What We Know About The Voynich Manuscript

Sravana Reddy*
Department of Computer Science
The University of Chicago
Chicago, IL 60637
sravana@cs.uchicago.edu

Kevin Knight
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
knight@isi.edu

Abstract

The Voynich Manuscript is an undeciphered document from medieval Europe. We present current knowledge about the manuscript's text through a series of questions about its linguistic properties.

1 Introduction

The Voynich manuscript, also referred to as the VMS, is an illustrated medieval folio written in an undeciphered script.

There are several reasons why the study of the manuscript is of interest to the natural language processing community, besides its appeal as a long-enduring unsolved mystery. Since even the basic structure of the text is unknown, it provides a perfect opportunity for the application of unsupervised learning algorithms. Furthermore, while the manuscript has been examined by various scholars, it has much to benefit from attention by a community with the right tools and knowledge of linguistics, text analysis, and machine learning.

This paper presents a review of what is currently known about the VMS, as well as some original observations. Although the manuscript raises several questions about its origin, authorship, the illustrations, etc., we focus on the *text* through questions about its properties. These range from the level of the letter (for example, *are there vowels and consonants?*) to the page (*do pages have topics?*) to the document as a whole (*are the pages in order?*).

* This work was completed while the author was visiting the Information Sciences Institute.

2 Background

2.1 History

From the illustrations – hairstyles and features of the human figures – as well as the shapes of the glyphs, the manuscript is posited to have been created in Europe. Carbon-dating at the University of Arizona has found that the vellum was created in the 15^{th} century, and the McCrone Research Institute has asserted that the ink was added shortly afterwards[1].

The exact history of the VMS is not established. According to Zandbergen (2010), the earliest owner that it can be traced to is Jacobus de Tepenec in Prague in the early 1600s. It is speculated that it was given to him by Emperor Rudolf II, but it is unclear how and from where the manuscript entered Prague.

The VMS appears to have circulated in Prague for some time, before being sent to Athanasius Kircher in Italy in 1665. It remained in Italy until 1912, when it was sold to Wilfrid Voynich, who brought it to America. It was then sold to the bookdealer Kraus, who later donated it to the Yale University library[2], where it is currently housed.

2.2 Overview

The manuscript is divided into *quires* – sections made out of folded parchment, each of which consists of *folios*, with writing on both sides of each folio (Reeds, 2002). Including blank pages and pages with no text, there are 240 pages, although it is believed that some are missing (Pelling, 2006). 225

[1] These results are as yet unpublished. A paper about the carbon-dating experiments is forthcoming in 2011.

[2] High-resolution scans are available at http://beinecke.library.yale.edu/digitallibrary/voynich.html

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 78–86,
Portland, OR, USA, 24 June 2011. ©2011 Association for Computational Linguistics

pages include text, and most are illustrated. The text was probably added after the illustrations, and shows no evidence of scratching or correction.

The text is written left to right in paragraphs that are left-aligned, justified, and divided by whitespace into words. Paragraphs do not span multiple pages.

A few glyphs are ambiguous, since they can be interpreted as a distinct character, or a ligature of two or more other characters. Different transcriptions of the manuscript have been created, depending on various interpretations of the glyphs. We use a machine-readable transcription based on the alphabet proposed by Currier (1976), edited by D'Imperio (1980) and others, made available by the members of the Voynich Manuscript Mailing List (Gillogly and Reeds, 2005) at http://www.voynich.net/reeds/gillogly/voynich.now. The Currier transcription maps the characters to the ASCII symbols A-Z, 0-9, and *. Under this transcription, the VMS is comprised of 225 pages, 8114 word types, and 37919 word tokens. Figure 1 shows a sample VMS page and its Currier transcription.

2.3 Manuscript sections

Based on the illustrations, the manuscript has traditionally been divided into six sections: (1) *herbal*, containing drawings of plants; (2) *Astronomical*, containing zodiac-like illustrations; (3) *Biological*, mainly containing drawings of female human figures; (4) *Cosmological*, consisting of circular illustrations; (5) *Pharmaceutical*, containing drawing of small containers and parts of plants, and (6) *Stars* (sometimes referred to as *Recipes*), containing very dense text with drawings of stars in the margins.

Currier (1976) observed from letter and substring frequencies that the text is comprised of two distinct 'languages', A and B. Interestingly, the Biological and Stars sections are mainly written in the B language, and the rest mainly in A.

Using a two-state bigram HMM over the entire text, we find that the two word classes induced by EM more or less correspond to the same division – words in pages classified as being in the A language tend to be tagged as one class, and words in B language pages as the other, indicating that the manuscript does indeed contain two different vocabularies (which may be related languages, dialects, or simply different textual domains). In Figure 2, we

Figure 1: Page *f81v* (from the Biological section).

(a) Scan of page

```
BAR ZC9 FCC89 ZCFAE 8AE 8AR OE BSC89 ZCF 8AN
OVAE ZCF9 4OFC89 OFAM FAT OFAE 2AR OE FAN
OEFAN AE OE ROE 8E 2AM 8AM OEFCC89 OFC89 89FAN
ZCF S89 8AEAE OE89 4OFAM OFAN SCCF9 89 OE FAM
8AN 89 8AM SX9 OFAM 8AM OPAN SX9 OFCC89 4OF9
FAR 8AM OFAR 4OFAN OFAM OE SC89 SCOE EF9 E2
AM OFAN 8AE89 OEOR OE ZCXAE 8AM 4OFCC8AE 8AM
SX9 2SC89 4OE 9FOE OR ZC89 ZCC89 4OE FCC89 8AM
8FAN WC89 OE89 9AR OESC9 FAM OFCC9 8AM OEOR
SCX9 8AII89

BOEZ9 OZ9PCC8 4OB OFCC89 OPC89 OFZC89 4OP9
8ATAJ OZC9 4OFCC9 OFCC9 OF9 9FCC9 4OF9 OF9EF9
OES9 F9 8ZOE98 4OE OE S89 ZC89 4OFC89 9PC89
SCPC89 EFC8C9 9PC89 9FCC2C9 8SC8 9PC89 9PC89
8AR 9FC8A IB*9 4OP9 9FC89 OFAE 8ZC89 9FCC89
C2CCF9 8AM OFC89 4OFCC8 4OFC89 ESBS89 4OFAE
SC89 OE ZCC9 2AEZQ89 4OVSC89 R SC89 EPAR9
EOR ZC89 4OCC89 OE S9 RZ89 EZC89 8AR S89
BS89 2ZFS89 SC89 OE ZC89 4OESC89 4OFAN ZX9 8E
RAE 4OFS89 SC9 OE SCF9 OE ZC89 4OFC89 4OFC89
SX9 4OF9 2OEFCC9 OE ZC89 4OFAR ZCX9 8C2C89
4OFAR 4OFAE 8OE S9 4OQC9 SCFAE SO89 4OFC89
EZCP9 4OE89 EPC89 4OPAN EZO 4OFC9 EZC89 EZC89
SC89 4OEF9 ESC8AE 4OE OPAR 4OFAE 4OE OM SCC9
8AE EO*C89 ZC89 2AE SPC89PAR ZOE 4CFS9 9FAM
OEFAN ZC89 4OF9 8SC89 ROE OE Q89 9PC9 OFSC89
4OFAE OFCC9 4OE SCC89 2AE PCOE 8S89 E9 OZC89
4OPC89 ZOE SC89 9ZSC9 OE SC9 4OE SC89 PS8 OF9
OE SCSOE PAR OM OFC89 8AE ZC9 OEFCOE OEFCC89
OFCOE 8ZCOE O3 OEFCC89 PC89 SCF9 ZXC89 SAE

OPON OEFOE
```

(b) Transcription in the Currier alphabet. Paragraph (but not line) breaks are indicated.

illustrate the division of the manuscript pages into the six sections, and show the proportion of words in each page that are classified as the B language.

For coherence, all our experimental results in the rest of this paper are on the B language (which we denote by VMS B) – specifically, the Biological and Stars sections – unless otherwise specified. These sections together contain 43 pages, with 3920 word types, 17597 word tokens, and 35 characters. We compare the VMS's statistical properties with three natural language texts of similar size: the first 28551 words from the English Wall Street Journal Corpus, 19327 words from the Arabic Quran (in Buckwalter transcription), and 18791 words from the Chinese Sinica Treebank.

3 The Letter

3.1 Are vowels and consonants represented?

If a script is alphabetic, i.e., it uses approximately one character per phoneme, vowel and consonant characters can be separated in a fully unsupervised way. Guy (1991) applies the vowel-consonant separation algorithm of (Sukhotin, 1962) on two pages of the Biological section, and finds that four characters (o, a, c, g) separate out as vowels. However, the separation is not very strong, and several words do not contain these characters.

Another method is to use a two-state bigram HMM (Knight et al., 2006; Goldsmith and Xanthos, 2009) over letters, and induce two clusters of letters with EM. In alphabetic languages like English, the clusters correspond almost perfectly to vowels and consonants. We find that a curious phenomenon occurs with the VMS – the last character of every word is generated by one of the HMM states, and all other characters by another; i.e., the word grammar is a^*b.

There are a few possible interpretations of this. It is possible that the vowels from every word are removed and placed at the end of the word, but this means that even long words have only one vowel, which is unlikely. Further, the number of vowel types would be nearly half the alphabet size. If the script is a syllabary or a logograph, a similar clustering will surface, but given that there are only 35 characters, it is unlikely that each of them represents a syllable or word. A more likely explanation is that the script is an abjad, like the scripts of Semitic languages, where all or most vowels are omitted. Indeed, we find that a 2-state HMM on Arabic without diacritics and English without vowels learns a similar grammar, a^*b^+.

3.2 Do letters have cases?

Some characters (f, b, p, v) that appear mainly at paragraphs beginnings are referred to 'gallows' – glyphs that are taller and more ornate than others. Among the glyphs, these least resemble Latin, leading to the belief that they are null symbols, which Morningstar (2001) refutes.

Another hypothesis is that gallows are uppercase versions of other characters. We define BESTSUB(c) to be the character x that produces the highest decrease in unigram word entropy when x is substituted for all instances of c. For English uppercase characters c, BESTSUB(c) is the lowercase version. However, BESTSUB of the VMS gallows is one of the other gallows! This demonstrates that they are not uppercase versions of other letters, and also that they are contextually similar to one another.

3.3 Is there punctuation?

We define punctuation as symbols that occur only at word edges, whose removal from the word results in an existing word. There are two characters that are only found at the ends of words (Currier k and l), but most of the words produced by removing k and l are not in the vocabulary. Therefore, there is most likely no punctuation, at least in the traditional sense.

4 The Word

4.1 What are the word frequency and length distributions?

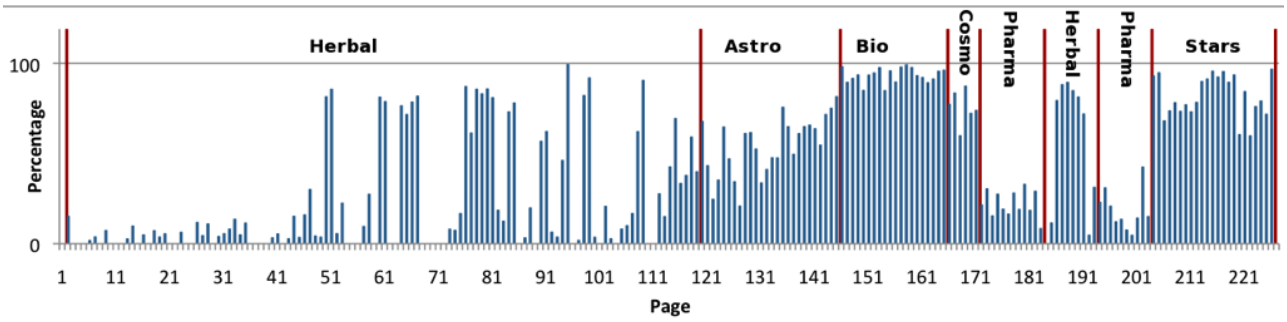
The word frequency distribution follows Zipf's law, which is a necessary (though not sufficient) test of linguistic plausibility. We also find that the unigram word entropy is comparable to the baseline texts (Table 1).

Table 1: Unigram word entropy in bits.

VMS B	English	Arabic	Chinese
9.666	10.07	9.645	10.31

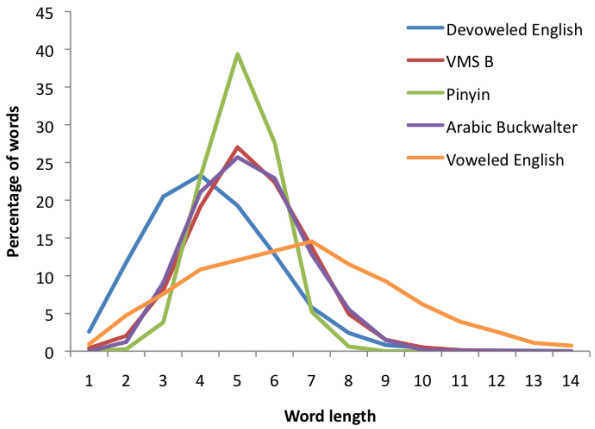
Several works have noted the narrow binomial distribution of word lengths, and contrasted it with

Figure 2: VMS sections, and percentage of word tokens in each page that are tagged as language B by the HMM.



the wide asymmetric distribution of English, Latin, and other European languages. This contributed to speculation that the VMS is not a natural language, but a code or generated by some other stochastic process. However, Stolfi (2005) show that Pinyin Chinese, Tibetan, and Vietnamese word lengths follow a binomial distribution, and we found (Figure 3) that certain scripts that do not contain vowels, like Buckwalter Arabic and devoweled English, have a binomial distribution as well.[3] The similarity with devoweled scripts, especially Arabic, reinforces the hypothesis that the VMS script may be an abjad.

Figure 3: Word length distributions (word types).



Landini (2001) found that the VMS follows Zipf's law of word lengths: there is an inverse relationship between the frequency and length of a word.

[3]This is an example of why comparison with a range of languages is required before making conclusions about the language-like nature of a text.

4.2 How predictable are letters within a word?

Bennett (1976) notes that the second-order entropy of VMS letters is lower than most European languages. Stolfi (2005) computes the entropy of each character given the left and right contexts and finds that it is low for most of the VMS text, particularly the Biological section, compared to texts in other languages. He also ascertains that spaces between words have extremely low entropy.

We measure the *predictability* of letters, and compare it to English, Arabic, and Pinyin Chinese. Predictability is measured by finding the probabilities over a training set of word types, guessing the most likely letter (the one with the highest probability) at each position in a word in the held-out test set, and counting the proportion of times a guess is correct. Table 2 shows the predictability of letters as unigrams, and given the preceding letter in a word (bigrams). VMS letters are more predictable than other languages, with the predictability increasing sharply given the preceding contexts, similarly to Pinyin.

Table 2: Predictability of letters, averaged over 10-fold cross-validation runs.

	VMS B	English	Arabic	Pinyin
Bigram	40.02%	22.62%	24.78%	38.92%
Unigram	14.65%	11.09%	13.29%	11.20%

Zandbergen (2010) computes the entropies of characters at different positions in words in the Stars section, and finds that the 1^{st} and 2^{nd} characters of a word are more predictable than in Latin or Vulgate, but the 3^{rd} and 4^{th} characters are less predictable.

It has also been observed that word-final characters have much lower entropy compared to most other languages – some characters appear almost exclusively at the ends of words.

4.3 Is there morphological structure?

The above observations suggest that words are made up of morpheme-like chunks. Several hypotheses about VMS word structure have been proposed. Tiltman (1967) proposed a template consisting of roots and suffixes. Stolfi (2005) breaks down the morphology into 'prefix-midfix-suffix', where the letters in the midfixes are more or less disjoint from the letters in the suffixes and prefixes. Stolfi later modified this to a 'core-mantel-crust' model, where words are composed of three nested layers.

To determine whether VMS words have affixal morphology, we run an unsupervised morphological segmentation algorithm, Linguistica (Goldsmith, 2001), on the VMS text. The MDL-based algorithm segments words into prefix+stem+suffix, and extracts 'signatures', sets of affixes that attach to the same set of stems. Table 3 lists a few sample signatures, showing that stems in the same signature tend to have some structural similarities.

Table 3: Some morphological signatures.

Affixes	Stems
OE+, OP+, null+	A3 AD AE AE9 AEOR AJ AM AN AR AT E O O2 OE OJ OM ON OR SAJ SAR SCC9 SCCO SCO2 SO
OE+	BSC28 BSC9 CCC8 COC8CR FAEOE FAK FAU FC8 FC8AM FCC FCC2 FCC9R FCCAE FCCC2 FCCCAR9 FCO9 FCS9 FCZAR FCZC9 OEAR9 OESC9 OF9 OR8 SC29 SC89O SC8R SCX9 SQ9
+89, +9, + C89	4OFCS 4OFCZ 4OFZ 4OPZ 8AES 8AEZ 9FS 9PS EFCS FCS PS PZ OEFS OF OFAES OFCS OFS OFZ

5 Syntax

5.1 Is there word order?

One of the most puzzling features of the VMS is its weak word order. Notably, the text has very few repeated word bigrams or trigrams, which is surprising given that the unigram word entropy is comparable to other languages. Furthermore, there are sequences of two or more repeated words, or repetitions of very similar words. For example, the first page of the Biological section contains the line 4OFCC89 4OFCC89 4OFC89 4OFC89 4OFCC89 E89.

We compute the predictability of a word given the previous word (Table 4). Bigram contexts only provide marginal improvement in predictability for the VMS, compared to the other texts. For comparison with a language that has 'weak word order', we also compute the same numbers for the first 22766 word tokens of the Hungarian Bible, and find that the *empirical* word order is not that weak after all.

Table 4: Predictability of words (over 10-fold cross-validation) with bigram contexts, compared to unigrams.

	Unigram	Bigram	Improvement
VMS B	2.30%	2.50%	8.85%
English	4.72%	11.9%	151%
Arabic	3.81%	14.2%	252%
Chinese	16.5%	19.8%	19.7%
Hungarian	5.84%	13.0%	123%

5.2 Are there latent word classes?

While there are very few repeated word bigrams, perhaps there are latent classes of words that govern word order. We induce ten word classes using a bigram HMM trained with EM (Figure 4). As with the stems in the morphological signatures, the words in each class show some regularities – although it is hard to quantify the similarities – suggesting that these latent classes are meaningful.

Currier (1976) found that some word-initial characters are affected by the word-final characters of the immediately preceding word. He concludes that the 'words' being syllables or digits would explain this phenomenon, although that is unlikely given the rarity of repeated sequences.

We redo the predictability experiments of the previous section, using the last m letters of the previous word to predict the first n letters of the current word. When $n > 2$, improvement in predictability remains low. However, when n is 1 or 2, there is a noticeable improvement when using the last few characters of the previous word as contexts (Table 5).

5.3 Are there long-distance word correlations?

Weak bigram word order can arise if the text is scrambled or is generated by a unigram process. Alternately, the text might have been created by inter-

Figure 4: Some of the induced latent classes.

(a) (b) (c)

Table 5: Relative improvement in predictability of first n word-characters using last m characters of previous word, over using no contextual information.

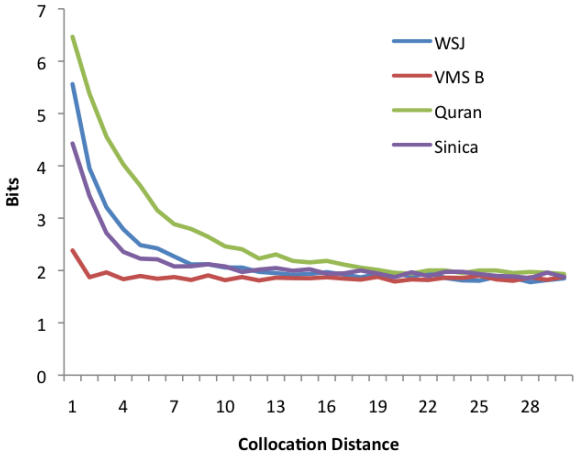
		VMS B	English	Arabic
Whole words		8.85%	**151%**	**252%**
$n=1$	$m=1$	**31.8%**	31.1%	26.8%
	$m=2$	30.7%	45.8%	61.5%
	$m=3$	29.9%	60.3%	92.4%
$n=2$	$m=1$	16.0%	42.8%	0.0736%
	$m=2$	12.4%	67.5%	14.1%
	$m=3$	10.9%	94.6%	33.2%

leaving the words of two or more texts, in which case there will be long-distance correlations.

Schinner (2007) shows that the probability of *similar* words repeating in the text at a given distance from each other follows a geometric distribution.

Figure 5 illustrates the 'collocationness' at distance d, measured as the average pointwise mutual information over all pairs of words w_1, w_2 that occur more than once at distance d apart. VMS words do not show significant long-distance correlations.

Figure 5: Long-range collocationness. Arabic shows stronger levels of long-distance correlation compared to English and Chinese. VMS B shows almost no correlations for distance $d > 1$.



6 The Page

6.1 Do pages have topics?

That is, do certain words 'burst' with a high frequency within a page, or are words randomly distributed across the manuscript? Figure 6 shows a visualization of the TF-IDF values of words in a VMS B page, where the 'documents' are pages, indicating the relevance of each word to the page. Also shown is the same page in a version of the document created by scrambling the words of the original manuscript, and repaginating to the same page lengths. This simulates a document where words are generated independent of the page, i.e., the pages have no topics.

To quantify the degree to which a page contains topics, we measure the entropy of words within the page, and denote the overall 'topicality' T of a document as the average entropy over all the pages. As a control, we compute the topicality T_{rand} of the scrambled version of the document. $1 - T/T_{rand}$ indicates the extent to which the pages of the document contain topics. Table 6 shows that by this measure, the VMS's strength of page topics is less than the English texts, but more than the Quran[4], signifying that the pages probably do have topics, but are not independent of one another.

6.2 Is the text prose?

Visually, the text looks like prose written in paragraphs. However, Currier (1976) stated that "the line

[4]We demarcate a 'page' to be approximately 25 verses for the Quran, a chapter for the Genesis, and an article for the WSJ.

Figure 6: TF-IDF visualization of page *f108v* in the Stars section.

(a) Original document, showing bursts

(b) Scrambled version – flatter distribution

Table 6: Strength of page topics in VMS and other texts, cropped to be of comparable length to the VMS.

	VMS B	English WSJ	English Genesis	Arabic Quran
T	7.5	6.3	6.6	7.7
T_{rand}	7.7	6.5	7.1	7.9
$1 - T/T_{rand}$	0.033	0.037	0.069	0.025

is a functional entity" – that is, there are patterns to lines on the page that are uncharacteristic of prose. In particular, certain characters or sequences appear almost exclusively at the beginnings or ends of lines.

Figure 7 shows the distribution of characters at line-edges, relative to their occurrences at word beginnings or endings,confirming Currier's observation. It is particularly interesting that lower-frequency characters occur more at line-ends, and higher-frequency ones at the beginnings of lines.

Schinner (2007) found that characters show long-range correlations at distances over 72 characters, which is a little over the average line length.

7 The Document

7.1 Are the pages in order?

We measure the similarity between two pages as the cosine similarity over bags of words, and count the proportion of pages P_i where the page P_{i-1} or P_{i+1} is the most similar page to P_i. We denote this measure by ADJPAGESIM. If ADJPAGESIM is high, it indicates that (1) the pages are not independent of each other and (2) the pages are in order.

Table 7 shows ADJPAGESIM for the VMS and other texts. As expected, ADJPAGESIM is close to zero for the VMS with pages scrambled, as well as the WSJ, where each page is an independent article, and is highest for the VMS, particularly the B pages.

Table 7: ADJPAGESIM for VMS and other texts.

VMS B	38.8%
VMS All	15.6%
VMS B pages scrambled	0%
VMS All pages scrambled	0.444%
WSJ	1.34%
English Genesis	25.0%
Arabic Quran	27.5%

This is a convincing argument for the pages being mostly in order. However, the non-contiguity of the herbal and pharmaceutical sections and the interleaving of the A and B languages indicates that larger chunks of pages were probably reordered. In addition, details involving illustrations and ink-transfer across pages point to a few local reorderings (Pelling, 2006).

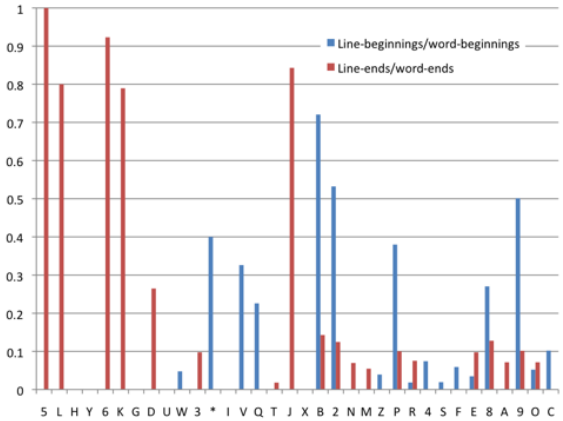
7.2 How many authors were involved?

Currier (1976) observed that the distinction between the A and B languages corresponds to two different types of handwriting, implying at least two authors. He claimed that based on finer handwriting analysis, there may have been as many as eight scribes.

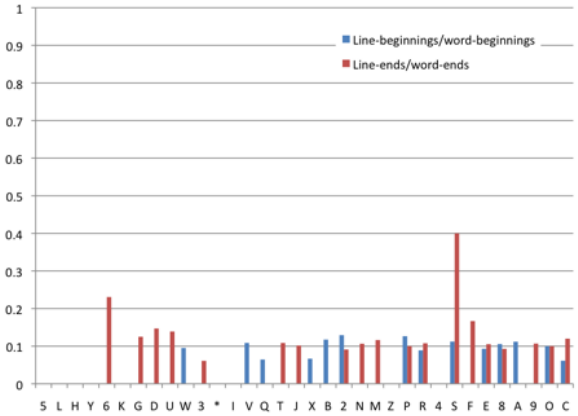
8 Latin, Cipher, or Hoax?

Claims of decipherment of the VMS script have been surfacing for several years, none of which are convincing. Newbold (1928) believed that microscopic irregularities of glyph edges correspond to anagrammed Latin. Feely in 1943 proposed that the script is a code for abbreviated Latin (D'Imperio, 1980). Sherwood (2008) believes that the words are coded anagrams of Italian. Others have hypoth-

Figure 7: Proportion of word-edge characters at line-edges for lines that span the width of the page. Characters are in ascending order of their total frequencies.



(a) Original document, showing biased distribution.



(b) Flat distribution when words within lines are scrambled.

esized that the script is an encoding of Ukrainian (Stojko, 1978), English (Strong, 1945; Brumbaugh, 1976), or a Flemish Creole (Levitov, 1987). The word length distribution and other properties have invoked decodings into East Asian languages like Manchu (Banasik, 2004). These theories tend to rely on arbitrary anagramming and substitutions, and are not falsifiable or well-defined.

The mysterious properties of the text and its resistance to decoding have led some to conclude that it is a hoax – a nonsensical string made to look vaguely language-like. Rugg (2004) claims that words might have been generated using a 'Cardan Grille' – a way to deterministically generate words from a table of morphemes. However, it seems that the Grille emulates a restricted finite state grammar of words over prefixes, midfixes, and suffixes. Such a grammar underlies many affixal languages, including English. Martin (2008) proposes a method of generating VMS text from anagrams of number sequences. Like the previous paper, it only shows that this method *can* create VMS-like words – not that it is the most plausible way of generating the manuscript. It is also likely that the proposed scheme can be used to generate any natural language text.

Schinner (2007) votes for the hoax hypothesis based on his observations about characters showing long-range correlations, and the geometric distribution of the probability of similar words repeating at a fixed distance. These observations only confirm that the VMS has some properties unlike natural language, but not that it is necessarily a hoax.

9 Conclusion

We have detailed various known properties of the Voynich manuscript text. Some features – the lack of repeated bigrams and the distributions of letters at line-edges – are linguistically aberrant, which others – the word length and frequency distributions, the apparent presence of morphology, and most notably, the presence of page-level topics – conform to natural language-like text.

It is our hope that this paper will motivate research into understanding the manuscript by scholars in computational linguistics. The questions presented here are obviously not exhaustive; a deeper examination of the statistical features of the text in comparison to a number of scripts and languages is needed before any definite conclusions can be made. Such studies may also inspire a quantitative interest in linguistic and textual typologies, and be applicable to the decipherment of other historical scripts.

Acknowledgments

We would like to thank the anonymous reviewers and our colleagues at ISI and Chicago for their helpful suggestions. This work was supported in part by NSF Grant 0904684.

References

Zbigniew Banasik. 2004. http://www.ic.unicamp.br/ stolfi/voynich/04-05-20-manchu-theo/alphabet.html.

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

Robert Brumbaugh. 1976. The Voynich 'Roger Bacon' cipher manuscript: deciphered maps of stars. *Journal of the Warburg and Courtauld Institutes*.

Prescott Currier. 1976. New research on the Voynich Manuscript: Proceedings of a seminar. Unpublished communication, available from http://www.voynich.nu/extra/curr_pdfs.html.

Mary D'Imperio. 1980. *The Voynich Manuscript: An Elegant Enigma*. Aegean Park Press.

Jim Gillogly and Jim Reeds. 2005. Voynich Manuscript mailing list. http://voynich.net/.

John Goldsmith and Aris Xanthos. 2009. Learning phonological categories. *Language*, 85:4–38.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*.

Jacques Guy. 1991. Statistical properties of two folios of the Voynich Manuscript. *Cryptologia*.

Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of COLING*.

Gabriel Landini. 2001. Evidence of linguistic structure in the Voynich Manuscript using spectral analysis. *Cryptologia*.

Leo Levitov. 1987. *Solution of the Voynich Manuscript: A Liturgical Manual for the Endura Rite of the Cathari Heresy, the Cult of Isis*. Aegean Park Press.

Claude Martin. 2008. Voynich, the game is over. http://www.voynich.info/.

Jason Morningstar. 2001. Gallows variants as null characters in the Voynich Manuscript. Master's thesis, University of North Carolina.

William Newbold. 1928. *The Cipher of Roger Bacon*. University of Pennsylvania Press.

Nicholas John Pelling. 2006. *The Curse of the Voynich: The Secret History of the World's Most Mysterious Manuscript*. Compelling Press.

Jim Reeds. 2002. Voynich Manuscript. http://www.ic.unicamp.br/ stolfi/voynich/mirror/reeds.

Gordon Rugg. 2004. The mystery of the Voynich Manuscript. *Scientific American Magazine*.

Andreas Schinner. 2007. The Voynich Manuscript: Evidence of the hoax hypothesis. *Cryptologia*.

Edith Sherwood. 2008. The Voynich Manuscript decoded? http://www.edithsherwood.com/voynich_decoded/.

John Stojko. 1978. *Letters to God's Eye: The Voynich Manuscript for the first time deciphered and translated into English*. Vantage Press.

Jorge Stolfi. 2005. Voynich Manuscript stuff. http://www.dcc.unicamp.br/ stolfi/voynich/.

Leonell Strong. 1945. Anthony Ashkam, the author of the Voynich Manuscript. *Science*.

Boris Sukhotin. 1962. Eksperimental'noe vydelenie klassov bukv s pomoscju evm. *Problemy strukturnoj lingvistiki*.

John Tiltman. 1967. The Voynich Manuscript, the most mysterious manuscript in the world. *NSA Technical Journal*.

René Zandbergen. 2010. Voynich MS. http://www.voynich.nu/index.html.