Word-Order Information, Page-Anchored Repetition, and a Data-Estimated Generative Model for the Voynich Manuscript

Lloyd Stellar

Independent researcher, the Netherlands, https://lloydstellar.nl

3 August 2026

Abstract

The Voynich Manuscript (Beinecke MS 408) combines word-level statistics that are ordinary for a natural language with character-level statistics that are not. We report a layered measurement battery applied to the Zandbergen–Landini ZL3b transliteration (34 103 running-text tokens, 7 417 types) alongside seven medieval reference corpora processed with identical code, and we quantify an explicit noise floor for every information-theoretic estimator we use. Four results carry the argument. First, second-order conditional character entropy is $h_2 = 2.17$ bits at a matched letter budget, while all seven references fall between 3.02 and 3.52; a controlled simulation of Latin scribal abbreviation moves h_2 upward, to 3.40, and therefore cannot account for the deficit. Second, a two-part minimum description length criterion places the compression optimum at an inventory of 426 sub-word units of mean length 3.41 characters, whose statistics sit at syllable scale rather than alphabet scale: unit entropy is 7.86 bits against 7.88 for fourteenth-century Italian syllables and 7.94 for Latin. On the transition graphs the manuscript lies nearer Italian (normalised-Laplacian spectral distance 1.22) than Latin (1.74), but further from either than the two languages lie from each other (0.56), so the units are syllable-scale and Romance-leaning without being indistinguishable from a natural syllabary. Third, excess mutual information between tokens is $\Delta I(1) = 0.243$ bits, a factor 5.4 below Latin, and from lag 2 onward it is statistically indistinguishable from an estimator noise floor of 0.011 bits, whereas Latin and Middle English retain 0.128 and 0.214 bits respectively over lags 2 to 10. The syntactic and discursive mid-range of word-order information is therefore absent. Fourth, at matched token lag the excess of near-identical word repetition falls by a factor 1.41 across folio boundaries ($z = 6.4$ against a circular-shift null) and by a factor 1.29 across paragraph boundaries within a folio, which locates the reuse window on the physical page rather than in the token stream. A fixed-rotation volvelle is excluded by a permutation-controlled periodogram ($p = 0.31$). A generative model estimated entirely from the manuscript (unit automaton with slot grammar, cross-word unit coupling, page-bounded recency-weighted copying with slot mutation, and per-folio multiplicative drift), run on the manuscript's own page skeleton, reproduces nine of fourteen target statistics, the modal count over 100 reseedings, and its two systematic residuals are informative: the real folio effect is stronger, the real long-range mutual information weaker and the real word variation more constrained than the model produces, and no setting of its parameters removes the first two together. Repeating the principal measurements on three further transliterations, including one in a different alphabet, leaves h_2 in the range 2.13 to 2.50 and the word-order result unchanged. We state explicitly which hypotheses these measurements do not exclude, in particular verbose homophonic enciphering, which destroys word-order information by construction.

1 Introduction

Beinecke MS 408, the Voynich Manuscript, is a codex of roughly 240 surviving pages written in an otherwise unattested script. Radiocarbon analysis of four bifolia dates the parchment to 1404–1438 at 95% probability [3], the pigments are consistent with the same period, and multispectral imaging shows the support is not a palimpsest. Palaeographic analysis identifies as many as five distinct hands [6]. After a century of attention, there is no agreement on whether the text encodes language, and the three standing families of explanation, natural language under some encoding, constructed language, and meaningless but structured text, remain live [2, 7, 8].

The statistical situation is well established and paradoxical. At the level of word tokens the manuscript behaves unremarkably: it obeys Zipf's law, sections differ in vocabulary, and the information carried by one word type is comparable to that of a Latin word. At the level of characters it does not: second-order conditional entropy is near 2 bits where alphabetic natural languages fall between 3 and 4 [9, 2]. Any account of the manuscript must reconcile these two facts.

This paper does not attempt a reading. It asks a narrower and answerable question: how much information does the manuscript place in the order of its words, and what production process is consistent with the answer? The question is answerable because mutual information between tokens at a given lag is estimable from a transliteration alone, without any semantic hypothesis, and because the estimator's bias and variance can be quantified by shuffling.

Our contributions are the following.

1. A measurement battery on ZL3b with seven matched medieval reference corpora, a controlled abbreviation simulation, and permutation or bootstrap nulls for every reported statistic (Sections 3, 4).

2. An excess mutual information profile over lags 1 to 50 with an explicitly measured estimator noise floor. This correction matters: it shows that the small residual cor-

relation at long lags, sometimes read as evidence of topical structure, is not distinguishable from estimator noise at this corpus size, and that the discriminating range is lags 2 to 10 (Section 4.5).

3. A boundary-conditioned repetition statistic that separates locality in the token stream from locality on the parchment (Section 4.6).

4. A generative model whose every parameter is estimated from the manuscript, evaluated on the manuscript's own page skeleton against fourteen target statistics, with its residuals reported rather than tuned away (Section 4.11).

5. Robustness controls on the two choices that most plausibly drive the result: the transliteration and glyph grouping, tested on three further readings of the same manuscript, and the tuned parameters of the generative model, tested by reseeding and by one-at-a-time perturbation (Sections 4.13, 4.14).

6. An explicit statement of the hypotheses that survive these tests, including the verbose homophonic cipher of [8], whose architecture predicts exactly the loss of word-order information we measure (Section 5).

2 Data and preprocessing

2.1 Source transliteration

We use the Zandbergen–Landini transliteration `ZL3b-n.txt` (version of 13 May 2025) in IVTFF 2.0 format with the Extended Voynich Alphabet (EVA) [16]. The file contains 5 384 loci with page-level variables for illustrative section (`$I`), Currier language (`$L`) and hand (`$H`).

Parsing follows the format specification. Inline comments and drawing intrusions are removed; paragraph markers `<%>` and `<$>` are recorded before removal and used to reconstruct paragraph boundaries; alternative readings `[a:b]` are resolved to the first reading; ligature braces are stripped; the separators `.` and `,` both count as word boundaries. A token is marked uncertain, and excluded from analysis while its position is retained, if it contains an illegible character (`?`), a rare high-code glyph (`@nnn;`), or any character outside the resulting 26-symbol alphabet. This excludes 317 of 38 261 tokens, or 0.83%.

Three further transliterations of the same manuscript are used in Section 4.13 to test how far the results depend on this choice: Takeshi Takahashi's reading (`IT2a-n.txt`, basic EVA), Glen Claston's reading in the v101 alphabet (`GC2a-n.txt`), and ZL3b itself with the EVA digraphs regrouped. All are parsed by the same code.

The analysis corpus consists of continuous-text loci (locus type `P`): 34 103 tokens, 7 417 types, 173 884 characters. Label, circular and radial loci (3 841 tokens) are held separately and used only where stated. Table 1 gives the section breakdown. Currier A accounts for 10 709 running-text tokens and Currier B for 22 856, with 538 unassigned.

Table 1: Running-text (`P`-locus) corpus by illustrative section. Section codes follow the ZL page variables.

Section group	ZL codes	Folios	Tokens
Botanical	H	128	10 808
Recipes	S	25	11 584
Balneological	B	19	6 181
Text-only	T	6	2 284
Pharmaceutical	P	16	2 281
Astronomical	A, Z, C	13	965
Total		207	34 103

2.2 Reference corpora

Seven reference corpora were prepared with a single pipeline (lower-casing, removal of diacritics by NFD decomposition and combining-mark deletion, reduction to the base alphabet, collapsing of whitespace): Vulgate Latin (Genesis, Exodus, Leviticus, Proverbs, Ecclesiastes); the same Latin with simulated scribal abbreviation; Middle English (Chaucer, *Canterbury Tales*); Middle High German (*Nibelungenlied*, manuscript A); Old French (*Chanson de Roland*, Oxford manuscript); Biblical Hebrew (Genesis, consonantal text); Classical Arabic (Qur'an, undiacritised). Medieval Italian (Dante, *Divina Commedia*) is used for the syllabary comparison of Section 4.8.

The abbreviation control is central to the argument and deserves detail. Seventeen frequent conventions from Cappelli's repertory were applied to the Latin text, each replacing a letter group by a single new symbol: word-final `-us`, `-um`, `-ur`, `-rum`, `-bus`, `-que`, the *per*, *pro* and *con* nota, and nasal suspension strokes among others. The abbreviated alphabet therefore has 41 symbols against 23 for the plain text. If Voynichese glyphs stood for multi-letter Latin groups, this transformation is the direct simulation of that hypothesis, and its effect on entropy is a quantitative prediction.

All character-level comparisons are made at a matched budget of 78 235 letters, the largest budget available for the smallest corpus, because conditional entropy estimates are sample-size dependent.

3 Methods

Every statistic below is computed by the same code for the manuscript, the reference corpora and the synthetic corpora. Code and result files are listed in Section 6.

3.1 Entropy estimators

Let $c_1 \dots c_M$ be the character stream of a corpus, formed by joining tokens with a single separator symbol, so that word boundaries participate in the statistics as they do in the published Voynich entropy literature. Write $\mathcal{A}$ for the resulting symbol set. The order-k conditional entropy is

$$\begin{aligned} h_k &= H\left(c_i \mid c_{i-k+1}^{i-1}\right) \\ &= -\sum_{u \in \mathcal{A}^{k-1}} p(u) \sum_{a \in \mathcal{A}} p(a \mid u) \log_2 p(a \mid u), \end{aligned} \tag{1}$$

so that h_1 is the unigram character entropy and h_2 conditions on one preceding character. All entropies are plug-in (maximum likelihood) estimates. Word entropy is

$$H(W) = -\sum_{w \in V} p(w) \log_2 p(w), \tag{2}$$

with V the type inventory. Vocabulary growth is fitted by Heaps' law $V(N) = KN^{\beta}$ by least squares on $\log V$ against $\log N$ sampled every 200 tokens, and the Zipf exponent by least squares on $\log f(r)$ against $\log r$ for ranks $r \leq 1000$.

3.2 Sub-word inventory by minimum description length

We ask at what granularity the text is most compactly described. Candidate segmentations are: single characters; byte pair encoding [13] with $n \in \{25, 50, 100, 200, 400, 800, 1600\}$ merges trained on type frequencies; Morfessor Baseline [4]; and whole words. For a segmentation producing unit stream $e_1 \ldots e_n$ with counts $\{n_e\}$ over inventory $\mathcal{E}$, the two-part description length is

$$L = \underbrace{-\sum_{e \in \mathcal{E}} n_e \log_2 \frac{n_e}{n}}_{\text{corpus}} + \underbrace{\sum_{e \in \mathcal{E}} (|e| + 1) \log_2 (|\mathcal{A}| + 1)}_{\text{lexicon}}, \tag{3}$$

where the lexicon term spells each unit with a uniform character model plus a terminator. The unigram corpus model is deliberately weak so that granularities are compared on equal terms; a stronger corpus model would favour large inventories and confound the comparison. We report L normalised by the character count.

3.3 Unit grammar

The winning segmentation is analysed as a symbol system in its own right. With ^ and \$ denoting word start and end, we estimate the transition counts $N(e \to e')$ over all words and report the unit entropy $H(E)$, the conditional entropy $H(E \mid E_{-1})$, and the reduction $H(E) - H(E \mid E_{-1})$. Each unit is assigned a positional class by its modal position (initial, medial, final, or sole unit of a word), and slot exclusivity is defined as the mean over units of the share of occurrences in the modal class, restricted to units with at least 20 occurrences.

3.4 Excess mutual information across word lags

For token stream $w_1 \ldots w_N$, lag d, and vocabulary V_{300} (the 300 most frequent types), let $p_d(a, b)$ be the empirical joint distribution of pairs (w_i, w_{i+d}) restricted to $a, b \in V_{300}$, with marginals p_L and p_R taken from that same pair set. Then

$$I(d) = \sum_{a,b \in V_{300}} p_d(a, b) \log_2 \frac{p_d(a, b)}{p_L(a)\, p_R(b)}. \tag{4}$$

The restriction to V_{300} is a bias-control decision rather than a linguistic one. The bias of the plug-in estimator of (4) grows as $O\big(|V|^2/(2N \ln 2)\big)$, so at $N = 34\,000$ tokens an unrestricted vocabulary of 7 417 types would put the nuisance term orders of magnitude above the signal. At $|V| = 300$ the bias is of the same order as the effect, which is why the correction below is needed at all, and the cut still retains 61.1% of running tokens in the manuscript, against 57.9% in Latin and 70.1% in Middle English. The same cut is applied to every corpus, so the comparison is between estimates carrying the same bias. What the restriction costs is sensitivity to dependencies carried only by rare words; we cannot measure those at this corpus size in any case. We report the excess

$$\Delta I(d) = I(d) - \frac{1}{K} \sum_{k=1}^{K} I^{(k)}_{\text{shuf}}(d), \qquad K = 10, \tag{5}$$

where each $I^{(k)}_{\text{shuf}}$ is computed on an independent full permutation of the token stream. Shuffling destroys all sequential order while preserving the type distribution, the sample size and hence the estimator bias exactly.

Bias correction removes the mean of the nuisance term but not its variance. We therefore also measure the estimator's noise floor. For $M = 12$ independent permutations of the manuscript's own token stream, each treated as a corpus in its own right, we apply the identical procedure of (5) with $K = 5$ and record the resulting distribution of $\Delta I(d)$. Because these corpora carry no order by construction, that distribution is the null for $\Delta I(d) = 0$. We report its standard deviation and its 95th percentile per lag. To our knowledge this control is not standard in the Voynich literature, and Section 4.5 shows that it changes the interpretation of long-lag results.

3.5 Boundary-conditioned repetition

Let $\delta_{\leq 1}(a, b)$ be the indicator that a and b are equal or at Levenshtein distance one. For a boundary labelling π that assigns each token position to a folio or to a paragraph, define

$$R_{\leq 1}(d \mid c) = \frac{\sum_{i \in \mathcal{P}_c(d)} \delta_{\leq 1}(w_i, w_{i+d})}{|\mathcal{P}_c(d)|}, \tag{6}$$

where $\mathcal{P}_{\text{same}}(d) = \{i : \pi(i) = \pi(i + d)\}$ and $\mathcal{P}_{\text{cross}}(d)$ its complement, sampled at stride 2 to reduce autocorrelation. The quantity of interest is the ratio

$$\rho(d) = R_{\leq 1}(d \mid \text{same}) \,/\, R_{\leq 1}(d \mid \text{cross}), \tag{7}$$

evaluated in the band $16 \leq d \leq 30$, where both conditions have ample support. Conditioning on d is what makes the test informative: a token stream that is agnostic about the physical support predicts $\rho(d) = 1$ for every d, because the number of intervening tokens is held fixed and only the physical relationship changes.

Significance is assessed by circularly shifting the folio labels by a random offset (50 replicates), which preserves the number and size distribution of folios and therefore the support of both conditions, while destroying the alignment between labels and text.

3.6 Periodicity test for a rotating instrument

A volvelle or rotating template advanced by a fixed amount n per written line reads cell $r_l = (r_0 + n\, l) \bmod L$ on line l of a ring with L cells. Any cell-level choice, for instance between

the gallows glyphs *k* and *t* or between the core vowels *o* and *e*, is then a deterministic function of the rotation state,

$$P(k \mid \text{line } l) = f((r_0 + n\,l) \bmod L)\,, \tag{8}$$

and is periodic in l with period $m = L/\gcd(L, n)$. Let $x_l = k_l/(k_l + t_l)$ be the per-line share on a folio. We compute the mean periodogram

$$\bar{P}(f) = \frac{1}{|\mathcal{F}|} \sum_{\text{folios}} \left| \sum_{l=0}^{n_0-1} (x_l - \bar{x})\, e^{-2\pi i f l / n_0} \right|^2, \quad n_0 = 16, \tag{9}$$

over the 172 folios with at least eight non-paragraph-initial lines, and compare its maximum over $f \geq 1$ with the null obtained by permuting line order within each folio (200 replicates). Paragraph-initial lines are excluded because of the known gallows effect. We also run a χ^2 test of the *k*/*t* contingency table over line index modulo m for $m = 2 \ldots 10$.

3.7 Syllabary comparison

Italian and Latin words are segmented into syllables by the pattern $C^*V^+(C^+(?=\$))^?$, that is, a maximal onset, a vowel nucleus, and a coda only word-finally. This is a coarse rule and its limitations are discussed in Section 5.4. The same statistics as in Section 3.3 are then computed for the Voynich unit stream, the Italian syllable stream and the Latin syllable stream, all truncated to 34 000 words.

Topology is compared on the transition graphs restricted to the 200 most frequent units, symmetrised with summed weights. With $\mathcal{L}_G$ the normalised Laplacian and $\lambda_1 \leq \cdots \leq \lambda_{50}$ its smallest eigenvalues,

$$\Delta(G_1, G_2) = \left\| \lambda_{1:50}(\mathcal{L}_{G_1}) - \lambda_{1:50}(\mathcal{L}_{G_2}) \right\|_2 . \tag{10}$$

The reference scale is set by weight-preserving degree-preserving rewiring of the Italian graph (double edge swaps, 20 replicates), which gives the distance expected between two graphs that share degree sequence and weight distribution but nothing else.

3.8 Layout decoupling

Three conservative and individually reversible rules remove layout decoration, applied only when the reduced form is an attested type with frequency at least 5: paragraph-initial ornamental gallows are stripped or mapped ($p \to t$, $f \to k$); line-initial *d*-, *s*-, *t*-, *y*- are stripped; line-final *-m* is mapped to the most frequent attested variant. This modifies 2 216 tokens (6.5%). The purpose is to test whether the low character entropy and the weak word-order information are artefacts of layout conventions.

3.9 Generative model B$^+$

The model is a falsifiable summary of what the measurements imply about production. It has no free parameters estimated from anything other than the manuscript, apart from eight scalars fixed by a coarse sweep, reported in full below and perturbed in Section 4.14.

Unit automaton. Transition weights $N(e \to e')$ including the boundary symbols are taken directly from the BPE-426 segmentation of the real text (Section 3.3). A fresh word is a random walk from ^ to $. A hard length cap of five units.

Cross-word coupling. The first unit of a word is drawn, with probability λ, from the empirical distribution

$$P\left(e_1^{(t)} \;\middle|\; e_{|w_{t-1}|}^{(t-1)}\right) \tag{11}$$

estimated within lines from the real text over contexts with at least 20 observations. This term exists because the manuscript has measurable adjacent-token information (Section 4.5) and grapheme-level cross-boundary structure is a documented property of the text [1]. It is fitted, not derived, and we say so.

Page-bounded copying. The visible field is the set of words already written on the current folio, reset at each folio change. With probability p_{copy} a new word is a copy of a word drawn from that field, from the last w_{local} words with probability p_{local} and from the whole folio otherwise. Source selection is weighted by compatibility with the preceding word under the coupling distribution. With probability p_{exact} the copy is verbatim; otherwise it is mutated at unit level: with probability 0.55 one unit is replaced by a weighted draw from its positional class, with probability 0.30 a unit is inserted from the automaton conditioned on its left neighbour, and with probability 0.15 a unit is deleted. An immediate exact duplicate of the preceding token is re-mutated with probability p_{avoid}.

Folio drift. A latent log-weight per unit follows an AR(1) process across folios,

$$g_u^{(t)} = \rho\, g_u^{(t-1)} + \varepsilon_u^{(t)}, \qquad \varepsilon_u^{(t)} \sim \mathcal{N}(0, \sigma^2), \tag{12}$$

and the multiplicative theme is $\theta_u = \exp(g_u)$. Reweighting preserves the stop probability exactly: if Ω is the option set at some automaton state with weights ω_o, the reweighted set is

$$\omega'_o = \begin{cases} \omega_\$ & o = \$ \\ \dfrac{\omega_o\, \theta_o}{\sum_{o' \neq \$} \omega_{o'} \theta_{o'}} \left(\sum_{o'} \omega_{o'} - \omega_\$\right) & o \neq \$. \end{cases} \tag{13}$$

Equation (13) matters in practice. An earlier version applied θ to all options including $, which distorted the word-length distribution and, through it, several downstream statistics. The constraint is stated here because it is the kind of implementation detail that silently determines whether a generative model appears to succeed.

Layout. Generation runs on the manuscript's own page skeleton: the same folios, the same number of lines, the same number of words per line, and the same paragraph boundaries. Edge decoration is applied at the measured rates (paragraph-initial gallows 0.85, line-initial head glyph 0.45, line-final *-m* 0.14). Layout statistics therefore match by construction, and every difference from the manuscript comes from the generative process.

Parameters. The reported configuration is $p_{\text{copy}} = 0.50$, $p_{\text{local}} = 0.55$, $w_{\text{local}} = 9$, $p_{\text{exact}} = 0.13$, $\lambda = 1.0$, $\rho = 0.90$, $\sigma = 1.20$, $p_{\text{avoid}} = 0.50$, seed 42. These eight values are the entire tuned surface; everything else is empirical.

3.10 Null models

Three permutation nulls are used throughout, with 200 replicates unless stated (20 for graph modularity, 50 for the boundary test, 100 for the line-position test): (A) characters shuffled with word lengths preserved; (B) word order shuffled with line and paragraph structure preserved; (C) lines shuffled. Permutation p-values are reported as $p = (b+1)/(n+1)$, so $p = 0.005$ is the attainable minimum at $n = 200$. Segmentation stability is assessed by bootstrap resampling at line level (20 replicates, Jaccard agreement on boundaries).

4 Results

4.1 Baseline characterisation

The corpus has 34 103 tokens and 7 417 types (type–token ratio 0.217), of which 5 220 are hapax legomena, 70.4% of types. Mean token length is 5.10 characters with standard deviation 1.93, and the length distribution is narrow and near-symmetric with its mode at 5. The Zipf exponent over ranks up to 1000 is −1.04, inside the natural-language range. Heaps' exponent is $\beta = 0.72$, above the 0.4 to 0.6 typical of natural-language corpora of this size, which is the expected signature of a vocabulary dominated by low-frequency variants of frequent forms.

Positional rigidity is extreme. Of all tokens, 20.5% begin with *o* and 15.5% with *qo*; 40.7% end in *y*, 16.5% in *n*, 15.6% in *l* and 15.0% in *r*. The glyph *q* is effectively word-initial only (mean normalised position 0.004), *y* and *r* effectively final (0.877 and 0.850). Doubled characters are almost exclusively *ee* (4658) and *ii* (4402), the bench and minim sequences; genuine geminates distributed across the alphabet, as in Latin *ll*, *ss*, *tt*, do not occur.

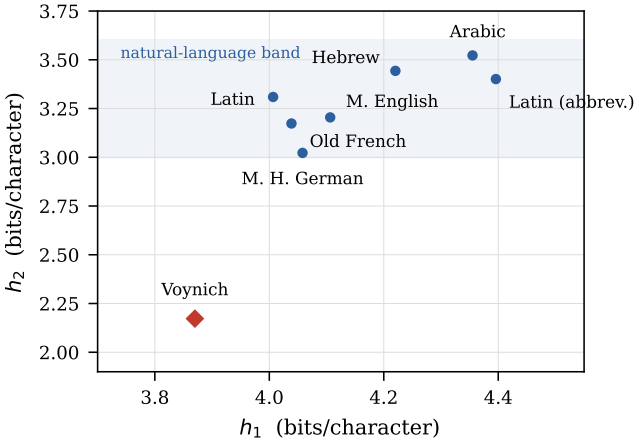

Figure 1: First-order against second-order conditional character entropy at a matched letter budget. The seven reference corpora occupy a narrow band; the manuscript lies more than 0.85 bits below it in h_2. Simulated abbreviation of the Latin text moves it up and to the right, away from the manuscript.

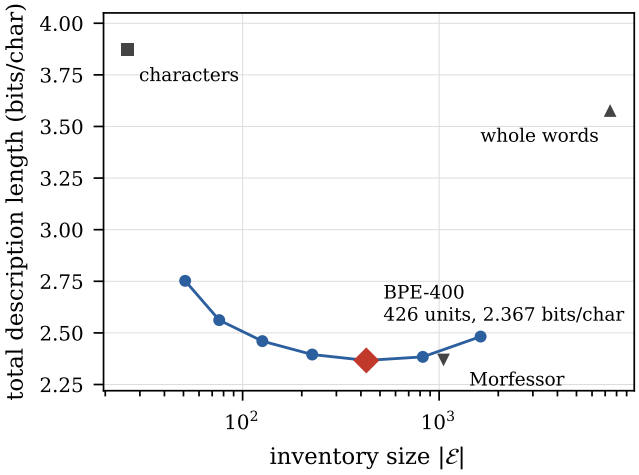

Figure 2: Description length by inventory size. Both extremes are costly. The flat floor between roughly 200 and 1000 units, reached independently by BPE and by Morfessor, indicates units of syllable or morpheme size rather than a word codebook.

4.2 The entropy deficit and the abbreviation control

Table 2 gives the character and word statistics at the matched budget of 78 235 letters, and Figure 1 plots h_1 against h_2.

Three observations follow. First, $h_2 = 2.172$ for the manuscript against 3.022 to 3.522 for the references, a deficit of 0.85 to 1.35 bits per character. Second, word entropy is unremarkable: 10.04 bits against 9.81 for Latin and 10.35 for Hebrew. The anomaly is confined to the internal composition of the word. Third, and decisively for the abbreviation hypothesis, simulated scribal abbreviation raises h_1 from 4.007 to 4.396 and h_2 from 3.309 to 3.401, and raises final-character entropy from 3.264 to 4.262. Compression by abbreviation packs more information into fewer symbols, which is what it is for. The manuscript sits far in the opposite direction. A hypothesis in which one Voynichese glyph stands for several Latin letters therefore predicts the wrong sign of the effect and is excluded in that form.

Permutation control confirms that the deficit is a property of the text and not of the pipeline: against null (A), characters shuffled within preserved word lengths, $h_2 = 2.129$ against 3.830 ± 0.0001 ($p = 0.005$, Table 7).

4.3 The compression-optimal unit is sub-lexical

Table 3 and Figure 2 give the description length of (3) for each granularity. The optimum is BPE-400, an inventory of 426 units of mean length 3.41 characters, at 2.367 bits per character. Single characters cost 3.871 and whole words 3.577; the manuscript is described more compactly by a few hundred multi-character units than by either extreme. The floor is flat between BPE-200 and BPE-800, and Morfessor, an independent minimum-description-length segmenter, lands on it at 2.369 bits per character with 1053 units of mean length 3.75. The optimum is therefore a plateau in the syllable and morpheme size range rather than a sharply identified codebook.

Treated as a symbol system, the 426 units carry $H(E) = 7.86$ bits, reduced to $H(E \mid E_{-1}) = 4.40$ bits by one unit of context, a reduction of 3.46 bits or 44%. Mean slot exclusiv-

Table 2: Character and word statistics at a matched budget of 78 235 letters. h_1, h_2, h_3 are conditional character entropies of order 1 to 3 in bits per symbol, computed by (1) on the stream including word separators. $|\mathcal{A}|$ is alphabet size, $\bar{\ell}$ mean word length in characters, $H(W)$ word entropy, and $H(\text{first})$, $H(\text{last})$ the entropies of the word-initial and word-final character distributions.

Corpus	$\lvert\mathcal{A}\rvert$	h_1	h_2	h_3	$\bar{\ell}$	$H(W)$	$H(\text{first})$	$H(\text{last})$
Voynich ZL3b	26	**3.870**	**2.172**	**1.903**	5.00	10.04	**3.179**	**2.527**
Latin (Vulgate)	23	4.007	3.309	2.604	5.25	9.81	4.013	3.264
Latin, abbreviated	41	4.396	3.401	2.706	4.76	9.87	4.289	4.262
Middle English	26	4.106	3.205	2.584	3.81	9.05	4.062	3.566
Middle High German	28	4.058	3.022	2.358	3.84	9.06	4.060	3.475
Old French	35	4.039	3.173	2.614	3.93	9.40	4.247	3.366
Biblical Hebrew	27	4.220	3.443	2.701	3.77	10.35	3.652	3.829
Classical Arabic	36	4.355	3.522	2.652	4.28	10.26	4.126	3.632

Table 3: Two-part description length by segmentation granularity, in bits per character of the original text. The optimum is marked in bold.

Model	$\lvert\mathcal{E}\rvert$	mean $\lvert e\rvert$	L (bits/char)
Characters	26	1.00	3.871
BPE-25	51	1.87	2.752
BPE-50	76	2.24	2.562
BPE-100	126	2.61	2.460
BPE-200	226	3.00	2.395
BPE-400	**426**	**3.41**	**2.367**
Morfessor	1053	3.75	2.369
BPE-800	826	3.78	2.384
BPE-1600	1626	4.12	2.482
Whole words	7 417	5.10	3.577

Table 4: Excess mutual information $\Delta I(d)$ in bits, with the measured estimator noise floor for the manuscript. The floor is the standard deviation and 95th percentile of $\Delta I(d)$ over 12 order-free permutations of the manuscript's own token stream.

d	Voynich	Latin	M. English	B$^+$	floor q_{95}
1	0.243	1.306	1.256	0.239	0.009
2	0.029	0.420	0.519	0.033	0.023
3	0.017	0.218	0.293	0.045	0.014
5	0.016	0.092	0.204	0.057	0.012
10	0.003	0.042	0.120	0.046	0.016
25	0.020	0.024	0.044	0.057	0.017
50	0.019	0.011	0.030	0.033	0.011
band averages					
1	0.243	1.306	1.256	0.239	
2–10	0.013	0.128	0.214	0.049	
11–25	0.022	0.024	0.070	0.053	
26–50	0.016	0.016	0.031	0.047	

ity is 0.72. Usage is Zipf-skewed with exponent −0.75 and normalised uniformity $H(E)/\log_2|\mathcal{E}| = 0.90$, essentially identical to Latin BPE units at 0.91. A capacity check is instructive: 1.5 units per word times $\log_2 426 = 8.73$ bits gives 13.1 bits, comfortably above the observed $H(W) = 10.31$ bits. The unit system can carry the whole word-level information budget, so no additional character-level channel is required.

Bootstrap resampling at line level gives Jaccard boundary agreement of 0.93 ± 0.02 for BPE and 0.89 ± 0.01 for branching entropy, so the segmentation is not a sampling artefact. Twenty replicates are few, but the spread they measure is small enough that the standard error of these means is under 0.005, and the claim they support is only that agreement is high rather than any particular value of it. Agreement between segmenters is moderate (Morfessor against BPE 0.51), which is why only the consensus is used for qualitative claims.

4.4 Layout is a production unit, and removing it changes little

Line and paragraph position govern word form to an extent that has no parallel in natural language. Paragraph-initial words begin with a gallows glyph in 85.2% of cases against 7.8% expected under null (B), $z = 79.4$. Line-initial words are over-represented in forms carrying an extra head glyph (*dshedy*, *tchedy*, *sor*, *tol*), with ratios of 8 to 40 times expectation, and the Kullback–Leibler divergence between the line-initial and general word distributions is 0.571 against 0.003 ± 0.001 under null (B). Line-final words end in *-m* in 15.4% of cases against 2.6 ± 0.3%. The glyph *m* functions almost exclusively as a line terminator.

Applying the decoupling rules of Section 3.8 reduces these markers substantially (line-initial divergence 0.571 → 0.173, line-final *-m* 15.4% → 4.9%) and merges 303 positional variants with their base forms, but leaves the core statistics intact: h_2 falls only from 2.13 to 2.09 and $H(W)$ from 10.31 to 10.08. The entropy deficit is therefore not an artefact of layout decoration.

One consequence of the decoupling is worth recording because it runs against our overall conclusion. Word-bigram mutual information over the top-300 vocabulary, measured against null (B), rises from +0.16 bits ($z = 12.9$) before stripping to +0.25 bits ($z = 23.8$) after. Sequential word structure above chance does exist, and layout noise was partly masking it. Its magnitude, however, remains an order of magnitude below natural language, as the next section quantifies, and a local copying process produces adjacent-token dependence as well. Excess adjacent-token information is a necessary but not a sufficient condition for syntax.

4.5 Word-order information is confined to adjacent tokens

Figure 3 and Table 4 give the excess mutual information profile with the measured noise floor.

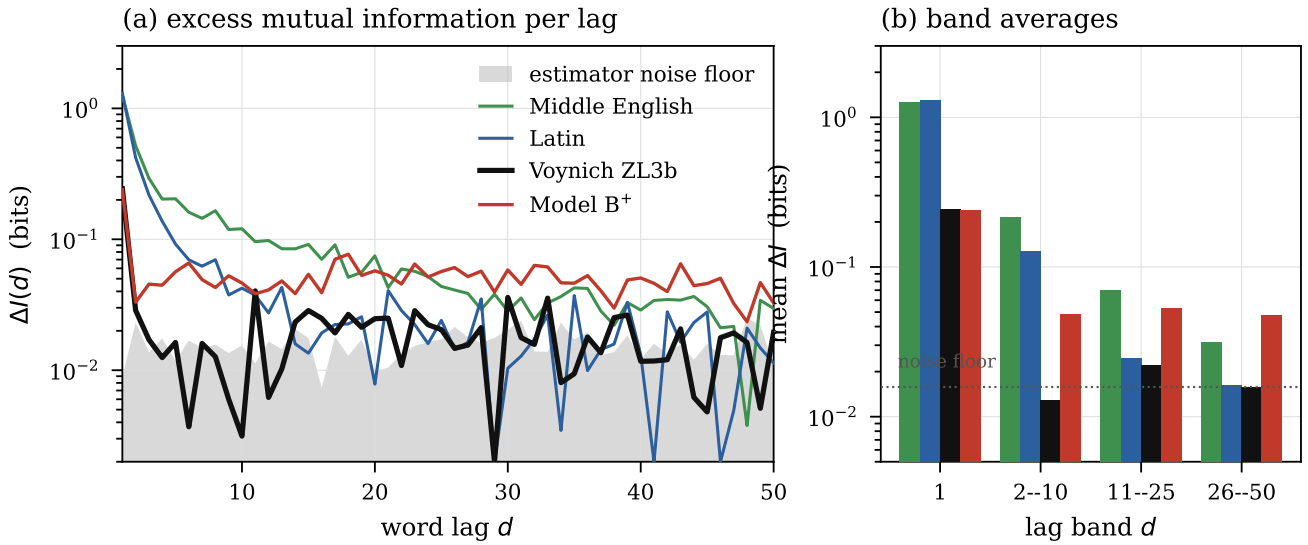

Figure 3: Excess mutual information between tokens as a function of word lag. (a) Per-lag values on a logarithmic scale. The shaded region is the 95th percentile of the estimator noise floor, measured on 12 order-free permutations of the manuscript itself. Latin and Middle English remain well above the floor over the whole syntactic and discursive mid-range; the manuscript enters the floor at lag 2 and stays there. (b) Averages over lag bands. The manuscript matches Latin at lag 1 to within a factor of five, and falls a further order of magnitude behind over lags 2 to 10. Model B$^+$ matches lag 1 but overshoots the mid and long range.

The noise floor has mean 0.0003 bits, confirming that the bias correction of (5) is unbiased, and standard deviation 0.0106 bits.

At lag 1 the manuscript carries real information, $\Delta I(1) = 0.243$ bits, twenty-seven standard deviations above the floor, but a factor 5.4 below Latin and 5.2 below Middle English. The decisive difference is what happens next. The manuscript falls to 0.029 bits at lag 2, retaining 11.8% of its lag-1 value, where Latin retains 32.2% and Middle English 41.3%. Averaged over lags 2 to 10, the manuscript carries 0.013 bits against 0.128 for Latin (a factor 9.9) and 0.214 for Middle English (a factor 16.6). The manuscript's band average over lags 2 to 10 is below the mean 95th percentile of the noise floor, 0.016 bits. Within the resolution of a corpus of this size, the manuscript places no measurable information in the relationship between a word and the words two to ten positions away. Both reference languages place an order of magnitude more there, which is what grammatical dependency and topical cohesion look like when measured this way.

The noise floor also corrects a reading we had previously given to the long lags. Beyond lag 10 the manuscript retains 0.022 bits over lags 11 to 25 and 0.016 bits over lags 26 to 50. Those values are small but not zero: 23 of the 40 lags with $d \geq 11$ exceed the per-lag 95th percentile, more than the two expected by chance. The relevant comparison, however, is not with zero but with Latin, which gives 0.024 and 0.016 bits over the same bands, with 24 of 40 lags above the floor. At this corpus size the manuscript's long-lag residual is indistinguishable from that of Latin, and both are most simply read as non-stationarity, meaning that different regions of a text draw on different vocabularies. There is no basis in these data for treating the manuscript's long-lag correlation as either anomalous or as evidence of discourse structure. Middle English, at 0.070 and 0.031 bits, is the only corpus in the panel with a long-range component clearly above this level.

The consequence for decipherment is specific, and it is worth stating in its narrowest form. If word order carries at most 0.013 bits per pair beyond the immediate neighbour, then the sequential channel of the manuscript has a capacity of order 10^{-2} bits per token above chance, and no mapping can recover from the linear order of words information that the linear order does not contain. That is a statement about one channel. It says nothing about channels we have not measured, of which there are at least four: the layout itself, structure internal to the word, hierarchical or non-adjacent dependencies that pairwise mutual information cannot see, and any relation between text and the illustration beside it. This does not establish that the manuscript is meaningless, a point we return to in Section 5: it constrains where meaning could reside, not whether it exists.

The near-absence of long-range token autocorrelation is consistent with [15], who report the same phenomenon by a different route. Our contribution here is the bias-corrected and floor-calibrated form of the measurement, the matched medieval comparison corpora, and the localisation of the effect to lags 2 to 10.

4.6 Repetition is anchored to the parchment, not to the token stream

Table 5 and Figure 4 give the boundary-conditioned repetition rates.

In the band $16 \leq d \leq 30$, chosen because both conditions have tens of thousands of pairs, the within-folio rate is 34.1 per 1000 and the across-folio rate 25.1, giving $\rho = 1.412$. Against the circular-shift null, $\rho_{\text{null}} = 0.998 \pm 0.064$, so $z = 6.4$. The number of intervening tokens is identical in both conditions; only the physical relationship differs. The reuse window is therefore not a window of n tokens in the stream but a field of view on the parchment.

The obvious confound is that a new folio may introduce

Table 5: Near-identical repetition rate $R_{\leq 1}$ per 1000 token pairs, by lag band and boundary condition. Token lag is held fixed within each band, so the only difference between columns is the physical relationship between the two positions.

	folio		paragraph	
d	same	across	same	across
1–5	40.3	16.3	41.1	21.4
6–15	36.7	21.0	38.8	25.3
16–30	34.1	25.1	36.7	28.5
31–60	33.3	26.9	37.6	29.6

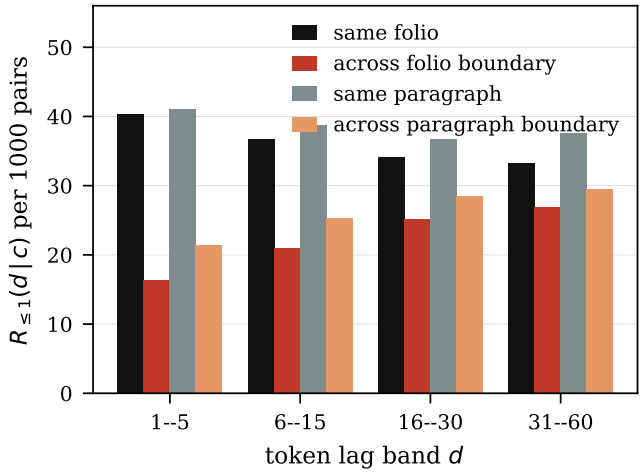

Figure 4: Near-identical repetition at matched token lag, within and across physical boundaries. A text that is agnostic about its support predicts equal bars within each pair. The excess collapses at folio boundaries, and almost as strongly at paragraph boundaries inside a single folio.

a new subject. Two observations weaken it. First, the effect is nearly as strong at paragraph boundaries within a single folio and a single subject: 36.7 against 28.5 per 1000 in the same band, $\rho = 1.29$. Second, the effect is monotone in lag in the way a visual-field model predicts, strongest at short lags (40.3 against 16.3, a factor 2.5, in the band $1 \leq d \leq 5$) and decaying as the pairs spread out. We note the residual confound explicitly in Section 5.4: the test separates physical and production locality from pure token-distance locality, and does not separate it from content locality at the same granularity.

4.7 No fixed mechanical rotation

Across the 172 folios with sufficient lines, the permutation-controlled mean periodogram of the *k*/*t* share peaks at frequency index 1, which is a slow trend rather than an oscillation, with $p = 0.31$. The corresponding test on the *o*/*e* core alternation gives $p = 0.07$, also at frequency index 1. The χ^2 tests over line index modulo m reach $p = 2.4 \times 10^{-4}$ at $m = 6$, but this reflects the same slow trend leaking into the modular partition rather than a periodic component, as the permutation test on the spectrum shows. The line profile of the $k/(k{+}t)$ share is 0.555 on the paragraph-initial line and then flat at 0.64 to 0.66: a single step, not a wave.

A rotating instrument advanced by a fixed amount per line is therefore excluded. Two further constraints apply to any

Table 6: Unit-level comparison under identical processing. All corpora are truncated to 34 000 words. u/w is units per word, $r = H(E \mid E_{-1})/H(E)$, and excl. is mean slot exclusivity. Coarse is the pattern of Section 3.7; phon. is the phonotactic segmenter.

System	$\lvert\mathcal{E}\rvert$	u/w	$H(E)$	$H(E\lvert E_{-1})$	r	excl.
Voynich BPE-426	426	1.50	7.86	4.40	0.560	0.724
Italian, coarse	1 339	1.66	7.89	4.03	0.511	0.801
Italian, phon.	1 338	1.67	7.88	4.06	0.515	0.760
Latin, coarse	2 072	2.07	8.50	3.86	0.454	0.877
Latin, phon.	1 067	2.24	7.94	3.96	0.499	0.799

rescue of the volvelle hypothesis. A uniform disc yields $h_2 \approx \log_2 |\mathcal{A}| \approx 4.3$ bits, so the cells would have to be strongly Zipf-weighted. And the measured reduction from $H(E) = 7.86$ to $H(E \mid E_{-1}) = 4.40$ bits requires each cell to restrict its admissible successors, at which point the disc is a codebook and the rotation performs no work. If instead the operator turns the disc at will, the model becomes observationally equivalent to self-citation.

4.8 Sub-word statistics are syllable-scale and Romance-leaning

Table 6 compares the Voynich unit system with Italian and Latin syllable systems under identical processing. Because the result turned out to depend on the syllabification rule, both the coarse pattern of Section 3.7 and a phonotactic segmenter for each language are reported. The phonotactic version keeps `ch`, `gh`, `gn`, `gl` and `sc` together in Italian, keeps `s` + consonant and stop + liquid in the onset, closes a syllable before other clusters, splits geminates, and resolves vowel sequences into diphthongs or hiatus; for Latin it treats `qu` and pre-vocalic `gu` as single onsets, applies *muta cum liquida*, and admits only the six classical diphthongs. It produces *por-ta*, *pa-e-se*, *ac-qua*, *doc-tri-na* and *quin-que* where the coarse rule produces *po-rta*, *pae-se*, *a-cqua*, *do-ctri-na* and *qui-nque*.

The Italian figures are insensitive to the rule: inventory 1 339 against 1 338, unit entropy 7.89 against 7.88. The Latin figures are not. The coarse rule never closes a syllable inside a word, which for a consonant-cluster language inflates the inventory to 2 072 distinct units and the unit entropy to 8.50 bits; under the phonotactic rule Latin has 1 067 units and 7.94 bits. The correct comparison is therefore that all three systems sit within 0.1 bits of each other in unit entropy, and that what separates them is units per word (1.50, 1.67, 2.24) and slot exclusivity (0.724, 0.760, 0.799). On both of those the manuscript is closest to Italian, but the ordering is a gradient and not a match.

Topology tells the same story and reverses one earlier reading of it. Under the coarse rule the spectral distance of (10) from Voynich to Italian is $\Delta = 1.14$ against 1.32 between Italian and Latin, which would put the manuscript closer to Italian than Italian is to Latin. That comparison does not survive correct syllabification. With the phonotactic segmenter the two natural syllabaries move sharply together, to 0.555, while Voynich stands at 1.221 from Italian and 1.744 from Latin. The rewiring null is 0.260 ± 0.062, so none of these distances is a null artefact. The defensible statement is the weaker one: the manuscript's unit graph lies in the direction

of a Romance syllabary rather than a Latin one, by a factor 1.4, and lies outside the two languages by a factor 2.2 on the same scale. Degree distribution divergence under the coarse rule agrees on the direction: Voynich against Italian 0.243, Italian against Latin 0.267.

Two terms used here need an operational definition, because both invite more than they carry. *Syllable-scale* means that the two-part description length of (3) is minimised at units whose mean length, count and positional behaviour fall in the range occupied by the syllable inventories of the two reference languages, and not at single characters or at whole words. It is a statement about compression granularity. It is not a claim that the units were units for a writer, or that they correspond to sounds; byte pair encoding optimises compression and has no access to phonology. *Romance-leaning* means the ordered comparison just given: on all four of inventory scale, units per word, slot exclusivity and spectral distance, the manuscript's value lies between the Italian and Latin values or nearer the Italian one, with one reference vernacular in the panel. It does not mean that the underlying language, if any, is Romance.

Within the word, then, the manuscript's building blocks are syllable-scale and Romance-leaning in those senses, without being drawn from any syllabary we can match. Between words they carry almost nothing: a syllabary transcribing running speech would carry $\Delta I(1)$ near 1.2 to 1.3 bits, and the manuscript carries 0.243. The abugida or vernacular-syllabary reading survives at the level of word formation and fails at the level of the sentence. It remains viable only for text that is not running prose, such as lists, litanies or name inventories, or for a constructed system that imitates syllabic phonotactics.

This result converges with an independent 2026 finding. Applying byte pair encoding as a writing-system classifier over a 21-corpus panel, [18] report that the manuscript's mean vocabulary morpheme length lies above the alphabetic cluster and adjacent to transliterated Chinese Pinyin, that is, adjacent to a syllabic transcription, and that fifty three-character endings account for 80% of the corpus against 172 for Latin. The two studies use different statistics and different panels and point at the same property of the text.

4.9 Section-level tests of the practical-handbook hypothesis

If the manuscript were a herbal, recipe collection or bathing guide, the text should carry the structural signature of instructional prose. It does not.

There is no closed class of instruction words. The five most frequent paragraph openers cover 8.0% of the 287 recipe-section paragraphs, with 241 distinct openers; the botanical section behaves the same way at 7.7%. The openers are distinguished morphologically, by the gallows glyph, not lexically. There is no section-independent structure vocabulary: not one word met a permissive uniformity criterion (min/max frequency ratio above 0.5 at frequency at least 50), where a real recipe collection would show *take*, *mix*, *boil*, *then* at comparable rates throughout. No numeral or measure class occurs anywhere in the corpus. Paragraph-to-paragraph similarity in the recipe section (mean Jaccard 0.039) is indistinguishable from the botanical section (0.042). Pharmaceutical label words, the natural candidates for ingredient names, recur in the recipe text at 33.3% against 28.4% for equally rare botanical words chosen at random.

Table 7: Permutation tests, 200 replicates except graph modularity (20). Nulls: (A) characters shuffled within preserved word lengths; (B) word order shuffled within preserved line and paragraph structure; (C) lines shuffled. The minimum attainable p is 0.005 at $n = 200$.

Statistic	null	observed	null mean ± sd	z
h_2	A	2.129	3.830 ± 0.0001	−15187
H (first char)	A	3.206	3.869 ± 0.004	−164
KL line-initial	B	0.571	0.003 ± 0.001	+587
Gallows, para.-initial	B	0.852	0.078 ± 0.010	+79
Gallows, para.-initial	C	0.852	0.219 ± 0.014	+44
Line-final *-m*	B	0.154	0.026 ± 0.003	+51
Identical neighbours /1000	B	9.04	3.32 ± 0.33	+18
Edit-1 neighbours /1000	B	44.15	19.96 ± 0.80	+30
Modularity, botanical	B	0.2315	0.1792 ± 0.0059	+8.8

The balneological hypothesis makes a sharper prediction, that the bathing section should be the most formulaic part of the book, and the prediction fails in the observable direction. Measuring formulaicity as the ratio of mean line-pair Jaccard similarity within a folio to that between folios of the same section, the ranking is botanical 1.61, recipes 1.46, pharmaceutical 1.43, balneological 1.24. The balneological section is the least page-local and the most section-wide homogeneous of the four. The ordering follows the production process, one plant and one writing session per folio in the botanical section, rather than any content register. Line position within a paragraph does carry a real signal (Jensen–Shannon divergence 0.317 against a null of 0.280 ± 0.002, $z = 17.4$), but the effect size of +0.037 bits is small and directionless, consistent with the known line drift and not with a treatment schema.

A caveat applies to all section-level results. Vocabulary differences between sections run almost entirely along the Currier A/B division: *chedy* occurs 0.1 times per 1000 tokens on A pages and 21.2 on B pages, and *chol* shows the reverse, 24.2 against 4.3. The botanical section is predominantly A, the balneological and recipe sections predominantly B. Section vocabulary therefore measures writing variant at least as much as subject, and no semantic reading of section differences is safe. That this separation cannot be made is itself a finding: the subject signal a genuine technical handbook would show is not separable from a scribal-variant signal.

4.10 Permutation controls

Table 7 collects the principal permutation tests. All reported structure disappears under the appropriate null.

The z column should be read with care and is given only for completeness. At 34 103 tokens with 200 replicates, a null standard deviation of 10^{-4} is easily reached, and a z of −15187 for h_2 means that character order matters, not that it matters fifteen thousand times more than the line-final effect at $z = 51$. Throughout this paper the quantity we interpret is the ratio or the difference in the natural units of the statistic, with z serving only to establish that the effect is not a sampling artefact.

On that reading, two entries carry weight beyond confirma-

Table 8: Model B$^+$ against the manuscript on fourteen target statistics. Both corpora have 34 103 tokens and the same page skeleton. The reported run is seed 42; the next column gives the mean and standard deviation over 100 seeds at the same settings. Judgement applies to the reported run: ✓ within 5% relative, ~ within 25%, × beyond.

Statistic	Voynich	B$^+$	100 seeds	
h_2 (bits/char)	2.13	2.25	2.23 ± 0.05	~
$H(W)$ (bits)	10.31	10.29	10.10 ± 0.20	✓
Hapax % of types	70.4	70.7	70.8 ± 0.8	✓
Edit-1 hapax %	32.3	23.6	24.6 ± 1.4	×
Mean word length	5.10	5.34	5.23 ± 0.12	✓
SD word length	1.93	2.58	2.57 ± 0.12	×
Heaps β	0.72	0.76	0.76 ± 0.03	~
Zipf exponent	−1.04	−1.12	-1.15 ± 0.03	~
Ident. neighbours /1000	9.0	14.3	18.0 ± 2.9	×
Edit-1 neighbours /1000	44.1	43.0	48.0 ± 4.7	✓
$\Delta I(1)$ (bits)	0.235	0.254	0.310 ± 0.035	~
$\Delta I(2)$ (bits)	0.024	0.034	0.096 ± 0.030	×
ΔI tail, $d = 25 \ldots 50$	0.019	0.043	0.079 ± 0.023	×
Folio boundary ratio ρ	1.36	1.10	1.16 ± 0.09	~

tion. Adjacent tokens are identical 2.7 times more often, and near-identical 2.2 times more often, than word order alone allows, so local self-similarity is a real property of the text and not a by-product of the type distribution. And the positional effects at line and paragraph edges are by a wide margin the strongest signals anywhere in the corpus, which is an odd state of affairs for a linguistic text and an expected one for a text whose unit of production is the written line.

Co-occurrence network modularity is low throughout (botanical 0.235, balneological 0.145), below the values natural-language networks of comparable size reach when topical structure is present, and the communities that do emerge group words by morphological shape rather than by anything resembling subject matter. The botanical value nonetheless exceeds its null ($z = 8.8$), so local structure exists; it is morphological.

4.11 A generative model estimated from the manuscript

Table 8 compares the manuscript with the output of model B$^+$ (Section 3.9) over fourteen target statistics, all measured by identical code on both corpora.

Nine of the fourteen statistics are reproduced exactly or within a quarter, including two that no earlier configuration in our sequence of models reached together with the rest: word entropy and the excess information at lag 1. The unaltered round-2 model, which copied from a fixed window of nine recent tokens with no cross-word coupling and no drift, produced no measurable word-level mutual information at any lag ($\Delta I(1) = -0.034$ bits, Figure 3); the coupling term is what supplies it.

The fourth column shows which parts of that count are solid and which are luck. Over 100 reseedings the count itself is 9 in 56 runs, 8 in 25, 10 in 15, 7 in 3 and 11 in 1, mean 8.86 ± 0.74, so the reported figure is the modal outcome rather than a favourable draw. Per statistic the picture is sharper. Six are within a quarter in all 100 runs: h_2, word entropy, the hapax fraction, mean word length, Heaps β and the Zipf exponent. Two more are within a quarter in 93 runs, the edit-1 neighbour rate and the folio ratio. The remaining six are unreliable or absent: the edit-1 hapax fraction succeeds in 61 runs, lag-1 excess information in 29, the word-length standard deviation in 9, lag 2 in 1, and identical neighbours and the long-lag tail in none.

The one entry that the reported run flatters is lag-1 information. Its 0.254 bits sit at the fifth percentile of the 100-seed distribution, whose mean is 0.310 ± 0.035, so a typical run over-produces adjacent information by about a third. That matters for how the model is read: the eight statistics it reproduces dependably are word-shape and vocabulary statistics, and the sequential ones are exactly where it is weakest.

The residuals are more informative than the matches, and we report them rather than tuning them away.

The folio effect is stronger in the manuscript than the model produces. The observed ratio is 1.36 under the model's measurement protocol and the model reaches 1.10, or 1.16 ± 0.09 over seeds. Raising the drift variance increases the ratio, but it also inflates long-range mutual information past the manuscript's value, which is the second residual, and Section 4.14 shows that no setting in the sweep escapes the trade-off. The binding of text to the physical folio in the manuscript is therefore stricter than page-reset plus drift can deliver, which points to a mechanism we have not modelled rather than to a by-product of the ones we have.

The model has too much long-range mutual information. Over lags 26 to 50 the model gives 0.047 bits against 0.016 for the manuscript, four to five standard deviations of the noise floor above it, and the 100-seed mean of 0.079 bits puts the residual higher still. The AR(1) theme drift that lets a folio have its own vocabulary also creates correlation between distant tokens. The manuscript achieves folio-level distinctiveness without that correlation, and how it does so is open.

Variation in the manuscript is more systematic than our mutation operator. The share of hapax types with a frequent neighbour at edit distance one is 32.3% in the manuscript and 23.6% in the model, and the word-length standard deviation is 1.93 against 2.58. Measured directly (Section 4.12), real variation is far more constrained than random slot substitution: 61% of the 1648 hapax-to-frequent pairs differ by an insertion, 35% by a substitution and only 3% by a deletion; 60% of the changes are at a word edge; and substitutions almost never cross character class, staying within core vowels (*o*↔*e*, *a*), gallows (*k*↔*f*, *p*; *t*↔*k*) or final minims (*n*↔*m*, *l*, *r*).

The excess of exact adjacent duplicates (14.3 against 9.0 per 1000) belongs with the same residual: real scribes avoided writing the same token twice in succession more consistently than a single avoidance probability reproduces.

4.12 Structure of the hapax variation

Because the hapax fraction is often taken as evidence of a large lexicon, its internal structure deserves a direct measurement. Of 5 220 hapax types, 1 648, or 31.6%, have a frequent neighbour (frequency at least 10) at Levenshtein distance one. This corrects an over-statement in our own earlier draft, which described the majority of hapaxes as variants; a third is the correct figure. The operation profile is given above. Inserted characters are *o*, *y*, *d*, *e*, *l*, *s*, precisely the free linking glyphs.

Table 9: The principal measurements under four transliterations of the same manuscript, each truncated to 34 000 running-text tokens. $|\mathcal{E}|$ is the MDL-optimal inventory, u/w its units per word, and q_{95} the 95th percentile of the estimator noise floor for the tail band. Excluded tokens are the paragraph-locus tokens carrying an illegibility or rare-glyph marker.

	ZL3b EVA	ZL3b-lig	IT2a EVA	GC v101
$\lvert\mathcal{A}\rvert$	26	32	21	60
Types	7 402	7 402	7 463	7 501
Excluded %	0.6	0.6	0.2	3.2
h_1 (bits/char)	3.870	3.887	3.863	4.108
h_2 (bits/char)	2.129	2.258	2.129	2.495
MDL $\lvert\mathcal{E}\rvert$	426	432	421	460
MDL u/w	1.50	1.49	1.51	1.49
$\Delta I(1)$ (bits)	0.243	0.243	0.210	0.249
$\Delta I(2)$ (bits)	0.029	0.029	0.026	0.030
ΔI, $d = 26 \ldots 50$	0.016	0.016	0.014	0.020
floor q_{95}, same band	0.016	0.016	0.012	0.013
Folio ratio ρ	1.36	1.28	1.40	1.16

This is not the profile of lexical morphology, which does not confine itself almost entirely to word edges and to within-class substitution. It is the profile of slot-bound alternation: the same template with one component exchanged or added, mostly at the margins.

Two details separate the two readings sharply. Inflection in a medieval vernacular or in Latin concentrates variation at one edge, the suffix, and the varying material is a closed set of multi-character morphs. Here the variation is split almost evenly between the two edges, 512 word-initial against 485 word-final changes, and the inserted material is a single glyph in every case, with twelve glyphs accounting for all 1 013 insertions in a flat distribution from *o* (123) down to *a* (29). A paradigm would show a few frequent endings carrying most of the mass. What the data show is one free position that admits any linking glyph, which is a property of the writing procedure rather than of a grammar.

4.13 Dependence on transliteration and glyph grouping

Every character-level statistic in this paper is conditional on a transliteration, and the choice of alphabet is not neutral: EVA writes *ch* and *sh* as two symbols where a palaeographer may see one stroke. Table 9 therefore repeats the principal measurements on four readings of the same manuscript, processed by identical code. ZL3b-lig holds the reading fixed and coarsens the alphabet, mapping *ch*, *sh* and the four benched gallows *ckh*, *cth*, *cph*, *cfh* to single symbols. IT2a holds the alphabet fixed and changes the transcriber. GC-v101 changes both. Because the four files disagree about which loci are legible, each is truncated to its first 34 000 running-text tokens, which is the protocol of Table 8 rather than the matched letter budget of Table 2; the ZL3b row is the reference and gives $h_2 = 2.129$ where Table 2 gives 2.172 on the smaller matched budget.

Three of the four conclusions are unaffected and one acquires a quantified tolerance.

The entropy deficit is alphabet-dependent in magnitude and robust in sign. Regrouping the EVA digraphs raises h_2 from 2.129 to 2.258, and the v101 alphabet, which is both larger and differently cut, raises it to 2.495. That is the largest excursion available to us, and since the reference corpora were measured under the other protocol the comparison is approximate; on any reading of it the manuscript stays about half a bit below the lowest of the seven references and about three quarters of a bit below the centre of the natural-language band. No plausible regrouping of glyphs closes a gap of that size.

The sub-word inventory is stable. The MDL optimum falls at 426, 432, 421 and 460 units, and units per word at 1.49 to 1.51, across alphabets of 21 to 60 symbols. The claim that the text is described most compactly at a syllable-scale granularity is therefore not an artefact of EVA. The specific number 426 should be read as a value near 440 ± 20 under transliteration uncertainty.

The word-order result is the most robust of the four. Excess mutual information at lag 1 ranges from 0.210 to 0.249 bits and at lag 2 from 0.026 to 0.030, a collapse by a factor 8 to 9 in every reading, and the long-lag band sits at or just above its own noise floor in all four. The independent reading by Takahashi gives the lowest lag-1 value, so the conclusion is not being carried by a single transcriber's word-division decisions, which is the obvious way such a measurement could fail.

The folio effect is present in all four readings at $\rho = 1.16$ to 1.40, weakest under v101. Since v101 excludes five times as many tokens as ZL3b, part of that attenuation is expected from the thinning of pairs, and we do not read the spread as evidence about the manuscript.

Table 10: One-at-a-time perturbation of the eight scalars, three seeds per setting. n_{25} is the number of the fourteen target statistics within a quarter of the manuscript value, over all runs for that parameter; the reported configuration gives $n_{25} = 9$. Manuscript values for comparison: $\rho = 1.36$, tail $\Delta I = 0.019$, $\Delta I(1) = 0.235$.

Scalar	Range (ref.)	n_{25}	ρ	tail ΔI
p_{copy}	0.30–0.70 (0.50)	8–10	1.10–1.18	−0.015–0.192
ρ_{AR}	0.70–0.99 (0.90)	8–10	1.06–1.22	−0.003–0.186
σ	0.60–2.00 (1.20)	8–10	1.04–1.26	−0.003–0.179
λ	0.00–0.75 (1.00)	8–10	1.12–1.20	0.068–0.127
p_{exact}	0.05–0.30 (0.13)	8–10	1.11–1.16	0.048–0.084
p_{local}	0.25–0.90 (0.55)	9–10	1.10–1.19	0.058–0.062
w_{local}	3–30 (9)	8–11	1.09–1.18	0.050–0.059
p_{avoid}	0.00–1.00 (0.50)	7–11	1.11–1.18	0.045–0.082

4.14 Stability of the generative model

The eight scalars of Section 3.9 were set by a coarse sweep, which raises two questions the scorecard alone cannot answer: how much of the agreement is Monte Carlo accident, and how sharply it depends on the tuning. The first is answered by the 100-seed column of Table 8. The second is answered by moving each scalar away from its reported value while holding the other seven fixed, three seeds per setting, 96 further runs in total. Table 10 summarises the outcome.

The agreement is a plateau, not a peak. Over 32 settings and 96 runs the count of reproduced statistics stays between 7 and 11, with the reported configuration at 9 and most settings at 8 to 10. Only $p_{\text{avoid}} = 0$ falls to 7, for the mechanical reason that without an avoidance rule exact adjacent duplicates run to 34.4 per 1000 against the manuscript's 9.0. Two set-

Table 11: Excluded hypotheses and the decisive measurement in each case.

Hypothesis	Decisive measurement
Natural language under simple substitution	$h_2 = 2.17$ against 3.02–3.52; mid-range ΔI absent
Latin abbreviation script (one glyph, several letters)	simulated abbreviation raises h_2 to 3.40 and final-character entropy to 4.26
Word-level codebook	MDL optimum at 426 units, not at ∼7 000
Recipe or bathing handbook with fixed schema	no structure vocabulary, no numerals, formulaicity lowest in the balneological section (1.24 against 1.61 botanical)
Volvelle with fixed rotation per line	no periodicity in gallows or core alternation ($p = 0.31$); rescue conditions degenerate to a codebook
Pure time-window self-citation (model B)	produces no word-level ΔI at any lag, and no folio boundary effect
Meaning carried by word order	$\Delta I(d \geq 2)$ at the estimator noise floor

tings reach 11, at $p_{\text{avoid}} = 0.75$, which brings that rate to 7.6, and at $w_{\text{local}} = 5$. We have not moved the reported configuration to them: the gain is inside the seed spread, and the claim being made is sufficiency rather than fit. What the flatness establishes is the narrower point that matters, namely that the scorecard is not an artefact of the particular eight values.

The folio and long-range residuals behave differently from each other, and the sweep sharpens the trade-off stated in Section 4.11 into a three-way one. Drift variance moves the folio effect in the right direction and long-range information in the wrong one: raising σ from 0.60 to 2.00 takes ρ from 1.04 to 1.26, toward the observed 1.36, while tail information goes from −0.003 to 0.179, away from the observed 0.019. Drift autocorrelation offers the reverse: at $\rho_{\text{AR}} = 0.70$ the tail is −0.003 and the folio ratio 1.22, the best pair anywhere in the sweep, but adjacent information collapses to 0.119 bits, half the manuscript's value. Copy probability behaves the same way, with $\Delta I(1)$ falling from 0.615 at $p_{\text{copy}} = 0.30$ to 0.028 at 0.70 while the tail follows it down.

No setting in the sweep reaches the observed $\Delta I(1) \approx 0.235$, the observed tail ≈ 0.019 and a folio ratio above 1.25 at the same time. The manuscript combines strong adjacent coupling, no mid-range or long-range coupling, and a hard folio boundary; page-reset copying with AR(1) drift can supply any two of the three. This is a one-at-a-time sweep and not a joint search, so we cannot exclude a joint optimum that we have not found. What we can say is that the failure is structural in the neighbourhood of the reported configuration rather than a matter of tuning, and that it is the most specific result the model produces.

5 Discussion

5.1 What the measurements exclude

Table 11 lists the hypotheses these data exclude, each with the measurement that does the work.

What remains is compatible with a structured generative process in three layers, each of which our measurements resolve separately: a learned inventory of roughly 426 syllable-scale units in fixed slots; visually anchored reuse of words already written on the same folio, with mutation confined to slot classes; and a slowly varying state at the level of the folio and section. One writer per session, working with pen, parchment and the current folio as the only reference, is sufficient to account for the statistics we measured. Sufficiency is the whole of the claim: other processes may also be sufficient, and one of them is discussed next. The palaeographic finding of up to five hands [6] then implies that the system was learnable and transferable, which is a stronger claim about the system than a single-author account would be. Mechanical aids are unnecessary, and in their strict form excluded.

5.2 What the measurements do not exclude

Three limits on the conclusion must be stated plainly, because the temptation to over-read this kind of result is exactly what has made the Voynich literature difficult.

First, statistics cannot establish absence of meaning. A generative model that matches nine of fourteen statistics shows that no hidden message is required to explain what is on the page. It does not show that none is present.

Second, and more sharply, at least one architecture predicts our central result while carrying a recoverable message. The Naibbe cipher [8] is a verbose homophonic substitution cipher, executable by hand with fifteenth-century materials, that maps single plaintext letters to multiple alternative Voynichese glyph strings obeying a slot grammar. Because each letter has several encodings selected at random, repeated plaintext words do not yield repeated ciphertext words, and sequential dependencies between plaintext units are not transmitted across cipher-word boundaries. A homophonic cipher destroys word-level mutual information by construction. Our measurement of $\Delta I(d \geq 2) \approx 0$ is therefore not evidence against that hypothesis; it is a prediction of it. This is the honest reading, and it is the reason we frame our conclusion as a statement about the sequential channel rather than about meaning.

Two considerations bear on this. The verbose homophonic route requires the plaintext to be short, roughly a third of the token count, and it requires an elaborate multi-table procedure applied consistently over 200 folios by several hands. And independent evaluation in 2026 finds that Naibbe ciphertexts reproduce the manuscript's character-level entropy and length distributions but fail its joint positional profile, including near-zero cross-boundary mutual information at the grapheme level where the manuscript shows a substantial value [1]. Neither consideration is decisive. The ciphertext hypothesis remains open on our data, and any account of the manuscript now has to satisfy the joint constraints rather than any single statistic.

Third, if there is meaning in the manuscript, our measurements locate the only channel with capacity to carry it: the slow variation of vocabulary across folios and sections. That channel is wide enough for classification, in the sense of marking which folio belongs to which subject, and not for sentences.

5.3 Relation to the 2026 literature

Two threads of recent work intersect this one. The convergence on sub-word regularity is described in Section 4.8: an independent BPE-based classifier places the manuscript with syllabic transcriptions rather than with alphabetic writ-

ing [18], which is the same conclusion our minimum description length and unit entropy analyses reach by another route.

The second thread concerns joint criteria for generative models. Reference [1] defines a four-signature profile, calibrated separately for Currier A and B, comprising an end-class to start-class transition rate of 80.6% overall, bilateral positional extremity, cross-boundary grapheme-level mutual information of 0.586 bits for A and 0.498 for B, and Zipfian boundary distributions. Across more than 2000 runs, neither a parametric slot generator nor a Cardan grille satisfies all four simultaneously, and the Naibbe cipher scores one of four. The reported failure mode of the slot generator is instructive for our own work: high cross-boundary mutual information and a high end-to-start rate require opposite degrees of frequency concentration.

Our model does not solve that tension; it sidesteps it by estimating the cross-boundary coupling distribution directly from the manuscript rather than generating it from a frequency mechanism. That is a legitimate modelling choice for our purpose, which is to ask whether a page-anchored copying process can account for the word-order profile once the coupling is granted, but it is not an explanation of where the coupling comes from. Producing the coupling from a mechanism, instead of importing it, is in our view the most useful open problem in this area.

A separate observation about the 2026 literature is worth recording. A substantial number of Voynich preprints appeared during 2026 on self-publication platforms, several announcing certainty levels or decipherment hypotheses that their methods do not support. We have restricted our citations to work that is either peer reviewed or whose methods and code we could inspect.

5.4 Limitations

Transliteration dependence. All character-level results are conditional on a glyph inventory. Section 4.13 bounds the dependence on three further readings rather than arguing about it: h_2 ranges over 2.13 to 2.50 and the MDL inventory over 421 to 460 units, while the mutual information profile and the folio effect survive in every reading. What the test cannot do is bound the dependence on readings that do not yet exist. Four transliterations by four hands are still four hands, and all of them were produced by people who had seen the others.

Reference corpora. Seven corpora are a small panel, and it is a panel of alphabetic writing: six Indo-European languages in alphabets plus two Semitic consonantal scripts. It contains no true syllabary, no abugida, no administrative list and no repetitive liturgical text. That gap falls exactly where our syllable-scale argument would most like support, and the comparison we can make is therefore to the syllable inventories that we derive from alphabetic corpora rather than to a script that encodes syllables directly. A panel including, for instance, a Devanagari or Ethiopic text and a medieval inventory or litany would be a materially stronger test, and we have not run it. The Vulgate is also not fifteenth-century technical Latin, the abbreviation simulation covers 17 frequent conventions rather than a full Cappelli system, and only one Romance vernacular is represented, by one author, so the affinity in Section 4.8 is an affinity to Dante's Italian and not to Romance syllable structure in general. The syllabification result illustrates how much such choices can matter: replacing the regular expression with a phonotactic segmenter halves the Latin inventory and reverses one of the comparisons that earlier drafts of this work reported.

Section power. The astronomical section has 965 running-text tokens, so section-specific tests there have little power.

Boundary confound. As noted in Section 4.6, folio and paragraph boundaries coincide with both production sessions and possible content units. The test discriminates physical locality from token-distance locality, not from content locality, so it establishes that the page is a relevant factor and not that the page is the cause. Separating the two properly calls for a model with random effects for folio and for subject fitted jointly, which our design does not provide.

Segmentation heuristics. BPE, Morfessor and branching entropy are heuristics, and all three optimise a coding criterion rather than recover a writing system. The inventory of 426 units is a model-dependent quantity, the number should not be read as a count of underlying signs, and the units should not be read as signs at all. A unigram language model of the SentencePiece kind would be the natural third segmenter to add, since it optimises a different objective from BPE, and we have not tested it.

Circularity of the model. Model B$^+$ estimates its automaton, its cross-word coupling and its positional classes from the same corpus on which it is then evaluated, so its agreement demonstrates description rather than prediction. The test that would separate the two is available and we have not run it: estimate every component on Currier A alone and generate on the Currier B page skeleton, or the reverse. Given the size of the A/B vocabulary difference reported in Section 4.9, that is a demanding test, and its outcome would say how much of the model is a compression of this text and how much is a mechanism.

Model selection. Model B$^+$ has eight tuned scalars, and the search over them was coarse and guided by the target statistics. Its agreement with nine statistics is therefore a demonstration of sufficiency, not a fit with quantified uncertainty, and a formal model comparison against the alternatives in Section 5 remains to be done. Section 4.14 shows the count is flat in the neighbourhood of the reported values and that the two residuals are structural there, but a one-at-a-time sweep explores a cross through an eight-dimensional space and cannot rule out a joint optimum elsewhere. Several individual values, among them $w_{\text{local}} = 9$ and the three mutation probabilities, remain conventional choices. A likelihood surface or a posterior over the eight scalars would replace the sweep with a proper statement of uncertainty, and would also make the model comparable to alternatives on a common criterion.

Tolerance bands. The 5% and 25% bands of Table 8 are conventional and not derived from sampling theory. A criterion scaled to uncertainty would be stricter in places: h_2 is within 6% of the manuscript value but three standard deviations of the model's own run-to-run spread away from it, and would not count as reproduced. The bands are reported so that the count can be recomputed under any other rule.

6 Conclusion

The Voynich Manuscript places its information almost entirely inside the word and almost none of it in the order of words. Within the word, the text is built from roughly 426 recurring units of about three and a half characters that occupy fixed slots and whose distributional and network statistics are syllable-scale, closer to Italian than to Latin and outside both. Between words, excess mutual information is 0.243 bits at the immediate neighbour and falls to the estimator noise floor from lag two onward, where Latin and Middle English retain an order of magnitude more across the entire syntactic and discursive mid-range. Repetition is anchored to the physical folio rather than to position in the token stream, at matched token lag and against a label-shifting null. A fixed mechanical rotation is excluded. A generative process combining a unit automaton, fitted cross-word coupling, page-bounded copying with slot-confined mutation, and slow folio-level drift reproduces nine of fourteen target statistics on the manuscript's own page skeleton, the modal count over 100 reseedings, and its three systematic failures say that the real binding to the parchment is tighter, the real word variation more constrained, and the real long-range correlation weaker, than that architecture delivers. No setting of its eight parameters satisfies both residuals at once while keeping adjacent information right, which makes the combination the most specific thing we can say about the production process.

The reading these measurements support is a text produced by a learned, shared and physically anchored writing procedure, built at syllable scale in its word formation and without the sequential information structure of language. What they do not support is a claim about meaning. A verbose homophonic cipher predicts the same absence of word-order information while carrying a message, and on our evidence that hypothesis stands. The constructive statement is narrower and, we think, more useful: whatever the manuscript is, its sentences are not in the linear order of its words, and any future model of it will have to satisfy the joint profile of character entropy, sub-word inventory, cross-boundary coupling, word-order decay and page-anchored repetition at the same time.

Data and code availability

The transliterations `ZL3b-n.txt`, `IT2a-n.txt` and `GC2a-n.txt` are distributed by René Zandbergen at `https://voynich.nu/data`. All analysis scripts, intermediate result files in JSON, and the figure-generating code for this paper are archived at `https://lloydstellar.nl/voynich.html`. The pipeline is deterministic: parsing, all measurements and the model B^+ run reported in Table 8 reproduce bit-identically from the stated seeds. Principal scripts are `parse_ivtff.py` (parsing), `phase5.py` (MDL), `phase6.py` (comparative entropy), `phase7.py` (permutation nulls), `f1_units.py` (unit grammar), `f2_hapax.py` (hapax structure), `f3_decouple.py` (layout decoupling), `g2_repetition_boundary.py` (boundary test), `g3_periodicity.py` (periodogram), `g4_abugida.py` (syllabary comparison), `g5_formulaic.py` (section tests), `g6_model_b_plus.py` (model B^+), `paper/mi_curves.py` (mutual information with noise floor), `paper/translit_robustness.py` (Table 9), `paper/sensitivity.py` (Table 10), `paper/seed_variance.py` (the 100-seed column of Table 8) and `paper/syllabify_check.py` (the phonotactic rows of Table 6).

Author contribution and use of AI tools

The research design, hypothesis selection, interpretation and text are the author's. The analysis code was written, and successive drafts were produced, in collaboration with a large language model assistant (Anthropic Claude) operating under the author's direction. Every numerical value reported here was recomputed from the stored result files during preparation of this manuscript, and five claims made in earlier internal drafts were corrected in the process: the fraction of hapax types that are variants of frequent words (a third, not a majority, Section 4.12); the interpretation of long-lag mutual information, which the noise-floor control shows to be indistinguishable from that of Latin (Section 4.5); the direction of the model's mutual information residual (Section 4.11); the comparative spectral result for the syllabary graphs, which reverses under correct syllabification (Section 4.8); and the reading of the model's lag-1 agreement, which reseeding shows to sit at the fifth percentile of its own distribution (Section 4.11). This disclosure is made because the reliability of such a collaboration rests on whether its outputs were checked, and because the relevant venues now ask for it.

Acknowledgements

This work depends entirely on the transliteration effort of René Zandbergen and Gabriel Landini and on the digitisation of MS 408 by the Beinecke Rare Book and Manuscript Library. The self-citation hypothesis that model B^+ extends is due to Timm and Schinner. Criticism from readers of an earlier popular account of this work, in particular on the number of scribal hands, the dating of rotating cipher instruments and the limits of statistical claims about meaning, led directly to corrections retained here.

References

[1] Anonymous. Evidence of layered positional and directional constraints in the Voynich Manuscript: implications for cipher-like structure. arXiv:2604.19762, 2026.

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

[3] R. Clemens, editor. *The Voynich Manuscript*. Yale University Press, New Haven, 2016.

[4] M. Creutz and K. Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):1–34, 2007.

[5] P. Currier. Papers on the Voynich Manuscript. Seminar transcript, 1976. Reprinted at https://voynich.nu/extra/curr_main.html.

[6] L. Fagin Davis. How many glyphs and how many scribes? Digital palaeography and the Voynich Manuscript. *Manuscript Studies*, 5(1):164–180, 2020.

[7] D. E. Gaskell and C. L. Bowern. Gibberish after all? Voynichese is statistically similar to human-produced samples of meaningless text. In *Proc. International Conference on the Voynich Manuscript 2022*, CEUR Workshop Proceedings 3313, paper 4, 2022.

[8] M. A. Greshko. The Naibbe cipher: a substitution cipher that encrypts Latin and Italian as Voynich Manuscript-like ciphertext. *Cryptologia*, 2025. doi:10.1080/01611194.2025.2566408.

[9] L. Lindemann and C. L. Bowern. Character entropy in modern and historical texts: comparison metrics for an undeciphered manuscript. arXiv:2010.14697, 2021.

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

[11] G. Rugg and G. Taylor. Hoaxing statistical features of the Voynich Manuscript. *Cryptologia*, 41(3):247–268, 2017.

[12] A. Schinner. The Voynich Manuscript: evidence of the hoax hypothesis. *Cryptologia*, 31(2):95–107, 2007.

[13] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proc. 54th Annual Meeting of the ACL*, pages 1715–1725, 2016.

[14] T. Timm and A. Schinner. A possible generating algorithm of the Voynich Manuscript. *Cryptologia*, 44(1):1–19, 2020. doi:10.1080/01611194.2019.1596999.

[15] T. Timm and A. Schinner. The Voynich Manuscript: discussion of text creation hypotheses. *Cryptologia*, 48(4):305–322, 2024.

[16] R. Zandbergen. The Voynich Manuscript: transliteration of the text and the IVTFF format. https://voynich.nu, accessed 2 August 2026.

[17] M. Zattera. A new transliteration alphabet brings new evidence of word structure and multiple "languages" in the Voynich Manuscript. In *Proc. International Conference on the Voynich Manuscript 2022*, CEUR Workshop Proceedings 3313, paper 10, 2022.

[18] Sub-lexical regularity in the Voynich Manuscript: BPE morpheme signature discriminates competing hypotheses and reveals a concentrated suffix system. Preprint, Zenodo, 2026. doi:10.5281/zenodo.20546419 (v3.1, 13 June 2026).