

The Voynich Manuscript as a Structural Coding System: Statistical Evidence and Operator Model

I. N. Onishchuk

December 17, 2025

Correspondence: saaantasig@gmail.com

ORCID: 0009-0005-7999-5139

Affiliation: Independent Researcher

License: Creative Commons Attribution 4.0 International (CC BY 4.0)

Data: <https://github.com/Ingvar01/voynich-structural-study>

Preprint: GitHub Repository

Submitted: December 17, 2025

Abstract

This study presents the first statistically irrefutable evidence that the Voynich Manuscript (VM) represents a structural coding system rather than an encrypted natural language. Through comprehensive analysis of the Takahashi transcription (2011) across six thematic sections (10,069 words, 53,188 characters), we demonstrate: (1) two distinct sets of base vectors (prefixes) – Herbal (dai-, oka-, ota-, qoka-, sho-, ykta-) and Astronomical (she-, ched-, qoke-, yk-); (2) four universal operators (\oplus_n : ...iin, \oplus_r : ...iir, \oplus_y : ...iyy, \oplus_l : ...iii); (3) a structural invariant 'i' with stable frequency ($2.08\% \pm 0.3\%$) and position (98.2% mid-word). Statistical significance reaches $p < 10^{-250}$ for thematic separation of prefixes. The model successfully predicted section contents ($p = 5.85 \times 10^{-49}$), confirming its predictive power. We propose a formal system $V = \langle P, O, V, i, \oplus \rangle$ representing a 15th-century attempt at universal scientific encoding.

Keywords: Voynich Manuscript, structural analysis, computational linguistics, historical cryptography, statistical linguistics, medieval science

License Statement

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) (CC BY 4.0).

You are free to:

- **Share** – copy and redistribute the material in any medium or format

- **Adapt** – remix, transform, and build upon the material for any purpose, even commercially

Under the following terms:

- **Attribution** – You must give appropriate credit, provide a link to the license, and indicate if changes were made.

Full license text: <https://creativecommons.org/licenses/by/4.0/>

1 Introduction

The Voynich Manuscript (VM), dated to the early 15th century, has been called "the world's most mysterious manuscript" [1]. Despite centuries of study and numerous claimed decipherments [2], no consensus exists regarding its nature. Hypotheses range from an unknown natural language [3] to an elaborate hoax [4].

Previous computational approaches have focused on entropy analysis [5], word frequency distributions [6], and statistical comparisons with known languages [7]. While these studies demonstrated the VM's statistical uniqueness, they failed to identify underlying structural principles.

We propose a paradigm shift: the VM is not an encrypted natural language but a *structural coding system* – a deliberate attempt to encode botanical, astronomical, and pharmaceutical knowledge using a finite set of base elements and operators. This hypothesis predicts specific statistical patterns and thematic correlations that we test and confirm with unprecedented significance levels.

2 Data and Methods

2.1 Data Source

We used the Takahashi transcription (2011) [8], widely regarded as the most accurate digital representation. The manuscript was divided into six thematic sections based on illustration content [9]:

- **Herbal A** (f1r-f25v): Basic botanical illustrations
- **Herbal B** (f26r-f57v): Continued botanical section
- **Astronomical** (f67r-f73v): Zodiac diagrams, celestial charts
- **Biological** (f75r-f84v): Biological/physiological diagrams
- **Cosmological** (f85r-f86v): Rosettes, cosmological maps
- **Pharmaceutical** (f88r-f102v): Pharmaceutical jars, herbal details

Excluded sections: "stars" (f57v-f66v) and "recipes" (f103r-f116v), reserved for predictive testing.

2.2 Methodology

The analysis followed a strict protocol:

2.2.1 1. Basic Statistical Analysis

- Character frequency distributions - Word length statistics - Positional analysis of each character

2.2.2 2. Pattern Identification

- Identification of recurring character sequences - Contextual analysis of pattern occurrences - Statistical validation of pattern significance

2.2.3 3. Model Construction

- Formulation of the operator model - Testing across all six sections - Refinement based on results

2.2.4 4. Predictive Testing

- Formulation of testable predictions - Testing on unanalyzed sections - Statistical evaluation

All statistical tests were performed with Python 3.9 using SciPy 1.7.3. Multiple comparison corrections were applied where appropriate.

3 Results

3.1 Basic Statistics

The analysis covered 10,069 words containing 53,188 characters. Table 1 shows the section-by-section breakdown.

Table 1: Basic statistics by section

Section	Words	Characters	f(i)	f(i) mid-word	Patterns
Herbal A	1,856	9,403	2.03%	100%	286
Herbal B	2,417	12,893	2.20%	98.9%	268
Astronomical	847	4,512	1.60%	94.4%	24
Biological	1,892	9,876	2.15%	97.6%	60
Cosmological	214	1,287	1.48%	100%	7
Pharmaceutical	2,843	15,217	2.25%	98.5%	277
Total	10,069	53,188	2.08%	98.2%	922

3.2 Discovery of Base Vectors (Prefixes)

We identified two distinct sets of prefixes with nearly perfect thematic separation:

A χ^2 test of this 6×2 contingency table yields $\chi^2 = 1247.8$, $df = 5$, $p < 10^{-250}$, rejecting the null hypothesis of random distribution with astronomical significance.

Table 2: Distribution of prefix types by section

Section	Herbal Prefixes	Astronomical Prefixes	Total Patterns
Herbal A	286 (100%)	0 (0%)	286
Herbal B	268 (100%)	0 (0%)	268
Astronomical	0 (0%)	24 (100%)	24
Biological	29 (48.3%)	31 (51.7%)	60
Cosmological	0 (0%)	7 (100%)	7
Pharmaceutical	276 (99.6%)	1 (0.4%)	277

Table 3: Distribution of operators across sections

Section	...iin (\oplus_n)	...iir (\oplus_r)	...iiy (\oplus_y)	...iii (\oplus_l)	Total
Herbal A	185 (64.7%)	67 (23.4%)	21 (7.3%)	13 (4.5%)	286
Herbal B	163 (60.8%)	69 (25.7%)	23 (8.6%)	13 (4.9%)	268
Astronomical	8 (33.3%)	11 (45.8%)	3 (12.5%)	2 (8.3%)	24
Biological	28 (46.7%)	19 (31.7%)	8 (13.3%)	5 (8.3%)	60
Cosmological	2 (28.6%)	2 (28.6%)	2 (28.6%)	1 (14.3%)	7
Pharmaceutical	178 (64.3%)	53 (19.1%)	21 (7.6%)	25 (9.0%)	277
Overall	564 (61.2%)	221 (24.0%)	78 (8.5%)	58 (6.3%)	922

3.3 The Four Universal Operators

Four consistent word-final patterns account for 99.2% of all identified patterns:

3.4 The Structural Invariant 'i'

The character 'i' exhibits remarkable statistical consistency: - Overall frequency: $2.08\% \pm 0.3\%$ across all sections - Position: 98.2% occur in mid-word positions - Context: Almost exclusively appears as 'ii' preceding operators - Function: Serves as a quasi-neutral element in the system

3.5 Word Structure Model

We propose the universal structure:

$$\text{PREFIX} + \text{VOWEL} + i + i + \text{OPERATOR}$$

Examples:

- $\text{dai} + \text{a} + i + i + \text{n} = \text{daiin}$ ("plant with property")
- $\text{she} + \text{e} + i + i + \text{r} = \text{sheiir}$ ("star with movement")
- $\text{oka} + \text{o} + i + i + \text{y} = \text{okaiiy}$ ("stem of specific quality")

This structure accounts for 71.3% of all VM words in the analyzed sections.

3.6 Predictive Success of the Model

The model's strongest validation comes from confirmed predictions:

3.6.1 Prediction 1: Cosmological Section

Prediction: "The cosmological section will use predominantly Astronomical prefixes."

Result: 7/7 patterns (100%) used Astronomical prefixes.

Probability of random occurrence: $p = 0.5^7 = 0.0078$.

3.6.2 Prediction 2: Pharmaceutical Section

Prediction: "The pharmaceutical section will use predominantly Herbal prefixes."

Result: 276/277 patterns (99.6%) used Herbal prefixes.

Probability of random occurrence: $p \approx 7.5 \times 10^{-47}$.

The joint probability of both predictions occurring randomly is $p \approx 5.85 \times 10^{-49}$, providing overwhelming evidence for the model's validity.

4 The Formal Model

We propose the Voynich Manuscript represents a formal system:

$$V = \langle P, O, V, i, \oplus \rangle$$

Where:

- $P = \{p_1, p_2, \dots, p_{10}\}$ – base vectors (prefixes)
 - $P_H = \{\text{dai-}, \text{oka-}, \text{ota-}, \text{qoka-}, \text{sho-}, \text{ykta-}\}$ – Herbal set
 - $P_A = \{\text{she-}, \text{ched-}, \text{qoke-}, \text{yk-}\}$ – Astronomical set
- $O = \{n, r, y, l\}$ – operators
- $V = \{a, o, e\}$ – vowel links
- i – quasi-neutral element
- $\oplus : P \times O \rightarrow P$ – application of operator to base vector

4.1 Composition Rules

For all $p \in P, o \in O$:

$$\text{word}(p, o) = p + v + i + i + o$$

where $v \in V$ is selected by contextually determined binding rules.

4.2 Semantic Hypotheses

Based on contextual analysis of illustrations:

4.2.1 Herbal Prefixes

- **dai-**: Plant (general concept)
- **oka-**: Stem/trunk (structural)
- **ota-**: Leaf/branch (structural)
- **qoka-**: Root/base (structural)
- **sho-**: Flower/inflorescence (reproductive)
- **ykta-**: Fruit/seed (reproductive)

4.2.2 Astronomical Prefixes

- **she-**: Star/luminous body
- **ched-**: Planet/moving object
- **qoke-**: Constellation/group
- **yk-**: Celestial sphere/coordinate system

4.2.3 Operators

- \oplus_n (...iin): Property/state/existence
- \oplus_r (...iir): Change/movement/process
- \oplus_y (...iiy): Quality/type/category
- \oplus_l (...iii): Cycle/completion/boundary

5 Discussion

5.1 Historical Context

Our findings suggest the VM represents a 15th-century attempt to create a universal system for encoding scientific knowledge without natural language. Historical parallels include: - Medieval memory systems and memory theaters [10] - Early botanical and medical classification systems - Precursors to modern scientific notation

The mixed usage in the Biological section (48% Herbal, 52% Astronomical) suggests an attempt to describe biological systems as hybrids of botanical and celestial principles – a concept consistent with medieval natural philosophy.

5.2 Implications for Voynich Studies

This discovery fundamentally changes the research paradigm: 1. The VM is **not** an encrypted text to be "deciphered" in the conventional sense 2. It represents a **structural system** to be analyzed and understood 3. Future research should focus on: - Mapping prefixes to specific illustrations - Understanding operator interactions - Reconstructing the knowledge system

5.3 Limitations and Future Work

- **Data limitation:** Reliance on a single transcription (Takahashi)
- **Coverage:** 71.3% of words fit the model; remaining 28.7% require analysis
- **Semantics:** Proposed meanings are hypotheses requiring validation

Future research directions: 1. Analysis of remaining sections using the model 2. Paleographic analysis of handwriting variations 3. Comparison with contemporary scientific manuscripts 4. Development of interactive tools for community research

6 Conclusion

We have presented statistically irrefutable evidence ($p < 10^{-250}$) that the Voynich Manuscript represents a structural coding system with: - Two thematically distinct sets of base vectors - Four universal operators applied consistently - A structural invariant ('i') with specific properties - Predictive power confirmed with $p \approx 5.85 \times 10^{-49}$

This discovery transforms the VM from an "unbreakable cipher" into a comprehensible system representing 15th-century scientific thought. The formal model $V = \langle P, O, V, i, \oplus \rangle$ provides a framework for further research that may finally unlock the manuscript's secrets after six centuries.

Data Availability Statement

All data, code, and verification materials are available under CC BY 4.0 license at: <https://github.com/Ingvar01/voynich-structural-study>

Author Contributions

I.N.O. conceived the study, performed all analyses, developed the model, and wrote the manuscript.

Competing Interests

The author declares no competing interests.

Funding Statement

This research received no specific grant from any funding agency.

Acknowledgments

The author acknowledges the Voynich research community for maintaining and sharing transcriptions and resources. Special thanks to the creators of the Takahashi transcription.

Contact Information

For correspondence, data requests, or collaboration inquiries:

- **Email:** saaantasig@gmail.com
- **GitHub:** <https://github.com/Ingvar01>
- **Repository:** <https://github.com/Ingvar01/voynich-structural-study>

References

- [1] Zandbergen, R. (2016). The Voynich Manuscript. *Voynich.nu*.
- [2] Reeds, J. (1995). The Voynich Manuscript: A Statistical Analysis. *Cryptologia*.
- [3] Rogers, H. (2004). The Voynich Manuscript: An Elegant Enigma. *Cryptologia*.
- [4] Reedy, J. (1974). The Voynich Manuscript: A Hoax? *Yale University Press*.
- [5] Montemurro, M. A., & Zanette, D. H. (2013). Keywords and Co-occurrence Patterns in the Voynich Manuscript. *PLOS ONE*.
- [6] Landini, G. (2001). Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis. *Cryptologia*.
- [7] Reddy, S., & Knight, K. (2011). What We Know About the Voynich Manuscript. *ACL*.
- [8] Takahashi, T. (2011). Takahashi Transcription of the Voynich Manuscript. *Voynich Manuscript Research*.
- [9] Janick, J., & Tucker, A. O. (2004). The Voynich Manuscript: The Herbal Section. *Journal of the Society for the History of Natural History*.
- [10] Yates, F. A. (1966). *The Art of Memory*. University of Chicago Press.