The Linguistics of the Voynich Manuscript

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

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

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

Veterans Affairs, Office of Academic Affiliation, West Haven, Connecticut; email: luke.lindemann@yale.edu

September 8, 2020

Abstract

The Voynich Manuscript is a 15th 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 we can make statistical arguments about the phonology, morphology, and structure of the document, even though the contents remain undecipherable.

1 THE VOYNICH MANUSCRIPT

[1]

1.1 Introduction

Manuscript 408 in Yale University's Beinecke Library[2] — 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 and Churchill, 2004). While some argue that the work is a hoax (modern or medieval) or gibberish, others have advanced claims for what language the text is written in, or the type of cipher that may have been used to encode the text. There is at present no consensus on the language underlying the manuscript.

In this review, we survey linguistic aspects of the Voynich manuscript and summarize issues around decipherment. That is, 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.

[1] This is a preprint of a article forthcoming in the *Annual Review of Linguistics*.

[2] High resolution images of all pages are available from `https://brbl-dl.library.yale.edu/vufind/Record/3519597`.

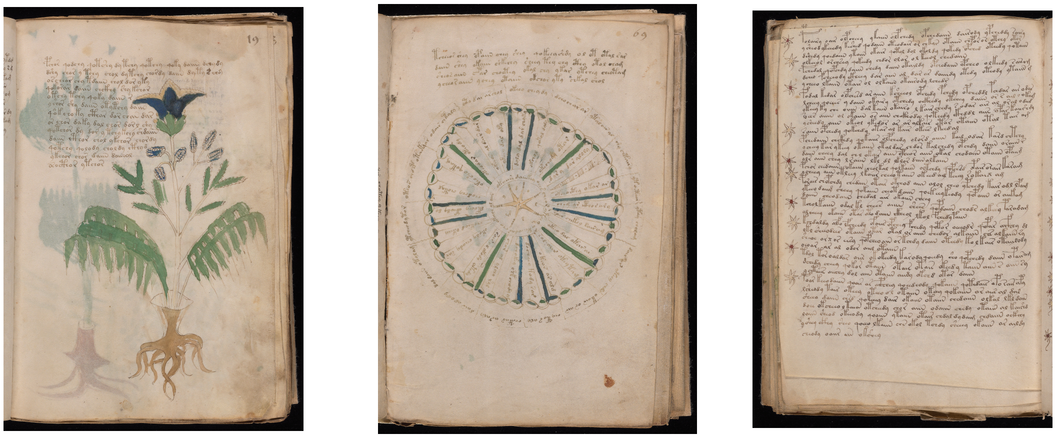


Figure 1: Pages 19r, 69v, and 105v of the manuscript, illustrating some of the subject themes. Credit: Beinecke Rare Book and Manuscript Library, Yale University

The manuscript itself is bound in vellum (the binding is old but not original). The Voynich Manuscript has 116 folios (that is, 232 pages), bound in 18 quires of varying size. Several pages are clearly missing. Thematically, the manuscript has 5 parts. The longest, covering over half the manuscript, contains drawings of plants (Folio 19r in Figure 1 is an example). The astrology section (e.g. 69v) contains astral charts, a zodiac, and other diagrams which may be related to astrology but which are not currently identifiable. The so-called 'balneological' section is almost unparalleled elsewhere; it is text illustrated by naked women in green water.[3] The fourth section shows illustrations of medical bottles and plant roots, while the fifth is unillustrated and consists of a set of paragraphs demarcated by stars (see folio 105v in Figure 1). Some pages are fold-outs. The text is written in iron-gall ink (typically 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 1911 or 1912. The villa housed a former Jesuit library and was sold in about 1874; presumably the manuscript was in the collection before that. The circumstances around the acquisition of the manuscript are unclear. Voynich was very circumspect at the time, and the manuscript does not appear to have the usual information about provenance that other manuscript purchases had, adding to the mystery surrounding the manuscript. Following Voynich's death, it passed to his wife, the novelist Ethel Voynich. The manuscript was donated to Yale's Beinecke Library in 1969 by H.P. Kraus. More information about these points can be found in Kennedy and Churchill (2004) and on the site `voynich.nu`.

The Voynich Manuscript may seem like a curious topic for a linguistics article, given that the language of the manuscript is unknown. We argue that it is an opportunity to address this problem by applying methods related to uncovering linguistic structure. That is, we use insights from typology, language documentation, and statistical arguments related to a wide variety of languages (ancient, medieval, and modern) to shed light on the properties

[3] Some 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 interpretable to anyone with a passing knowledge of childbirth, but the same cannot be said of the Voynich Manuscript's.

of the linguistic system underlying “Voynichese”. Moreover, much work that attempts to decipher the manuscript is claimed to do so on linguistic principles, and should be evaluated on linguistic principles as well (see §5 below).

In §1 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 §2 we give an overview of the script and phonology. In §3 we discuss the morphology (word-internal structure), and §4 discusses syntax and discourse structure of the Voynich manuscript. §5 discusses some of the current theories as to the language underlying the manuscript.

In our investigations of the Voynich manuscript, we do not take a traditional decipherment approach as had been used by early cryptogrpahic 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 which manipulate strings in various ways. While this is very unlikely to give us a direct language match, it does provide us with useful information about how sensitive tests are to both morphology and encipherment methods. This 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 see Lindemann and Bowern (2020).

In this article we concentrate on what we know about the language of the manuscript, and not about its thematic content or the historical context of its composition, though both are important. We also do not address the history of the manuscript following its composition or construction.[4]

We hope this overview will ground some of the discussions about the manuscript and guide plausibility of theories, as well as giving an example of the ways in which it can be useful to consider language documentary methods in ancient languages and cryptographic problems.

1.2 Natural language or gibberish

The first question about the manuscript for an article like this 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 rather unlike many other types of text (cf. Barlow, 1986). The linguistic status of the Voynich Manuscript has been the subject of controversy. At the same time as authors such as Reddy and Knight (2011) and Hauer and Kondrak (2016) work from principles of decipherment, others have claimed to show that the Voynich “language” is not language at all. Rugg (2004); Rugg and Taylor (2016) suggests that the language could be generated

[4] We recognize, however, that the circumstances of composition are very relevant to the language. For example, if MS 408 is a 15th century copy of a possibly much earlier work, that changes the languages which may likely underlie the manuscript text.

by a 15th Century cryptographic techniques, which could produce *either* enciphered text or gibberish. Rugg (2004) introduces explanation of the Cardan grille technique, while Rugg and Taylor (2016) suggests that the material is more compatible with hoax than enciphered language. Daruka (2020) and Timm and Schinner (2020) likewise comes to the conclusion that the Voynich Manuscript is a hoax and contains gibberish, though created by different means than those suggested by Rugg. As shown by Montemurro et al. (2013), Landini (2001), Reddy and Knight (2011), and Sterneck and Bowern (2020), however, the gibberish account does not explain the higher level document structure that we discuss further in §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.

We are not, at this stage, persuaded by any of the arguments in favor of Voynich being gibberish. In essence, because gibberish is by nature random, it should not display any of the higher level organizational properties that The Voynich Manuscript displays (as summarized here in §3.3 and §4). The Voynich Manuscript is highly unusual and non-language like at the character level. For measures that look above the word to line and paragraph, as well as in the distribution of words across the manuscript, it looks like a natural language. This strongly implies that the manuscript is encoded natural language rather than gibberish, since the measures used to track the paragraph structure are very unlikely to be directly manipulated and so 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 probably be ruled out. It is very unlikely to be a simple substitution cipher, for example, simply because it should have been deciphered by now if so. Conversely, polyalphabetic ciphers are unlikely both because they were probably not used in the early 15th century, and because a polyalphabetic cipher would lead to identical words being encoded differently in different parts of the manuscripts. We would not see the same (or similar) words being the most frequent on every page (see §3).

Finally, the manuscript could be in a constructed language ('conlang'). To our knowledge, the most extensive pre-modern conlang is the Turkish, Persian and Arabic-based language Balaibalan (Haeberl, 2015; Koc, 2005). Balaibalan was a Sufi esoteric language, represented in 3 manuscripts dating from roughly 1580 but likely a collective effort at composition over many years. There are two other well attested ancient conlangs, Latin and German-based *Lingua Ignota*, created by Hildegard von Bingen (Higley, 2007) in the 12th Century, and Enochian, an "angelic" English-like language invented in the 16th Century by Edward Kelly and John Dee (Laycock, 2001). All three conlangs are heavily based on natural languages and consist of embedding made-up roots 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 be anomalous at the root level.

1.3 Background to the manuscript

In the discussion that follows we assume some points which are still debated (to our minds unproductively) in Voynich studies. First, we proceed on the basis of the carbon dates of the vellum (and overall appearance and views of experts in medieval studies) that the manuscript is a genuine medieval object (Clement, 1997; Harkness, 2016) and not a modern forgery. Those who assume 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, yet been invented.

We also consider it unlikely that the The Voynich Manuscript is an ancient 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 with much less expense. Moreover, people who assume that the manuscript is a medieval gibberish hoax massively underestimate the amount of effort required to produce sustained language-like nonsense.[5]

The physical codex dates from 1404–1438 (Harkness, 2016) based on carbon dating but we do not know if 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. She 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 Currier (1976); Davis (2020); Reddy and Knight (2011), and much other work, we assume that there are two "languages" in the manuscript, referred to 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 them are because of different underlying languages or varieties (though this is possible). The main differences between them are in word frequency (which we discuss below). Certain character sequences are very common in Language A (*ol* and *or*) and relatively uncommon in Language B, and vice versa (*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' hands and Currier's languages, with scribe 1 writing in Language A and Scribes 2, 3, 4 and 5 in Language 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 instances of other orthographies: the Occitan month names in the Zodiac pages (*f*70v–73v), the partially obscured phrase in Latin script at the end of the manuscript (*f*116v), or the now invisible signature of de Tepenec on the first page (cf. Skinner et al., 2017). We note that page numbers were added after the composition of the manuscript but those numbers do not feature in our discussion here.

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

For the purposes of examining properties of Voynichese, we use a conventionalized 1:1 mapping between Voynich characters and ascii characters known as EVA or ‘Extended Voynich Alphabet’ (in Takahashi’s digital transcription; see Zandbergen 2020 and Table 1 below). The purpose of the transliteration is to make it possible to investigate the text using statistical methods. However, 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. At present, however, the Takahashi transcription is the only possibility for textual analysis using computational methods. 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 further Clemens and Graham 2007), numerals, and characters that have no counterpart in other manuscripts except perhaps as ornamental flourishes (see Cappelli 1899, plate x and Figure 2). Some Voynich characters are clearly part of the Latin alphabet (including , , ,). Others closely resemble numbers: cf. , , . Others appear to have no parallels in other scripts, including , , , . These characters are known as “gallows” characters in writings on the Voynich script (because of their superficial appearance to gallows). The best known parallel of these characters at present are the ornaments described in Cappelli (1899, Plate IV). While the ornamental characters in Cappelli are ligatures created from joining digraphs together, no such assumption can be made about the Voynich gallows characters at present. The other type of character is known as a “bench” , because it looks something like a bench.

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

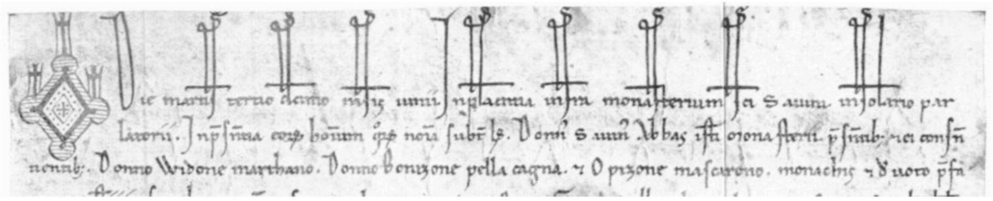


Figure 2: Gallows-like characters as in Cappelli (1899, Tavulo IV)

There are 21 or 22 common glyphs, plus approximately that many rarer character forms. The total number of glyphs in the dataset 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 each use different characters. We use EVA here because it is both the most widely used and the basis for the machine readable transliterations on which our statistic work in following sections depends.

EVA (an acronym for the ‘Extensible Voynich Alphabet’ (formerly European Voynich Alphabet) 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 to ASCII characters were based on Glen Claston's assumptions about what characters in the Latin alphabet the Voynich characters might be closest to (Zandbergen, 2020). For example, is transliterated in EVA as *d*. The transliteration equivalents are heavily influenced by European/Latinate considerations. For example, the sequence is represented as *qo* in large part because never appears in the text except before , which is reminiscent of the distribution of *q* never appearing except before *u* in Latin. But there is no other independent reason for assuming that Voynich is Latin script *q*.

trans	glyph	trans	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	

Table 1: List of most common Voynich characters

In our opinion, EVA probably under-differentiates characters (grouping together two variants of , for example). It also creates digraphs from characters that may be better understood as a single glyph. EVA ch , for example, is comprised of two distinct components: and (rendered in ascii as *ch*), even though is never found separately, and is otherwise identical to . It is, however, the transcription system which is used in the machine-readable version of the Voynich Manuscript, so we adopt it here pending a thorough review of the transcription system.

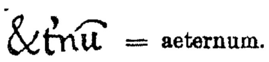
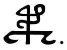


Figure 3: Cappelli (1982, 18): apostrophe used as an abbreviation character; compare Voynich

Some characters occur both in combination and separately in the manuscript. The "benches" character combines with the "gallows" characters to form , , and . The characters and appear to be distinguished only by the plume over the bench. This plume 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 display some underlying principle of regularity in the script.

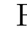
Finally, note that although the EVA transliterations makes Voynich material easy to pronounce, it has no basis beyond some similarity with some of the letters in the VMS script. For example, using gallows transcriptions as *f*, *k*, *t*, and *p* is purely a convention.

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, nor what the relationship between the orthography and the phonology (proper) is. However, we can draw some conclusions about the representation of the linguistic system in comparison to orthographic representations of other languages.

2.3.1 Vowels vs consonants

Both Reddy and Knight (2011) and Guy (1991) discuss whether the Voynich writing system is an alphabet, containing both vowels and consonants, or an abjad (representing consonants only). One test for this uses the Sukhotin algorithm (Guy, 1991). The algorithm 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 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 4 vowels: *a*, *o*, *e*, and *i* (that is, , , ,). It is worth noting that 3 of these characters (, ,) appear similar to characters which represent vowels in the Latin alphabet. The fourth, is the most common character in the Voynich alphabet.

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 instead , , , , . That is, is no longer identified as a vowel, and three characters that are almost exclusively word final are identified (, ,). Secondly, different characters are identified between the two Voynich Languages, A and B.

Reddy and Knight (2011) argue that Voynichese shows more properties of an abjad than an alphabetic script. They argue this in part from the induction of character classes from 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 the word versus the rest of the word. However, this result could be driven primarily by the characters which are only found word finally (such as , , , and). That is, the result may not indicate the Voynich script is an abjad, but rather that there are positional variants for character forms. Furthermore, though the Sukhotin algorithm does pick 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 Voynichese, almost all words in MS408 have one of the characters picked out by the Sukhotin algorithm as vowels.[6]

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, or "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 to each other and to Voynichese.

Bennett (1976) noted 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 (1):

(1)

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

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

In fact, the Voynich text in the EVA transcription is significantly lower than 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.[7] In Figure 4,

[6] Furthermore, character entropy, as discussed in §2.3.2 following, is usually higher for abjads than for alphabetic scripts. For Voynichese, the entropy is *lower*.

[7] 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 conditional entropy. All the lowest conditional entropy languages are written with alphabets.

Voynichese (A, B, and full sample with rare characters included) is much lower than any of the natural language samples.

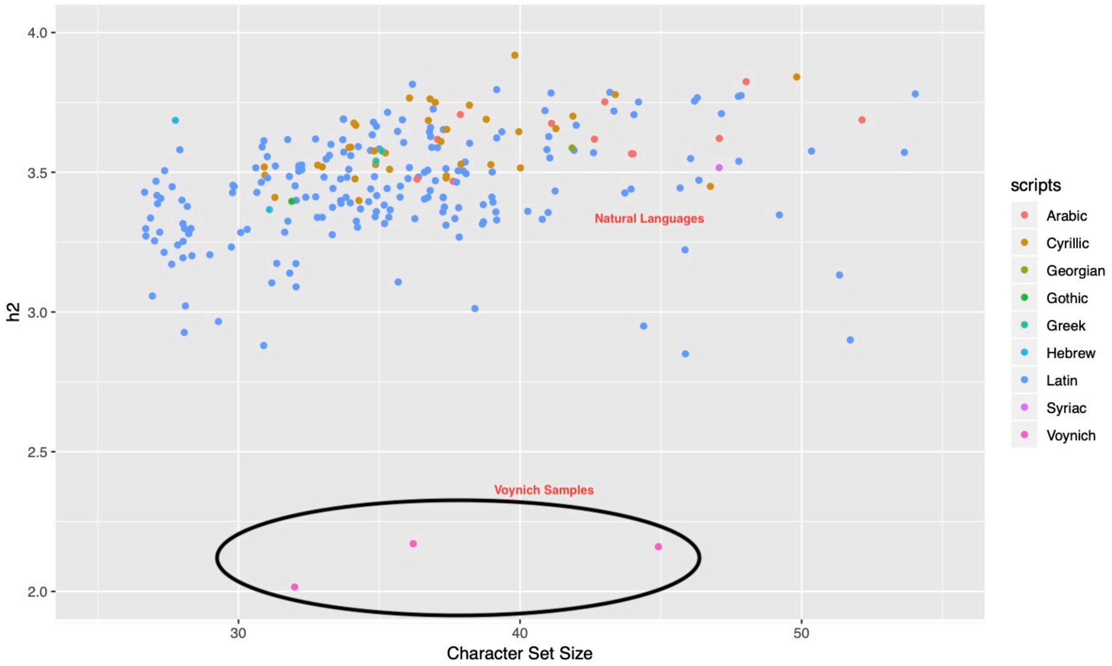


Figure 4: Conditional Character Entropy vs number of characters for Voynichese and other languages

The entropy of Voynichese is unlike any other language or script. Plausible manipulations of the script were investigated, including various shorthand abbreviations and devoweling the script. These do affect the character entropy, but not to the extent that would be required to bring Voynichese 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 based on dividing characters into whether they occur in the first or second half of the alphabet, or sorting all characters in the 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 4and o are treated as separate characters, 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 (3.4 for the abreviated text of the Latin *Secreta Secretorum*, rather than 3.2 for the plain text version of the same work).

Thus in summary, Voynichese has much lower conditional entropy than other texts and languages to which it was compared. It is not a function of script transliteration. However, it 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 other traditions consist of running text without spaces to indicate word boundaries. The Voynich manuscript contains spaces which separate the text into word-sized chunks. If we make the assumption that these chunks do in fact represent words, then we can investigate the morphology of Voynichese. That is, we can see if there is evidence for internal structure to Voynich words such as prefixes and suffixes which could 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.

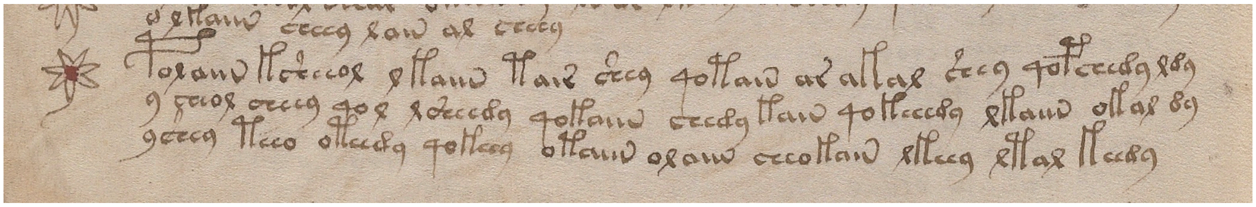


Figure 5: Example of a Voynich paragraph with word breaks. Credit: Beinecke Rare Book and Manuscript Library, Yale University

There is sometimes difficulty in telling where the breaks are. For example, in Figure 5 the last two characters of the middle line, , are more set off from the characters that come before them compared to the word directly below, which ends in the same sequence. We could therefore either treat as a separate (short) word or as part of the preceding word with an unusual gap. Despite ambiguities like this, however, there are consistent patterns about word structure which can be deduced with the assumptions that visual spaces are word breaks.

3.2 Structure in Voynich Words

Many Voynich characters and character combinations are restricted to certain parts of the word. Tiltman (1950) (quoted by D'Imperio, 1978) proposed that Voynich words consist of three separate "fields", with particular symbols occurring at either the beginning, middle or end of the word. The subsequent proposals and analyses of Roe (1997), Stolfi (2005), Reddy and Knight (2011), and Palmer (2014) differ in complexity and coverage, but they maintain this basic notion that there are separate fields for particular characters and character combinations.

Here are a few of the most common character combinations in each field:

1. Prefixes:[8] *qo-* , *o-* , *y-* , *ch-* , *sh-* , *d-*

[8] The first three prefixes are usually followed by a gallows character (*p*, t , k , f), whereas the others are not. Some words instead begin with a "bench-and-gallows" combination (cph , cth , ckh , cfh).

2. Roots/Midfixes: *p* , *t* , *k* , f , e , *ee* , *o* , *a*
3. Suffixes: *-y* , *-dy* , *-l* , *-r* , *-m* , *-iin* , *-in*

This structure is similar in all words in all sections of the manuscript and in both Language A and Language B. There is some minor variation in the frequency of particular affixes between A and B. Most significantly, Language B has a higher frequency of both the *qo-* ( ) prefix and the *-dy* ( ) suffix (they are about two times and three times more common, respectively).[9]

Unlike other textual traditions, many characters and character combinations are found exclusively in one of the three character fields. This is the ultimate source of the unusually low conditional character entropy (h2) in Voynich. The text is highly predictable because certain letters only occur in certain parts of the word 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 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 beginning (*q*) or end (*m* , *g* , *r* , *n*) of words. The closest example to this phenomenon in other texts are abjads like Arabic and Hebrew which contain a small number of glyph variants that only occur word finally. The distributions of characters in bigrams are given in Figure 6.

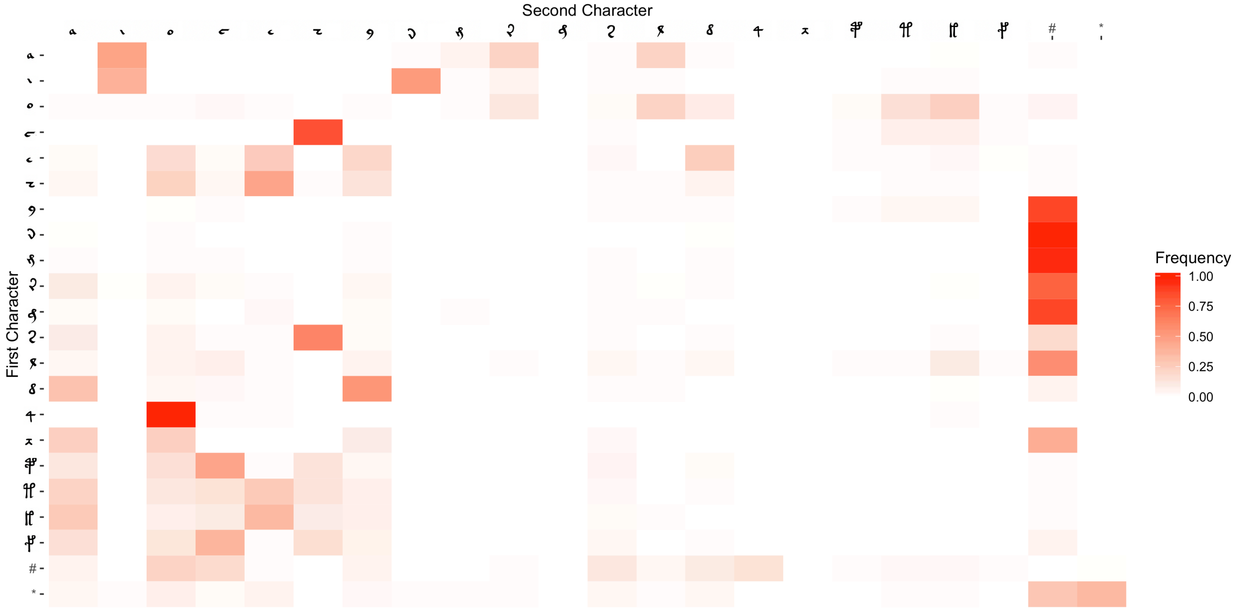


Figure 6: Frequency of each Voynich Character given the previous character

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* is

[9] Furthermore, as opposed to running text, labels tend to lack the *qo-* prefix, which may be evidence that it encodes a grammatical property like inflection or agreement.

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) noted that certain frequent word forms recur throughout the Voynich Manuscript. However, the most common words differ between Language A and Language B. Figure 2 gives the ten most common words in Voynich A and Voynich B, along with their proportional frequencies.

A			B		
daiin		4.5%	chedy		2.1%
chol		2.5%	ol		1.8%
chor		1.6%	shedy		1.8%
s		1.4%	aiin		1.5%
dy		1.1%	daiin		1.4%
shol		1.0%	qokeedy		1.3%
sho		0.9%	qokain		1.2%
chy		0.9%	qokedy		1.2%
cthy		0.9%	qokeey		1.1%
ol		0.9%	chey		1.0%
TOTAL		15.7%	TOTAL		14.5%

Table 2: Most common Voynich words in Languages A and B

While 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 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 and Knight (2011) note that the distribution of words in the Voynich manuscript follows Zipf's Law. This is a power law that relates the frequency of a word with 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. Figure 7 compares the distribution curves of the first 100 words in Voynich and four other languages, with frequency 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* and *shedy* represent the same word, as they are distinguished only by whether there is a plume stroke over the bench character. If we make

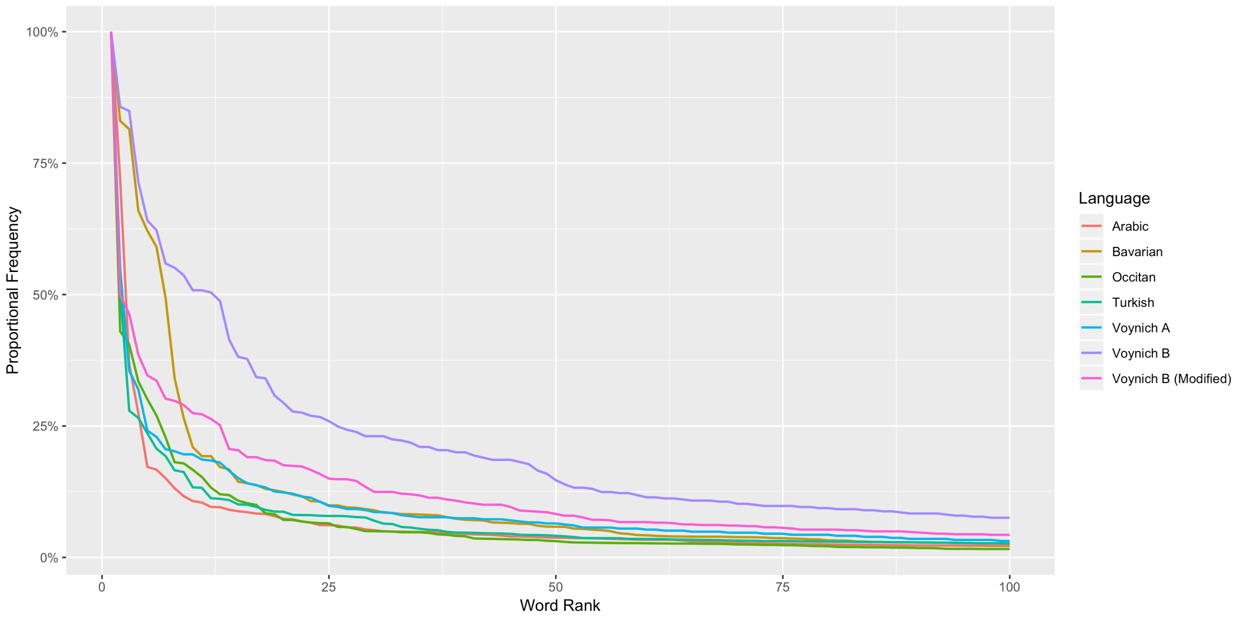


Figure 7: Word Frequency Distribution of Voynich and Selected Languages

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. But the word distribution does not look highly unusual in comparison to natural languages, which we might expect if the text is naively created gibberish. As Reddy and Knight comment, a Zipfian distribution "is a necessary (though not sufficient) test of linguistic plausibility."

Furthermore, if we believe Voynich to be 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 which rotate alphabets will flatten the frequency of the most common lexical items, since those words will be enciphered differently on different pages. Whatever the method used to encipher the The Voynich Manuscript languages, it is most likely to be 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*, accounts for 4.5% of the words in that text, while the most common word in Voynich B, *chedy*, 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'/'of'. In Voynich A, *daiin* is never found at the beginning of a paragraph, which may suggest that it is a connective.

The proportional frequency of the top ten most common words together is also within the range we find with natural languages. For Voynich A this is 15.7% and for Voynich B this is 14.5%. Figure 8 shows this for 101 languages in seven different language families. The red lines show one standard deviation from the mean.

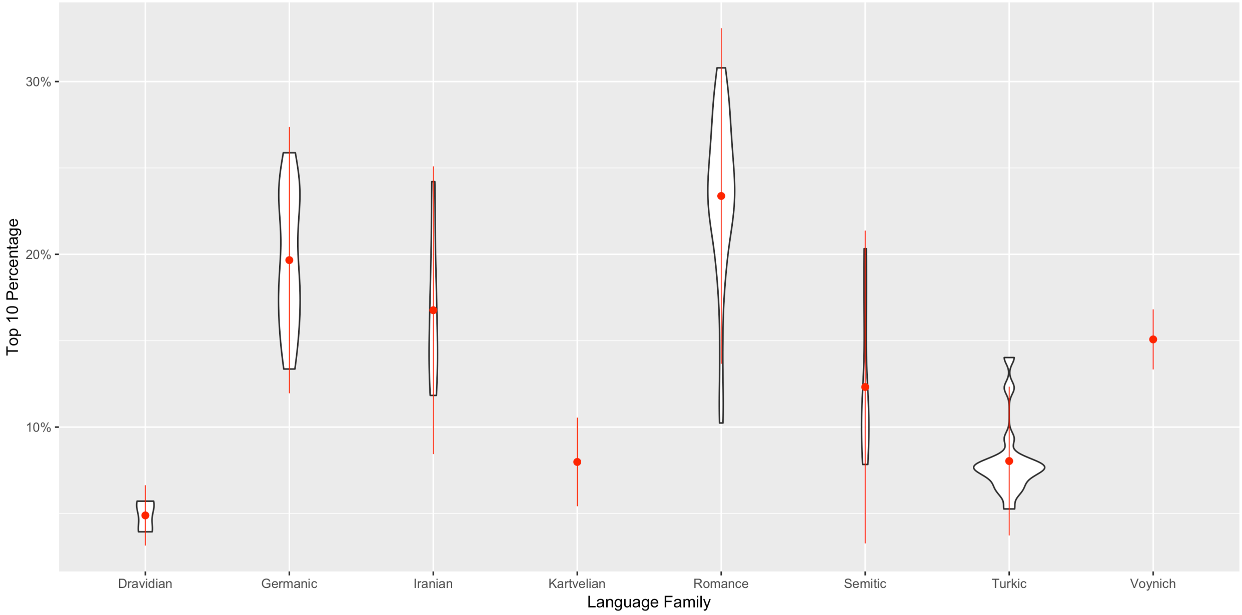


Figure 8: Proportional frequency of the ten most frequent words by Language Family

The Voynich languages are within the range of each of these families, and closest to the average for Semitic, Iranian, and Germanic. Note that there is an inverse correlation between this statistic and morphological complexity. The percentage will tend to be lower with languages that have many words with complex morphological structure, as with languages in the Turkic, Kartvelian, and Dravidian families. It will tend to be higher in languages with less morphological complexity, as 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 (MATTR)

Another statistic which 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) introduced the Moving Average Type Token ratio (MATTR) index, which takes the average TTR over a set word window, as a way to measure lexical diversity irrespective of text length. Gheuens (2019) examined MATTR in Voynichese compared with a sample of language texts, and concluded that 2000 is an ideal window. We have also found 2000 words to be a good window for comparing MATTR in language families.

Figure 9 shows the MATTR summaries across language families. Here the families with higher morphological complexity, like Dravidian and Kartvelian, have higher MATTR ratios.

The Voynich texts are once again in the medium range, and closest to Iranian, Germanic, and Romance.

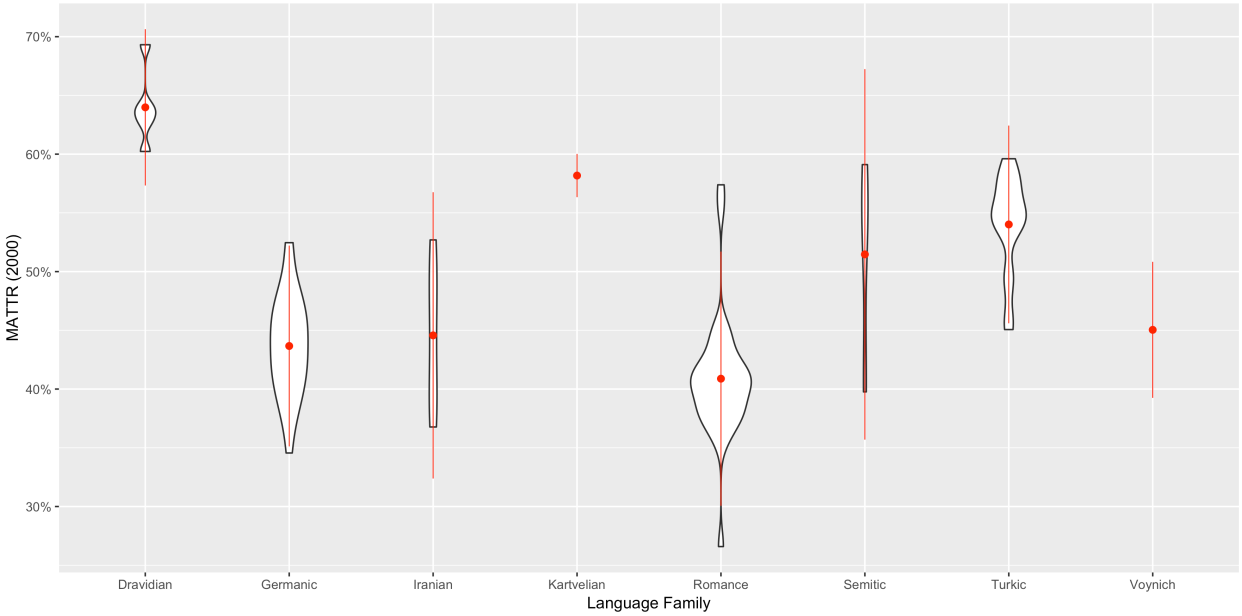


Figure 9: MATTR by Language Family

The MATTR and proportional frequency measures provide distinct and largely complementary evidence that Voynichese represents a language with medium morphological complexity. Voynich most closely resembles the averages for Germanic and Iranian, and least resembles those for Turkic, Dravidian, and Kartvelian. As with the Zipfian word distribution, we find Voynich to be well within the expected values for natural language texts, and far from random 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 Voynich. They are also largely independent of character-level parsing and transcription issues. Assuming 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 text that appears as labels on drawings and running text. In the running text there are word combinations and Voynich "phrases" which give evidence of syntactic structure. The text is written line-by-line, which may roughly equate to a sentence (Currier, 1976). Blocks of running text are separated into paragraphs, which could indicate a shift in topic. Above this level, each folio can be examined separately. The distinction between Voynich A and Voynich B is made at the folio level, as are (with one exception) the Voynich hands (Davis, 2020). Finally, there are the different sections of the manuscript which are distinguished by subject matter as evident by the illustrations or diagrams depicted on the page.

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) argued 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 the 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 certain types of words will tend to appear 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 like 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 this case we should expect to find pairs of similarly patterning Voynich words which 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 the paragraph, which usually begins with a gallows character. 85% of the paragraphs in the text begin with one of , , , . These "gallows-initial" words are (1) otherwise fairly infrequent; and (2) have the same structure as normal Voynich words except that they are preceded by a gallows character. John Grove first hypothesized 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 appear to be part of the word itself. They simply mark the beginning of the paragraph. Furthermore gallows-initial words, when they do appear elsewhere, usually begin with *k* or *f* rather than *p* or *t* . This may suggest sub-ordering of elements within the paragraph itself.

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 appears to only occur 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 which usually appear at the end of the last word of the line, particularly *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, however, particularly with the word *daiin*, which is common in every position except paragraph-initially.[10] The table 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 B are more commonly marked by the *p* gallows.

	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 ()

Table 3: Most common Voynich words by Position

All of these observations lead to generalizations which appear to be typographical rather than linguistic in nature. Voynich writing does not appear to have any conventional punctuation symbols, but rather 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 that words fit together in a hierarchical structure, and generalizations about word and phrase combinations can explicate this structure. This has been studied less systematically than character and word-level patterns in the Voynich Manuscript, where the focus has been on character interpretation or hypotheses about whether Voynich represents a natural language. Here we can simply present a few observations and their potential implications for syntactic structure.

Stolfi (2005) points to the repetitive nature of the Voynich text as evidence that it is not meaningful. Take for example the following line from Voynich B:

(2) keedy qokeedy qokey okar otar dar dar dy.

[10] *daiin* is never found as the first word of the paragraph, and there are only two possible examples of *pdaiin*.

This repetitiveness 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 the 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%.[11]

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

The next generalization holds for both languages, and involves the fourth most common word overall, *aiin* . This word is usually preceded by a short one or two-letter word (e.g. *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 are 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 multi-word structure involves gallows characters, which are most commonly preceded by *o-* (e.g. *okeedy, otaiin, opchy, ofchedy* , , ,). These words are prevalent on labels, and they occur with roughly the frequency we expect to find nouns in the text. Furthermore, there are four gallows characters, two common (*k/t*) and two uncommon (*p/f*), which led us to 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-* to be feminine singular, *ot-* to be masculine singular, *op-* to be feminine plural, and *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 born out. All roots pattern roughly the same, and most are found with every possible combination: *okaiin, otaiin, opaiin, ofaiin* , , , . Therefore, the hypothesis that the elements in question represent an article classification similar to Romance is untenable.

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

[11] However, the average for most relevant language families is somewhat smaller: Germanic is 0.37%, Romance is 0.36%, Iranian is 0.25%, and Semitic is 0.36%.

4.3 Topic

Reddy and 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 and Knight (2011) show that the 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 the words which are most likely to contribute to topics in texts. That is, they identify words which are more uniformly distributed throughout the The Voynich Manuscript and compare them to those which tend to cluster. Those which tend to cluster are more likely to provide information about the subject matter of the pages. Their method also returns an overall similarity between the pages with herbal and pharmacological illustrations, suggesting 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 The Voynich Manuscript and conclude that the Manuscript is most likely natural language thematic content.

Sterneck and Bowern (2020) further investigated topic modeling within the Voynich text and the relationships between Languages, scribal hands, and thematic material (as deduced from the illustrations). They used TF-IDF[12] weighted counts using 40 word chunks of text within each page; they use Nonnegative Matrix Factorization (NMF) topic clustering to cluster Voynich pages, and to compare those clusters to other types of structure in the document.

Using different methods from Amancio et al. (2013) and Reddy and Knight (2011), they were able to recover both general thematic topics and identify correlations between topics and hands within thematic sections. That is, the pages that Davis (2020) identified as being written by a different scribe also tend to emerge as a different 'topic' in the TD-IDF analysis. Figure 10 illustrates the topic, hand, and section clustering.

This 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 MS408. The manuscript was undecipherable even in the 17th Century. Athanasius Kircher[13] (Neal, 2017) thought it was likely written in the Glagolitic script ("Illyrian"), and

[12] TF-IDF stands for Term Frequency - Inverse Document Frequency is a statistic used to cluster text based on the frequency of words in the text itself (the TF) as compared to the frequency of that word in the document as a whole. It allows us to group together texts based on distinctive words. Because Voynich pages are texts of different lengths, Sterneck and Bowern (2020) normalized the text length for each page.

[13] 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.* Neal's 2017 is as follows: "Finally, I can let you know that the other sheet which appeared to be written in the same unknown

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, it is not an unreasonable guess. Reddy and Knight (2011) suggest that aspects of the script is reminiscent of an abjad, which would suggest a Semitic language (since all abjads currently known are used to write Semitic languages). This was the inspiration for Hauer and Kondrak (2016), who claim that the manuscript is in an enciphered and anagrammed Hebrew; their solution, however, is not generally accepted.

In this section we survey some of the theories regarding the language underlying Voynichese. We make no attempt to be comprehensive, and all have substantial conceptual problems.[14]

Current Voynich language theories fall into a number of 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. Secondly, these theories typically present very small amounts of data. For example, Gibbs' (2017) Latin "translation" was published with a single line, as was Hauer and Kondrak (2016). When they discuss the data, they focus almost entirely on the lexicon, ignoring morphology and syntax. This 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 Romance, 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) suggested that Latin is most likely, simply given the status of Latin as the language of learned discourse at the time.

Gibbs (2017) suggests that the manuscript is a type of abbreviated Latin. That is, he decodes the plant pages as a set of recipes based on medieval medicinal shorthand. Only two lines have been published, and nothing about it an 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 doesn't generalize to the parts of the manuscript which are not about plants. In short, it is deeply unconvincing.

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." Both are given at http://www.voynich.nu/letters.html

[14] Other claims include Greek, Estonian, and various mixed language hypotheses. Skinner et al. (2017, 31–32) discusses and dismisses several other theories.

5.2 Hebrew

The main work arguing for Hebrew underlying the Voynich Manuscript is Hauer and Kondrak (2016), following an earlier suggestion by Reddy and Knight (2011) that the Voynich script is an abjad. Hauer and 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 (that is, an abjad) and that there is possible anagramming within words. In order 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 and Kondrak (2016) attempted to decipher the first ten pages of the manuscript based on the anagrammed dictionaries they created. They were not, however, 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 the interpretation of the plant images, by Janick and Tucker (2018a,b). They 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 (Tucker and Janick, 2019, 183) that the picture on page *9v* is a better fit to the American *Viola bicolor* than the European *Viola tricolor*.

The details of the pictures notwithstanding, we consider this argument a non-starter due to the carbon dating of the manuscript. The manuscript is simply too early for a Mesoamerican origin to be plausible. If there is an error in the carbon dating, it's likely to be older, not younger. The idea that the illustrations solidly reflect Europe of the early 15th century, but the plants from Mexico in 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.

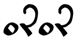
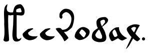
The linguistic arguments are also poorly developed. No direct comparisons with Nahuatl are made in Tucker and Janick (2018), while in the later work (e.g. Tucker and Janick, 2019) there are some superficial and unsystematic comparisons which take no account of Nahuatl grammar.

5.4 Bax's unknown language

The final language to be considered here is Stephen Bax's approach and proposed decipherment of 5 words. Before his untimely death in 2017, Bax made available an 80 page document with his progress and thoughts towards decipherment (Bax, 2014).

Bax's technique was 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 where they get an idea about which language underlies the Voynich manuscript. They then use various methods to find support for that assumption.

Bax, on the other hand, proceeds from labels and key terms, trying 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 on Aldebaran, the constellation Taurus (on *68v*) *d/toari* , and two plant labels: *oror*  "juniper" (on *15v*; cf. Arabic *arar*) and possibly "coriander", which Bax reads as *kooratu* .

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 more odd features of the manuscript encoding, such as the very 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 the text on a page is the label for the page. This may be a reasonable assumption, but on *15v* the first word is not but . It is also perhaps problematic that the illustration on 15v, as Bax notes, doesn't 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 these arguments are either proven or even particularly promising.

Our work argues that the character level metrics show Voynich 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.

This therefore implies 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 which have these properties, and in addition create predictability in the writing system. A further fruitful area of analysis may be the textual variation between the different hands, or the same hands in different portions of the manuscript.

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

Acknowledgements, general annotations, funding. Thanks to the members of the Yale Undergraduate class "The mystery of the Voynich Manuscript" (Spring, 2018 and Fall, 2019), Ray Clemens and Lisa Fagin Davis.

References

Amancio, Diego R., Eduardo G. Altmann, Diego Rybski, Osvaldo N. Oliveira and Luciano F. da Costa. 2013. Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript. *PLoS ONE* 8(7). e67310–e67310. doi:10/f48qmr. URL http://dx.plos.org/10.1371/journal.pone.0067310.

Barlow, Michael. 1986. the Voynich Manuscript - By Voynich? *Cryptologia* 10(4). 210–216. doi:10/bz7md2. URL http://dblp.uni-trier.de/db/journals/cryptologia/cryptologia10.htmlBarlow86a.

Bax, Stephen. 2014. A proposed partial decoding of the Voynich script. *University of Bedfordshire* .

Bennett, William Ralph. 1976. *Scientific and engineering problem-solving with the computer*. Prentice Hall PTR.

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

Cappelli, Adriano. 1982. *The elements of abbreviation in medieval Latin paleography*. University of Kansas libraries.

Cheshire, Gerard Edward. 2018. Linguistically dating and locating the origin of Manuscript MS408. University of Bristol, Science Survey 2.

Clemens, Raymond and Timothy Graham. 2007. *Introduction to manuscript studies*. Cornell University Press.

Clement, Richard W. 1997. Medieval and Renaissance book production. *Library Faculty and Staff Publications* 10. URL http://digitalcommons.usu.edu/lib$_p$ubs.

Currier, Captain Prescott H. 1976. Papers on the Voynich manuscript. URL http://www.voynich.nu/extra/img/curr$_m$ain.pdf.

Daruka, István. 2020. On the Voynich manuscript. *Cryptologia* 0(0). 1–37. doi: 10.1080/01611194.2019.1706063.

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

D'Imperio, Mary E. 1978. The Voynich manuscript: an elegant enigma. Tech. Rep. 9780894120381. URL http://oai.dtic.mil/oai/oai?verb=getRecordmetadataPrefix=htmlidentifier=ADA070618.

Gheuens, Koen. 2019. Type-token ratio. URL https://herculeaf.wordpress.com/2019/05/04/type-token-ratio/. Examining lexical diversity in the Voynich manuscript.

Gibbs, Nicholas. 2017. Voynich manuscript: The solution. *The Times Literary Suppplement* .

Guy, Jacques B. M. 1991. Statistical properties of two folios of the Voynich manuscript. *Cryptologia* 15(3). 207–218. doi:10/d85gfc. URL `http://www.tandfonline.com/doi/abs/10.1080/0161-119191865876`.

Haeberl, C. G. 2015. Balaybalan language. *Iranicaonline.org* URL `http://www.iranicaonline.org/articles/balaybalan-language`.

Harkness, Deborah E. 2016. *The Voynich manuscript.* Beinecke Rare Book & Manuscript Library in Association with Yale University Press.

Hauer, Bradley and Grzegorz Kondrak. 2016. Decoding Anagrammed Texts Written in an Unknown Language and Script. *Transactions of the Association for Computational Linguistics* 4. 75–86. doi:10/ggdfd6. URL `https://www.mitpressjournals.org/doi/abs/10.1162/tacl`$_{a0}$`0084`.

Higley, S. 2007. *Hildegard of Bingen's unknown language: An edition, translation, and discussion.* Springer. Google-Books-ID: LZGJDAAAQBAJ.

Janick, Jules and Arthur O. Tucker. 2018a. Cryptological Analyses, Decoding Symbols, and Decipherment. In *Unravelling the Voynich codex*, 245–262. Springer. doi:10.1007/978-3-319-77294-3$_1$1.

Janick, Jules and Arthur O. Tucker. 2018b. *Unraveling the Voynich Codex*, Fascinating Life Sciences. Cham: Springer International Publishing. doi:10.1007/978-3-319-77294-3. URL `http://link.springer.com/10.1007/978-3-319-77294-3`.

Kennedy, Gerry and Rob. Churchill. 2004. *The Voynich manuscript: The mysterious code that has defied interpretation for centuries.* Orion publishing company.

Koc, M. 2005. *Bâleybelen: İlk yapma dil.* Istanbul.

Landini, Gabriel. 2001. Evidence of linguistic structure in the Voynich manuscript using spectral analysis. *Cryptologia* 25(4). 275–295. doi:10/dhtvfp.

Laycock, Donald C. 2001. *The complete Enochian dictionary: A dictionary of the Angelic language as revealed to Dr. John Dee and Edward Kelley.* Weiser Books.

Lindemann, Luke and Claire Bowern. 2020. Voynich manuscript linguistic analysis. *arxiv.org* .

Montemurro, Marcelo A., Damián H. Zanette, F Menczer, E Munoz and AM Somoza. 2013. Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis. *PLoS ONE* 8(6). e66344–e66344. doi:10/m2t. URL `http://dx.plos.org/10.1371/journal.pone.0066344`.

Neal, Philip. 2017. Voynich sources. Http://philipneal.net/voynichsources/.

Reddy, Sravana and Kevin Knight. 2011. What we know about the Voynich manuscript. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 78–86. URL http://dl.acm.org/citation.cfm?id=2107647.

Rugg, Gordon. 2004. An elegant hoax? A possible solution to the Voynich manuscript. *Cryptologia* 28(1). 31–46. doi:10/bf6pwp. URL http://www.tandfonline.com/doi/abs/10.1080/0161-110491892755.

Rugg, Gordon and Gavin Taylor. 2016. Hoaxing statistical features of the Voynich Manuscript. *Cryptologia* 1–22. doi:10/bq22. URL https://www.tandfonline.com/doi/full/10.1080/01611194.2016.1206753.

Shannon, Claude E. 1949. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Skinner, Stephen, Rafał T Prinke and René Zandbergen (eds.). 2017. *The Voynich manuscript: The world's most mysterious and esoteric codex*. ReadHowYouWant.

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

Sterneck, Rachel and Claire Bowern. 2020. Topic modeling in the Voynich manuscript. URL arxiv.org.

Stolfi, Jorge. 2005. Voynich Manuscript Stuff. http://www.dcc.unicamp.br/stolfi/voynich. [Online; accessed 24-April-2018].

Tiltman, John H. 1950. The Voynich manuscript: The most mysterious manuscript in the world. URL http://www.mgh-bibliothek.de/dokumente/b/b043214.pdf.

Timm, Torsten and Andreas Schinner. 2020. A possible generating algorithm of the voynich manuscript. *Cryptologia* 44(1). 1–19. doi:10/gg8ff6.

Tucker, Arthur O. and Jules Janick. 2018. Origin and Provenance of the Voynich Codex. In *Unravelling the Voynich codex*, 3–39. Springer. doi:10.1007/978-3-319-77294-3_1.

Tucker, Arthur O. and Jules Janick. 2019. *Flora of the Voynich codex: An exploration of aztec plants*. Springer Nature.

Zandbergen, Rene. 2020. Voynich ms - text analysis - transliteration of the text. Http://www.voynich.nu/transcr.html.

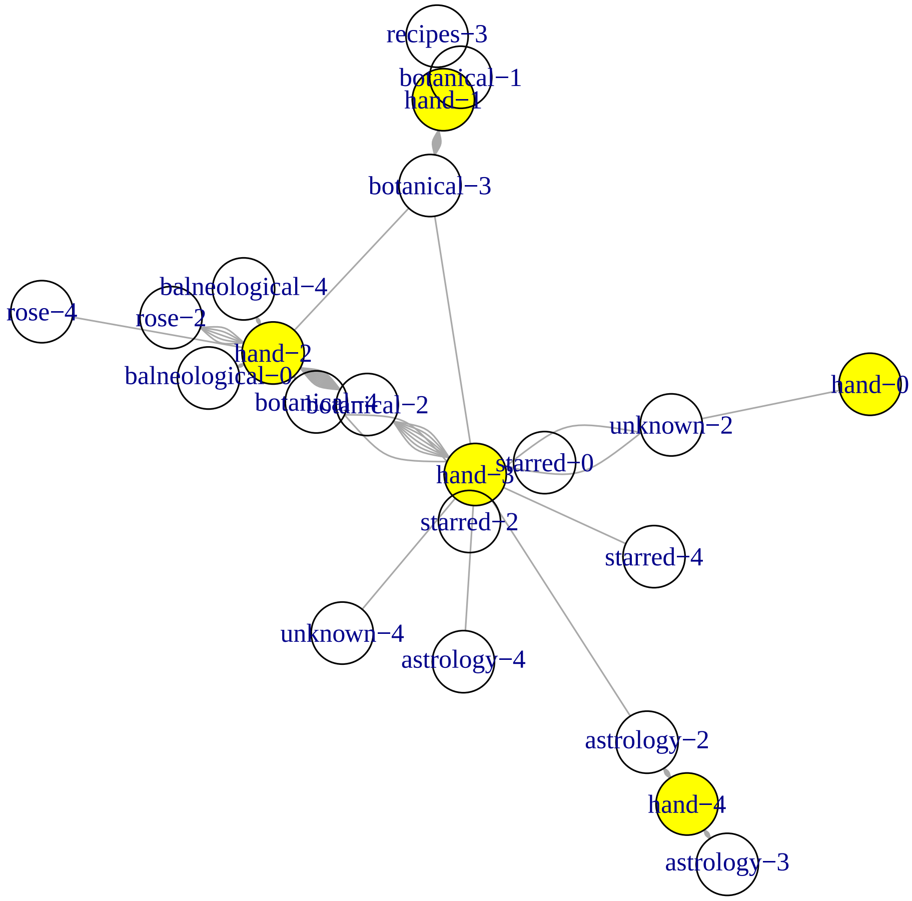


Figure 10: Topic: TF/IDF generated topics matching manuscript hands and illustrated subjects. Figure 10 illustrates a network analysis of the manuscript hands (that is, the scribes writing the pages; in yellow circles) and the thematic sections of the manuscript (in clear circles). The hand numbers are Davis'; the subject numbers are the topics derived from the TF/IDF analysis. The links 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.