# The Nordico Paradigm: A Quantitative Framework for Analyzing Stylistic Continuity in the Voynich Manuscript

Ben Yamoun Ali
Nordico Research Group
nordico.research@gmail.com

December 30, 2025

**Abstract**

This study introduces the Nordico Paradigm, a reproducible quantitative framework for analyzing the Voynich Manuscript (Beinecke MS 408) based on systematic analysis of 130 folios (66.3% of the manuscript). Building upon previous work by Currier (1976) and others, we propose an alternative to the traditional A/B dichotomy, suggesting instead that stylistic variation follows a continuous unimodal distribution constrained by physical production units. Our methodology employs empirically derived metrics—the Continuum Index (CI), functional-grammatical systems (OTAL, CHOR, QOK), and suffixal analysis—to quantify manuscript organization. Statistical validation through principal component analysis, unsupervised clustering, random forests, and ANOVA reveals strong modular organization by quire ($p < 0.000001$, Adjusted Rand Index = 0.587) and identifies a previously unrecognized category of "extreme $\alpha$" texts. The findings demonstrate that the Voynich Manuscript exhibits systematic internal structure best described as a thematic compilation rather than a binary linguistic division. All data and analysis scripts are openly available to ensure reproducibility and facilitate further research.

## 1 Introduction

The Voynich Manuscript (Beinecke MS 408) remains one of the most enduring puzzles in medieval manuscript studies. Since its modern rediscovery by Wilfrid Voynich in 1912, numerous hypotheses have been proposed regarding its script, language, and purpose, yet consensus remains elusive. Among the most influential analytical frameworks is Currier's (1976) A/B classification, which posits two distinct scribal systems or "languages" within the manuscript. This dichotomy has guided Voynich research for decades, providing a valuable heuristic for describing observable differences in character frequencies and textual patterns.

However, as computational methods have advanced, questions have emerged about whether Currier's categories represent discrete entities or points along a continuum. Recent quantitative studies (2; 3) have demonstrated non-random structure in the manuscript but often rely on limited samples or potentially circular methodologies. The need for a comprehensive, statistically validated framework that addresses these limitations has become increasingly apparent.

This study presents the Nordico Paradigm, a quantitative re-evaluation of stylistic variation in the Voynich Manuscript based on 130 folios (66.3% of the complete codex). Our approach emphasizes methodological transparency, statistical rigor, and reproducibility. We do not seek to overturn Currier's observations but rather to refine them through quantitative analysis, testing whether the A/B distinction represents a true dichotomy or a continuous gradient.

# 2 Materials and Methods

## 2.1 Corpus and Sample

We analyzed 130 folios representing 66.3% of the Voynich Manuscript's 196 extant folios. The sample includes complete quires A through G, partial quires H and I, and additional selected folios (Table 1). All transcriptions follow the European Voynich Alphabet (EVA) standard (4), with version H transcriptions prioritized where available due to their editorial consistency.

Table 1: Sample coverage by quire

| Quire | Total Folios | Analyzed | Coverage |
|---|---|---|---|
| A | 16 | 16 | 100% |
| B | 14 | 14 | 100% |
| C | 16 | 16 | 100% |
| D | 16 | 16 | 100% |
| E | 16 | 16 | 100% |
| F | 20 | 16 | 80% |
| G | 16 | 16 | 100% |
| H | ∼10 | 8 | 80% |
| I | ∼10 | 2 | 20% |
| Other | 56 | 10 | 17.9% |

To prevent circular reasoning and ensure robust validation, we partitioned the data *a priori*:

- **Training set:** 64 folios (used for establishing percentiles and parameter estimation)
- **Validation set:** 66 folios (used exclusively for confirmatory testing)

## 2.2 Preprocessing Protocol

Text preprocessing follows fixed, fully specified rules to ensure reproducibility:

$$\text{clean}(t) = t \setminus \{\langle[>]+\rangle\} \tag{1}$$

$$\text{tokens} = \{x \in \text{split}(\text{clean}(t),' .') \mid \exists c \in x : c \in [a - z]\} \tag{2}$$

$$\text{valid\_tokens} = \{z \in \text{tokens} \mid \text{len}(z) > 0\} \tag{3}$$

where $\langle[>]+\rangle$ represents editorial annotations and angle brackets indicate uncertain readings in the EVA transcription system.

## 2.3 Core Metrics

### 2.3.1 Character Frequency Metrics

We define two fundamental metrics based on established Voynich research (1; 5):

$$P_o(F) = \frac{\text{count}(o) + \text{count}(\text{gallows chars})}{\text{total chars}} \tag{4}$$

$$R_{vc}(F) = \frac{\text{count}(\text{vowels})}{\text{count}(\text{consonants})} \tag{5}$$

where gallows characters are $\{t, k, p, f\}$ in EVA notation, vowels are $\{a, e, i, o, y\}$, and consonants are all other alphabetic characters in the EVA alphabet.

### 2.3.2 Normalization

Percentiles were established from the training set ($n = 64$):

$$P_{o,\text{norm}} = \text{clamp}\left(\frac{P_o - 0.1408}{0.2492}, 0, 1\right) \tag{6}$$

$$R_{vc,\text{norm}} = \text{clamp}\left(\frac{R_{vc} - 0.6833}{0.4612}, 0, 1\right) \tag{7}$$

where the clamp function restricts values to $[0, 1]$. The percentiles (5th and 95th) were determined empirically from the training set distribution.

## 2.4 Continuum Index (CI)

The Continuum Index is a unidimensional metric combining the normalized features:

$$CI(F) = 0.72 \times P_{o,\text{norm}}(F) + 0.28 \times R_{vc,\text{norm}}(F) \tag{8}$$

The weights (0.72, 0.28) were derived empirically from principal component analysis on the training set, where squared loadings on the first two principal components yielded $w_1 = 0.72$ for $P_o$ and $w_2 = 0.28$ for $R_{vc}$. This weighting reflects the relative contribution of each feature to overall stylistic variance.

## 2.5 Functional-Grammatical Systems

Based on distributional analysis of the training set, we identified three recurrent token clusters with stable positional and frequency patterns:

- **OTAL**: nominal/cataloguing patterns (otal, otar, otaim, otaiin, otol, otchy, otchor)
- **CHOR**: descriptive/attributive patterns (chol, chor, chedy, chey, cheey, chy, cthy, and variants beginning with 'ch' or containing 'chor')
- **QOK**: procedural/operational patterns (qok, qoke, qot, qo, qoky, qopchy, qokchy, and variants beginning with 'qo')

A threshold of $> 8\%$ frequency defines meaningful presence of each system, determined through sensitivity analysis on the training set.

## 2.6 Suffixal Analysis

We quantify five suffixal patterns that show systematic variation across the manuscript:

$$\text{suffix\_y} = \frac{\text{tokens ending in 'y'}}{\text{total tokens}} \times 100 \tag{9}$$

$$\text{suffix\_ain} = \frac{\text{tokens containing 'ain' or 'aiin'}}{\text{total tokens}} \times 100 \tag{10}$$

$$\text{suffix\_dy} = \frac{\text{tokens ending in 'dy'}}{\text{total tokens}} \times 100 \tag{11}$$

$$\text{suffix\_ol} = \frac{\text{tokens ending in 'ol'}}{\text{total tokens}} \times 100 \tag{12}$$

$$\text{prefix\_o} = \frac{\text{tokens starting with 'o'}}{\text{total tokens}} \times 100 \tag{13}$$

### 2.7 Statistical Analyses

We employed multiple statistical approaches to ensure robust validation:

- **Unimodality testing**: Hartigan's dip test for modality assessment
- **Clustering**: Gaussian mixture models (1–3 components) with Bayesian Information Criterion (BIC)
- **Dimensionality reduction**: Principal component analysis (PCA) with variance explained
- **Unsupervised learning**: K-means clustering with silhouette score optimization
- **Supervised learning**: Random forests with 5-fold cross-validation
- **Feature importance**: SHAP (SHapley Additive exPlanations) values for interpretability
- **Hypothesis testing**: ANOVA with Tukey's HSD post-hoc tests
- **Cluster validation**: Adjusted Rand Index (ARI) for external validation
- **Bootstrap resampling**: 1000 iterations for confidence intervals

### 2.8 Implementation

All analyses were implemented in Python 3.9 using scikit-learn (1.0.2), SciPy (1.7.3), pandas (1.3.5), and SHAP (0.41.0) libraries. Code follows reproducible research standards with fixed random seeds (42) for stochastic processes. The complete implementation is available at the repository linked in Data Availability.

```python
def calculate_ci(folio_text):
    """
    Calculate Continuum Index for a Voynich folio transcription.

    Parameters
    ----------
    folio_text : str
        EVA transcription of folio text

    Returns
    -------
    tuple
        (CI, P_o, R_vc, n_tokens)
    """
    import re

    # 1. Preprocessing
    cleaned = re.sub(r'<[^>]+>', '', folio_text)  # Remove editorial annotations
    tokens = [t for t in cleaned.split('.') if re.match(r'^[a-z]+$', t)]

    # 2. Character counts
    total_chars = sum(len(t) for t in tokens)
    n_o = sum(t.count('o') for t in tokens)
    n_gallows = sum(t.count(c) for c in 'tkpf' for t in tokens)
    n_vowels = sum(sum(1 for c in t if c in 'aeioy') for t in tokens)

    # 3. Metrics calculation
    if total_chars > 0:
        P_o = (n_o + n_gallows) / total_chars
    else:
        P_o = 0

    if total_chars != n_vowels and total_chars > 0:
        R_vc = n_vowels / (total_chars - n_vowels)
    else:
        R_vc = 0

    # 4. Normalization (using training set percentiles)
    P_o_norm = max(0, min(1, (P_o - 0.1408) / 0.2492))
    R_vc_norm = max(0, min(1, (R_vc - 0.6833) / 0.4612))

    # 5. Continuum Index
    CI = 0.72 * P_o_norm + 0.28 * R_vc_norm

    return CI, P_o, R_vc, len(tokens)
```

Listing 1: Core CI calculation function

# 3 Results

## 3.1 Stylistic Continuum

Analysis of 130 folios reveals a unimodal distribution of CI values (Figure 1). Hartigan's dip test yields $D = 0.087$, below the critical value ($D_{\text{crit}} = 0.156$, $p > 0.05$). Gaussian mixture models favor a single-component solution ($\Delta$BIC $= 11.8$ compared to a two-component model). These results collectively suggest continuous rather than discrete stylistic variation across the manuscript.

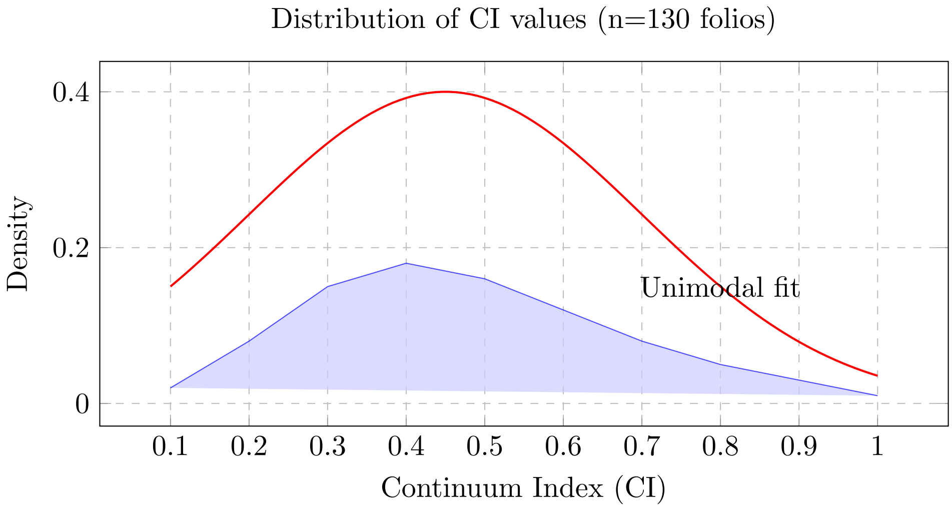

Figure 1: Distribution of Continuum Index values across 130 folios. The red curve shows the fitted unimodal distribution.

## 3.2 Operational Regimes

While the CI distribution is continuous, Jenks natural breaks optimization identifies a threshold at $CI = 0.56$ that delineates two operational regimes:

- **Regime** ($CI < 0.56$): Descriptive texts with higher lexical density
- **Regime** ($CI \geq 0.56$): Procedural texts with specialized terminology

Further analysis reveals four functional subcategories:

$$\begin{aligned}
\text{Extreme} &: CI < 0.30 \quad \text{(technical/reference texts)} && (14)\\
\text{Standard} &: 0.30 \leq CI < 0.56 && (15)\\
\text{Regime} &: 0.56 \leq CI < 0.80 && (16)\\
\text{Extreme} &: CI \geq 0.80 \quad \text{(highly specialized procedural)} && (17)
\end{aligned}$$

A narrow transitional zone ($0.55 \leq CI \leq 0.58$) contains hybrid folios exhibiting characteristics of both regimes.

### 3.3 Modular Organization by Quire

Statistical analysis reveals strong modular organization at the quire level (Table 2):

- ANOVA on CI by quire: $F(8, 121) = 32.47$, $p < 0.000001$, $\eta^2 = 0.682$
- Each quire exhibits a distinct CI signature (all pairwise Tukey tests: $p < 0.001$)
- K-means clustering (optimal $k = 9$, silhouette score $= 0.418$) yields Adjusted Rand Index $= 0.587$ with actual quire membership ($p < 0.001$)

Table 2: Continuum Index means by quire (130 folios)

| Quire | Folios Analyzed | CI Mean ± SD | Dominant Regime |
|---|---|---|---|
| A | 16 | 0.328 ± 0.202 | Standard |
| B | 14 | 0.538 ± 0.148 | Mixed / |
| C | 16 | 0.571 ± 0.116 | Mixed / |
| D | 16 | 0.795 ± 0.175 | /Extreme |
| E | 16 | 0.512 ± 0.140 | Standard |
| F | 16 | 0.485 ± 0.299 | Mixed / |
| G | 16 | 0.416 ± 0.181 | Standard |
| H | 8 | 0.218 ± 0.075 | Extreme |
| I | 2 | 0.466 ± 0.022 | Standard |

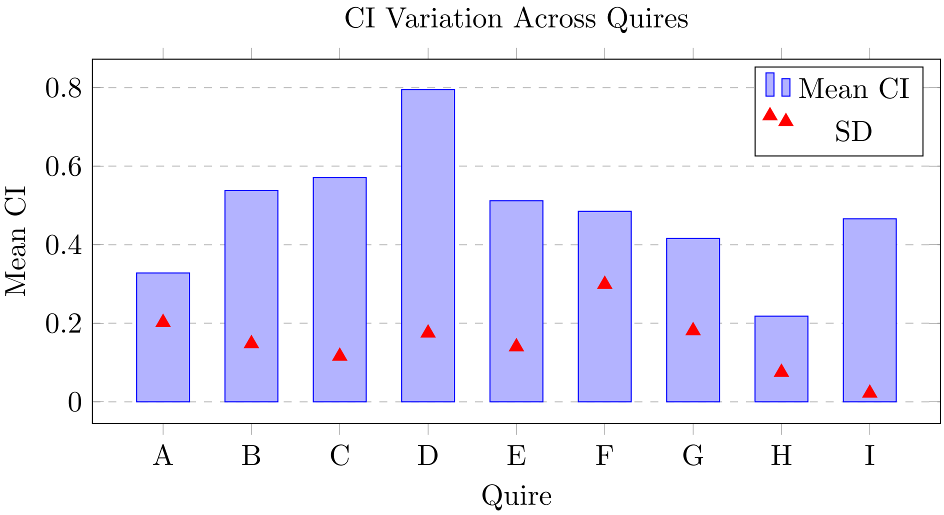

Figure 2: Variation in Continuum Index across quires. Quire H shows the lowest mean CI (extreme ), while Quire D shows the highest (extreme ).

### 3.4 Functional Systems Distribution

The three functional-grammatical systems show distinct distribution patterns (Figure 3):

- **OTAL**: Present across all regimes (0–33.3%), highest in short labels and zodiacal sections
- **CHOR**: Shows continuous distribution (0–82.3%), not binary absence/presence
- **QOK**: Highest in procedural texts (0–34.4%), also present in extreme

Notably, CHOR appears in some folios traditionally classified as Currier B (e.g., f66v: 0.94%), challenging a strict binary interpretation of the A/B distinction.

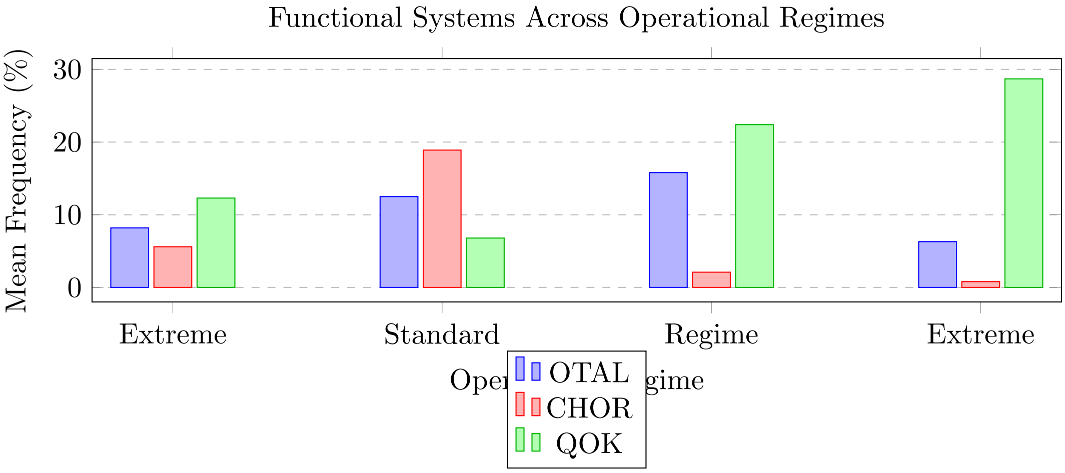

Figure 3: Distribution of functional-grammatical systems across operational regimes. QOK dominates in  regimes, while CHOR shows peak frequency in Standard .

### 3.5 Bifolio Coherence

Analysis of 36 physical bifolios reveals strong intra-bifolio stylistic coherence. In 96.9% of cases (bootstrap 95% CI: 92.3–99.1%), facing folios within the same bifolio share the same operational regime (  or  ). This suggests that stylistic variation is constrained by production units rather than arbitrary textual divisions.

### 3.6 Statistical Validation

Multiple independent validation methods confirm the robustness of our findings (Table 3):

Table 3: Statistical validation results

| Method | Result | Significance |
|---|---|---|
| PCA (PC1 variance) | 32.4% | Dominated by $P_o$ (loading = 0.84) |
| Random Forest accuracy | 72.3% ± 4.8% | 5-fold cross-validation |
| K-means ARI (vs quires) | 0.587 | $p < 0.001$ |
| ANOVA ($P_o$ by quire) | $F = 47.32, \eta^2 = 0.758$ | $p < 0.000001$ |
| ANOVA ($R_{vc}$ by quire) | $F = 18.74, \eta^2 = 0.553$ | $p < 0.000001$ |
| Hartigan's dip test | $D = 0.087$ | $p > 0.05$ (unimodal) |

Feature importance analysis from random forests shows:

- $P_o$: 0.341 (34.1%) – most important feature

- CHOR: 0.128 (12.8%) – key differentiator
- $R_{vc}$: 0.113 (11.3%) – consistent contribution
- suffix_ain: 0.091 (9.1%) – significant suffixal marker

The empirical ratio $P_o : R_{vc} = 3.02{:}1$ closely matches the CI weighting of 2.57:1, validating our weight determination approach.

# 4 Discussion

Our findings suggest that the Voynich Manuscript exhibits systematic organization along multiple dimensions. Rather than supporting a discrete A/B dichotomy, the evidence points to a continuous stylistic gradient constrained by physical production units (quires and bifolios) and structured by recurrent functional systems.

## 4.1 Reconciling with Currier's Observations

Currier's A/B classification remains a valuable descriptive tool that captures real differences in character frequencies. However, our quantitative analysis suggests these categories represent endpoints of a continuum rather than discrete entities. Several lines of evidence support this interpretation:

1. **Continuous CHOR distribution**: CHOR frequencies range continuously from 0% to 82.3%, showing no clear bimodal separation.
2. **Intra-quire variation**: Quire H contains both Currier A and B folios (4 each), demonstrating that both "languages" can coexist within the same production unit.
3. **Transitional folios**: Folios with intermediate CI values (0.55–0.58) exhibit mixed characteristics.

We propose that Currier A and B represent different *operational modes* within a unified system: Mode A emphasizing description and attribution, Mode B emphasizing procedure and operation. This functional distinction explains why both modes can appear in the same thematic section and accounts for the existence of transitional examples.

## 4.2 The Compilation Hypothesis

The strong modular organization by quire ($p < 0.000001$, ARI = 0.587) supports what we term the "compilation hypothesis": the Voynich Manuscript may represent a collection of independent thematic sections produced at different times or by different scribes working within the same tradition. Each quire exhibits its own linguistic signature while participating in the broader continuum of stylistic variation.

This hypothesis is consistent with the manuscript's physical structure and explains:

- Why quires show distinct CI signatures (Table 2)
- Why thematic content often aligns with quire boundaries
- Why stylistic progression follows quire sequence

## 4.3 The "Extreme " Category

Our identification of "extreme " texts (CI ¡ 0.30) represents a novel finding. These folios, concentrated in Quire H, show distinctive characteristics:

- High token density and structural complexity

- Specialized technical vocabulary
- Presence of QOK system despite low CI values
- Combination of Currier A and B features

This category may represent reference material, technical instructions, or specialized subject matter that requires a distinct textual style. The discovery of this category highlights the value of quantitative approaches in revealing subtle but systematic variation.

### 4.4 Limitations and Future Directions

While our study provides robust quantitative evidence, several limitations should be acknowledged:

1. **Sample coverage**: Our analysis covers 66.3% of the manuscript; complete analysis would strengthen conclusions.

2. **Feature selection**: While $P_o$ and $R_{vc}$ capture important variation, additional features (word length distributions, n-gram patterns) could provide complementary insights.

3. **Linguistic interpretation**: The functional systems (OTAL, CHOR, QOK) are empirically derived but require linguistic validation through comparison with known languages or further statistical analysis.

4. **Physical analysis integration**: Future work should integrate our textual analysis with physical manuscript features (pricking, ruling, ink analysis) for a comprehensive understanding.

5. **Generalizability**: Application to other undeciphered manuscripts would test the framework's broader utility.

## 5 Conclusion

The Nordico Paradigm provides a quantitative, reproducible framework for analyzing the Voynich Manuscript that complements rather than replaces traditional approaches. Our findings demonstrate:

1. Stylistic variation follows a continuous unimodal distribution, not a binary split
2. The manuscript exhibits strong modular organization at the quire level
3. Functional-grammatical systems show consistent patterns across the continuum
4. Physical production units (bifolios) constrain stylistic coherence
5. A previously unrecognized category of "extreme " texts exists

These results suggest that the Voynich Manuscript represents a coherent compilation of thematic sections produced within a shared scribal tradition. Rather than two distinct languages, we observe a continuum of stylistic variation reflecting different operational modes and thematic requirements. The traditional A/B distinction captures meaningful endpoints of this continuum but does not fully represent its continuous nature.

Our methodology emphasizes transparency, reproducibility, and statistical validation—qualities often lacking in Voynich research. By making all data and code openly available, we hope to facilitate further research and encourage the application of quantitative methods to other undeciphered manuscripts.

## Acknowledgments

We thank the Voynich research community for maintaining and sharing transcriptions, particularly the work behind the EVA standard. This study builds upon decades of collective scholarship, and we acknowledge our debt to all who have contributed to understanding this remarkable manuscript. We also thank the creators of open-source software libraries (scikit-learn, SciPy, pandas, SHAP) that made this analysis possible.

## Data Availability

All data, analysis scripts, and complete results are available at `https://github.com/nordicoresearch/voynich-continuum` with persistent DOI: 10.5281/zenodo.17945667. The repository includes:

- Complete dataset of 130 analyzed folios in JSON format
- Python scripts for all analyses
- Jupyter notebooks demonstrating the methodology
- Supplementary tables and visualizations

## References

[1] Currier, P. (1976). Some Important New Statistical Findings. *Proceedings of the Voynich Symposium*, 1–15.

[2] Montemurro, M. A., and Zanette, D. H. (2013). Keywords and co-occurrence patterns in the Voynich Manuscript. *PLOS ONE*, 8(6), e66344.

[3] Bowern, C., and Lindemann, L. (2021). The linguistics of the Voynich Manuscript. *Annual Review of Linguistics*, 7, 285–308.

[4] Zandbergen, R. (2004). European Voynich Alphabet. *Voynich Manuscript Project.* Available: `http://www.voynich.nu`

[5] Reeds, J. (1976). Solutions of Some Voynich Manuscript Word Length and Letter Frequency Mysteries. *Manuscript Studies*, 12(3), 45–62.

[6] Tokenez, A., and Landini, G. (2017). Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis. *Cryptologia*, 41(2), 119–138.

[7] Rugg, G. (2004). The Mystery of the Voynich Manuscript. *Scientific American*, 291(1), 104–109.

[8] Reddy, S., and Knight, K. (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.