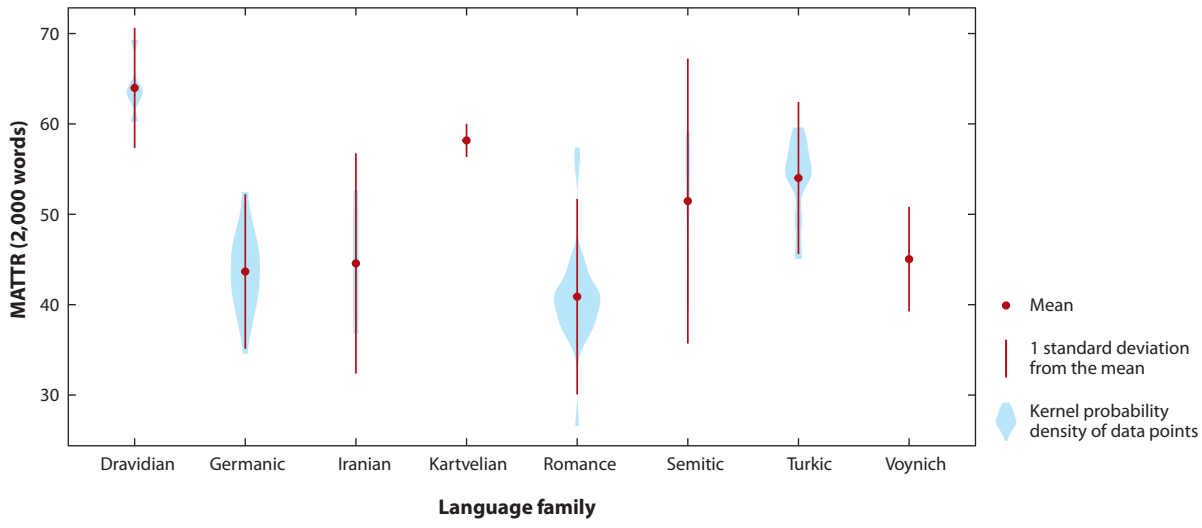


Figure 9

Moving-Average Type–Token Ratio (MATTR) using a window of 2,000 words, calculated for 101 languages in seven language families.

gibberish. If the Voynich text is meaningless, its creators mimicked natural language in a sophisticated way.

These measures are useful for our purposes because they allow us to narrow down possible languages without knowing the meaning of any of the words in Voynichese. They are also largely independent of character-level parsing and transcription issues. If we assume that the transcription style is consistent, the word distributions will be identical even if there are incorrect assumptions about character boundaries and the relationships between characters and phonemes.

4. GENERALIZATIONS ABOVE THE WORD LEVEL

The Voynich Manuscript text can be examined at multiple levels above the unit of the word. One major division is between the running text and the text that appears as labels on drawings. The running text includes word combinations and Voynich phrases that give evidence of syntactic structure. A single line may roughly equate to a sentence (Currier 1976). Blocks of running text are separated into paragraphs, which could indicate topic shifts. Above this level, each folio can be examined separately. The distinction between Voynich A and Voynich B is made at the folio level; with one exception, the Voynich hands also are distinguished at this level (Davis 2020). Finally, the different sections of the manuscript are distinguished by subject matter, as evident by the illustrations and diagrams.

Just as there is clear evidence of structure and patterning at the character level and at the word level, there are patterns at each of these higher levels of structure. These highest levels have received the least amount of attention by researchers.

4.1. Line and Paragraph

Currier (1976) argues that the line itself should be treated as a meaningful unit because certain Voynich characters and character combinations tend to be found at either the beginning or the end of a line. There are also certain characters that precede the first word in a paragraph.

One hypothesis is that these patterns originate in the underlying linguistic structure of the text. If we assume that a line of text roughly corresponds to a sentence, then words that are more likely to occur at the beginning or end of a sentence will be found at the edges of the line. For example, the first word of a line might consist of a noun or definite article. If Voynichese represents a verb-final language, then we might see certain character combinations that are uniquely associated with verbal morphology, such as tense and agreement, appearing more frequently in the last word of the line.

Another hypothesis is that these patterns are typographical in nature. In other words, the same word will be written differently depending on where it appears in the line. In that case, we would expect to find pairs of similarly patterning Voynich words that occur in different places along the line. Indeed, Montemurro et al. (2013) suggest that some similar words have affinities in the text in that they have similar patterns of occurrence.

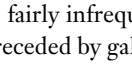
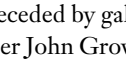
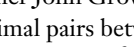
The clearest example of this phenomenon is at the level of the paragraph, which usually begins with a gallows character. Approximately 85% of the paragraphs in the text begin with , , , or . These gallows-initial words (*a*) are otherwise fairly infrequent and (*b*) have the same structure as normal Voynich words except that they are preceded by gallows characters. Stolfi (2000) discusses the hypothesis, attributed to Voynich researcher John Grove, that gallows-initial words were variants of other words because of the many minimal pairs between common paragraph-initial words and the most frequent words in the entire text: *tchor/chor*  , *pol/ol*  , and *tchedy/chedy*  . In other words, the gallows characters do not seem to be part of the words themselves; they simply mark the beginnings of paragraphs. Furthermore, gallows-initial words, when

Table 3 Most common Voynich words by position

Paragraph-initial	Line-initial	Line-final	Everywhere
pol ( )	daiin ( )	daiin ( )	daiin ( )
tchedy ( )	saiin ( )	dy ( )	ol ( )
polaiin ( )	dain ( )	dam ( )	chedy ( )
tol ( )	sol ( )	am ( )	aiin ( )
pchedar ( )	sor ( )	dal ( )	shedy ( )

they do appear elsewhere, usually begin with *k* or *f* rather than *p* or *t*. This pattern may suggest subordering of elements within the paragraphs themselves.

There is a similar but less robust pattern associated with the beginning of each line. The first word is somewhat more likely to begin with *s-* . This may be another orthographic variant, but it seems to occur only with words that otherwise begin with *o-* or *a-* . Thus, *aiin* , *ol* , and *or* are replaced with *saiin* , *sol* , and *sor* .

There are also characters that usually appear at the end of the last word of the line—in particular, *m* and the infrequent character *g* . It is plausible that *m* and *g* are variant forms of the word-final glyphs *-iin* and *-y* . For example, some of the most common words in line-final position include *dam* and *am* , which appear to be counterparts of the very frequent words *daiin* and *aiin* . Similarly, there are several minimal pairs of *-g* and *-y* words between line-final position and elsewhere: *g/y* , *alg/aly* , *dairodg/dairody* , and *arg/ary* . However, if this is an orthographic convention, it is not applied in a consistent manner: The forms *-iin* and *-y* are also found line-finally, albeit somewhat less frequently.

These generalizations are evident from **Table 3**, which shows the five most common words in each position. There are some exceptions, in particular with the word *daiin* , which is common in every position except paragraph-initially (*daiin* is never found as the first word of the paragraph, and there are only two possible examples of *pdaiin*). **Table 3** combines word counts from Voynich A and Voynich B. These overall patterns are found in both languages, although there are slight differences. For example, paragraphs in Voynich B are more commonly marked by the *p* gallows.

All of these observations lead to generalizations that seem typographical rather than linguistic in nature. Voynich writing does not appear to have any conventional punctuation symbols; rather, it uses character variants and appended characters to structure the text in a way that is similar to punctuation. A comprehensive linguistic analysis needs to take seriously the possibility that, for example, *paiin* , *saiin* , *aiin* , and *am* are all positional variants of the same word.

4.2. Phrases

Syntax describes the ways in which words fit together in a hierarchical structure, and generalizations about word and phrase combinations can explicate this structure. Syntax has been studied less systematically than character- and word-level patterns in the Voynich Manuscript. In this section, we simply present a few observations and their potential implications for syntactic structure.

Stolfi (2000) points to the repetitive nature of the Voynich text as evidence that it is not meaningful. An example can be found in the following line from Voynich B:

keedy qokeedy qokey okar otar dar dar dy.

The repetitiveness of this is at least partly the result of the relatively limited set of character combinations and the predictable structure of Voynich words. Full reduplication, in which an entire word is repeated, is also common in Voynich. However, it is still within the realm of plausibility for natural language texts. In Voynich A, each word has a 0.84% chance of repeating, while in Voynich B that chance is 0.94%. The range among the samples in our language corpus is 0.02–4.8%, with an average of 0.63% (however, the average for most relevant language families is somewhat smaller: 0.37% for Germanic, 0.36% for Romance, 0.25% for Iranian, and 0.36% for Semitic).

There are a few generalizations about multiword structures that may provide evidence of syntactic structure in the manuscript. The first holds for Voynich B but not for Voynich A. A word that begins with *qo-* qo is usually preceded by a word that ends with *-y* y (e.g., *shedy qokeedy*, *ody qokaiin*, *qokeedy qokedy* shedy qokeedy, ody qokaiin, qokeedy qokedy). This pattern might indicate some form of agreement or a compound verb structure.

The next generalization, which holds for both languages, involves the fourth most common word overall, *aiin* aiin. This word is usually preceded by a short one- or two-letter word (e.g., *ar aiin*, *or aiin*, *s aiin* ar aiin, or aiin, s aiin). Short words tend to be the most common words in natural language texts, but the most common Voynich words have four or five letters. The short words may represent articles or prepositions, although identification with parts of speech cannot be accomplished at this stage.

Another possible multiword structure involves gallows characters, which are most commonly preceded by *o-* o (e.g., *okeedy*, *otaiin*, *opchy*, *ofchedy* okeedy, otaiin, opchy, ofchedy). These words are prevalent on labels, and they occur at roughly the frequency at which we expect to find nouns in the text. Furthermore, there are four gallows characters—two common (*k/t* k t) and two uncommon (*p/f* p f)—and one might hypothesize that they represent a two-by-two article classification similar to that of many Romance languages.

Given their frequency of occurrence (and the preponderance of feminine nouns in medieval philosophical texts), we would expect *ok-* ok to be feminine singular, *ot-* ot to be masculine singular, *op-* op to be feminine plural, and *of-* of to be masculine plural. This hypothesis predicts that most "roots" (i.e., *-eedy*, *-aiin*, *-chy*, *-chedy*) will be associated with only one common gallows and one uncommon gallows symbol. Relatively few words will take both masculine and feminine marking. However, this prediction is not borne out. All roots show roughly the same pattern, and most are found with every possible combination: *okaiin*, *otaiin*, *opaiin*, *ofaiin* okaiin, otaiin, opaiin, ofaiin. Therefore, the hypothesis that the elements in question represent an article classification similar to that of Romance languages is untenable.

If these gallows sequences do represent articles, the different gallows characters might change on the basis of the underlying root, as we find the definite article assimilating phonologically to the noun in Arabic speech (although this is not expressed in Arabic writing). We do find certain constraints on what can follow gallows characters. For example, *p/f* p f are never followed by *e* e and almost never by *i* i or *l* l.

4.3. Topic

Reddy & Knight (2011), Montemurro et al. (2013), and Amancio et al. (2013) discuss Voynich topic modeling. That is, they use techniques from automatic text summation or keyword identification to group together similar pages of the manuscript. Reddy & Knight (2011) show that the Voynich Manuscript has a number of properties that are consistent with natural language and inconsistent with a hoax. For example, the pages that are nearest neighbors in topic modeling tend to be adjacent to one another in the manuscript.

Montemurro et al. (2013) use techniques from information theory to identify which words are most likely to contribute to topics in texts. That is, they identify words that are more uniformly distributed throughout the Voynich Manuscript and compare them with those that tend to cluster. Those that tend to cluster are more likely to provide information about the subject matter of the pages. Montemurro and colleagues' method also returns an overall similarity between the pages with herbal and pharmacological illustrations, which suggests that the illustrations in each part of the text are relevant to the linguistic text in each section. Amancio et al. (2013) also evaluate the discourse properties of the Voynich Manuscript and conclude that the manuscript most likely consists of natural language thematic content.

Sterneck & Bowern (2020) further investigate topic modeling within the Voynich text and the relationships between Voynich A and Voynich B, scribal hands, and thematic material (as deduced from the illustrations). They use TF-IDF-weighted counts[8] using 40 word chunks of text within each page; they also use nonnegative matrix factorization topic clustering to cluster Voynich pages and to compare those clusters with other types of structure in the document. Using different methods from those of Amancio et al. (2013) and Reddy & Knight (2011), Sterneck & Bowern (2020) have been able to recover general thematic topics and identify correlations between topics and hands within thematic sections. That is, the pages that Davis (2020) identifies as being written by a different scribe also tend to emerge as a different topic in the TF-IDF analysis. **Figure 10** illustrates the topic, hand, and section clustering. This result suggests that different scribes may have used different encipherment strategies or written about different subjects.

5. LINGUISTIC IDENTIFICATION

Finally, we briefly survey the general theories that have been advanced as to what language underlies the Voynich Manuscript. The manuscript was undecipherable even in the seventeenth century. Athanasius Kircher[9] thought it was likely written in the Glagolitic script ("Illyrian"; see **http://www.voynich.nu/letters.html**), and indeed there are certain similarities but not more than one might expect given the common origin of Glagolitic and Latin scripts (see also Bennett 1976). Others have assumed that the Voynich Manuscript is in either Latin or a Romance language, and given the widespread use of Latin as a lingua franca in Europe throughout this period, that assumption is not unreasonable. Reddy & Knight (2011) suggest that aspects of the script are reminiscent of an abjad, the use of which would suggest a Semitic language (since all currently known abjads are used to write Semitic languages). This idea inspired Hauer & Kondrak's (2016) claim that the manuscript is in an enciphered and anagrammed Hebrew; their solution, however, is not generally accepted.

---

[8]TF-IDF (term frequency–inverse document frequency) is a statistic used to cluster text according to the frequency of words in the text itself (the TF) compared with the frequency of a given word in the document as a whole. It allows one to group together texts based on distinctive words. Because Voynich pages are texts of different lengths, Sterneck & Bowern (2020) normalize the text length for each page.

[9]The source of this observation is a letter from Kircher to Theodor Moretus (a mathematician) in March 1639. The relevant passage is "Alterum denique folium quod ipsi ignoto characteri genere scriptum uidebatur *illyrico idiomate*, charactere quem D. Hieronymi uulgo uocant, impressum sciat; utunturque eodem charactere hic Romae in missalibus alijsque sacris libris illyrico sermone imprimendis" (emphasis ours). Philip Neal's translation is as follows: "Finally, I can let you know that the other sheet which appeared to be written in the same unknown script is printed in the Illyrian language in the script commonly called St Jerome's, and they use the same script here in Rome to print missals and other holy books in the Illyrian language." This letter text and Neal's translation are available at **http://www.voynich.nu/letters.html**.

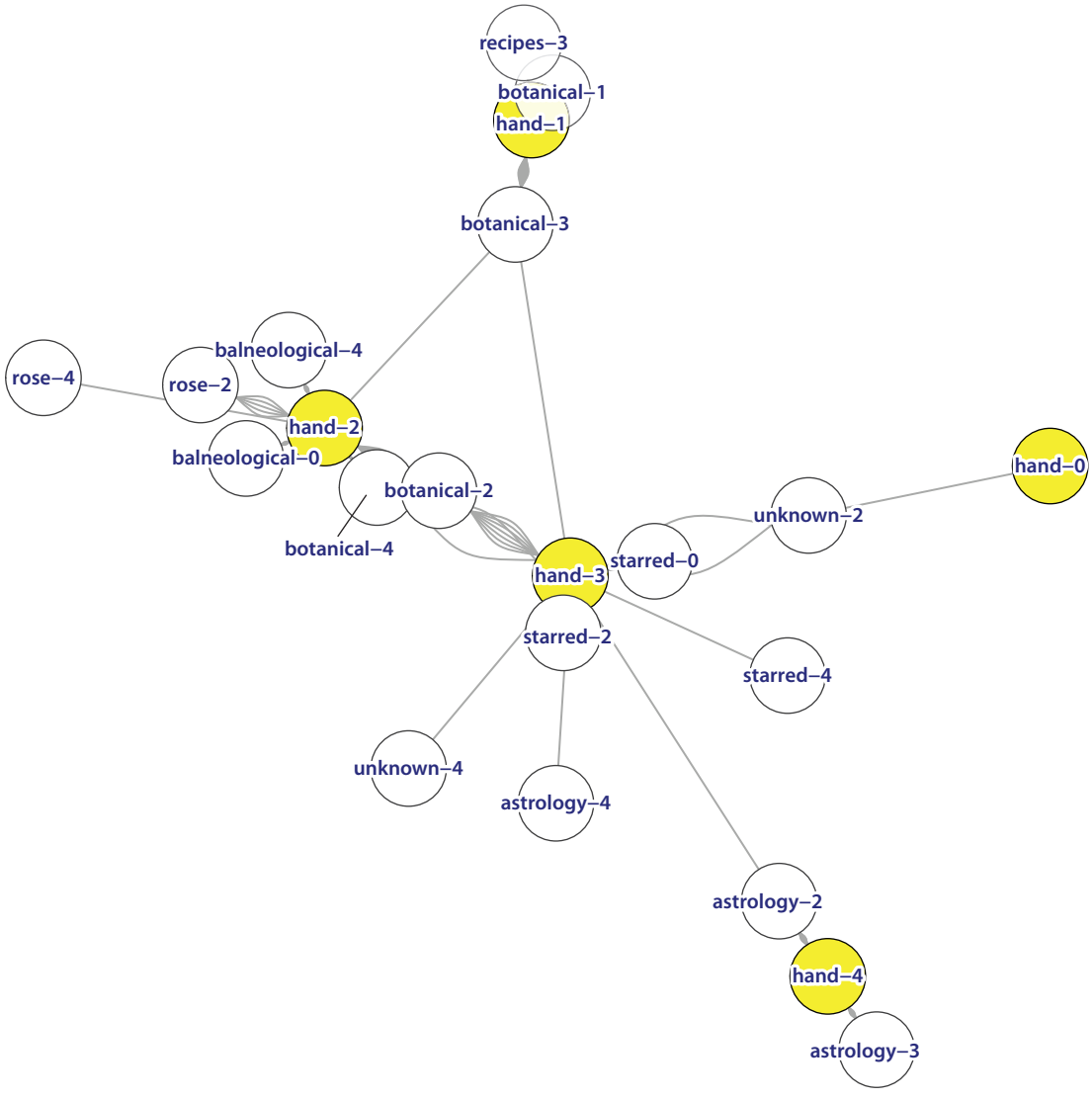


Figure 10

Network analysis of the manuscript hands (i.e., the scribes writing the pages; *yellow circles*) and the thematic sections of the manuscript (*open white circles*). The hand numbers are Davis's (2020); the subject numbers are the topics derived from the TF/IDF (term frequency–inverse document frequency) analysis. The links, illustrated by gray lines that represent the number of pages, show the association between hands and sections. For example, hand 4 is solely responsible for the astrology–3 topic; hand 3 contributes to astrology, starred paragraphs, and several botanical sections. However, where the association is very close, the circles overlap and the number of pages is less clear. Data from Sterneck & Bowern (2020).

In this section, we survey some of the theories regarding the language underlying Voynichese. We make no attempt to be comprehensive, and all of these theories have substantial conceptual problems.[10]

[10]Other claims not discussed in this review include Greek, Estonian, and various mixed-language hypotheses. Skinner et al. (2017, pp. 31–32) discuss and dismiss several other theories.

Current Voynich language theories fall into several common methodological traps. They "decode" based on a hunch. It is striking how many claimed decipherments emphasize how the solution came to the authors or stood out to them from cursory examination of the text. They exhibit a strong confirmation bias, omitting any information that does not fit the theory they propose. These theories also tend to present very little supporting evidence. For example, Gibbs's (2017) Latin "translation" was published with a single line, as was Hauer & Kondrak's (2016). When they discuss the data, they focus almost entirely on the lexicon and ignore morphology and syntax. This approach is problematic if the presumption is that the manuscript was written by authors fluent in the language.

5.1. Latin or Romance

Several previous claims are based around Voynichese being a Romance language, most likely a Latin cipher but possibly a vernacular variety. Cheshire (2018) argues that the language is "Proto-Romance," though it is unclear whether he intends the ancestral language of contemporary Romance languages or a lingua franca based on Romance sources. D'Imperio (1978) suggests that Latin is most likely, simply given the status of Latin as the language of learned discourse at the time.

Gibbs (2017), suggesting that the manuscript is a type of abbreviated Latin, decodes the plant pages as a set of recipes based on medieval medicinal shorthand. Only two lines have been published, and nothing about it can be regarded as convincing. Gibbs does not use abbreviations in the way that medieval Latin writers abbreviate; the Latin is itself not grammatically correct, and it does not generalize to the parts of the manuscript that are not about plants. In short, this analysis is deeply unconvincing.

5.2. Hebrew

The main work arguing for Hebrew underlying the Voynich Manuscript is by Hauer & Kondrak (2016), following an earlier suggestion by Reddy & Knight (2011) that the Voynich script is an abjad. Hauer & Kondrak (2016) assume that the Voynich Manuscript is written in a monoalphabetic substitution cipher; they also allow the possibility that it is written in a consonantal script (i.e., an abjad) and that there may be anagramming within words. To create an encryption key, the authors compare words by the frequency of repeated symbols within words (e.g., a word such as 'seems' has two *s* characters, two *e*'s, and one *m*). A substitution cipher based on optimized frequency matching with 380 languages suggests that the language with the closest distribution of anagrammed word patterns is Hebrew.

Hauer & Kondrak (2016) have attempted to decipher the first 10 pages of the manuscript on the basis of the anagrammed dictionaries they created. However, they have not been able to produce any sentences "that were grammatically correct or semantically consistent," either for Hebrew or other languages with pattern matches in the anagram dictionary (such as abjad Latin). We therefore also consider the Hebrew hypothesis not proven at best, and more accurately unconvincing.

5.3. Nahuatl

Another recent suggestion is based on Janick & Tucker's (2018a,b) interpretation of the plant images in the manuscript. Janick & Tucker point out a number of similarities between plants in Mexico and those in the Voynich Manuscript. They also claim that some plants are a better match for Mesoamerican plants than European ones because of details in the drawings. For example, they argue that the picture on folio 9v is a better fit to the American *Viola bicolor* than the European *Viola tricolor* (Tucker & Janick 2019, p. 183).

The details of the pictures notwithstanding, we consider this argument a nonstarter because of the carbon dating of the manuscript. The manuscript is simply too old for a Mesoamerican origin to be plausible. If there is an error in the carbon dating, the manuscript is likely to be older, not younger. The idea that the illustrations solidly reflect the Europe of the early fifteenth century but the plants from Mexico in the 1550s is simply implausible, especially when we know that the plant illustrations, with cubed roots and biologically impossible details, are unlikely to be intended as faithful representations.

The linguistic arguments are also poorly developed. No direct comparisons with Nahuatl are made in Tucker & Janick's (2018) earlier work, while in the later work (e.g., Tucker & Janick 2019) there are some superficial and unsystematic comparisons that take no account of Nahuatl grammar.

5.4. Bax's Unknown Language

Before his untimely death in 2017, Stephen Bax proposed the decipherment of five words. Bax did not name the language but suggested the terms were from a language spoken in Europe or the Middle East. Bax made available an 80-page document with his progress and thoughts toward decipherment (Bax 2014).

Bax's technique is rather different from the other linguistic approaches described above. Other authors have relied on what we might call inspiration; that is, they speak of an "aha" moment in which they get an idea about which language underlies the Voynich Manuscript. They then use various methods to find support for their intuition.

By contrast, Bax proceeds from labels and key terms and tries to use patterns in the label names to infer readings of the script. He assumes (for the purposes of decoding) that the grapheme and phoneme systems are isomorphic.

Bax's hypotheses proceed based on the reading of the label near the Pleiades and the constellation Taurus (folio 68v) *d/toari* ſoaꝛg, and two plant labels: *oror* oꝛoꝛ 'juniper' (folio 15v; cf. Arabic *arar*) and possibly 'coriander', which Bax reads as *kooratu* ſſccꝛoſaꝛ.

If Bax's provisional decipherments are correct, the language of the Voynich Manuscript is probably Indo-European (or at least in contact with languages of the region). However, Bax has nothing to say about some of the odder features of the manuscript encoding, such as the unusual conditional entropy values. His decipherment technique, like many others, takes the script at face value. Bax's use of labels relies on the assumption that the first word of text on a page is the label for the page. This may be a reasonable assumption, but on folio 15v the first word is not oꝛoꝛ but ſſoꝛoꝛ. It is also perhaps problematic that the illustration on folio 15v, as Bax notes, does not look at all like a juniper tree. If the phonemic equivalences put forward by Bax generalize to the rest of the manuscript, we should be able to read it.

6. SUMMARY

In summary, none of the arguments discussed in this review are proven or even particularly promising. Our work argues that the character-level metrics show Voynichese to be unusual, while the word- and line-level metrics show it to be regular natural language and within the range of a number of plausible languages. The higher structure of the manuscript itself is completely consistent with natural language and is very unlikely to be manufactured.

These observations imply that the script is not structure-preserving in that the graphemes are not one-to-one, but they do encode words in a regular orthography. Future work should focus on ciphers that have these properties and that also create predictability in the writing system. A

further fruitful area of analysis may be the textual variation among the different hands or among the same hands in different portions of the manuscript.

SUMMARY POINTS

1. Voynichese is most likely a natural language.
2. It is probably neither a simple substitution cipher nor a polyalphabetic code.
3. All current language claims are clearly problematic.
4. Nonetheless, there is a lot we can say about the language. We have shown that it patterns with languages with some morphology, but not extensive morphology, for example.

FUTURE ISSUES

1. An accurate transcription of the manuscript with a better understanding of the writing system is a high priority.
2. Analysis of scribal abbreviations and a better sense of how these affect text metrics is needed.
3. A better sense of the linguistic variation between hands is clearly called for.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank the members of the Yale undergraduate course "The Mystery of the Voynich Manuscript" (Spring 2018 and Fall 2019), Raymond Clemens, and Lisa Fagin Davis.

LITERATURE CITED

Amancio DR, Altmann EG, Rybski D, Oliveira ON, da Costa LF. 2013. Probing the statistical properties of unknown texts: application to the Voynich Manuscript. *PLOS ONE* 8:e67310

Barlow M. 1986. The Voynich Manuscript—by Voynich? *Cryptologia* 10:210–16

Bax S. 2014. *A proposed partial decoding of the Voynich script*. Work. Pap., Univ. Bedfordshire, Luton, UK

Bennett WR. 1976. *Scientific and Engineering Problem-Solving with the Computer*. Englewood Cliffs, NJ: Prentice Hall

Cappelli A. 1899. *Dizionario di abbreviature latine ed italiane usate nelle carte e codici specialmente del medio-evo*. Milan: Ulrico Hoepli

Cappelli A. 1982. *The Elements of Abbreviation in Medieval Latin Paleography*, transl. D Heimann, R Kay. Lawrence: Univ. Kans. Libr.

Cheshire GE. 2018. *Linguistically dating and locating the origin of Manuscript MS408*. Work. Pap., Univ. Bristol, Bristol, UK

Clemens R, ed. 2016. *The Voynich Manuscript*. New Haven, CT: Beinecke Rare Book Manuscr. Libr./Yale Univ. Press

Clemens R, Graham T. 2007. *Introduction to Manuscript Studies*. Ithaca, NY: Cornell Univ. Press

Clement RW. 1997. *Medieval and Renaissance book production*. Work. Pap., Utah State Univ., Logan. **https://digitalcommons.usu.edu/lib_pubs/10/**

Currier PH. 1976. Papers on the Voynich manuscript. *The Voynich Manuscript*, ed. R Zandbergen. **http://www.voynich.nu/extra/curr_main.html**

Daruka I. 2020. On the Voynich manuscript. *Cryptologia*. **https://doi.org/10.1080/01611194.2019.1706063**

Davis LF. 2020. How many glyphs and how many scribes? Digital paleography and the Voynich Manuscript. *Manuscr. Stud.* 5:164–80

D'Imperio ME. 1978. *The Voynich manuscript: an elegant enigma*. Tech. Rep., Natl. Secur. Agency/Cent. Secur. Serv., Fort George G. Meade, MD

Gheuens K. 2019. Type-token ratio. *The Voynich Temple Blog*, Apr. 5. **https://herculeaf.wordpress.com/2019/05/04/type-token-ratio/**

Gibbs N. 2017. Voynich manuscript: the solution. *Times Literary Supplement*, Sep. 8. **https://www.the-tls.co.uk/articles/voynich-manuscript-solution/**

Guy JBM. 1991. Statistical properties of two folios of the Voynich manuscript. *Cryptologia* 15:207–18

Häberl CG. 2015. Bālaybalan language. *Encyclopedia Iranica*. **https://www.iranicaonline.org/articles/balaybalan-language**

Hauer B, Kondrak G. 2016. Decoding anagrammed texts written in an unknown language and script. *Trans. Assoc. Comput. Linguist*. 4:75–86

Higley S. 2007. *Hildegard of Bingen's Unknown Language: An Edition, Translation, and Discussion*. New York: Palgrave Macmillan

Janick J, Tucker AO. 2018a. Cryptological analyses, decoding symbols, and decipherment. See Janick & Tucker 2018b, pp. 245–62

Janick J, Tucker AO, eds. 2018b. *Unraveling the Voynich Codex*. Cham, Switz.: Springer Int.

Kennedy G, Churchill R. 2006. *The Voynich Manuscript: The Mysterious Code That Has Defied Interpretation for Centuries*. Rochester, VT: Inn. Tradit. 3rd ed.

Koç M. 2005. *Bâleybelen Muhyî-i Gülsçnî: ilk yapma dil*. Istanbul: Klasik

Landini G. 2001. Evidence of linguistic structure in the Voynich manuscript using spectral analysis. *Cryptologia* 25:275–95

Laycock DC. 2001. *The Complete Enochian Dictionary: A Dictionary of the Angelic Language as Revealed to Dr. John Dee and Edward Kelley*. Boston: Weiser

Lindemann L, Bowern C. 2020. Character entropy in modern and historical texts: comparison metrics for an undeciphered manuscript. arXiv:2010.14697 [cs.CL]

Montemurro MA, Zanette DH, Menczer F, Munoz E, Somoza A. 2013. Keywords and co-occurrence patterns in the Voynich manuscript: an information-theoretic analysis. *PLOS ONE* 8:e66344

Reddy S, 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*, pp. 78–86. Stroudsburg, PA: Assoc. Comput. Linguist.

Rugg G. 2004. An elegant hoax? A possible solution to the Voynich manuscript. *Cryptologia* 28:31–46

Rugg G, Taylor G. 2016. Hoaxing statistical features of the Voynich Manuscript. *Cryptologia* 41:247–68

Shannon CE. 1949. *The Mathematical Theory of Communication*. Urbana: Univ. Ill. Press

Skinner S, Prinke RT, Zandbergen R, eds. 2017. *The Voynich Manuscript: The World's Most Mysterious and Esoteric Codex*. Sydney, Aus.: ReadHowYouWant

Stallings D. 1998. Understanding the second-order entropies of Voynich text. *Ixoloxi.com*, May 11. **http://ixoloxi.com/voynich/mbpaper.htm**

Sterneck R, Bowern C. 2020. *Topic modeling in the Voynich manuscript*. Work. Pap., Yale Univ., New Haven, CT

Stolfi J. 2000. *A grammar for Voynichese words*. **https://www.ic.unicamp.br/~stolfi/EXPORT/projects/voynich/00-06-07-word-grammar/**

Tiltman JH. 1967. *The Voynich Manuscript: the most mysterious manuscript in the world*. Talk presented at Baltimore Bibliophiles Meeting, Baltimore, MD, Mar. 4. **https://www.mgh-bibliothek.de/dokumente/b/b043214.pdf**

Tucker AO, Janick J. 2018. Origin and provenance of the Voynich Codex. See Janick & Tucker 2018b, pp. 3–39

Tucker AO, Janick J. 2019. *Flora of the Voynich Codex: An Exploration of Aztec Plants*. Cham, Switz.: Springer