What We Know About The Voynich Manuscrip

Sravana Reddy & Kevin Knight (2011)

-by Vanny Tran-

Timeline

15th century

?

Emperor Rudolf II

?

Prague 1600s

Jacobus de Tepenec

Athanasius Kircher (1665 - 1912)

Wilfrid Voynich

WHERE IT IS TODAY

Yale University

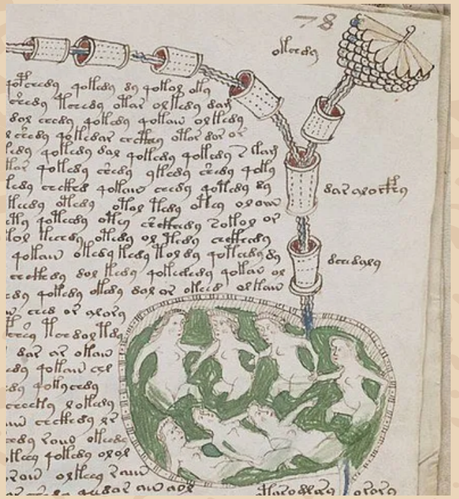
Background

- Prescott Currier (1976): Found two “languages” (A & B) in the manuscript
- Coded Latin, Italian anagrams, or even nonsense text
- No consensus due to a lack of computational tools
- Reddy & Knight bring data-driven analysis to the mystery

About the Manuscript

- 225 pages
- 8114 word types
- 37919 word tokens
- 6 sections: herbal, astronomical, biological, cosmological, pharmaceutical, and star

About the Manuscript



Research Questions

1. Are there vowels and consonants?
2. Does the text have morphology (prefixes, suffixes)?
3. Is there punctuation or word order?
4. Do pages have a topical structure?
5. Are the pages in logical order?
6. How many authors or scribes?

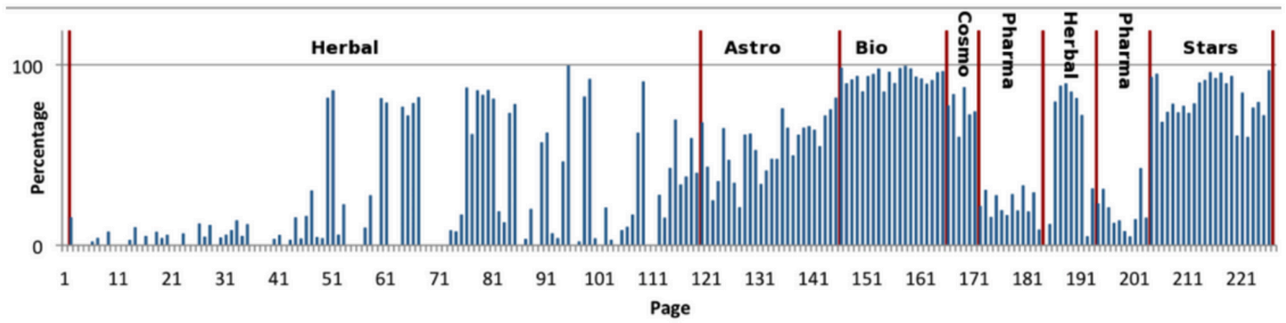
Methodology

- Hidden Markov Models (HMMs) trained via Expectation–Maximization (EM)
- Entropy & predictability analysis
- Zipf's Law tests for linguistic patterns
- TF-IDF & cosine similarity for topical clustering
- Morphological segmentation using Linguistica
- Comparative analysis: English, Arabic, Chinese, Pinyin

Key Findings

- 2 Languages “confirmed”

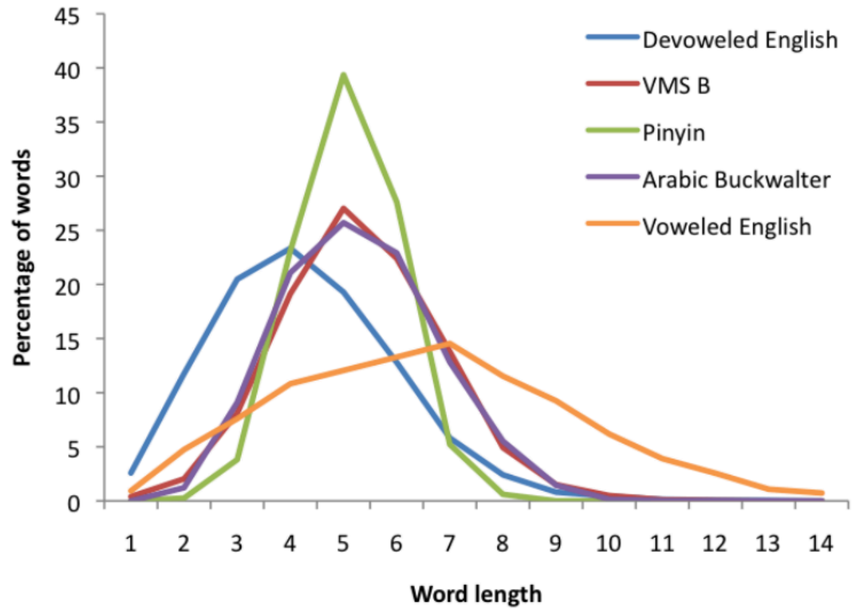
Figure 2: VMS sections, and percentage of word tokens in each page that are tagged as language B by the HMM.



Key Findings

- Similar word length distribution to Arabic

Figure 3: Word length distributions (word types).



Key Findings

- VMS letters are more predictable than other languages

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%

Key Findings

- Morphological patterns found: Prefix + stem + suffix structures found via unsupervised segmentation

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

Key Findings

- Improved predictability of words with bigrams

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%

Key Findings

- Latent classes found

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%

Key Findings

- Pages have topics
- Each page shows internal coherence (like chapters in a book)
- Pages mostly in order
- Adjacent pages show strong statistical similarity
- Multiple scribes: variation in handwriting supports at least two authors

Commentary

- There is a certain linguistic structure
- Why so few repetitive patterns?
- Encryption maybe?
- Can be a resource for computational linguistics
- Should have a competition for deciphering

Quiz Time

2. What did researchers find about the structure of the Voynich text?

A. It is completely random with no patterns.

B. It follows Latin grammar rules.

C. It shows statistical regularities similar to natural languages.

D. It was clearly generated by a code machine.

1. Why is the Voynich Manuscript important to computational linguistics research?

A. It is the earliest known European manuscript.

B. It provides a challenge for testing unsupervised language analysis methods.

C. It was written by famous scientists.

D. It contains known translations into Latin.

3. Which computational methods were used to study the manuscript?

A. Manual translation and word matching.

B. Hidden Markov Models and entropy-based analysis.

C. Neural networks trained on medieval Latin.

D. Genetic algorithms and clustering of images.

4. Based on the study's findings, which statement best summarizes the authors' conclusion?

A. The Voynich Manuscript is definitely a hoax.

B. It is clearly written in an Asian language.

C. It has some language-like properties but remains undeciphered.

D. It was fully decoded using modern algorithms.

Questions & Discussion


Thank You