How the Voynich Manuscript was written

Matthew D. Schwartz*
Department of Physics, Harvard University

Abstract

The Voynich manuscript is an illustrated book of the early fifteenth century in a script no one has been able to read. Since Rugg, and then Timm and Schinner, showed that text with many of its statistics can be produced mechanically, quantitative study has come to favor a writing procedure over a cipher or an unknown language. Here we make that picture precise with quantitative tools: information theory, efficient methods for mixture-model computation using exact arithmetic, and solvers validated on planted test cases. Readings and text generators become probability models, fitted on half a scribe's pages and judged on the rest. In six candidate languages, readings by simple or verbose substitution, consonantal writing, abbreviation, codebooks and the Naibbe cipher all fail on withheld pages. Such failures exclude a reading only where the test detects disguised real text of the same kind, so far only for scripture. Knowing a word barely helps predict the next, less than in any comparison text, glossaries included, yet the edge glyphs of adjacent words are coupled as strongly as in a recipe collection. The scribes reused their words, occasionally copied a recent one, and coined others by letter habit; this keyless procedure reproduces nearly all summary statistics of both principal hands. No proposed generator, ours included, reproduces that coupling. Two weak traces of content remain: distinctive label words and a faint dependence of word choice on subject. The low cost of the writing and the genres the pictures imitate suggest a book made to be looked at, not read.

1 Introduction

The Voynich manuscript (Beinecke MS 408, Yale University) is a small vellum codex, written in a looping script known from no other document, that no one has been able to read [D'Imperio, 1978, Clemens, 2016]. Its pages show largely unidentified plants, star charts, bathing women, and rows of pharmacy jars, with isolated label words beside many of the drawings (Figure 1). About 240 pages survive, fluently written, and the vellum has been radiocarbon-dated to the early fifteenth century [Hodgins, 2009]. William Friedman and John Tiltman, among the best cryptanalysts of the twentieth century, worked on it [Friedman, 1959, Tiltman, 1967]. A reading is announced roughly once a decade, from Roger Bacon's scientific shorthand [Newbold, 1928] to a lost proto-Romance language [Cheshire, 2019], and none has survived scrutiny [Manly, 1931, Bowern and Lindemann, 2021].

Much is nonetheless known about the text's statistics. The pages fall into two statistical "languages" that follow the handwriting [Currier, 1976], in which Fagin Davis [2020] distinguishes five scribes, though Timm [2026b] argues for one. Characters are more predictable than in any European language tried [Bennett, Jr., 1976], and words have a rigid structure of prefix,

*Work done with Anthropic, but results and views are not endorsed by Anthropic.

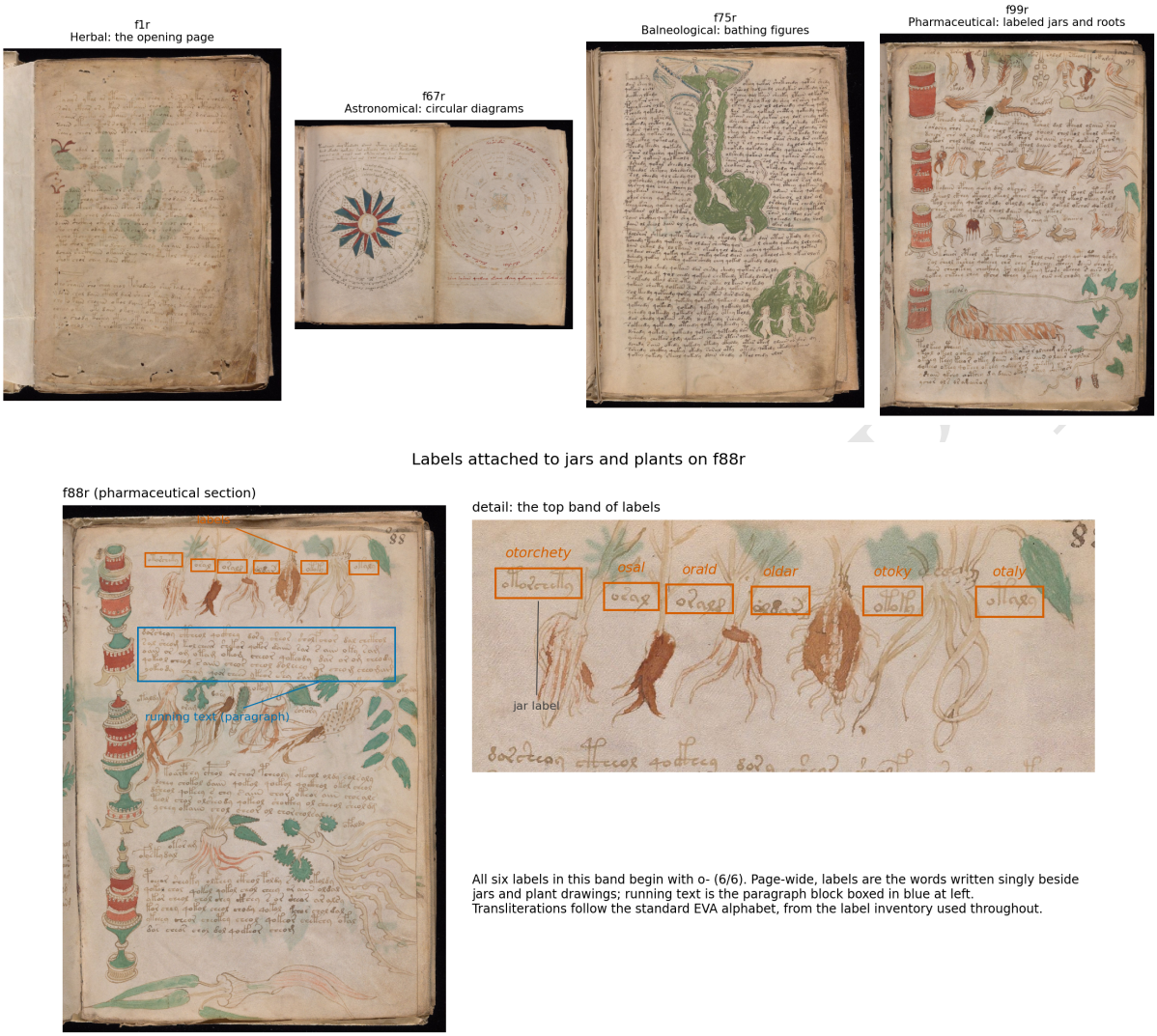


Figure 1: Four pages of Beinecke MS 408 and the pharmacy page f88r. (a) Top, one page from each principal illustrated section: the opening page of the herbal section; a circular astronomical diagram; bathing figures in green pools and streams; and pharmacy jars with labeled roots. All four pages are written in the same looping script, which has never been read. (b) Bottom, the pharmacy page f88r. Highlighted: the six isolated labels of the top band, beside the jars and plant parts, all beginning with the glyph transliterated o-. The page carries fifteen labels, thirteen of them o-initial, counted again in the label tests below; the boxed running text is shown for contrast. Images: Beinecke MS 408, Yale University (public domain).

core, and suffix. Each word constrains the next only weakly [Reddy and Knight, 2011], by under one percent of token entropy, less than in any prose control [Rozanova and Temerev, 2026]. Yet most word-level statistics are compatible with natural language [Montemurro and Zanette, 2013, Amancio et al., 2013, Bowern and Lindemann, 2021]. Text with many of these statistics can be produced mechanically, by a grille over a table of word fragments or by copying and modifying earlier words [Rugg, 2004, Timm and Schinner, 2020, 2024], a process that Timm [2026a] now describes as one writer's drifting habit. Greshko [2025] argues that a cipher remains viable and

Bowern and Lindemann [2021] that a natural language is most likely, but much of the recent quantitative work points to text produced by such a procedure and carrying no message [Gaskell and Bowern, 2022, Timm and Schinner, 2024, Timm, 2026a].

Announced readings have come from wide searches over keys, languages, and rearrangements, which find language-like output in almost any text [Manly, 1931]. Friedman, Tiltman, Rugg, and Timm asked instead what process produced the text; the fullest answer is the self-citation generator of Timm and Schinner [2020]. The self-citation account was argued from the statistics it reproduces; it was not, so far as we know, tested against text withheld from fitting or against matched controls. Readings have been judged by how language-like their output is, and generators by the statistics they reproduce [Bowern and Gaskell, 2022, Parisel, 2026b, Ou, 2026]; no such study has fitted a hypothesis on some pages and scored it on the rest, or weighed readings and generators on one scale.

Here we make the picture emerging from that work precise by putting readings and generators to the same test: each is fitted on half of a scribe's pages and judged on the rest. Every mechanical reading system, paired with one of six candidate languages, becomes a probability model for the glyph sequence, as does every proposed generator. A model's Bayesian evidence, the probability the model gives the text once its parameters are integrated out, is computed exactly for models of word and glyph counts, by closed forms and recursions that Schwartz and Sen [2026] developed for mixture models and Schwartz [2026] applied to authorship. Each solver must first recover a known text disguised by the mechanism under test; a family of readings is excluded only as far as that test's measured power allows.

Our first result is negative: wherever the test detects a disguised text of the same kind and rejects false readings, none of these reading systems survives on withheld pages, in any language, reading order, or parsing we searched. The test has that power only for some ciphers of scripture-like prose; for other registers its criterion, calibrated on scripture, rejects even correct readings. We do not attempt to test free anagramming, since at this text length it can fit almost any target. That choice leaves untested the reading of the text as Bacon's shorthand [Newbold, 1928, Manly, 1931], which rests on free anagramming; the fixed-rule anagramming of Hauer and Kondrak [2016] is tested, inconclusively for Hebrew and Arabic. The tests also suggest why readings keep being announced: searched freely, the text yields several bits per word of spurious signal.

Our second result concerns structure: knowing a word barely helps predict the next, whatever the model class, and by less than in any comparison text, recipes, glossaries and a botanical catalog included. Yet the edge glyphs of adjacent words are coupled as strongly as in some recipe collections and that catalog. A keyless procedure of word reuse, a little local copying, coinage by letter habit, and line-position and word-length habits reproduces nearly all the summary statistics of both principal hands within its sampling noise, but not that edge coupling, which no proposed generator yields. Local copying of the kind Timm and Schinner [2020] propose is a minor component, and copying from further back cannot be told from reuse. We also find two weak traces of attention to the pictures. Why the book was made can be argued only from the object, as in Section 5.

2 The manuscript and its readers

Beinecke MS 408 is a codex of about 240 vellum pages. Some 102 leaves survive, several as multi-panel foldouts, and at least 14 leaves are lost. Radiocarbon dating places the vellum between 1404 and 1438 with 95% confidence, and the iron-gall ink and mineral pigments fit that period [Hodgins, 2009, McCrone Associates, Inc., 2009]. The vellum shows no trace of erased earlier text, and the script and drawings flow around each other, evidently made as one project.

2.1 The codex and its history

The book is conventionally divided by its illustrations (Figure 1). The herbal section of full-page plant drawings is the largest, 129 of the 227 text pages. The others are astronomical and zodiac diagrams (31 pages), the balneological section of bathing figures (19), labeled jars and plant parts (16), star-marked paragraphs usually read as recipes (25), and seven text-only pages. The script, Voynichese, runs left to right in words, lines, and paragraphs and occurs in no other surviving document. Its basic alphabet has roughly two dozen glyph shapes, though inventories from about two dozen to over forty have been defended [Lindemann and Bowern, 2020]; our tests presuppose no particular division into characters.

The fluent writing, with very few corrections, indicates that the scribes were doing something familiar. Currier showed in the 1970s that the pages divide into two statistical "languages" or dialects, Currier A and B, differing in word and glyph frequencies [Currier, 1976], and Lisa Fagin Davis distinguishes five scribal hands [Fagin Davis, 2020, 2022]. Character-based classifiers largely reproduce her attribution on withheld pages for the three hands tested [Farrugia et al., 2022], whereas Timm [2026b] argues for one scribe whose hand drifted.

An erased inscription on the first folio names Jacobus Horcicky de Tepenec of Rudolf II's Prague court, ennobled as "de Tepenec" only in 1608 [Clemens, 2016, Zandbergen]. Georg Baresch of Prague described the unreadable book to Athanasius Kircher in 1639 [Baresch]. Jan Marek Marci sent it to Kircher in 1665 or 1666 [Marci], reporting at second hand that Rudolf had paid 600 ducats for it and that Roger Bacon was thought its author. Rudolf's inventories have no entry for it [Zandbergen]. The rest of this paper therefore treats the imperial purchase as an unverified report. Writing as old as the vellum, as the evidence indicates [McCrone Associates, Inc., 2009], excludes both Bacon (thirteenth century) as author and a forgery by John Dee or Edward Kelley (sixteenth century). Wilfrid Voynich bought the book from the Jesuits in 1912; Yale's Beinecke Library has held it since 1969 [Beinecke Rare Book & Manuscript Library]. Several quires are misbound in a later binding [Clemens, 2016], which our page-order tests allow for.

2.2 Readers, statistics and hypotheses

In 1921 William Romaine Newbold announced that the manuscript was Bacon's notebook, written in a microscopic shorthand hidden in the ink strokes [Newbold, 1928]. John Manly dismantled the reading in 1931: the shorthand was cracks in the drying ink, and the readings came from anagramming so free that Manly produced his own counter-readings [Manly, 1931]. William F. Friedman led study groups on the manuscript from the 1940s. His conclusion appeared in a 1959 journal footnote as an anagram, "I put no trust in anagrammatic acrostic cyphers, for they are of little real value — a waste — and may prove nothing — finis". Its

solution, revealed after his death, was that the text is an early constructed language and not a cipher [Friedman, 1959]. John Tiltman's 1967 assessment cataloged the rigid structure of Voynichese words and endorsed no solution [Tiltman, 1967], and D'Imperio's 1978 monograph surveys that era's work [D'Imperio, 1978].

William Bennett found that the text's conditional character entropy is well below that of European languages [Bennett, Jr., 1976], a result since confirmed across hundreds of languages and historical texts [Lindemann and Bowern, 2020]. Besides finding the two dialects, Currier saw that they change with the hand and that the line behaves as "a functional entity" [Currier, 1976]. Stolfi quantified Tiltman's word structure as a prefix–midfix–suffix decomposition [Stolfi]; Montemurro and Zanette found language-like long-range structure in the text's informative words [Montemurro and Zanette, 2013], and Amancio et al. [2013] found its word statistics mostly compatible with natural language. Each word depends only weakly on the one before it, as Reddy and Knight showed by cross-validated prediction of the next word [Reddy and Knight, 2011] and Caruana et al. [2022] by comparing word-pair order with shuffled text. Rozanova and Temerev [2026] measured this dependence as 0.79% of token entropy, against 2 to 10% in prose controls and 15.81% in a botanical catalog, and found that the edge glyphs of adjacent words share 0.2 bits, more than in any prose control. Bowern and Lindemann's review judged Voynichese most likely a natural language under some non-one-to-one encoding, arguing that gibberish would not explain its higher-level structure [Bowern and Lindemann, 2021]. However, Gaskell and Bowern showed that forty-two volunteers writing gibberish reproduce the low-level statistics [Gaskell and Bowern, 2022], leaving untested the higher-level structure, which we measure directly. The same authors showed that simple re-encodings of Latin, such as writing one sound with two letters, can lower character entropy to Voynich levels [Bowern and Gaskell, 2022]. Quantitative work, ours included, relies on transliterations maintained by volunteer scholars [Zandbergen, 2022].

In 2004 Rugg proposed a mechanism for the hoax hypothesis: sliding a sixteenth-century Cardan grille over a table of word fragments mass-produces text with many Voynichese properties [Rugg, 2004, Rugg and Taylor, 2017]. Zandbergen found the word-length distribution consistent with a grille-like method but called his analysis inconclusive [Zandbergen, 2021], and Schinner argued statistically for mechanical generation [Schinner, 2007]. Timm's 2014 study, *How the Voynich Manuscript was created*, whose title ours echoes, reported that "it was possible to reveal the text generation method," a copy-and-modify process inferred from the network of similarly spelled words [Timm, 2014, 2016]. Timm and Schinner formalized this process as the "self-citation" generator, which reproduces several of the text's statistics [Timm and Schinner, 2020]. Answering later critiques, they noted that the algorithm "does not reproduce some of the more subtle statistical observations" and concluded that "most likely the minimal representation of the VMS [the manuscript] is the manuscript itself" [Timm and Schinner, 2024]. Timm [2026a] restates it as the work of one writer whose habits drifted continuously, and attributes the apparent link between vocabulary and illustration to the order of writing.

Hauer and Kondrak's language-identification models ranked Hebrew first among source languages, with substitution followed by within-word anagramming as the mechanism, a hedged 2016 result that 2018 press coverage inflated into a decipherment [Hauer and Kondrak, 2016]. In 2019 a university announced a proto-Romance solution and withdrew the announcement within days [Cheshire, 2019]. Greshko's Naibbe cipher of 2025, a hand-executable verbose (one-to-many) homophonic substitution, matches the manuscript on many statistics at once and is offered as evidence that the ciphertext hypothesis remains viable [Greshko, 2025]. Rozanova and

Temerev [2026] find that Naibbe ciphertext and a reimplementation of the self-citation generator both reproduce the low character entropy and weak word order but not the edge-glyph coupling or the many words that occur once. Three further studies score generators on whole-text statistics [Parisel, 2026b, Ou, 2026, Izumo, 2026]. Neither a slot grammar nor a Cardan grille shows all of Parisel's positional signatures, none of eighteen generators meets all of Ou's criteria, and a mechanical glyph-pair device matches the text no better than a bigram null. None of this work scores a proposed generator on pages withheld from its fitting, or scores a reading and a generator by one measure. Below we do both for every proposal above that can be made mechanical, free anagramming excepted.

3 Data, models and text structure

Every statistic below depends on how glyphs divide into characters, which pages belong to which hand and dialect, and which texts serve for comparison. We fix these choices here, together with the method of model comparison.

3.1 Transliteration and comparison corpora

All counts are taken over the Zandbergen–Landini (ZL) transliteration [Zandbergen] in the EVA alphabet, whose letters name glyph shapes without any claim about sound; one glyph can become two or three letters. The corpus is the 227 pages that bear text, roughly 38,000 word tokens, Latin-script marginalia excluded. The counts use two parsings, a *folded* one in which the frequent glyph groups *iin*, *ii*, *ee*, *ch* and *sh* are single units and an *atomic* one in which every EVA character is separate. The folded parsing is primary; no test below addresses error in the transcription itself. Two prior classifications are taken as given: Currier's two statistical "languages" [Currier, 1976], here dialects A and B, and the five scribal hands distinguished by Fagin Davis [2020, 2022]. The hand attribution is disputed [Timm, 2026b]; our per-hand results are conditional on it and do not test it.

The comparison texts form two sets, each cut to the manuscript's page and line layout. The reading tests of Section 4 use scripture in six candidate languages (Vulgate Latin, Hebrew with the vowel points removed, a modern Arabic Bible, and Italian, German and English Bibles). The production comparisons of Section 5 use a medieval Latin verse herbal (Macer's *De viribus herbarum*, machine-read from a printed edition, with some transcription noise), Leningrad-codex Hebrew, the medical Arabic of Ibn Sina's *Qanun*, and Dante's *Inferno*. For word order we add 23 further texts, including technical Latin prose, recipes, glossaries and a botanical catalog. The reading tests are thus calibrated on scripture alone and control neither for genre within a language nor for list structure. Such texts serve only the word-order comparison and, enciphered, the power measurements of Section 4.

3.2 Model comparison and validation

To compare two accounts of how a text was produced, we write each as a probability model M with parameters θ (letter frequencies, table entries) and compute its *evidence*, the probability it assigns to the text with θ integrated out:

$$Z_M \;=\; P(\text{text} \mid M) \;=\; \int P(\text{text} \mid \theta, M)\, P(\theta \mid M)\, d\theta. \tag{1}$$

The evidence penalizes flexibility automatically, since a model that could produce almost any text assigns little probability to the one observed. For multinomial and Markov models of counts with Dirichlet priors the integral has a closed form; mixtures, over a blending weight (Eq. (3) below) or over the ways a scribe may reuse, copy or coin a word, reduce to finite sums and forward recursions. We evaluate both exactly, in rational arithmetic where floating point would lose the answer [Schwartz and Sen, 2026, Schwartz, 2026]. The same calculation scores a proposed reading and a proposed generator of meaningless text on one scale. We keep three kinds of number apart: the evidence itself, for the page accounts below and the whole-corpus comparisons of Section 5; held-out predictive scores under a language model, in closed form but with nothing integrated, for the reading tests; and held-out cross-entropies of fitted generators, for the writing procedures. Where a ranking by evidence depends on the prior we give both rankings.

Evidence ratios are expressed per token: on a text of N tokens the advantage of M_1 over M_2 is

$$b(M_1, M_2) \;=\; \frac{1}{N}\,\log_2 \frac{Z_{M_1}}{Z_{M_2}} \tag{2}$$

bits per word, and a preference is called *decisive* when the Bayes factor Z_{M_1}/Z_{M_2} exceeds 100. A tally of die rolls favors a die loaded toward sixes over a fair one only modestly at 60 rolls and 18 sixes and overwhelmingly at 600 rolls and 180 sixes, because Eq. (1) averages over every degree of loading instead of taking the most favorable one.

A reading hypothesis, that the text is language L disguised by mechanism T, passes when its decode of the pages withheld from fitting behaves like L at two levels. The decode's character statistics must match those of L, and a word-level statistic of the decode under the word model of L (for codebooks, a word-order margin) must reach a threshold calibrated on the comparison corpus of L.

Throughout, models are fitted on one half of the relevant pages and evaluated on the other (the test half), with the mechanism families, languages, statistics and thresholds fixed before any manuscript page was evaluated. Each solver and generative model first had to recover a known answer from real Latin, Hebrew or Arabic disguised by the mechanism under test, or from text written by a known procedure. None was allowed to classify the undisguised comparison languages as procedurally generated, and a family of readings is excluded only at the strength this recovery demonstrates (Table 1). Three of these synthetic checks fell short of a criterion set for them: the identification of fused glyphs, the naming of the mechanism family on out-of-genre text, and a rank margin in the full-width verbose test. One word-level threshold was widened after its own Hebrew control missed it, before any manuscript scoring. We validated three components only after the fact: the variant-family term of the vocabulary model, the extension of the affix comparison to the minor hands, and the sensitivity check of the label-reuse test.

3.3 Hands and dialects

Reduced to its word counts, each page has an evidence under three accounts, its words drawn from dialect A's frequency profile θ_A, from B's θ_B, or from a blend with prior $p(w)$ on the weight w:

$$\begin{aligned} Z_A &= P(\text{page} \mid \theta_A), \qquad Z_B = P(\text{page} \mid \theta_B), \\ Z_{\text{mix}} &= \int_0^1 P\big(\text{page} \mid w\,\theta_A + (1-w)\,\theta_B\big)\, p(w)\, dw. \end{aligned} \tag{3}$$

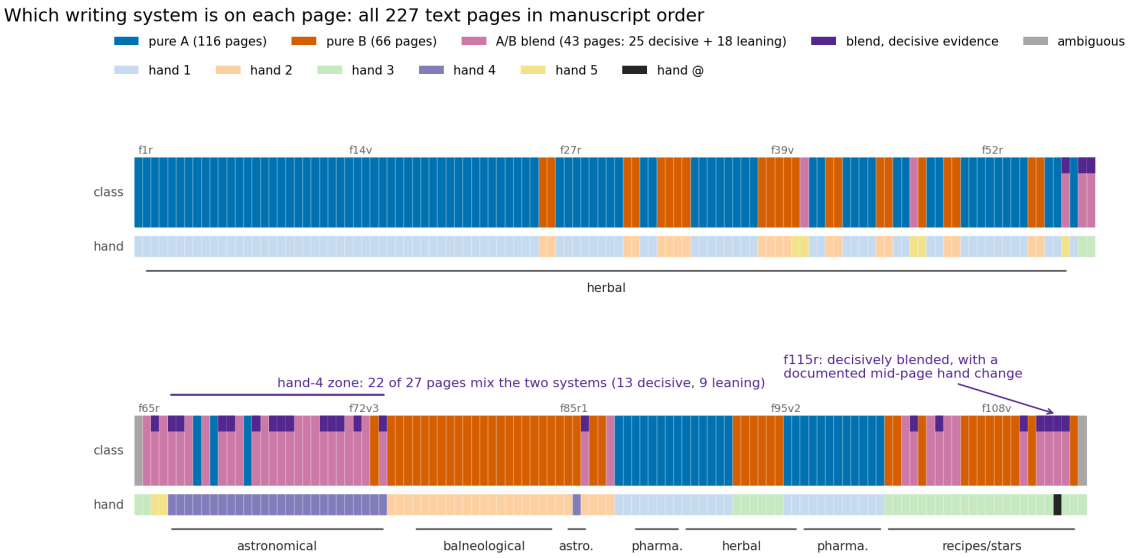


Figure 2: Per-page composition map. Only Scribe 4's pages, all in the astronomical section, classify page by page as A, B or blend (see text). The 227 text pages in manuscript order, in two rows. Upper band of each row: the dialect of each page by exact evidence, Eq. (3) (pure A, pure B, or blend, with decisively blended pages marked). Lower band: scribal hand, numbered after Fagin Davis [2020]. Brackets mark the illustrated sections. The blend class of the legend marks any preference for the mixture, whereas the counts given in the text are decisive preferences (Bayes factor exceeding 100).

A page is called decisively A or B when Z_A/Z_B or its inverse exceeds 100, decisively blended when Z_{mix} exceeds the larger of Z_A and Z_B by that factor, and *pure* when Z_{mix} does not exceed it at all. Every page is evaluated against profiles fitted on reference pages (Scribe 1's Currier-A and Scribe 2's Currier-B pages) other than itself; folio 26v, a herbal page of about a hundred words, favors B over A by a factor exceeding 10^{35} and is pure B. Pages have been classified probabilistically before [D'Imperio, 1979, Reddy and Knight, 2011, Steckley and Steckley, 2024, Parisel, 2026a]; here each also receives a posterior over the blending weight.

In the composition map (Figure 2), Scribe 1 writes dialect A on all 113 of that hand's pages. Scribes 2, 3 and 5 write dialect B (46 of 46 pages, 31 of 33 with two ambiguous, and 7 of 7). Only Scribe 4's pages divide: of that hand's 27 pages, 13 classify as A and 12 as B, and 13 show decisive evidence for the mixture over both pure profiles (these counts overlap); folio 115r, where the hand changes partway down, is counted apart.

In control experiments on Latin texts of known composition (Appendix B) the page classifier separated two kinds of medieval Latin without error, at evidence strengths comparable to the manuscript's, and left single corpora undivided. A and B are thus two generating profiles, dialects in that sense only. The classifier also preferred the mixture on the pages of an unrelated third text about as often as on the manuscript's astronomical pages (77% in both cases). A preference for the mixture cannot by itself distinguish pages that blend the two dialects from pages written in some third profile. Within the herbal section alone, where the subject is constant, dialect still varies with the hand, so the dialects appear to belong to the writers.

Fagin Davis [2020] reads Scribes 2 to 5 as all writing dialect B; only Scribe 4's pages disagree, and all 27 of them are label-heavy astronomical, cosmological and zodiac pages. A third profile,

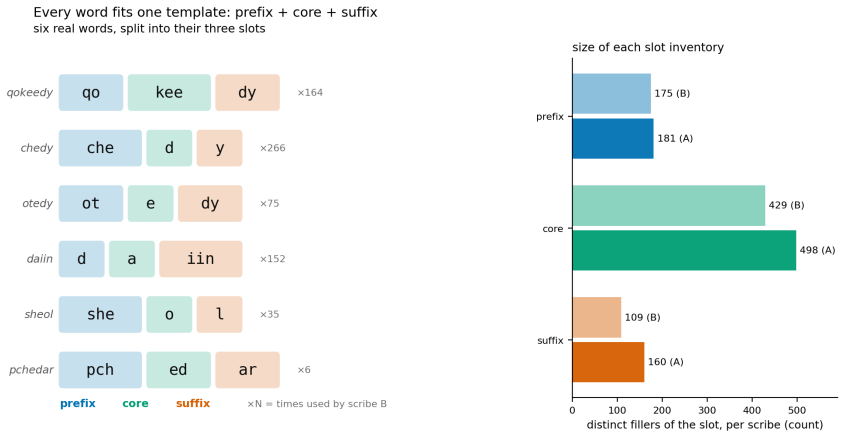
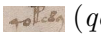


Figure 3: Anatomy of Voynichese words. Every word fills at most three slots, prefix, core and suffix, each from a small inventory. Left: six frequent words split into their slots, with the number of times Scribe 2 uses each ($\times N$). Right: the number of distinct fillers of each slot for the two principal hands (Scribe 1: 181, 498, 160; Scribe 2: 175, 429, 109; counts include the empty slot); A and B in the panel denote Scribe 1, who writes dialect A, and Scribe 2, who writes dialect B.

fitted for each of these pages on the other 26, has higher evidence than A, B and their blend on 19 of the 27, and a three-way blend of A, B and that profile is preferred to the two-way blend on all 27. Scribe 4's five pages that classify as pure fit their dialect worse than 89 to 100% of its reference pages. Scribe 3's star pages and Scribe 5's herbal pages, all nearer B than A, fit B about as poorly on average, whereas the herbal, pharmaceutical and balneological reference pages fit normally. Dialects A and B as fitted are thus the profiles of Scribe 1's herbal and pharmaceutical pages and of Scribe 2's herbal and balneological pages. Other sections fit both poorly and, where they hold enough text, prefer further profiles, although two of Scribe 4's pages still prefer a reference dialect to their own section's profile. All hands and sections nonetheless share one word grammar and, where a hand's pages suffice to show it, largely the same affixes.

3.4 Word and line structure

Voynichese words have a rigid anatomy of optional prefix, core and optional suffix, each from a small inventory [Tiltman, 1967, Stolfi, Zattera, 2022]. We split each hand's word types by one fixed rule: a word's first two and last two folded units (fewer in short words) are its prefix and suffix, the rest its core (Figure 3). The common words  (*qokedy*),  (*chedy*),  (*otedy*) are built this way, *chedy* dividing as *che-d-y* and *otedy* as *ot-e-dy*. Thousands of words are built from a few hundred such parts, each confined to one position with very rare violations, and word lengths cluster around five or six glyphs, more tightly than in any comparison language.

Within words the glyphs *a, e, o, y* alternate with the rest. Sukhotin's vowel-finding algorithm [Sukhotin, 1962, Guy, 1991] separates this vowel-like class on the two principal hands at 38 to 63 standard deviations depending on hand and alphabet (EVA or the alternative v101), close to the 68 of our Latin control and far above the 18 of vowelless Hebrew. Yet the class found on one half of the manuscript differs from that on the other, whereas the Latin control's halves agree. At the character level, the low conditional glyph entropy [Bennett, Jr., 1976, Lindemann and Bowern, 2020, Bowern and Lindemann, 2021] excludes any encoding that cannot produce glyph sequences this predictable. Line-initial and line-final words also differ from mid-line words

[Currier, 1976]. Across a word break the last unit of one word predicts the first unit of the next, by about 0.12 and 0.21 bits per word on the two principal hands' withheld pages against at most 0.11 in the four comparison languages; Rozanova and Temerev [2026] measure this coupling at 0.197 bits. Any account of how the text was produced must reproduce these features, in five hands sharing one word grammar.

4 Tests of proposed reading systems

Each reading hypothesis pairs a candidate language L with a disguising mechanism T and predicts that undoing T on pages not used in fitting yields text with the character and word statistics of L. All were fitted on half of a hand's pages under the most favorable settings we could find and tested on the rest.

4.1 The search space

A century of proposed readings falls into five mechanism families: one-to-one substitution, one-to-many or "verbose" substitution (a group of glyphs standing for one letter), consonantal writing with the vowels dropped, expansion of medieval Latin abbreviations, and codebook or numbered-dictionary schemes. We add the Naibbe cipher [Greshko, 2025] and composites of these families.

The candidate languages were Latin, Hebrew, Arabic, Italian, German and English; reading order and character parsing were searched jointly with the mapping. Free anagramming, the mechanism of Newbold's reading [Newbold, 1928], is not tested. At this text length it admits almost any answer, as Manly showed [Manly, 1931], so it makes no prediction that scoring on unfitted text could contradict, and nothing below bears on it. Anagramming by a fixed rule within each word, the final stage of Hauer and Kondrak's method [Hauer and Kondrak, 2016], is by contrast testable, and we test it here.

None of the 130 hypotheses of Table 1, each scored by held-out prediction under the language model, meets the word-level criterion of Section 3.2, nor, save an unselective anagram acceptance treated below, does any mapping from its lower rows. A failure on the manuscript counts against a family only insofar as the criterion accepts a true reading of that family. We measured this acceptance rate on 22 texts in the six languages (scripture; herbal, medical, technical and narrative prose; recipes), disguised by up to sixteen encodings at four levels of transcription noise, five trials each. On scripture at up to 2% noise the criterion typically accepts, in three to five trials of five, a correct substitution reading of Latin, Italian or Arabic and a fixed-rule anagram reading in four or five of the languages. The same holds for abbreviated Latin and for homophonic readings with up to fifty glyphs for Arabic and eighty for Hebrew. Outside scripture it accepts a correct reading in about one trial in a hundred at zero noise and none at 2% or more. Self-citation text [Timm and Schinner, 2020], a matched control containing no reading, passed in none of 6,520 trials. The exclusions below therefore hold for scripture-like plaintext at low noise and carry no weight for herbal, medical, recipe or narrative text in any language.

4.2 Substitution and consonantal readings

Simple substitution, the oldest family of proposals [D'Imperio, 1978], includes Cheshire's "proto-Romance" [Cheshire, 2019] and the substitution stage of Hauer and Kondrak's Hebrew hypoth-

Family (languages)	Hyp.	Check on enciphered real text	Manuscript: nearest approach
One-to-one substitution (five languages) [D'Imperio, 1978, Hauer and Kondrak, 2016, Cheshire, 2019]	20	key recovered on out-of-genre Latin, 18 of 23 units; text read. Power: scripture (Latin, Italian, Arabic) 5 of 5, 4 to 5 at 2% noise; Hebrew 0 of 5, 5 of 5 at 2% noise; other registers 0 of 5 (one narrative text 3 of 5)	Hebrew, 2.8 bits/word short
Verbose substitution, up to four glyph groups (five languages)	20	20 of 22 units recovered; Latin read. Power (two-glyph groups): scripture 0 to 4 of 5 by language; other registers 0 of 5 (one narrative text 2 of 5)	Hebrew, 1.1 bits/word short
Verbose substitution, full table width, 250 units (Latin)	4	Latin table recovered, map accuracy 0.81; Hebrew-direction table not recovered (0.62 against the 0.70 required), so that direction untested	best decode fits Latin worse than a self-citation control
Consonantal, vowels dropped [Hauer and Kondrak, 2016, Reddy and Knight, 2011]	24	recovered	Hebrew, 2.1 bits/word short
Latin abbreviation [D'Imperio, 1978]	16	recovered. Power: scripture 4 of 5; other registers 0 of 5 (one narrative text 3 of 5)	Latin, 5.7 bits/word short
Codebook or numbered dictionary, three table sizes (five languages) [Friedman, 1959, D'Imperio, 1978]	30	table structure identified and head of table recovered; individual word types read at 0.17–0.32 accuracy	best order margin 0.28 bits/word (Latin, 240-word table, Scribe 1)
Naibbe cipher, published mapping (Latin, Italian) [Greshko, 2025]	4	published tables: 0.97 of words recovered. Freshly randomized tables of the same design, six languages: no demonstrable recovery by the wide-table solver in 1,370 trials (no power)	Italian, 1.5 bits/word short (published mapping only)
Composite: consonantal skeleton with verbose groups (Arabic, Hebrew, Latin)	12	Arabic encoding recovered, true language ranked first	all twelve fail
Composite: consonantal skeleton with positional letter variants	0	glyph entropy stays near 3.7 bits against the manuscript's 2.0–2.2 for any decodable table	excluded before scoring
Reading orders, ten reorderings (Hebrew, Arabic)	40	optimizer against matched controls	identity order fits best
Character parsing solved jointly with the mapping (Hebrew, Arabic)	12	mis-parsed Hebrew read at word accuracy 0.87; 2 of 4 fused glyph pairs identified	no movement toward any language
Anagramming by a fixed rule within the word [Hauer and Kondrak, 2016]		criterion accepts disguised scripture in five languages, 4 to 5 of 5; German 2 of 5	the four European languages 2.2 to 5.0 bits/word short; uninformative for Hebrew and Arabic (see text)
Free anagramming [Newbold, 1928]		not testable at this text length [Manly, 1931]	not scored

Table 1: Reading systems tested. Apart from a non-specific Arabic anagram (see text), none of the systems run on the manuscript yields a reading of it that meets the word-level criterion. Rows above the rule are scored by held-out predictive statistics under the language model on the test half of each principal hand's pages, one hypothesis (Hyp.) per combination of mechanism, language, parsing and hand, 130 in all; rows below it are optimizer searches against matched control texts. The third column gives the result of running the same solver on real text enciphered by the row's mechanism. Its "power" entries give the number of trials, out of five per language at zero transcription noise unless stated otherwise, in which the acceptance criterion accepts the true reading of a test text of the stated register enciphered by that mechanism; families without an entry have not had this rate measured. "Bits/word short" is the nearest hypothesis's margin below the word-level threshold; rows without a margin state the reason instead. Citations name the proposals each family covers.

esis [Hauer and Kondrak, 2016]. Consonantal readings, vowels omitted as in a Semitic script, are suggested by the abjad-like grammar of Reddy and Knight's two-state model [Reddy and Knight, 2011]. No substitution or consonantal hypothesis passes on the manuscript in any language, on either principal hand or either parsing (Table 1). Even at the character level only three of the forty consonantal and abbreviation hypotheses pass, and those three fail the word-level statistic by two to three bits per word.

On out-of-genre Latin enciphered by substitution the solver recovered the key nearly in full and read out the Latin, though its family-identification step ranked the verbose family above the true one. The word-level criterion was calibrated on biblical prose. On the herbal, medical and recipe texts, in five languages, it rejects a correct substitution reading in every trial. For the Latin herbal and medical texts the solver recovers most of the key, yet the word statistic lies typically 2.7 bits per word beyond the scriptural window, a shift comparable to the margins in Table 1. The statistic therefore says "not scripture" rather than "not a language", and cannot by itself exclude prose of another genre. Within these limits we find no substitution or consonantal reading of the manuscript in the languages tested. On both principal hands the fixed-rule anagram test rejects Latin, Italian, German and English readings by 2.2 to 5.0 bits per word. It does not reject a Hebrew reading at the character or word level, and it accepts an Arabic one, but it also accepts an Arabic reading of most disguised German and English herbal and recipe texts. It is therefore uninformative for both Semitic languages.

4.3 Verbose ciphers

A verbose cipher is the natural mechanical form of Bowern and Lindemann's suggestion that Voynichese may encode a natural language in some non-one-to-one way [Bowern and Lindemann, 2021]. It would explain the low glyph entropy directly: Bowern and Gaskell [2022] showed that writing single sounds of Latin as glyph pairs lowers character entropy to the Voynich level. We searched the family at two table widths. A narrow solver, allowing at most four glyph groups that fuse with their successors, reads disguised Latin but fails on the manuscript in all five languages tested, the nearest miss being Hebrew at 1.1 bits per word (Table 1). That exclusion is only as firm as the measured power for two-glyph groups, which is sporadic even on scripture. Wider tables, of fifty to five hundred units, are the form a proponent of verbose encipherment would defend, and need a different solver, an expectation–maximization search over a segmental hidden Markov model with 250 latent multi-glyph units.

On real Latin disguised by a 250-unit verbose table at manuscript length, the wide-table solver recovered the unit inventory and mapping at 0.81 map accuracy. On both principal hands of the manuscript, by contrast, its best decode fits Latin worse than a self-citation control text generated from the manuscript itself. The full-width search was run for Latin only: the corresponding Hebrew-direction synthetic table was not recovered at this width, so that direction is untested, and Arabic, Italian, German and English were not searched at full width. The solver also takes the manuscript's spaces as unit boundaries, so space-free verbose parsings, other table widths, and stacks of a wide table with a second transform remain open. Within the two searches run, prefix tables of up to four groups and one 250-unit table, verbose encipherment of Latin is excluded at both the narrow and the full-width settings, and of the remaining languages at the narrow setting only.

4.4 Abbreviation and codebook readings

Reading Voynichese glyphs as the contraction signs of medieval scribal Latin is among the proposals D'Imperio surveys [D'Imperio, 1978]. All sixteen abbreviation hypotheses fail, and the nearest is 5.7 bits per word short, the widest margin of any family. Codebook and numbered-dictionary schemes are the closest mechanical form of Friedman's hypothesis of an early constructed language [Friedman, 1959, D'Imperio, 1978]. We tested them at three table sizes in five languages, with and without null glyphs and homophones.

All thirty codebook hypotheses fail: their word-level statistic, a word-order margin, reaches at best 0.28 bits per word (Latin, 240-word table, Scribe 1), short of the criterion. On enciphered text the solver identified the table's structure and recovered its head but read only a minority of individual word types correctly. The synthetic check thus validates the identification of table structure rather than a reading, and the manuscript result is limited in the same way. We find no evidence that the text is a codebook encipherment of any of the five languages at the table sizes tested. A codebook whose structure our solver cannot identify on synthetic text is not addressed.

4.5 Composite mechanisms and the Naibbe cipher

Two composite families were built on a consonantal skeleton, since the consonantal hypotheses came closest at the character level. The first overlays the skeleton with verbose multi-glyph groups; its solver reads disguised Arabic, but on the manuscript all twelve of its hypotheses fail. The second writes the skeleton with position-dependent letter variants, which add no redundancy: glyph entropy stays near 3.7 bits, against the manuscript's 2.0 to 2.2, for any decodable table, and that alone excludes the family.

The Naibbe cipher [Greshko, 2025] is a verbose homophonic substitution executed by hand with card-drawing tables; we implemented it from the published tables. On the manuscript the published mapping fails the word-level criterion in both Latin and Italian, the nearest 1.5 bits per word short. This result concerns the published mapping; the broader claim of Greshko [2025], that a verbose homophonic cipher of this class remains viable, is addressed by the family results above. They do not exclude the class: on text in the six languages enciphered with freshly randomized tables of the same design, some from Greshko's own generator, the wide-table solver showed no demonstrable recovery in 1,370 trials. Its criterion accepted a Hebrew reading on most of them whatever the true language, so that solver supports no statement about such a cipher in either direction. The published Naibbe ciphertext shares the manuscript's weak word order but differs from it otherwise: it lacks the coupling between adjacent words' edge glyphs and the many once-occurring word types [Rozanova and Temerev, 2026], and in our measurements it misses 9 and 13 of the next section's 22 statistics on the two hands.

4.6 Reading order and character inventory

A reading could also fail because the text is read in the wrong order or divided into the wrong characters. We reran the solvers under ten alternative orders (right to left, columnar, alternating by line, interleaved). The identity order fits best, and no reordering moves the word-level statistics toward any language. Every earlier attack, ours included, also fixed a division of the script into characters before solving. Here the solver chooses parsing and mapping together, fusing or splitting glyph pairs freely. Tested on a Hebrew text enciphered with a composite

mapping and then corrupted by fusing letter pairs into single symbols, it returned, without being told which pairs were fused, a decode recognizable as the original Hebrew at a word accuracy of 0.87. It identified only two of the four fused pairs exactly where three were required, so on the manuscript its results constrain segmentation and reading more firmly than individual glyph identities. On the manuscript, freeing the parsing in the same way moves none of the word-level statistics toward any language (Table 1). So the readings' failure does not seem an artifact of the transliteration alphabet.

4.7 The decipherment illusion

Modern decipherment attempts are optimizations: searches over many mappings for the one whose output looks most like language. We built such optimizers, which largely recover the true mapping when one exists, and ran them on the manuscript unconstrained. They succeed by the usual standard: every run finds mappings whose output scores above control texts, reproducibly under new page splits. Two observations show that this success is spurious. First, the decoded "Hebrew" and "Arabic" are not language-shaped. They use eight to ten of the twenty-odd available letters, and their three commonest letters cover 61–78% of all letter occurrences, compared with 33–38% in genuine Hebrew and Arabic. Their vocabularies contain 70 to 115 distinct words per 220 tokens, against roughly 160 to 180 in genuine text. Second, the optimizers decode the same manuscript pages equally well into Hebrew and into Arabic, which could not happen if either decode were a genuine reading.

Figure 4 shows the optimizer's advantage on the manuscript over a self-citation control text [Timm and Schinner, 2020] searched with the same freedom, and its collapse when the output must also be language-shaped, with letter and vocabulary statistics in the genuine ranges above. In the Hebrew test, run on Scribe 1's pages, the advantage falls from 4.1 to 0.26 bits per word. In the Arabic test, on Scribe 2's pages, it falls from 6.3 to 2.6, and the residue is carried by a decode that fails the shape requirements on the test half, so it does not establish a language-shaped solution either. The constraint does not impair recovery of genuinely enciphered text. Thus between three and four bits per word of the unconstrained solver's advantage appear to arise solely from the search's reward for language-like output and to carry no information about the text.

The reward for language-like output is not peculiar to our optimizers. Manly recognized this mechanism in Newbold's method [Manly, 1931], and Hauer and Kondrak met a version of it in their own experiments [Hauer and Kondrak, 2016], but we are not aware of an earlier measurement of the spurious signal it produces. That signal may account for much of the history of announced readings, from Newbold's anagrams to recent language-identification algorithms. Without a control text given the same freedom and a test on held-out pages, any search of a large mapping space for language-shaped output would gain a few bits per word from the text's genuine but non-linguistic structure.

5 The production of the text

Since none of the proposed readings holds on pages it was not fitted to, we ask instead how the text was made. The models below are fitted and tested on the decisively pure pages of the two principal hands, Scribe 1's 112 pages of dialect A and Scribe 2's 43 of dialect B, together 155 of the 227 text pages and about 55% of the text. Four further pages of these hands in

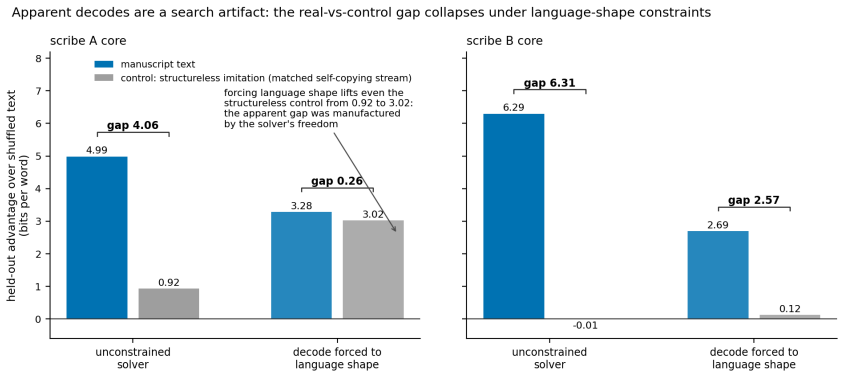


Figure 4: The decipherment illusion. The advantage an optimizing solver gains on the manuscript over a matched control text collapses when its output must be language-shaped. Bars give the advantage of the solver's best decode over shuffled text (bits per word) for the manuscript (blue) and for a self-citation control text (gray), with the solver unconstrained (left pair in each panel) and with its output required to be language-shaped (right pair). Left panel ("scribe A core"), Hebrew test on Scribe 1's dialect-A pages; right panel ("scribe B core"), Arabic test on Scribe 2's dialect-B pages, where the residual advantage is carried by a decode that fails the shape requirements on the test half.

the astronomical section, which may follow a third profile (Section 3.3), are set aside. The three minor hands and the mixture-preferring pages are classified but not separately fitted, and statements about them below are marked as such.

5.1 Word order and vocabulary

We sort the proposed writing processes into five classes. In one, the habit chain, each word depends only on the word before it. The other four are procedures proper: a template filled from lists of parts, the table and grille of Rugg [2004], Rugg and Taylor [2017], the local self-citation of Timm [2014], Timm and Schinner [2020], and a recurrence rule as in the Book of Soyga [Reeds, 2006]. Each class is written as a working generator, fitted on one half of a hand's pages and scored on the other by held-out cross-entropy, beside a bag of words (a memorized vocabulary, order ignored) and two letter models (Appendix B, which also lists the comparison texts). For the bag, the chain and the copying process we also compute the evidence of Eq. (1) on all of a hand's pages. On synthetic text written by each class, this comparison identifies the generating class in five cases of five and classifies no comparison language as generated.

Two statistics separate the manuscript from the four comparison languages (Figure 5), both differences of the cross-entropy H_M, in bits per token, of one half of a hand's pages under a model M fitted on the other half. The word-order gain is

$$G_{\text{order}} \;=\; H_{\text{vocab}} - H_{\text{chain}}, \qquad H_M \;=\; -\frac{1}{N}\,\log_2 P_M(w_1 w_2 \cdots w_N \mid \text{layout}), \tag{4}$$

where $w_1 \cdots w_N$ are the tokens of the test half with their line layout given, "vocab" is the bag of words and "chain" the habit chain, so that G_{order} measures, on withheld pages, how much the identity of a word helps predict the next. The vocabulary surplus

$$G_{\text{lex}} \;=\; H_{\text{letter}} - H_{\text{vocab}} \tag{5}$$

uses the second-order letter model, which must also place the word boundaries.

The manuscript's weak word order has been measured before [Reddy and Knight, 2011, Caruana et al., 2022] and, concurrently with this work, by Rozanova and Temerev [2026]. In the four comparison languages, laid into the manuscript's page templates, G_{order} is 0.15 to 1.06 bits per token; on the two principal hands it is 0.010 to 0.022 (Table 2, block i). We laid 23 further texts into the same templates: technical, herbal and medical Latin prose, Latin and vernacular recipe collections, glossaries, an index and a botanical catalog, narrative Latin, vernacular herbals, and the volunteer gibberish of Gaskell and Bowern [2022]. No richer estimator we fitted, word-class and affix-class bigrams included, gives either hand more than 0.022, and each hand's largest held-out gain lies below that of every one of these texts, with the botanical catalog showing the most order, as Rozanova and Temerev [2026] found. The page-bootstrap intervals separate the manuscript from every control text for Scribe 1 and from all but one for Scribe 2: the most type-rich stretch of Pliny's enumerative books on trees, at 0.039 bits per token against Scribe 2's 0.019. No control text matched to the manuscript's vocabulary sparsity comes within an order of magnitude. Near the full fitting sample the manuscript's gain grows by thousandths of a bit per doubling, the other controls' by hundredths or tenths.

The information that adjacent words do share sits at their edges. The last unit of a word predicts the first unit of the next with a held-out gain of 0.122 bits for Scribe 1 and 0.215 for Scribe 2, an order of magnitude above G_{order} and above every prose control except the recipes (at most 0.110). Some vernacular recipe collections and the botanical catalog show as much coupling or more (up to 0.31). This coupling is a lower bound on the information adjacent words share. Apart from the edge glyphs, a word's identity tells the fitted models almost nothing about the next, whereas in every meaningful control specific word pairs carry the gain. A diffuse dependence between hidden word classes of about half a bit per word pair would, however, produce the manuscript's numbers and escape these estimators at this sample size, as synthetic texts show. Each scribe's word frequencies also drift between the halves of his pages [Timm, 2026a], a drift that fitting and testing on alternating pages absorbs.

In the comparison languages G_{lex} is 1.3 to 3.0 bits per token. On the manuscript it is negligible on the atomic parsing and only 0.15 to 0.18 on the folded parsing: on the raw glyphs the manuscript's words are nearly as predictable from letter habits alone as from a memorized vocabulary. By the evidence of Eq. (1) on all of a hand's pages, the real word order is decisively preferred to within-page shuffles for Scribe 2 under every prior tried, whereas for Scribe 1 the sign of that contrast depends on the prior (Table 2, block ii).

Which of the bag and the chain has the higher evidence likewise depends on the prior: a flat Dirichlet prior on the chain's transition rows favors the bag, penalizing a table the text barely uses. A hierarchical prior that shrinks each row strongly toward the running word frequencies instead favors the chain on both hands and on each half of either hand's pages, by at most a few hundredths of a bit per token, and favors the bag again once the words are shuffled within pages. Independent of the prior, the table-and-grille generator, as parametrized here, fits worse than the bag on both hands, and local copying is a real component on both, decisively better than its shuffled control. In the complete procedure below, where reuse competes for the same tokens, the copy weight is 2.4% for Scribe 1 and 5.8% for Scribe 2. Scribe 1's weight cannot be separated from page-to-page variation in word frequencies or persistence of a prefix class, which alone yield fitted weights of 2–4% on synthetic text written without copying; Scribe 2's exceeds them. Such a fit tests copying only within the local window of Timm and Schinner [2020]: copying from much further back is statistically the same process as reuse of a growing personal stock and is returned as zero however wide the window. Hence the broader, drifting self-citation

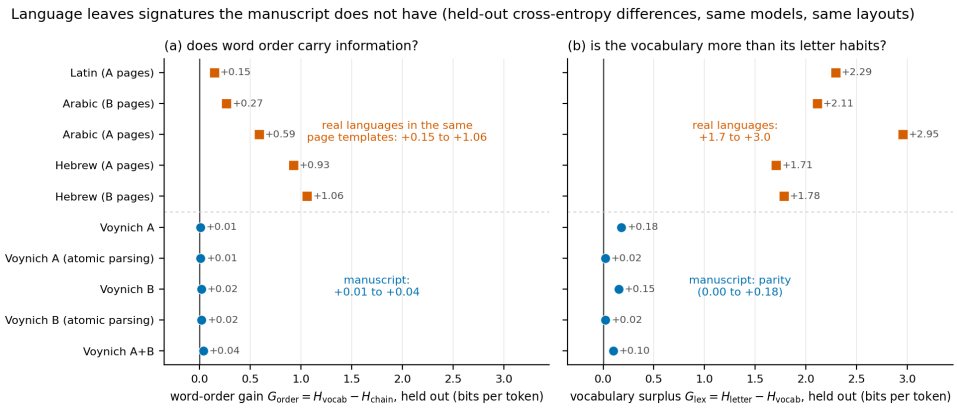


Figure 5: Two statistics that separate the manuscript from the comparison languages. In real languages the identity of a word helps predict the next and a memorized vocabulary predicts the text far better than spelling habits do; in the manuscript both effects nearly vanish. (a) The word-order gain G_{order} of Eq. (4), the improvement in predicting the next word from knowing the previous one, in bits per token, for Latin, Arabic and Hebrew laid into the manuscript's page templates (squares) and for the manuscript (circles: the two principal hands under the folded and the atomic parsing, and the pooled fit of both hands). (b) The vocabulary surplus G_{lex} of Eq. (5), the gain of a memorized word list over a second-order letter-habit model, for the same corpora. A and B denote Scribes 1 and 2 and the page sets of their hands; per-hand values are in Table 2. Italian, the fourth comparison language, is not plotted; with it the vocabulary-surplus range of Table 2 begins at 1.3 rather than the 1.7 marked in panel (b). Both panels plot differences of held-out cross-entropies.

of Timm and Schinner [2024] and Timm [2026a] is not excluded by any statistic of this kind. Naibbe ciphertext [Greshko, 2025] and genuine output of Timm and Schinner's generator [Timm and Schinner, 2020], scored the same way, share the manuscript's weak word order (gains up to 0.007 and 0.039) but otherwise show the opposite profile: local copying preferred to the bag and a copy rate above either hand's (Table 2). The self-citation process in its published form therefore does not appear to be the manuscript's, although copying remains a minor component of the procedure recovered below.

5.2 The writing procedure

Pages were composed with two or three small decisions per word (Table 2 and Figure 6). Most words are *reused*: roughly three quarters of either hand's test-half tokens are frequency-weighted draws from the scribe's own vocabulary. A few are *copied* from one of the preceding twenty or so words, mostly on the same page: about one token in fifty for Scribe 1 and one in twenty for Scribe 2, whose copies are nearly all verbatim. The rest, a quarter of Scribe 1's tokens and under a fifth of Scribe 2's, are *coined* afresh, following the scribe's letter habits so closely that a letter model predicts them as well as the vocabulary does.

Two smaller regularities complete the procedure: a preference for particular words at line boundaries and a short-range momentum in word length. The first is among the positional effects documented by Feaster [2022]; both are shared by Gaskell and Bowern's volunteer gibberish [Gaskell and Bowern, 2022]. The boundary words appear to be distinct words rather than ordinary words modified at the line's edge: we find no pairs of modified and unmodified forms beyond chance.

	Scribe 1 (A)	Scribe 2 (B)	Comparison
(i) Predictive comparisons on the test half of each hand's pages			
Word-order gain G_{order}, bits/token (folded, atomic)	0.010, 0.012	0.019, 0.022	four languages 0.15–1.06
Largest gain over the bag, any model class fitted, bits/token (folded; 95% page-bootstrap interval)	0.0102 [0.0045, 0.0161]	0.0190 [0.0073, 0.0292]	below all 23 further control texts; nearest 0.039
Edge-glyph coupling (last unit of a word predicting the first unit of the next), bits (folded, atomic)	0.122, 0.112	0.215, 0.204	prose $\leq$0.110; some recipes and the catalog up to 0.31
Vocabulary over letter-habit model G_{lex}, bits/token (folded, atomic)	0.18, 0.02	0.15, 0.02	languages 1.3–3.0
Complete procedure over reuse-and-copy baseline, bits/token, two further page splits per hand	+0.14 to +0.19		
Summary statistics violated by generated pages, of 22 (expected under the fitted model itself; 95th percentile)	2 (0.7; 2)	3 (0.8; 2.5)	controls 6–11; Timm–Schinner output 7, 5; Naibbe 9, 13
(ii) Shuffle contrasts and copying (all pages of each hand); composition of the test half			
Word-order contrast against 20 within-page shuffles, total $\log_{10}$ evidence	$+20 \pm 16$	$+240 \pm 15$	
Local copying against its shuffled control, total $\log_{10}$ evidence	$+195 \pm 24$	$+408 \pm 52$	
Local copy rate under the copying model alone, fraction of tokens	0.28	0.38	Timm–Schinner generator 0.42–0.48
Existing type (reuse), share of test-half tokens	72.9%	76.7%	
Local copy of one of the preceding $\sim$20 words	2.0%	4.9%	
New type (coinage); fitted copy weight at convergence	25.1%; 2.4%	18.4%; 5.8%	
(iii) The fitted inventories (all fitted pages of each hand)			
Pages fitted; word types (folded parsing)	112; 3,155	43; 2,389	
Affix overlap between the two hands (Jaccard)	prefixes 0.65, suffixes 0.48; cores 0.23		two halves of one hand: cores 0.23–0.28
Core types in one-edit variant families; joined by single edits	60%; 90%	57%; 89%	
Odometer ordering statistic of core inventory	0.55 and 0.60		written-out sheet 0.97
In-word glyph entropy of running text, bits (folded): test pages; pages generated with a vocabulary drawn from the construction model as first fitted	2.41; 3.45		

Table 2: Production statistics of the two principal hands. Knowing the previous word adds almost nothing to prediction of the next, although the edge glyphs of adjacent words are coupled more strongly than in the four comparison-language texts, and a memorized vocabulary improves only slightly on a letter-habit model (block i); pages are composed mainly by reuse, with a little local copying that differs between the hands (block ii); the two hands' affixes overlap far more than their cores, as do those of two halves of either hand's pages, and the cores show no enumeration order (block iii). The bits-per-token rows of block (i) are differences of held-out cross-entropies of fitted models on the test half of each hand's pages. In the last row of block (i) the counts in parentheses are the mean and 95th percentile of the same count when the procedure is applied to pages simulated from the fitted model itself, and the Comparison column gives the counts for control models and for output of two published generators scored in the same way. In block (ii) the two contrasts and the copy rate use all pages of each hand on the atomic parsing with flat Dirichlet priors, and $\pm$ gives one standard deviation over shuffle draws; Scribe 1's word-order contrast is clearly positive under sparser priors on the counts and is $+0.0 \pm 13.0$ on the test half alone. The three shares are posterior probabilities of the three branches of the complete procedure, averaged over the test-half tokens, and the copy weight is the fitted mixture parameter. Block (iii) uses both halves together. "Folded" and "atomic" are the two parsings of the transliteration. "Comparison" gives the range over the comparison texts or the named control. The terms of block (iii) are defined in Section 5.3: variant families are sets of cores in which each member differs from another by one glyph in one position. The written-out sheet is a synthetic enumeration (Figure 7b).

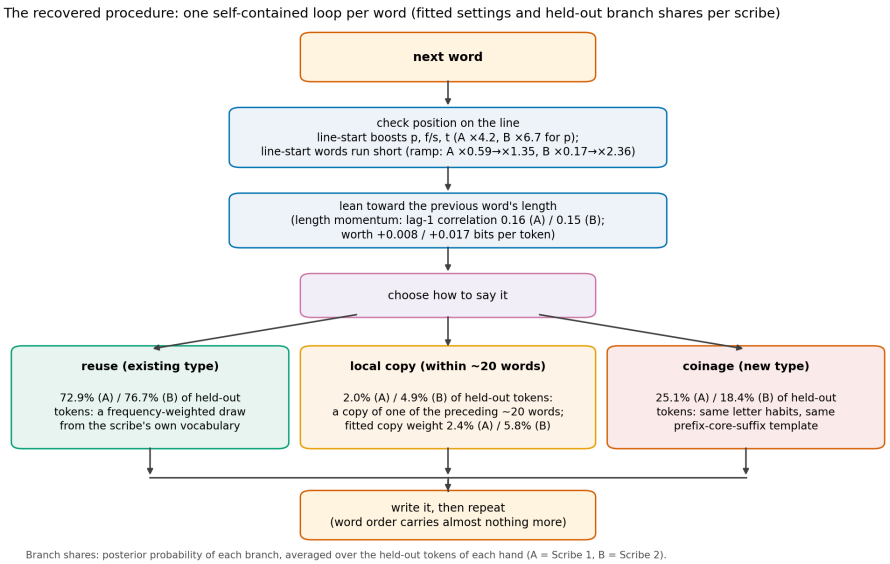


Figure 6: The recovered writing procedure. Each word is written by one pass through the loop: a check of the position on the line, an adjustment toward the previous word's length, then one of three branches (reuse of a word from the scribe's own vocabulary, a copy of a recent word, or a fresh coinage by letter habit), none of which involves a key. The loop is drawn with fitted settings of the two principal hands (A and B denote Scribes 1 and 2). The shares printed in the three branch boxes are posterior probabilities of the branches averaged over the test-half tokens, and the copy weight is the fitted mixture parameter at convergence (Table 2, block ii). The secondary settings (line-start factors, length correlation and its predictive value) are fitted values not used elsewhere in the paper.

Refitted jointly, the line-position and length terms improve on reuse, copying and coinage alone by +0.14 to +0.19 bits per token on two further page splits per hand. On twelve further splits none of the four procedures proper fits best. The test of generated pages counts, out of 22 summary statistics, those on which a model's pages fall outside the manuscript's own cross-page range. That count is two and three for the complete model on the two hands, against six to eleven for unigram, token-shuffle and randomized-vocabulary controls. Pages simulated from the fitted model itself reach counts this high with probability 0.16 and 0.06 (Table 2), so the counts do not reject the model. The unigram and token-shuffle controls are rejected ($P \leq 0.02$), and so is the procedure without its line-position and length terms on Scribe 2's pages. The complete model nonetheless misses several statistics outside the 22: by a whole-page measure of the edge-glyph coupling, a shuffle-corrected mutual information rather than the held-out gain above, generated pages carry 0.009 and 0.001 bits against 0.130 and 0.243 on the real ones. Adjacent near-repeats are too rare, and on Scribe 2's pages once-used types are too few and consecutive lines begin with the same glyph too often. No keyless generator yet proposed produces the edge coupling; Timm and Schinner's output and the Naibbe ciphertext lack it as well [Rozanova and Temerev, 2026], and a slot generator and Rugg's grille miss the in-sample word-boundary signatures of Parisel [2026b]. Appendix C prints folio 26v beside a page written by the fitted procedure.

5.3 Edge units and cores

Each scribe's vocabulary has three properties that jointly match no comparison corpus we could construct. Affixes and cores combine nearly independently, filling a dense grid (98% of the top 10×10 combinations occur for Scribe 1). Eighty-five percent of word types lie within one glyph edit of another, and 84–85% form one connected component of that one-edit network. This network is Timm's word grid [Timm, 2014, 2016] and Zattera's slot grammar [Zattera, 2022]. The number of types grows smoothly through the text, as in genuine Timm–Schinner output [Timm and Schinner, 2020]. Latin, Hebrew and Italian vocabularies decomposed by word grammars chosen to mimic this joint profile still fail to match it.

The affix inventories of both hands are small two-slot sets, filling grids of roughly 20×20 first and second units at 21–35% occupancy (Figure 7a and Appendix A). A grid or template fits these inventories better than a chain or letter-habit process, but the same holds for the word-edge inventories of every comparison language. Nothing in the inventories indicates whether they were ever written down.

The two hands' prefix and suffix inventories overlap at Jaccard 0.65 and 0.48, and their core vocabularies at only 0.23. Two halves of one hand's pages, however, give the same contrast (cores 0.26–0.28 for Scribe 1's halves and 0.23–0.25 for Scribe 2's), as do Scribes 3 and 4 against the principal hands and every comparison language cut into two samples. Any two samples of one productive spelling system, two scribes or two halves of one scribe's work alike, thus split into widely shared edge units around sample-specific cores, and the split does not indicate that materials passed between writers.

The core inventories show no sign of having been written out as tables (Figure 7b). A statistic built to detect the odometer-like order of a list copied from a written grid takes 0.97 on a synthetic enumeration sheet and 0.55 and 0.60 on the two hands' cores, near its value for an unordered list. Instead, the cores match strings coined from each scribe's letter habits: a little under 60% of each core vocabulary falls into families of one-position variants (variant families), and single edits join about 90% of it into one network. One family in the core inventory is *kcheo, kheo, pcheo, scheo, ssheo, lsheo, theo*: seven words on one skeleton, each differing from a neighbor by a single glyph in a single position. A family's members do not appear to have been coined together. Apart from a slight same-page excess for Scribe 1, their first occurrences lie no closer together in the text than those of unrelated cores, whereas whole words one edit apart, mostly affix variants of a core already on the page, do first appear near one another. The cores are thus coinages of a letter habit with strong positional structure, from which the vowel-like alternation of Voynichese also arises.

A generative model of the vocabulary, in which each new type is coined by letter habit or derived from an earlier one by a single edit, has two known limits. It reproduces the vocabulary's architecture (the occupancy, the edit network, the variant families), but as first fitted it draws vocabularies less glyph-regular than the scribes' (in-word glyph entropy 3.45 bits against 2.41 on Scribe 1's test pages). The generated text shown here therefore uses the fitted vocabularies. Conditioning edits, affix-core joins and frequencies on letter habit at the level of single units largely closes that gap. By the same measure over whole manuscripts, synthetic text written from such vocabularies reaches 2.563 and 2.194 bits against the real hands' 2.487 and 2.094, and misses a median of 4 and 4.5 of the 22 statistics. The other limit is that one likelihood comparison among candidate construction processes preferred the edit-derivation process on all eight real-language vocabularies on which it was run as a control, where no such process

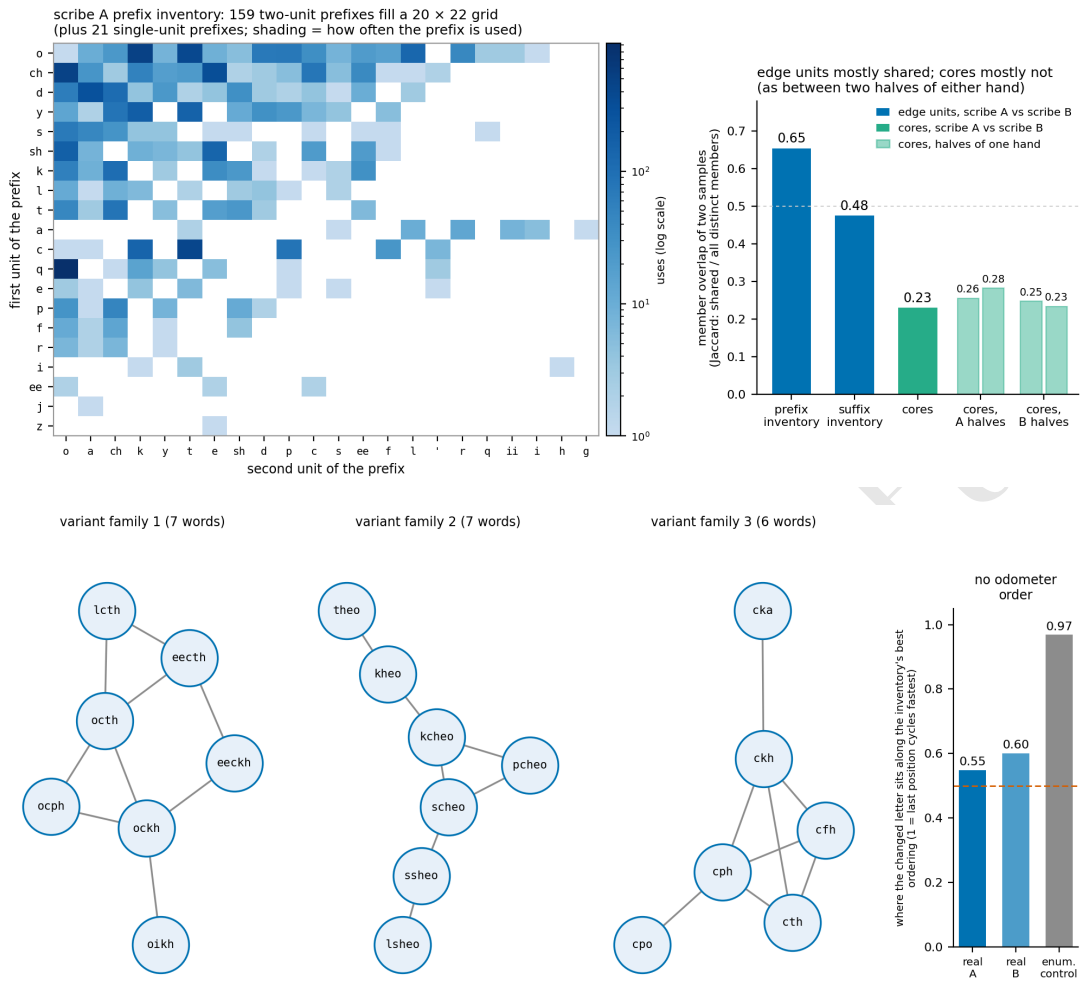


Figure 7: Affix inventories and core variant families. (a) Top: Scribe 1's prefix inventory drawn as a two-slot grid, first unit by row and second unit by column (the 20 first units heading at least one two-unit prefix; 159 filled combinations; shading gives how often each prefix is used, on a logarithmic scale), with the largest rows o-, ch-, d-, y-, s- at the top. Side panel: member overlap (Jaccard index, shared types over all distinct types) between the two principal hands' prefix inventories, suffix inventories and core vocabularies (dark bars), and between the core vocabularies of two halves of each hand's pages (light bars, two partitions per hand). Word beginnings and endings are largely shared between the hands and the middles are not, as between two halves of either hand's pages, so the split does not indicate a shared written table. (b) Bottom: three variant families from Scribe 1's core inventory drawn as graphs, each edge joining two cores that differ by one glyph in one position. Inset: the odometer ordering statistic, which is high for a list copied out from a written enumeration (the synthetic control, gray bar) and lies at the dashed line when the position that changes between neighbors is unrelated to list order; the two hands' real cores lie near the dashed line. A and B denote Scribes 1 and 2.

operated; it is therefore used only to rank processes within a family.

Quantity	Value
Label tokens (primary; all label positions)	864; 1,164
o-initial share: labels; line-initial words on label pages; null matched on scribe, section and position; paragraph-initial running text	59%; 42%; 32%; 18%
o-initial excess over matched nulls	11–17 standard deviations (within each of the three principal label hands 11–15; absent in the smallest hand)
Labels never occurring in running text	44%
Label types on more than one page: observed; expected under label-shuffle and running-text nulls	75; 85 and 111
Label reuse within a pictured-object class, against matched null	at or below null in all 7 object classes with adequate counts (stars included)
Sensitivity of the label-reuse test	synthetic nomenclature detected at 20% of repeat-mention mass ($z = 2.3$), missed at 15%
Section-conditioned word choice, full corpus	+63 bits; permutation $p \leq 1/101$
Restricted corpus, labels included; labels removed	+37 bits, quire-aligned $p \leq 1/201$; +47 bits
Per page; under harshest page-condition controls	$\approx$0.2–0.3 bits; $p \approx 0.05$
By section: astronomical; balneological; herbal (129 pages)	+109 bits ($z = 2.6$); +127 ($z = 2.9$); −31 ($z = 1.2$)
Other channels: first, penultimate and last unit against contiguous null	$z = -0.01$, 1.3, 1.4 (all within the contiguous null)
Word length; second unit and core	fails under restriction; fail quire-aligned nulls

Table 3: Tests for meaning. Two weak traces of attention to the pictures survive the null controls: the labels form a distinct, o-initial register whose words recur across pages less often than chance would have it (upper block), and the choice of words on a page depends faintly on the illustrated section, in the smaller sections only (lower block). Nulls are matched on scribe and page order throughout; z values are against the stated null ensemble and p values are empirical, bounded by the number of permutation draws.

5.4 Tests for meaning

A generated text can still carry meaning in labels, indices or content-keyed word choices. We tested for it with scribe identity and page order controlled throughout; every positive had to survive nulls that destroy the claimed effect by construction: contiguous fake sections, reshuffled labels and, because sections, hands and binding are entangled [Clemens, 2016], fake sections aligned to the quires. Two components survive these controls, both weak (Table 3).

The first component is a *label register*: the 864 primary label tokens written beside the drawings are made differently from running text (Figure 8). On the pharmacy page f88r (Figure 1b) the six labels of the top band read  (*otorchety*),  (*osal*), (*orald*), (*oldar*),  (*otoky*), (*otaly*). All six begin with the glyph transliterated o-, and across the page's three label bands the count is thirteen of fifteen (*dary* and *sorory* are the exceptions). Corpus-wide, about three labels in five are o-initial, against one word in three under a null matched on scribe, section and position.

Nothing we can measure indicates that the labels name the pictured objects. A shared nomenclature would make label words recur with their objects; instead the labels recur across pages less often than chance predicts, 75 label types on more than one page against 85 to 111 expected under the two null models. Within the pictured-object classes of the transliteration's locus codes (zodiac star, pharmaceutical container and the like), whole-token reuse of labels

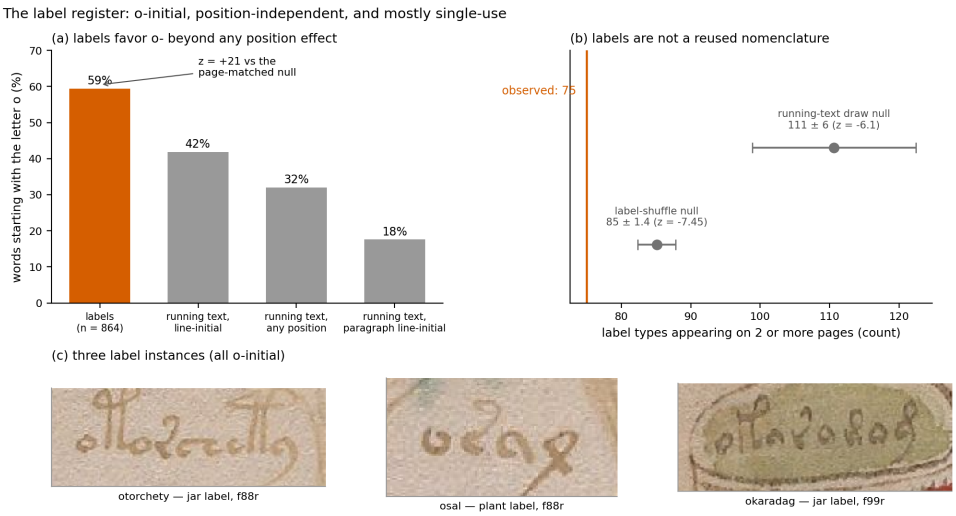


Figure 8: The label register. Labels favor an initial o- beyond any position effect, and they recur across pages less often than a nomenclature of recurring objects would. (a) Share of o-initial words among the primary labels, among line-initial words of running text on the labels' own pages, among running-text words under the null matched on scribe, section and position, and among paragraph-initial running text. The z value printed in the panel is against a page-matched null; the excess quoted in Table 3 is against the nulls matched on scribe, section and position. (b) Number of label types found on two or more pages, observed (vertical line) against the label-shuffle and running-text nulls (null mean ± two standard deviations); the observed count lies below both. (c) Three label crops from folios 88r and 99r. Manuscript images: Beinecke MS 408, Yale University (public domain).

is at or below a matched null in all seven classes with adequate counts. Even the two zodiac constellations drawn twice share no label word between their pages, a check that rests on the transliterator's constellation identifications. The test has a measured sensitivity: a synthetic nomenclature carrying a fifth of the repeat-mention mass is detected and one carrying 15% is not, so a sparser naming system could have escaped it. And the transliteration carries no cross-page identity for individual stars or plants, so a catalog in which each individual is named once and never again is untested.

The second component is a faint *subject tinting* of vocabulary: the mix of word types on a page shifts with the illustrated section beyond what scribe identity and page order explain (Figure 9). Section-linked word use has been reported before [Montemurro and Zanette, 2013, Reddy and Knight, 2011, Sterneck et al., 2021] and attributed to the hand [Gropper, 2026] or to drift in production order [Timm, 2026a] rather than to subject; the test here therefore conditions on scribe and on a window of manuscript order. We examined seven channels: which words a page uses, how long they are, and which units fill each of five positions in the word (Table 3).

Only word choice exceeds both the contiguous and the quire-aligned nulls, and the effect is small, roughly 0.2 to 0.3 bits per page, surviving removal of the label tokens and weakening to marginal significance under the harshest page-condition controls. With seven channels tried, the one survivor is the strongest of seven and not an independent confirmation; and under quire conditioning the raw evidence gain is negative, so the claim rests on the comparison with the nulls. The word-choice effect comes from the astronomical and balneological sections. The herbal section, the manuscript's largest at 129 pages, shows no coupling of its own, so the tint is probably a property of the smaller sections and not of the whole book. Tests for section-specific

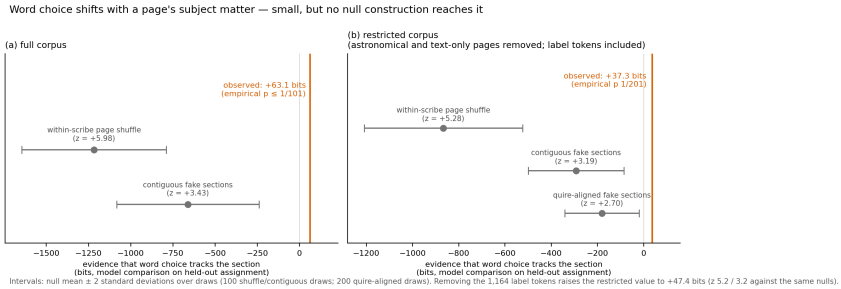


Figure 9: Subject dependence of word choice against three null ensembles. Evidence, in bits, that the mix of word types on a page tracks the illustrated section beyond scribe identity and page order (vertical line), against null ensembles in which the effect is destroyed by construction: within-scribe page shuffles, contiguous fake sections and, in (b), fake sections aligned to the quires (null mean ± two standard deviations over draws). (a) Full corpus; (b) restricted corpus with the astronomical and text-only pages removed and the label tokens included, the corpus on which the quire-aligned null was run. The observed value lies outside every null band, and least far outside the quire-aligned one.

grammar and for dependence of dialect on content are negative, and beyond the label register and the subject tint we detect no encoded content.

5.5 A reconstruction of the production

The measurements of word order, writing procedure, vocabulary and labels combine into one account of the book's production. It rests on the two principal hands, extends to the three minor hands largely by inference, and inherits the limits stated in those sections. The account is therefore a proposal, the production process most consistent with the statistics we have measured.

The five hands that Fagin Davis [2020] distinguishes share one word grammar and, where measurable, largely the same edge units. Word frequencies otherwise follow the hand and the illustrated section, and any two samples of pages, from two hands or from two halves of one hand's work, share most edge units and fewer of their cores. These vocabulary statistics therefore cannot distinguish several writers trained in one spelling habit from one writer whose habit drifted [Timm, 2026a,b], and they do not test the hand attribution itself. Scribe 2 copies a recent word more than twice as often as Scribe 1, nearly always verbatim, and coins new words less often. Writing pages in this way requires practice, as the fluent penmanship and the very few corrections suggest, but not literacy in any particular language, knowledge of ciphers, or apparatus such as the table and grille of Rugg [2004], Rugg and Taylor [2017].

It may help to picture the working routine that this procedure implies, on the understanding that the scene is illustrative and carries no evidence of its own. A scribe fills the space around a finished drawing from his vocabulary, varying or copying recent words and coining a label-like one beside each drawn part. The faint subject tinting of word choice and the freshly coined labels suggest the book was written onto finished drawings or at least with them in view. Apart from these two effects we find no dependence between words beyond the coupling of adjacent word edges, no stable names that could carry reference, and no mapping to any of the languages and mechanisms examined above, each within the sensitivity of the corresponding test.

5.6 Purposes, parallels and possible refutations

Why the book was made can only be inferred, and the interpretation of the historical record offered below, although it draws on the measurements, is less secure than the results above. The text was cheap to add: at ordinary scribal rates of 500 to 1,000 words an hour it represents some 40 to 80 hours of writing, little beside preparing a hundred leaves of vellum and executing the drawings. The balance of effort thus suggests an illustrated object with text added to complete its appearance.

The book's sections correspond to the secret-knowledge genres of its century [Williams, 2003, Eamon, 1994, Kauffmann, 1959, Siraisi, 1990, Carey, 2004]. Such a book might be sold to a collector of marvels, Jean de Berry for instance [Meiss, 1967, Guiffrey, 1894–1896], or shown as a practitioner's credential and never examined. Either use needs only a viewer who leafs through without reading. Such books are also made for their own sake: the *Codex Seraphinianus* [Serafini, 1981, Stanley, 2010] is 254 pages of deliberately meaningless script around impossible pictures, drawn over three years by one artist and published in 1981. An older parallel is the Enochian writing of John Dee and Edward Kelley, though statistical comparisons disagree on whether the manuscript resembles it [Daruka, 2021, Boxer, 2022].

Each alternative purpose seems to require something the text lacks, at least within the reach of our tests. An early constructed language, the possibility Friedman came to favor [Friedman, 1959], needs syntax, and the running text shows almost none, less even than the glossaries, index and catalog compared above. A secret nomenclature needs stable names, but labels of recurring pictured objects repeat no more often than chance. Glossolalia or a devotional writing practice might yield a vocabulary coined by habit, but hardly label-like words beside each pharmacy jar. A cipher needs a mapping, and across a century of proposals [Newbold, 1928, Hauer and Kondrak, 2016, Greshko, 2025], every mapping we could make mechanical fails on withheld text or passes unselectively. Among the interpretations that the transliterated text allows us to test, the one most consistent with the production evidence is a book of apparent meaning made deliberately, with a method, for an audience meant to be impressed by it rather than to read it.

The reconstruction is an inference from the statistics and could be refuted in several ways. A confirmed reading of any passage would overturn it, provided the reading passed where every past claim failed: on withheld text, against matched controls, with language specificity. So would a passage that predicted the imagery in detail, beyond the labels and the faint subject tinting. A natural language that matched the manuscript on the word-order and vocabulary comparisons would count against it too, though no text compared above comes close at the manuscript's vocabulary sparsity. Codicological evidence that the text was not written onto finished drawings, such as a production sequence or an erased under-text, would also refute it.

6 Conclusions

For four hundred years readers have asked what the Voynich manuscript says. We have asked instead, as others since Friedman have, what process produced its text. The questions are connected: a keyless procedure with no source text leaves nothing to decipher. On the first question, what the text says, our results are negative within a stated space. Where the test is selective, every reading that we could make mechanical fails to produce language-like output on test pages, in six candidate languages under both parsings. Where the test passes true readings of the same kind (substitution, fixed-rule anagramming, abbreviation and narrow homophony

of scripture in some languages at low transcription noise), the failures are the manuscript's own. Elsewhere the criterion rejects even correct readings or, for wide homophonic tables and the Naibbe class, has no power, and the anagram test is uninformative for Hebrew and Arabic. Such readings are constrained only by the character-level results and the production evidence. The space is limited in three respects: free anagramming cannot be tested at this text size and lies outside it [Newbold, 1928]; only Latin was searched at full verbose-table width; and the word-level criterion was calibrated on biblical prose only. An unconstrained search for the most language-like reading yields several bits per word of spurious signal, which may explain the succession of announced decipherments.

On the second question, how the text was produced, our measurements sharpen a picture that has been emerging since the work of Rugg and of Timm and Schinner. A keyless page-writing procedure reproduces nearly all the summary statistics of the two principal hands, within its sampling noise, but not the coupling between adjacent words' edge glyphs, which published generators also lack. Each scribe's vocabulary joins a small set of edge units, found in every hand, to a larger stock of cores, most of which occur in only one sample of pages, whether the samples are two hands or two halves of one hand's work. This split therefore does not show how the writing was organized. Each scribe composed pages by reusing his vocabulary, with some coining, a little local copying, and no detectable dependence between words beyond that edge coupling. Local copying, central to Timm and Schinner's generator [Timm, 2014, Timm and Schinner, 2020], is a minor component of ours; copying from further back cannot be told from reuse, so their broader self-citation [Timm and Schinner, 2024, Timm, 2026a] is not excluded. Rugg's table and grille [Rugg, 2004], as parametrized here, fits the text worse than reuse alone. Two weak traces of content survive controls for scribe and page order: a distinct label register and a faint subject dependence of word choice in the smaller sections. We read these traces as signs that the writers tracked the pictures, but neither trace shows that the text describes them.

Our reconstruction has clear limits and assumes the five-hand attribution of Fagin Davis [2020]. It is fitted on two of the five hands and extended to the others largely by inference, and even on the two fitted hands the complete model misses a few of the summary statistics it is scored on. Reference confined to single, unrepeated objects cannot be tested from a transliteration, and meaning carried by something other than word identity and word order would be invisible to every test used here. Why the book was made requires a further step of inference. The low cost of the writing and the genres the pictures imitate suggest a book made to be looked at and believed, as a credential or for sale to a collector. The reconstruction is falsifiable, most directly by a confirmed reading of any passage or by codicological study beyond the published collation [Clemens, 2016]. The same comparison, exact evidence and prediction on withheld text against controls of known construction, applies to other disputed texts, such as suspected forgeries and undeciphered scripts.

A Affix inventories

The four tables give the largest rows of the prefix and suffix inventories of the two principal hands, the inventories compared in Section 5.3. The inventories are compiled from the decisively pure pages of each hand, Scribe 1's 112 pages of dialect A and Scribe 2's 43 of dialect B, the test half included. Units follow the folded parsing of the ZL transliteration: *iin*, *ii*, *ee*, *ch* and *sh* are single units and every other character stands alone. Each inventory is the set of distinct prefixes (or suffixes) found when that scribe's word types are split into prefix, core and suffix;

frequencies are not shown.

Each row holds the two-unit affixes that share a first unit, with their second units in alphabetical order. We show the eight largest rows of each inventory. In the first table the largest are the rows of *o-* (19 members), *ch-* (16), *d-* (14), *y-* (11) and *s-* (11). Single-unit affixes (21 prefixes and 25 suffixes for Scribe 1, 19 and 24 for Scribe 2) have no second unit and are omitted.

Arranged as two-slot grids, Scribe 1's full prefix and suffix inventories measure 23 first units by 22 second units and 28 by 23, and Scribe 2's 19 by 23 and 24 by 15. In that order, 159 of 506 cells are filled (occupancy 0.31), 134 of 644 (0.21), 155 of 437 (0.35) and 84 of 360 (0.23). The fraction of two-unit members lying in rows with at least two members is 0.9874, 0.9776, 0.9935 and 0.9405, against 0.9997, 0.9945, 0.9999 and 0.9767 expected if the same number of cells were filled at random. The rows are thus, if anything, slightly emptier than chance, and the grid is only a way of displaying the inventory. The row count in a heading can be smaller than the number of first units, since a unit attested only as a single-unit affix heads no row. Scribe 1's prefixes have three such units among the 23, hence the 20 rows of the first heading and of Figure 7a.

The cross-scribe overlap is the Jaccard index of the two full inventories, single-unit affixes included: the affixes attested for both scribes as a fraction of those attested for either. It is 0.65 for prefixes and 0.48 for suffixes, against 0.23 for the cores. Two halves of one scribe's pages give the same contrast: the first and last halves of Scribe 1's pages overlap at 0.683, 0.522 and 0.256, and those of Scribe 2's at 0.644, 0.509 and 0.248, so the pattern does not distinguish two writers from one. The core inventories are not tabulated here. The tables show which affixes each scribe used and nothing about whether, or in what form, the scribes ever wrote them down.

Scribe 1, prefix rows (top 8 of 20 rows; 159 two-unit members across all rows):

first unit	second units attested
o	*a, c, ch, d, e, ee, f, i, ii, k, l, o, p, q, r, s, sh, t, y*
ch	*', a, c, ch, d, e, ee, f, k, l, o, p, s, sh, t, y*
d	*a, c, ch, d, e, ee, k, l, o, p, s, sh, t, y*
s	*a, c, ch, e, ee, f, k, o, q, sh, y*
y	*a, c, ch, d, f, k, o, p, s, sh, t*
k	*a, c, ch, d, e, ee, o, s, sh, y*
sh	*a, c, d, e, ee, f, k, o, t, y*
l	*a, ch, d, k, o, p, s, sh, t*

Scribe 1, suffix rows (top 8 of 24 rows; 134 two-unit members across all rows):

first unit	second units attested
o	*d, f, g, iin, j, l, m, n, p, r, s, t, y*
e	*a, b, d, g, m, o, p, r, s, t, u, y*
a	*g, iin, j, l, m, n, r, s, t, y*
r	*a, d, e, g, iin, l, n, o, r, y*
ee	*b, d, e, g, o, p, s, u, y*
d	*a, d, g, l, o, s, sh, y*
i	*a, g, l, m, n, r, s*
ch	*', a, d, o, s, y*

Scribe 2, prefix rows (top 8 of 19 rows; 155 two-unit members across all rows):

first unit	second units attested
o	*a, c, ch, d, e, ee, f, h, i, ii, k, l, p, q, r, s, sh, t, y*
ch	*', a, c, ch, d, e, ee, f, k, l, o, p, s, t, y*
y	*c, ch, d, e, f, k, l, o, p, q, s, sh, t*
l	*a, ch, d, f, k, l, o, p, r, s, sh, t*
sh	*a, c, ch, d, e, ee, k, l, o, sh, t*
a	*c, d, i, ii, iin, k, l, m, r, sh*
d	*a, ch, e, ee, k, l, o, sh, y*
q	*c, e, ee, f, k, o, s, y*

Scribe 2, suffix rows (top 8 of 22 rows; 84 two-unit members across all rows):

first unit	second units attested
e	*d, f, g, o, r, s, t, y*
o	*d, g, iin, l, m, r, s, y*
i	*l, m, n, o, r, s, y*
a	*iin, l, m, n, r, s*
d	*e, iin, l, r, s, y*
ee	*d, g, l, o, s, y*
ch	*d, l, s, x, y*
h	*d, e, o, s, y*

B Models, cross-entropies and validation tests

Every model compared in Section 5 is a normalized probability distribution over the token sequence of a page, given the page's line layout. Two families of comparison are used, and they involve differently specified models. In the first, the bag of words, the habit chain and the copying process of Table 2, block (ii), are compared by the evidence of Eq. (1) on all pages of each hand. Here the vocabulary is closed, every frequency and transition row has a symmetric Dirichlet prior, the chain's context is reset at each line, and a word counts as copied when it lies within one glyph edit of a word earlier in its line or in the line before. In the second, eight model classes are fitted on alternate pages of each hand and scored by their cross-entropy on the remaining pages. Here the vocabulary is open, new spellings being drawn from a glyph-bigram model, the chain's context runs across line breaks within a page, and copying is a fitted mixture over the preceding twenty words. The table below and block (i) of Table 2 report the second family; the numbers are held-out cross-entropies of fitted generators, not evidences. The 23 further texts of Section 5, laid into the same two page templates for the word-order comparison only, are: Pliny's *Natural History*, books XII to XIX and, separately, books XX to XXVII, with the index of contents that forms book I; Celsus, *De medicina*; Macer's herbal in a clean edition; Isidore, *Etymologies* XVII; Sacrobosco, *De sphaera* with *De arte numerandi*; Caesar, *Gallic War*; the recipe collections of Apicius, the *Liber de coquina* with its *Tractatus*, the *Antidotarium Nicolai*, the Anonimo Toscano, the Anonimo Veneziano, Maestro Martino, the *Buch von guter Speise* and the *Forme of Cury*; the glossaries *Alphita* and *Sinonoma Bartholomei*; Linnaeus's *Species Plantarum*, a catalog ordered by class; Culpeper's herbal; Konrad von Megenberg's *Buch der Natur*; Cennini's *Libro dell'arte*; and the volunteer gibberish of Gaskell and Bowern [2022], many writers' samples joined. Long texts enter as up to three stretches chosen for vocabulary richness, short ones are joined or laid into a shortened template, and three (Megenberg, the *Antidotarium*,

Alphita) are machine-read with some transcription noise. Pages written by a fitted generator are also compared with the manuscript directly on 250 summary statistics of the text, with the manuscript's own page-to-page variation as the yardstick. In the main text and Table 2, that comparison uses ten broad statistics and a list of 24 finer ones, and the counts printed are over the 22 finer statistics that are numeric for both the real and the generated pages. A statistic counts as missed when the generated value lies at least four of the manuscript's page-to-page standard deviations from the real value (the deviation floored at 2% of the real value) and differs from it by more than 10%.

B.1 Model classes and the cross-entropy comparison

Three of the eight model classes are reference models with no writing procedure. In *vocabulary reuse*, the bag-of-words or reuse-only baseline of Section 5.2, each token is an independent, frequency-weighted draw from a memorized vocabulary. The two *letter models* have no vocabulary, only glyph-transition habits of order one and two. They must also encode where the word boundaries fall, which the token models receive as given. Hence the comparison is biased against the letter models. The *habit chain* is vocabulary reuse with one word of context, carried across line breaks within a page, and is therefore grouped here with the reference models, although the main text counts it among the five writing processes.

The remaining four classes (template, table-and-grille, copy-with-edits and wheel) are writing procedures, each implemented as a working generator. The *template* generator fills a fixed word skeleton, prefix then core then tail, from short lists of parts. In the *table-and-grille* procedure of Rugg [2004] a cut card is slid over a table of fragments and the exposed cells are copied out. We parametrize the grille generator here as a three-column fragment table with lazily advancing row pointers, fitted by expectation–maximization. A generator using Rugg and Taylor's own tables and placement rule [Rugg and Taylor, 2017] has not been scored. *Copy-with-edits*, after Timm and Schinner [Timm, 2014, Timm and Schinner, 2020], is line-window copying: a token repeats, with small changes, a word from nearby earlier lines. The *wheel* or *recurrence* generator derives each entry from its neighbors by rule, as in the Book of Soyga tables [Reeds, 2006].

Each class was fitted separately to the decisively pure pages of the two principal hands, with the minor hands excluded. These pages were selected with each page included in its own dialect's reference profile; when each is rescored with itself left out (Section 3.3), 104 of the 112 and 34 of the 43 remain decisively pure and the rest lean toward the blend. The pages of each hand were split alternately in page order, even positions for fitting and odd positions for testing, with hyperparameters tuned within the fitting half alone. The split gives 56 and 56 pages (5,554 and 5,195 tokens) for Scribe 1 and 22 and 21 pages (5,587 and 4,921 tokens) for Scribe 2. The fitting halves contain 1,992 and 1,582 distinct word types in the folded parsing. The table below gives each fitted model's cross-entropy on the test pages in bits per token. Lower values are better.

	folded parsing		atomic parsing	
model class	Scribe 1	Scribe 2	Scribe 1	Scribe 2
vocabulary reuse (baseline)	11.829	10.626	11.934	10.689
habit chain (word Markov)	11.819	10.607	11.922	10.668
template/slot generator	15.474	15.508	17.285	16.810
table-and-grille	12.638	11.654	12.745	11.661
copy-with-edits	12.136	11.048	12.469	11.389
wheel/recurrence	16.367	15.903	17.818	16.937
letter model, order 1	12.734	11.568	13.095	11.914
letter model, order 2	12.009	10.778	11.956	10.712

On both hands and both parsings the habit chain ranks first and the memorized vocabulary second, and the differences between their cross-entropies are, up to rounding, the word-order gains G_{order} of Eq. (4) printed in Table 2. Copy-with-edits is the best-placed procedural class in all four fits. Its cross-entropy is 0.318 bits per token above the habit chain's for Scribe 1 and 0.441 for Scribe 2 on the folded parsing. The page-bootstrap 95% intervals, from ten thousand resamples of the test pages, are 0.258–0.378 and 0.405–0.471. On the atomic parsing the margins are 0.546 and 0.722 (0.479–0.616 and 0.683–0.757). Page by page, the habit chain codes 50 of 56 and 21 of 21 test pages in fewer bits than copy-with-edits on the folded parsing, and 53 of 56 and 21 of 21 on the atomic.

On the folded parsing the order-2 letter model's cross-entropy exceeds the memorized vocabulary's by 0.181 and 0.152 bits per token. These two values are the vocabulary surplus of Eq. (5). On the atomic parsing the excess is 0.022 and 0.023 bits per token, and the two models are at parity. On twelve further random page splits, three per hand and parsing, a non-procedural model had the lowest cross-entropy every time: the habit chain on ten, and the order-2 letter model on two atomic-parsing splits of Scribe 2's pages. Unlike the alternating split, these random splits do not systematically place the two sides of a leaf, or facing pages, in opposite halves. Copy-with-edits was the best-placed procedural class on every split, 0.270–0.722 bits per token above the best non-procedural model across all sixteen fits (the four in the table and the twelve splits). The four page splits on which Table 2 compares the complete procedure with the reuse-and-copy baseline are, for each hand, the original partition with the fitting and test halves exchanged and a partition into blocks of page pairs.

In the first family of comparisons, by evidence on all pages of each hand, the ranking of the bag of words and the habit chain depends on the prior over the chain's transition rows. Under a symmetric Dirichlet prior with unit concentration on every row the bag has the higher evidence on both hands. The evidence integrates over every setting of the chain's table of word-to-word transition probabilities, and since word order in this text is nearly uninformative almost none of those settings predicts the pages better than the bag does. Under a hierarchical prior that shrinks each row strongly enough toward the running word frequencies, the chain has the higher evidence on the whole page set and on each half of the pages separately. At the best-fitting strength of shrinkage the factor on the whole set is $10^{58.1}$ for Scribe 1 and $10^{122.0}$ for Scribe 2 when the chain's context is carried across line breaks within a page, or $10^{136.0}$ for Scribe 2 when it is reset at each line. On the same pages with their words shuffled within the page, the bag keeps the higher evidence. Per token, the chain's advantage under the hierarchical prior is 0.0180 bits for Scribe 1 and 0.0386 to 0.0430 for Scribe 2, against held-out gains of 0.010 and 0.019 in Table 2. The held-out cross-entropies of the second family, computed from one fitted table, carry no integration penalty and rank the chain ahead of the bag in all four fits of the table

above.

B.2 The rule-classification test

The affix and core inventories were assigned to construction-rule classes by a classifier, which we tested on 48 synthetic inventories whose generating rule was hidden from it. Each of the four rule classes generated test inventories at the six real inventory profiles (two hands by three word slots) in two independent draws. In *grid enumeration* a written grid with an ordered unit alphabet per position is read out in odometer fashion, with the traversal left incomplete. In *template filling* a fixed alternation skeleton is filled from small sets. In the *edit chain*, one component of the variant-family model fitted to the cores, each new member is a one-unit edit of a recently written member. In *letter habit*, members are coined from unit-level letter statistics alone, with no construction rule.

The classifier is a random forest over inventory features that ignore the order in which members first appear, together with the inventory's closed-form evidence under each rule class. Trained on 96 labeled synthetic inventories of the same design, it reached a cross-validated accuracy of 0.989, against 0.552 for classification by closed-form evidence alone. It identified 47 of the 48 test inventories correctly (0.979). By class the counts are 12 of 12 for grid, template and chain, and 11 of 12 for letter habit, the one miss being a letter-habit suffix inventory classified as a chain. By slot they are 16 of 16 prefix, 15 of 16 suffix and 16 of 16 core inventories.

Where the class was identified, the grid's position count and the template count were recovered exactly. However, the chain's substitution share was recovered only to within an absolute error of 0.37. The edit-operation mix of the core fits was therefore estimated from the ordering analysis of Section 5.3 instead. The test covers the four pure classes only. The variant-family model fitted to the core inventories, a composite of letter habit and edit chain, is not among them. Grid and template coincide on affix inventories of unit length two, and there the classifier identifies only which units each affix inventory combines.

B.3 Controls and checks

In the page accounts of Eq. (3) the profiles θ_A and θ_B are the word frequencies of the pooled reference pages of each dialect (Scribe 1's Currier-A pages and Scribe 2's Currier-B pages) over the manuscript's 8,089 word types, smoothed by a symmetric Dirichlet prior of unit concentration and then held fixed. The prior $p(w)$ on the blending weight is uniform, and Z_{mix} is computed exactly by expanding the product over tokens in powers of w. A reference page is scored against profiles compiled without it. In the Latin control of Section 3.3 the two kinds of medieval Latin were a verse herbal and prose tales, and the single corpora left undivided were random splits of each. The unrelated third text, the recipes of Apicius, on whose pages the mixture was preferred about as often as on the manuscript's astronomical pages, contained no mixture by construction.

On the 31 pages of the astronomical section, which include all 27 of Scribe 4's, a three-way blend of A, B and a third profile compiled from the other astronomical pages is preferred to the two-way blend on 30 of the 31, by a combined factor of $10^{242.9}$. On Scribe 4's 27 pages alone the third profile, compiled for each page from the other 26, is preferred to the best of A, B and their blend on 19 pages, 17 of them decisively, by a combined factor of $10^{88.4}$. The three-way blend is preferred to the two-way blend on all 27, 24 decisively, by $10^{215.6}$, and with the label words removed, or under a second transliteration, still on 26 of the 27. When the third profile is

integrated over with a Dirichlet prior instead of compiled, it is preferred with the concentration estimated from the reference pages (0.178) and not with unit concentration, so this comparison is sensitive to the prior. The two pages of Scribe 4 that keep a reference dialect against their own group's profile are the second panel of folio 68r (A) and the Rosettes foldout (B). Scribe 3's star pages likewise prefer a profile of their own to the best of A, B and their blend on 19 of 24.

Folio 115r, where the hand changes partway down the page, is counted apart from the five hands' 226 pages. Classified as a whole page it prefers the mixture to either pure dialect, and a change-point test favors a homogeneous mixture over a switch of dialect at the hand boundary by nearly sixteen orders of magnitude. We measured the test's power directly: a simulated full switch on pages of this length is detected in every one of 2,000 trials. Finally, recomputing the statistics of Section 5 under a perturbed re-reading of the transcription left the ranking of generators and the subject dependence of vocabulary unchanged. That recomputation does not test error in the transcription itself.

C A real page and a generated page

Folio 26v is a dialect-B page in Scribe 2's hand. The writing procedure of Section 5.2 was fitted to Scribe 2's hand on one half of his 43 pages and tested on the other half. Folio 26v belongs to the test half. Its full text is printed below in the ZL transliteration used throughout this paper, with the manuscript's line breaks. Below it is a page with the same number of words on each corresponding line, written by the complete model of the procedure with Scribe 2's fitted settings. That model combines a preference for particular words at line boundaries, reuse from the scribe's own vocabulary, copying from nearby earlier text, coining by letter habit, and a short-range momentum in word length.

The reused words on the generated page are drawn from Scribe 2's observed vocabulary. Thus the generated page shows only how words are reused, copied, coined and placed under the fitted procedure. The model of how the vocabulary itself was constructed, from affix inventories and coined cores, plays no part in producing it. Neither page was selected for its resemblance to the other: the real page is the first of this hand's test pages in manuscript order, and the generated page is a single random draw for the same layout.

One line shows what the fitted procedure reproduces. Line 5 of the real page,

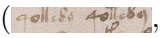
reads, in transliteration, *deeol cheody qoteedy qokody qotedy qotedy opchedy ofchy chs ar*. It has a run of near-identical *qot-/qok-* words, shared *-edy* tails and an exact repetition ( , *qotedy qotedy*). The fifth line of the generated page reads *oky qokam qeeey oky ykain sheey qokam chedy chey shedy*, with the same word skeletons, repetition pattern and tails as the real line. The same procedure with its vocabulary replaced by a random draw produces *syqy shay drshch qol aehee qiat qtd*, which is visibly unlike the manuscript's text. Nothing in the generator was tuned to these lines; the resemblance comes from fitting the procedure to other pages.

Figure 10 shows the opening lines of both pages beside lines from the randomized-vocabulary control. The pages and the figure are illustrations only. We compare generated and real pages quantitatively in Table 2 by counting, out of 22 summary statistics, those on which generated pages fall outside the manuscript's own range; for Scribe 2, the complete model misses three, a count that pages simulated from the fitted model itself reach with probability 0.06, so that

it does not by itself reject the model. In printing folio 26v, as throughout our analysis, we excluded any token containing a glyph that the transliteration marks as unreadable or codes only by number (308 tokens across the manuscript). The printed text may therefore lack a token that is present in the raw transliteration.

Folio 26v (real):

pchedar qodary daiiin pcheety s air shedy ypchedy ypchdy qopy shdy
saraiir chekedy qokedy otedy sar y etedy qokedy or ai'he alys chedy
pchdar opar dar cheeol ofchdy otedy odar chedy ytedy okchdy g
yckheody qokedy deey saldy okedor or cheos oraiin okeo chekaiin
deeol cheody qoteedy qokody qotedy qotedy opchedy ofchy chs ar
toeedy keody shedy dar chedy sches or cheeky dar chey cheky yteedy
pchedy dar cheoet chy sair chees odaiiin chkeeey ykey sheey
tchedy okeeos cheeos ysaiin okcheey keody s aiin cheeos qokes ory
ysheey okeshy shody peshy todydy

Generated by the recovered procedure:

pdshey chedal qeedy r dy qokain sheotaiin r kal y dal
pchedy lokar shes sheckhy alol qokain dychedy olsheedy qokykal qokeey qokaiin sain
sain chedy dainy oiir okaiin aiin y ar qokar qokam qedaiokaiin
olkchdy qokain odaiin aral dy shedy otedy cheykain kor sheol
oky qokam qeeey oky ykain sheey qokam chedy chey shedy
shedy otedy qol lshedary dair shedy dar thedy ry kchy chkain qokar
qokaiin rchedy ky aiin y ol y checkhy shear lched
daiin daiin chckhy s qokeedy otedy shedy ykeedy ol dar soky
olcheedy sheey ytain lkeey olkhedy

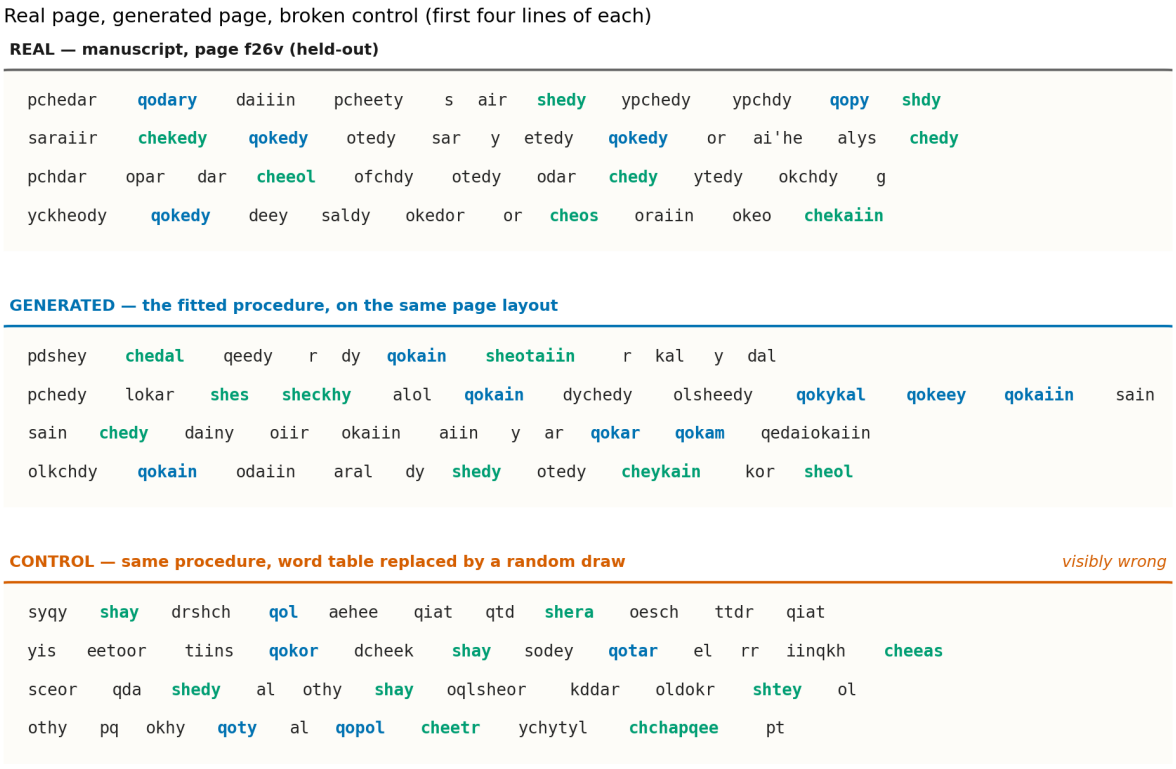


Figure 10: Real lines, generated lines and a control. The real page's mix of word skeletons and its runs of near-identical words recur in lines written by the fitted procedure but not in a control whose vocabulary is a random draw. The top panel gives the first four lines of folio 26v in ZL transliteration, and the middle panel the first four lines of the page written in the same layout by the complete fitted procedure of Section 5.2. The bottom panel gives four lines from that procedure with its vocabulary replaced by a random draw. Words opening with *qo-* (blue) and *ch-*/*sh-* (green), Scribe 2's two commonest word beginnings, are set in bold color in every panel. Folio 26v is in the hand of Scribe 2, who writes dialect B and is the "scribe B" of the figure's legend. The comparison is an illustration, and the quantitative test of generated against real pages is the count of violated statistics in Table 2.

References

Diego R. Amancio, Eduardo G. Altmann, Diego Rybski, Osvaldo N. Oliveira Jr., and Luciano da F. Costa. Probing the statistical properties of unknown texts: Application to the Voynich manuscript. *PLoS ONE*, 8(7):e67310, 2013. doi: 10.1371/journal.pone.0067310.

Georgius Baresch. Letter to Athanasius Kircher, 27 April 1639. Archives of the Pontificia Università Gregoriana, Rome, APUG 557, fol. 353; transcription and translation at https://www.voynich.nu/letters.html.

Beinecke Rare Book & Manuscript Library. Cipher manuscript (Beinecke MS 408). Yale University, digitized facsimile. https://collections.library.yale.edu/catalog/2002046; catalog description at https://pre1600ms.beinecke.library.yale.edu/docs/pre1600.ms408.htm.

William Ralph Bennett, Jr. *Scientific and Engineering Problem-Solving with the Computer*. Prentice-Hall, Englewood Cliffs, NJ, 1976.

Claire L. Bowern and Daniel E. Gaskell. Enciphered after all? Word-level text metrics are compatible with some types of encipherment. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022. Paper 6, https://ceur-ws.org/Vol-3313/paper6.pdf.

Claire L. Bowern and Luke Lindemann. The linguistics of the Voynich manuscript. *Annual Review of Linguistics*, 7:285–308, 2021. doi: 10.1146/annurev-linguistics-011619-030613.

Alexander Boxer. Fingerprinting gibberish: A quantitative comparison of the Voynich and Sloane MS 3188. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022. Paper 1.

Hilary M. Carey. Astrological medicine and the medieval English folded almanac. *Social History of Medicine*, 17(3):345–363, 2004. doi: 10.1093/shm/17.3.345.

Andrew Caruana, Colin Layfield, and John Abela. An analysis of the relationship between words within the Voynich manuscript. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022. Paper 8.

Gerard Cheshire. The language and writing system of MS408 (Voynich) explained. *Romance Studies*, 37(1):30–67, 2019. doi: 10.1080/02639904.2019.1599566. Not retracted as of July 2026; University of Bristol withdrew its press announcement (May 2019).

Raymond Clemens, editor. *The Voynich Manuscript*. Yale University Press, New Haven, 2016. With essays incl. P. Zyats et al., "Physical Findings"; ISBN 9780300217230.

Prescott H. Currier. Papers on the Voynich manuscript. In M. E. D'Imperio, editor, *New Research on the Voynich Manuscript: Proceedings of a Seminar (Washington, D.C., 30 November 1976)*. 1976. Transcribed by J. Guy and J. Reeds (1992); online at https://www.voynich.nu/extra/curr_main.html.

István Daruka. On the Voynich manuscript. *Cryptologia*, 45(1):44–80, 2021. doi: 10.1080/01611194.2019.1706063.

M. E. D'Imperio. *The Voynich Manuscript: An Elegant Enigma*. National Security Agency/Central Security Service, Fort George G. Meade, MD, 1978. Declassified; full text at https://archive.org/details/VoynichManuscriptAnElegantEnigma.

Mary E. D'Imperio. An application of PTAH to the Voynich manuscript. Technical report, National Security Agency, 1979. Undated, ca. 1979; declassified 2009 (DOCID 3562200, FOIA #58174).

William Eamon. *Science and the Secrets of Nature: Books of Secrets in Medieval and Early Modern Culture*. Princeton University Press, Princeton, 1994.

Lisa Fagin Davis. How many glyphs and how many scribes? Digital paleography and the Voynich manuscript. *Manuscript Studies*, 5(1):164–180, 2020. doi: 10.1353/mns.2020.0011.

Lisa Fagin Davis. Voynich paleography. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022. Keynote 2.

Kevin Farrugia, Colin Layfield, and Lonneke van der Plas. Demystifying the scribes behind the Voynich manuscript using computational linguistic techniques. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022. Paper 5.

Patrick Feaster. Rightward and downward grapheme distributions in the Voynich manuscript. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022. Paper 12.

William F. Friedman. Anagrammed conclusion on the Voynich manuscript, 1959. Published as an anagram in a footnote, Philological Quarterly (January 1959), plaintext deposited with the editor; theory (synthetic a priori language) documented in D'Imperio [1978], Section 6.5.

Daniel Gaskell and Claire Bowern. Gibberish after all? Voynichese is statistically similar to human-produced samples of meaningless text. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022.

Michael A. Greshko. The Naibbe cipher: A substitution cipher that encrypts Latin and Italian as Voynich manuscript-like ciphertext. *Cryptologia*, pages 1–37, 2025. doi: 10.1080/01611194.2025.2566408. Online first; open access.

Jonathan Gropper. What the Voynich manuscript says: Nothing, 2026. Preprint, Zenodo doi:10.5281/zenodo.22152069 (article, scripts, generated texts); also SSRN 7367158.

Jules Guiffrey, editor. *Inventaires de Jean duc de Berry (1401–1416)*. Ernest Leroux, Paris, 1894–1896. 2 vols.

Jacques B. M. Guy. Vowel identification: an old (but good) algorithm. *Cryptologia*, 15(3):258–262, 1991. doi: 10.1080/0161-119191865920. English exposition of Sukhotin's 1962 algorithm.

Bradley Hauer and Grzegorz Kondrak. Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computational Linguistics*, 4:75–86, 2016. doi: 10.1162/tacl_a_00084.

Greg Hodgins. Accelerator mass spectrometry radiocarbon determination of the Voynich manuscript parchment. University of Arizona AMS Laboratory; announced 8 December 2009, presented 2011–2012, 2009. Vellum dated 1404–1438 (95% probability), four samples. No peer-reviewed article exists; results summarized in Clemens [2016] and at https://www.voynich.nu/extra/carbon.html.

Takashi Izumo. Reproducing selected low-order structure of the Voynich manuscript with a material ars combinatoria: A 24×24 shelf-and-tablet proof of concept, 2026. SSRN preprint 7300521, doi:10.2139/ssrn.7300521.

C. M. Kauffmann. *The Baths of Pozzuoli: A Study of the Medieval Illuminations of Peter of Eboli's Poem*. Bruno Cassirer, Oxford, 1959.

Luke Lindemann and Claire Bowern. Character entropy in modern and historical texts: Comparison metrics for an undeciphered manuscript, 2020. arXiv:2010.14697.

John Matthews Manly. Roger Bacon and the Voynich MS. *Speculum*, 6(3):345–391, 1931. doi: 10.2307/2848508.

Johannes Marcus Marci. Letter to Athanasius Kircher, 19 august 1665, accompanying the manuscript. Travels with Beinecke MS 408; transcription and translation at https://www.voynich.nu/letters.html.

McCrone Associates, Inc. Materials analysis of the Voynich manuscript. Technical Report Project MA47613, Report to the Beinecke Rare Book and Manuscript Library, Yale University, 1 April 2009. Iron gall inks; azurite blue; text and drawings likely contemporaneous. Formerly at http://beinecke.library.yale.edu/sites/default/files/voynich_analysis.pdf.

Millard Meiss. *French Painting in the Time of Jean de Berry: The Late Fourteenth Century and the Patronage of the Duke*. Phaidon, London, 1967. 2 vols.

Marcelo A. Montemurro and Damián H. Zanette. Keywords and co-occurrence patterns in the Voynich manuscript: An information-theoretic analysis. *PLoS ONE*, 8(6):e66344, 2013. doi: 10.1371/journal.pone.0066344.

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

Yang Ou. A statistical Turing test for the Voynich manuscript: Evaluating eight quantitative linguistic properties across nine generative model families, 2026. Research Square preprint, doi:10.21203/rs.3.rs-9755825/v1.

Christophe Parisel. A quantitative confirmation of the Currier language distinction, 2026a. arXiv:2604.25979.

Christophe Parisel. Evidence of layered positional and directional constraints in the Voynich manuscript: Implications for cipher-like structure, 2026b. arXiv:2604.19762.

Sravana Reddy and Kevin Knight. What we know about the Voynich manuscript. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 78–86, Portland, OR, 2011. Association for Computational Linguistics.

Jim Reeds. John Dee and the magic tables in the Book of Soyga. In Stephen Clucas, editor, *John Dee: Interdisciplinary Studies in English Renaissance Thought*, volume 193 of *International Archives of the History of Ideas*, pages 177–204. Kluwer Academic Publishers (Springer), Dordrecht, 2006. doi: 10.1007/1-4020-4246-9_10. Analysis first circulated 1996–1998.

Liudmila Rozanova and Alexander Temerev. A glyph is not a letter, a token is not a word, a space is not a space: What the units of Voynichese are not, 2026. arXiv:2608.17096.

Gordon Rugg. An elegant hoax? A possible solution to the Voynich manuscript. *Cryptologia*, 28(1):31–46, 2004. doi: 10.1080/0161-110491892755.

Gordon Rugg and Gavin Taylor. Hoaxing statistical features of the Voynich manuscript. *Cryptologia*, 41(3):247–268, 2017. doi: 10.1080/01611194.2016.1206753.

Andreas Schinner. The Voynich manuscript: Evidence of the hoax hypothesis. *Cryptologia*, 31 (2):95–107, 2007. doi: 10.1080/01611190601133539.

Matthew D. Schwartz. Bayesian evidence for hypotheses of mixed and multiple authorship, 2026. Companion paper, https://bootloops.ai/authors.pdf.

Matthew D. Schwartz and Subhabrata Sen. Exact Bayesian evidence for mixture models: from two components to infinity, 2026. Companion paper, https://bootloops.ai/mixtures.pdf.

Luigi Serafini. *Codex Seraphinianus*. Franco Maria Ricci, Milan, 1981. 2 vols.

Nancy G. Siraisi. *Medieval and Early Renaissance Medicine: An Introduction to Knowledge and Practice*. University of Chicago Press, Chicago, 1990.

Jeffrey Stanley. To read images not words: Computer-aided analysis of the handwriting in the *Codex Seraphinianus*. Master's thesis, North Carolina State University, 2010. URL https://repository.lib.ncsu.edu/items/2f692480-0d8c-49f0-814f-57cbb4fb4fc2. M.S. thesis. Reports Serafini's May 2009 statement to the Oxford University Society of Bibliophiles that the Codex script is asemic.

Andrew Steckley and Noah Steckley. Subtle signs of scribal intent in the Voynich manuscript. In Michelle Waldispühl and Beáta Megyesi, editors, *Proceedings of the 7th International Conference on Historical Cryptology (HistoCrypt 2024)*. Tartu University Library, 2024. Also arXiv:2404.13069.

Rachel Sterneck, Annie Polish, and Claire Bowern. Topic modeling in the Voynich manuscript, 2021. arXiv:2107.02858.

Jorge Stolfi. Voynichese word-structure notes: prefix–midfix–suffix factorizations and a word grammar. https://www.ic.unicamp.br/~stolfi/voynich/. See Notes 008, 009, 030, 057, 058; accessed 28 July 2026.

B. V. Sukhotin. Eksperimental'noe vydelenie klassov bukv s pomoshch'ju elektronnoj vychislitel'noj mashiny. In *Problemy strukturnoj lingvistiki*, pages 198–206. 1962. As cited in Guy (1991).

John H. Tiltman. The Voynich manuscript: "The most mysterious manuscript in the world". Address to the Baltimore Bibliophiles, 4 March 1967; expanded paper privately circulated, Baltimore, 1968, 1967. Declassified copy released by NSA (FOIA, 2002); https://www.nsa.gov/portals/75/documents/news-features/declassified-documents/tech-journals/voynich-manuscript-mysterious.pdf.

Torsten Timm. How the Voynich manuscript was created, 2014. arXiv:1407.6639.

Torsten Timm. Co-occurrence patterns in the Voynich manuscript, 2016. arXiv:1601.07435.

Torsten Timm. The challenge of analyzing a dynamic text: Why the Voynich manuscript resists conventional interpretation. *Cryptologia*, pages 1–18, 2026a. doi: 10.1080/01611194.2026.2693462. Online first, 7 September 2026; preprint lingbuzz/009784.

Torsten Timm. One hand, five labels: A critical examination of the five-scribe hypothesis for the Voynich manuscript, 2026b. Zenodo preprint, doi:10.5281/zenodo.19169450; supplementary dataset "Archetype: Voynich Palaeography", doi:10.5281/zenodo.18827108.

Torsten Timm and Andreas Schinner. A possible generating algorithm of the Voynich manuscript. *Cryptologia*, 44(1):1–19, 2020. doi: 10.1080/01611194.2019.1596999.

Torsten Timm and Andreas Schinner. The Voynich manuscript: Discussion of text creation hypotheses. *Cryptologia*, 48(4):305–322, 2024. doi: 10.1080/01611194.2023.2225716.

Steven J. Williams. *The Secret of Secrets: The Scholarly Career of a Pseudo-Aristotelian Text in the Latin Middle Ages*. University of Michigan Press, Ann Arbor, 2003.

René Zandbergen. The Voynich manuscript. https://www.voynich.nu/. Accessed 28 July 2026; history and letters pages.

René Zandbergen. The Cardan grille approach to the Voynich MS taken to the next level, 2021. arXiv:2104.12548.

René Zandbergen. Transliteration of the Voynich MS text. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022. Keynote 1.

Massimiliano Zattera. A new transliteration alphabet brings new evidence of word structure and multiple "languages" in the Voynich manuscript. In *Proceedings of the 1st International Conference on the Voynich Manuscript (VOY2022), CEUR Workshop Proceedings Vol. 3313*, Qrendi, Malta, 2022.