Supplementary Material:
The Voice But Not the Song: A Shorthand Hypothesis and the Statistical Fingerprint of the Voynich Manuscript

Matthew Ruckman

April 2026

Contents

S1 Phase-by-Phase Narrative (Approach 2)

Approach 2 proceeds through 83 computational phases with strict selectivity gates. Key milestones are summarized below; the full codebase with all phase outputs is at `https://github.com/mruckman1/voynich_2`.

Script characterization (Phases 1–5): Stroke decomposition confirmed syllabary-like positional constraints. Fingerprinting matched Latin. A/B split confirmed. 2,328 paradigms discovered ($z = 178.4$, selectivity $1.47\times$).

Distributional and cipher attacks (Phases 6–9): Illustration-constrained decoding blocked. Noun candidates cluster $5.38\times$ above random. Homophonic, nomenclator, polyalphabetic models ruled out.

Hypothesis discrimination (Phase 10): Constructed script won (score 4.0, margin 2.5). Entropy parallelism with Latin ($r = 0.999$).

Phonetic decoding (Phases 11–16): 14-cell grid (11.1%) $\rightarrow$ 25 stroke triples (19.4%) $\rightarrow$ medieval dictionary expansion (35.4%) $\rightarrow$ modifier detection (43.6%).

Honesty diagnostics (Phase 17): NO-GO verdict. Null corpora achieve 37.6%.

Tachygraphic identification (Phases 18–19): Three-way degeneracy confirmed, then resolved (4/8 tests pass under corrected null; entropy cosine 0.820). Simplified Naibbe rejected on entropy shift (cosine -0.843); the generalized Naibbe (Greshko 2026) is addressed separately in Section **??** of the main paper and eliminated by the cross-boundary MI and frequency–connectivity diagnostics rather than by entropy shift alone.

Failed operationalization (Phases 20–23): Table construction, paleographic comparison, first-syllable extraction, statistical inversion all failed.

Signal isolation (Phases 28–36): 16.5% genuine signal identified. Sequential structure confirmed ($z = 14.78$ conservative minimum). 10K dictionary reveals 51 signal words and 15.0% net signal.

Macaronic identification (Phase 37): Italian selectivity $5.45\times$ vs Latin $1.30\times$. 22 Italian-only signal words.

Z-score audit (Phases 39–42): $z = 319.76$ for a Venetian dictionary was a measurement artifact; corrected $z = -0.47$. All bigram z-scores inflated 3–70$\times$; six of seven retain $z > 2.0$ under symmetric recomputation.

Optimization landscape (Phases 43–46): MaxSAT: 500+ solutions within 1%. Table confirmed (composite 0.985). 89 recipes extracted. Page order consistent.

External methods (Phases 49–53): Language model + edit-distance-1 expansion invalidated (selectivity $1.10\times$). Size-matched language ID: Italian #1. CC bigram $z = 21.0$. Bridge search yields 22 word identifications. Paradigm constraints: $z = 0.02$ (null-failing). Variable-length encoding confirmed.

Entropy shift generalization and Currier prediction (Phase 55): Extended ranking to 13 mechanisms. Cardan grille scores $+0.49$–0.59 (cleanly eliminated). Currier cross-boundary prediction confirmed: tachygraphic simulation $1.284\times$ vs observed $1.450\times$ ($z = 24.9$); Schinner produces only $1.044\times$ (indistinguishable from null).

Costamagna compatibility (Phase 56): 10/10 structural questions match. All 21 T_{P15} syllable values attested; 5 modifier stroke types correspond to 5 documented coda consonants; 3 ambiguous triples = 3 shared-sign pairs.

CVC coda analysis (Phases 57–75): Modifier characters reinterpreted as Costamagna's coda consonant markers (hook$\rightarrow$*n*, sigmoid$\rightarrow$*s*, vertical$\rightarrow$*t*). Connector and descender revised to null (non-phonetic). Recalibrated coherence $p = 0.006$. Bootstrap grammar validation $p < 0.002$.

Computational investigations (Phases 62–76): Visual sign comparison via 7 methods finds directional support but domain gap prevents definitive validation. Word boundary segmentation fails; EVA token boundaries are structurally essential. 63% of corpus tokens decode with zero error. 340 extracted recipes.

Self-citation elimination (Phase 77): Timm–Schinner self-citation algorithm tested across 540 corpora. Entropy shift cosine -0.153 (anticorrelated). Cross-boundary MI $1.036\times$ (null level). Both tests eliminate it.

CVC T1 permutation validation (Phase 78): 1,000 random CV tables; real 331 IDs vs null 209.6±32.0 ($z = 3.79$, $p = 0.002$).

Known-properties stress test (Phase 79): 7 sub-tests against Voynich litmus tests; 3 explained, 3 partial, 1 limitation.

Wildcard consistency check (Phase 80): 301/316 identifications fully decoded; 15 with wildcards show 0/5 triple agreement, 8 conflicts.

Exception DOF audit (Phase 81): 29 DOF vs 328 constraints; model over-determined.

Decode traces (Phase 82): Complete pipeline traces for *chedy*, *daiin*, *qokeedy*.

Cross-language comparison (Phase 83): German 15 signal words vs Latin 25; coherence (not count) discriminates.

Syllabary coverage (Phase 84): 14.4% actual coverage of Latin text with 21 confirmed values; 37% theoretical maximum for 21 optimally-chosen syllables; 47% projected ceiling with unresolved triples and CVC codas.

German-optimized table (Phase 85): T_{PG} produces 20 signal words vs. T_{P15}'s 28; 0/3 coherence vs. 3/3; language identification is framework-independent.

Self-similar words (Phase 86): Corpus 10.25% vs. Latin reference 9.69%; 99.7% consecutive-character artifacts; cited examples (*dydydy*, *olol*) are low-frequency or decode to attested forms.

Entropy floor simulation (Phase 87): Basic CV encoder produces $H_6 = 0.619$ vs. observed 0.978; 39.4% gap closure; shape cosine 0.634. Residual attributable to allographs, compounds, modifiers.

S2 Dialect Identification Battery

Eight experiments tested the decoded vocabulary against five northern Italian dialects. Fisher's combined $p = 0.019$ across all 8 experiments. The contradiction is linguistically coherent: morphological experiments (3, 6, 7) favor Tuscan; phonological experiments (1, 2) favor Gallo-Italic.

Experiment 1 (Degemination): 6 of 8 words with geminate Latin etyma show simplified consonants (*bela* < *bella*, *sene* < *senna*). Winner: Emilian ($z = -2.1$, weight 0.5).

Experiment 2 (Lenition): Mixed voicing patterns. Winner: Ligurian ($z = 0.7$, weight 1.0).

Experiment 3 (Articles/pronouns): *ci*, *si*, *tu* match Tuscan standard forms. Winner: Tuscan ($z = 3.7$, weight 1.0). Strongest single experiment (4/4 gates, $3.27\times$ selectivity).

Experiment 4 (Pharmaceutical regionalism): *diasene* found but insufficient for regional localization. Winner: Venetian ($z = 0.2$, weight 0.0). Failed selectivity gate.

Experiment 5 ("co" syntax): Inconclusive ($z = 0.7$, weight 0.0).

Experiment 6 (Verb morphology): *dice*, *dico* match Tuscan verb forms. Winner: Tuscan ($z = -0.8$, weight 1.0).

Experiment 7 (Simulation): *di* frequency best matches Tuscan distribution ($z = 3.2$, weight 1.0).

Experiment 8 (Zodiac): No dialectal signal in zodiac sections. Winner: Lombard ($z = -1.0$, weight 0.0).

Composite scores: Ligurian 0.248, Lombard 0.235, Venetian 0.213, Tuscan 0.192, Emilian 0.112 (all 95% CIs overlapping). Verdict: DIALECT_INDETERMINATE (agreement 40%, below 60% threshold).

Table S1: Complete decode traces for Table 6 identifications. The Pattern column shows confirmed characters as-is and unresolved characters as “?, reflecting the wildcard pattern submitted to the dictionary.

EVA	Chars	Roles	Decoded CVC	Pattern	Matched
`otol`	o, t, ol	S, S, S	*ratene*	`ra??ne`	*ratione*
`oty`	o, t, y	S, S, C	*rate*	`ra??`	*rabidi*
`qopchedy`	qo, p, ch, e, dy	S, S, S, S, C	*togacora*	`????cora`	*stercora*
`ytol`	y, t, ol	S, S, S	*ditene*	`di??ne`	*diasene*
`chotar`	ch, o, t, ar	S, S, S, C	*corates*	`cora??s`	*coralli*
`chkain`	ch, k, a, i, n	S, S, S, S, C	*codeladin*	`code??din`	*codex*
`otcham`	o, t, ch, a, m	S, S, S, S, C	*ratecolat*	`ra??co??t`	*radicom*
`chtol`	ch, t, ol	S, S, S	*cotene*	`co??ne`	*commune*
`shty`	sh, t, y	S, S, C	*sete*	`se??`	*secundi*

S3 Complete Decode Traces

Table S1 shows the complete EVA → decoded pipeline for all word-level identifications referenced in the main paper Table 6.

Table S4: Sensitivity to *rabidi* removal.

Metric	With	Without
Catalog entries	22	17
Distinct words	9	8
Token coverage	32.0%	31.6%
Verdict	ROBUST	

Table S2: Passage decode: f54r (first 35 of 104 tokens).

#	EVA	Decoded CVC	Roles
1	podaiin	*garadin*	SSSC
2	shodal	*seradit*	SSSC
3	qopchol	*togacone*	SSSS
4	cfheody	*corara*	SSSC
5	opcheol	*ragacorane*	SSSSS
6	chocphy	*coraco*	SSSC
7	ytodaiin	*diteradin*	SSSSC
8	otchey	*rateco*	SSSC
9	otchey	*rateco*	SSSC
10	qockhodal	*toradit*	SCSSC
11	sockhody	*serara*	SSCSC
12	sar	*ses*	SC
13	daiir	*diladidine*	SSSSS
14	cthody	*cora*	SSC
15	otal	*ratet*	SSC
16	ekchody	*radecora*	SSSSC
17	ol	*ne*	S
18	s	*se*	S
19	or	*ne*	S
20	y	*di*	S
21	ytchey	*diteco*	SSSC
22	dam	*dilat*	SSC
23	tor	*tes*	SC
24	ockhol	*rane*	SCS
25	shokchy	*seradeco*	SSSSC
26	kolchom	*denecorat*	SSSSC
27	s	*se*	S
28	chey	*co*	SC
29	ctheotol	*coraratene*	SSSSS
30	sar	*ses*	SC
31	sh	*se*	S
32	okeodain	*raderaradiladin*	SSSSSSSC
33	chokey	*corade*	SSSC
34	korare	*dessra*	SCCS
35	ckhos	*coras*	SSC
	… (69 more tokens)		

Table S5: Signal words for Latin_10K ($\sigma > 2.0$).

Word	Real	σ	Sel.
du	183	44.5	7.3×
ni	360	44.0	6.2×
cos	378	31.6	3.2×
ne	1478	30.6	2.8×
dis	363	25.7	2.8×
bet	266	25.2	3.2×
hi	11	24.5	9.2×
den	46	21.4	2.4×
se	652	19.6	1.4×
con	74	18.1	4.9×
do	9	17.1	15.0×
codi	252	16.2	2.3×
tot	116	14.7	2.0×
bene	143	14.5	2.5×
nes	208	13.1	1.9×
sene	247	10.3	2.1×
des	63	10.2	3.6×
ton	29	10.1	4.8×
fa	10	8.2	6.2×
ten	37	7.8	3.2×
net	79	7.2	1.5×
mi	3	5.3	7.5×
sera	735	4.5	1.2×
teras	15	4.5	2.1×
ha	11	2.6	1.5×

Table S3: Passage decode: f57v (first 35 of 175 tokens).

#	EVA	Decoded CVC	Roles
1	`dairal`	*diladinet*	SSSSC
2	`v`	*hi*	S
3	`saly`	*setdi*	SCS
4	`soeos`	*serararas*	SSSSC
5	`vs`	*his*	SC
6	`ar`	*ne*	S
7	`okees`	*raderaras*	SSSSC
8	`o`	*ra*	S
9	`d`	*di*	S
10	`soefchees`	*seraratecoraras*	SSSSSSSC
11	`lg`	*nedo*	SS
12	`sos`	*seras*	SSC
13	`okey`	*rade*	SSC
14	`defo`	*diratera*	SSSS
15	`f`	*te*	S
16	`o`	*ra*	S
17	`rkedam`	*nederadilat*	SSSSSC
18	`sh`	*se*	S
19	`ofol`	*ratene*	SSS
20	`sar`	*ses*	SC
21	`ddal`	*didit*	SSC
22	`yty`	*dite*	SSC
23	`s`	*se*	S
24	`y`	*di*	S
25	`daiir`	*diladidine*	SSSSS
26	`otey`	*rate*	SSC
27	`dshdy`	*di*	SCC
28	`dkals`	*didetse*	SSCS
29	`otypchchy`	*rategacoco*	SSCSSSC
30	`a`	*la*	S
31	`r`	*ne*	S
32	`opaiin`	*ragan*	SSC
33	`dal`	*dit*	SC
34	`karody`	*desra*	SCSC
35	`vrokeey`	*hisradera*	SCSSSC
	… (140 more tokens)		

Table S6: Signal words for Italian_10K ($\sigma > 2.0$).

Word	Real	σ	Sel.
ne	1478	30.6	2.8×
se	652	19.6	1.4×
ben	345	18.9	4.4×
con	74	18.1	4.9×
do	9	17.1	15.0×
bene	143	14.5	2.5×
fa	10	8.2	6.2×
ha	11	2.6	1.5×

Table S7: Signal words for German_10K ($\sigma > 2.0$).

Word	Real	σ	Sel.
bes	267	65.9	4.4×
ne	1478	30.6	2.8×
den	46	21.4	2.4×
sen	174	19.8	3.7×
se	652	19.6	1.4×
ben	345	18.9	4.4×
nen	229	16.6	3.4×
bene	143	14.5	2.5×
nes	208	13.1	1.9×
des	63	10.2	3.6×
ten	37	7.8	3.2×
set	74	7.2	2.0×
net	79	7.2	1.5×
tes	55	7.1	2.9×
ha	11	2.6	1.5×

S4 Extended Passage Decodes

Tables S2 and S3 show token-by-token decodes for the two example passages. Roles: S = syllabic, C = coda marker.

S5 Wildcard Conflict List

The 8 cross-identification wildcard conflicts and the per-triple consistency summary have been moved to Appendix H of the main paper (Tables 15–16).

S6 Complete Signal Vocabulary

The 70 unique CV signal words comprise 51 Latin-10K words plus 22 Italian-only words minus 3 overlaps (*dise*, *cu*, *dedi*). Of these, 42 are among the 56 permutation-validated signal words (•); 29 did not clear $\sigma > 2.0$ under the permutation test's independently generated null corpora (○). The 15 permutation-validated words not in the CV tables emerged only under the merged-dictionary (19K) methodology, where the combined Latin-Italian dictionary produced different null-corpus hit patterns. These 15 words include: *dicu*, *diga*, *corali*, *cora*, *bela*, *dice*, *cose*, *nera*, *co*, *la*, *ha*, *te*, *ra*, *sera*, *dico*.

The full CV signal tables (Tables 8–9 in the main paper appendix) and the CVC signal vocabulary (76 words) are documented in the main paper and at `https://github.com/mruckman1/voynich_2` in `docs/signal-vocabulary.md`.

S7 Sensitivity Analyses

S7.1 Rabidi Removal

Five of 22 word-level identifications map to *rabidi* ("of the fierce/raging"), appearing across 60 folios. Removing all 5 entries has negligible impact: corpus coverage drops by 0.4 percentage points, *Circa Instans* overlap drops from 88.9% to 87.5%, and all 20 morphological paradigms are preserved. The remaining 8 distinct Latin words form a coherent pharmaceutical vocabulary independent of *rabidi*. Verdict: ROBUST.

S7.2 Tachygraphic Grid Search

The entropy shift discriminator was tested across 24 configurations (grid search over consonant classes and vowel variants). The preferred region is 5 consonant classes $\times$ 4 vowel variants, independently confirmed by the Costamagna catalog. The discriminator's sensitivity to parameter choice was assessed by computing cosine similarity across all 24 configurations; the preferred region produces cosine ≥ 0.80 while all other regions produce cosine < 0.60. The 24-configuration search is not optimization against the Voynich but exploration of the structural parameter space of tachygraphic systems.

S8 Costamagna Plate Specimen

A digitized specimen of Costamagna's 1953 plates (*Il sistema tachigrafico sillabico usato dai notai medioevali italiani*) is permanently available at:

`https://mattruckman.com/papers/costamagna-1953-specimen/`

The plates are also available in the Approach 2 repository at `data/GL.S.III.MISC.12/`, which includes 40 high-resolution plate photographs from the Biblioteca Marucelliana (Florence), 239 individual sign crops with syllable-value metadata, and a structured catalog (`costamagna_1953_catalog.json`).

S9 Cross-Language Signal Details

The cross-language signal comparison (Phase 83) tested the same T_{P15} assignment table against Latin, Italian, German, and Hebrew 10K-word dictionaries. The full per-word signal lists are in the tables generated from this section's data. Hebrew produced 0 signal words (selectivity $0.74\times$, below null). The 9 words shared between Latin and German are: *bene*, *den*, *des*, *ha*, *ne*, *nes*, *net*, *se*, *ten*. All are short functional words present in both language families.

S10 Entropy Floor Simulation (Phase 87)

To test whether the tachygraphic mechanism quantitatively predicts the Voynich's distinctive H_6 elevation, we constructed a basic CV encoder: Latin plaintext (subsampled to Voynich text length) was converted to syllables and each syllable mapped to a two-character code via a randomly-seeded 25-triple assignment, without allographic variation, compound signs, or modifier-based coda encoding. Entropy at context orders H_0–H_6 was computed on the encoded output and compared to the Voynich and to natural Latin. Twenty instantiations (varying seed) were run to estimate variance.

Results: the basic CV simulation produces $H_6 = 0.619$ bits across all 20 seeds (variance is zero because the encoder is deterministic given the input text; the seed controls only the random mapping of syllables to output characters, which does not affect entropy), above Latin's 0.386 and closing 39.4% of the gap to the Voynich's 0.978. The entropy shift shape cosine across all seven context orders is 0.634. The simulation correctly predicts the direction of entropy elevation (encoded output has higher H_6 than plaintext) and substantial magnitude, but does not reach the Voynich's observed level.

The 0.359-bit residual is attributable to encoding features omitted from the basic simulation. T_{P15} includes allographic variation (three minim characters mapping to *mi*, reflecting stroke-count variation), compound signs (*qo* treated as a single unit), and modifier-based CVC coda formation. The full parameterized model used for the entropy shift discriminator (Figure 1, cosine $= 0.820$) includes these features and achieves a substantially better fit. A complete quantitative decomposition of the residual—showing how much each feature contributes to closing the remaining gap—is left to future work.

S11 Self-Similar Words Analysis (Phase 86)

Words like *dydydy*, *olol*, and *oror* contain internal character repetitions that might appear inconsistent with a syllabic encoding. We quantified this property across the full corpus under a self-similarity metric defined as: a word is self-similar if any character or character sequence of length ≥ 2 repeats consecutively within the word.

Corpus-wide self-similarity rate: 10.25% of tokens (1,665 types, 3,716 tokens). Decomposition by repeat type: 99.7% consecutive-character artifacts (doubled characters such as *ee*, *dd*, *ii*); 0.3% short-sequence repeats of length 2–3 (e.g., *olol*, *aral*); 0.0% triple-or-longer repeats of any sequence.

Reference comparison: a Latin reference corpus (subsampled to the same token count, same metric) shows 9.69% self-similarity. A tachygraphic simulation produces 9.75%. The Voynich's rate is not anomalous relative to natural-language baselines.

Specific examples: *dydydy* appears 2 times in the corpus; it is not a representative structure. *olol* appears 15 times and decodes under T_{P15} to *nene* (an attested form in both Latin and Italian). *oror* appears 8 times and decodes identically to *nene*, confirming allographic equivalence between *olol* and *oror*.

Verdict: the self-similarity property, quantitatively examined, does not distinguish the Voynich from natural-language baselines.

S12 German-Optimized Assignment Table (Phase 85)

The cross-language comparison in Section 12.1 tests a fixed assignment table (T_{P15}) against multiple dictionaries. A stronger test fixes the framework and re-optimizes the assignment against a different language. We constructed T_{PG} by applying the same 15-iteration optimization procedure used for T_{P15} to a German medical dictionary (*Buch der Natur*, ~890K characters). The 25 stroke triples, 5-onset $\times$ 6-nucleus grid structure, modifier classification, and all other elements of the stroke-feature framework were held constant; only the syllable values assigned to each triple were re-optimized against German.

T_{PG} results: 22.4% raw dictionary hit rate against a 10K German medical dictionary; 20 signal words at $\sigma > 2.0$ under the signal isolation pipeline; coherence tests fail on 0/3 criteria (no German verb paradigm emerges, no pharmaceutical register equivalent, no Romance function-word kit—the last fails by construction since the kit is Romance-specific).

T_{P15} results under identical pipeline parameters: 28 signal words; coherence 3/3.

Cross-comparison table:

Table × Dict	Signal	Sel.	Coh.
T_{P15} × Latin 10K	28	1.36×	PASS
T_{P15} × German 10K	29	1.50×	FAIL
T_{PG} × German 10K	20	1.57×	FAIL
T_{PG} × Latin 10K	11	1.65×	—

Interpretation: equivalent optimization effort against German produces fewer signal words (20 vs. 28) and categorically worse coherence (0/3 vs. 3/3). The stroke-feature framework is derived from the Voynich's internal visual structure and is language-agnostic. The result that Latin-Italian produces genuine signal where German does not reflects a property of the manuscript, not of the optimization target.

S13 Syllabary Coverage Analysis (Phase 84)

The 21 confirmed CV syllable values in T_{P15} cover 14.4% of a reference Latin corpus. We contextualize this figure against theoretical maxima and historical syllabaries.

Theoretical maximum for 21 well-chosen syllables. Latin syllable frequency is Zipfian. The 21 most frequent CV syllables in the reference corpus cover 37.0% of text. Our 14.4% is below this theoretical maximum, indicating the confirmed values are not purely the highest-frequency Latin syllables, but are within the same order of magnitude.

Projected ceiling with unresolved triples. If the 13 unresolved stroke triples each contribute a new CV syllable value, and the CVC coda system (currently 3 phonetic codas $\times$ 21 base syllables $= 84$ CVC types) is included, the model's theoretical coverage ceiling is approximately 47% of Latin text.

Historical syllabary comparison.

System	Signs	Top-21	Full	Ambig.
Linear B	87	72%	95%	33%
Cypriot	55	75%	98%	27%
Hiragana	46	82%	100%	0%
Costamagna	228	65%	92%	20%
Voynich (confirmed)	21	14%	14%	0%
Voynich (projected)	34	14%	47%	0%

Linear B achieves approximately 72% coverage of Greek text with ∼87 signs. The gap between 47% (our ceiling) and 72% (Linear B) reflects three factors: (a) Linear B was historically optimized for Greek, while our confirmed values emerge from computational convergence on Voynich internal evidence; (b) Linear B includes signs for common syllables that may not have direct analogs in our confirmed Voynich glyph set; (c) the 13 unresolved triples may not all resolve to useful new CV values—some may be variant forms of existing confirmed triples, reducing the effective inventory.

Verdict: the inventory limitation is real and represents the primary barrier to connected readable text. A naive $21/90 = 23\%$ calculation overstates the gap; Zipfian frequency distribution and the CVC coda system together bring the projected ceiling to approximately 47%. Closing the remaining gap to historical-syllabary levels requires external evidence constraining the unresolved triples or computational methods not yet developed.