A Statistical Evaluation of the Hebrew Cipher Hypothesis for the Voynich Manuscript

Antenore Gatta
antenore@simbiosi.org

February 2026

Abstract

We present a systematic computational evaluation of the hypothesis that the Voynich Manuscript (Beinecke MS 408) encodes Hebrew consonantal text via the European Voynich Alphabet (EVA) transcription. Starting from a monoalphabetic substitution mapping of 19 EVA characters to 19 of the 22 Hebrew consonants — derived through frequency analysis, allograph detection, digraph resolution, and positional splitting — we decode the full manuscript corpus (37,025 tokens, 7,861 types) and evaluate the output against Hebrew lexicons at three tiers of coverage (6.5K biblical, 45K curated, 491K corpus-attested forms).

The mapping produces a statistically significant signal: the decoded text matches Hebrew lexical forms at rates 1.7–3.4 times above random baselines ($z = 3.6$–4.4, $p < 0.005$) across all lexicon tiers, survives permutation testing for botanical anchors ($p = 0.017$) and domain vocabulary ($p = 0.004$), and outperforms a synthetic null model on match rate, Zipfian gloss distribution, and Hebrew bigram plausibility.

However, the decoded text does not produce readable Hebrew. Best passages yield incoherent word sequences. The signal concentrates in one of five identified scribal hands (Hand 1, 86 herbal pages) and in paragraph text rather than figure labels (24.7% vs 13.2%, $z = -8.10$). Alternative hypotheses (homophonic Naibbe cipher, Judeo-Italian substrate, Currier language split) are tested and largely rejected. Independent validation by a Hebrew-specialized language model (DictaLM) confirms 19.4% of matched types as genuine Hebrew words, with token-weighted confirmation at 56.6%; the signal concentrates in approximately 200 high-frequency forms that produce a real-to-random ratio of 118×. We conclude that the mapping captures genuine structural correspondence between EVA text and Hebrew consonantal morphology, but falls short of decipherment. The nature of this correspondence — partial cipher, structural mimicry, or coincidental phonotactic alignment — remains an open question.

Keywords: Voynich Manuscript, Hebrew cipher, computational cryptanalysis, monoalphabetic substitution, statistical linguistics

1 Introduction

The Voynich Manuscript (Yale, Beinecke Rare Book and Manuscript Library, MS 408) is an illustrated codex of approximately 240 pages, written in an undeciphered script and dated by radiocarbon to the early 15th century (1404–1438 CE). The script, conventionally transcribed using the European Voynich Alphabet (EVA) system (Landini, 2001), employs approximately 20 distinct characters and produces a corpus of roughly 37,000 word tokens and 8,000 types. The manuscript's sections — herbal, astronomical, zodiac, balneological, pharmaceutical, and text — suggest a practical or encyclopedic work.

Proposed decipherments range from natural languages (Latin, Italian, Hebrew, Nahuatl) to constructed or meaningless text (Rugg, 2004). Statistical analyses have consistently shown that the text exhibits word-frequency distributions, entropy values, and clustering properties

compatible with natural language (Montemurro and Zanette, 2013; Amancio et al., 2013), though these properties can also arise from certain non-linguistic generative mechanisms (Timm and Schinner, 2014).

Recent work by Davis (2020) identified five distinct scribal hands in the manuscript, correlating with Currier's earlier observation of two statistical "languages" (A and B) (Currier, 1976). Greshko (2025) demonstrated that a verbose homophonic cipher using historically plausible 15th-century materials (dice and playing cards) can produce ciphertext with statistical properties matching those of the Voynich text.

In this paper, we evaluate the hypothesis that EVA characters map to Hebrew consonants via monoalphabetic substitution, read right-to-left. This hypothesis draws on the observation that Hebrew consonantal writing — consisting of 22 consonant letters with vowels typically unwritten — produces short, high-entropy word forms structurally similar to EVA tokens. Rather than claiming a decipherment, we frame the evaluation as a statistical test: *does the proposed mapping produce output that matches Hebrew lexical forms at rates significantly above what random or structural baselines would predict?*

We organize 35 systematic investigations across mapping derivation (§2.2), multi-tier lexicon validation (§2.3), permutation testing (§3.3), alternative hypothesis testing (§3.5), scribal variation (§3.6), text structure (§3.7), and a meta-analysis comparing our results against 15 published studies (§5).

2 Data and Methods

2.1 Corpus

The EVA transcription used is the Takahashi (H) layer from the `LSI_ivtff_0d.txt` file (Landini, 2001). The IVTFF format encodes page structure, line numbers, scribal hand assignments, Currier language labels (A/B), illustration-based section codes (H=herbal, S=astronomical, Z=zodiac, B=balneological, P=pharmaceutical, T=text, C=cosmological), and — critically for our analysis — layout type via unit codes: paragraph (`P0`, `P1`...), label (`Lz`, `L0`...), circular (`Cc`), and title (`Pt`).

After parsing, the corpus contains 37,025 word tokens (7,861 types) across 225 pages, distributed as 33,684 paragraph words, 1,022 label words, 2,286 circular-text words, and 22 title words. The zodiac section (Z) contains zero paragraph text — all 1,322 words appear as labels under figures or in circular bands.

2.2 Mapping Derivation

The EVA→Hebrew mapping was derived through a multi-phase process:

1. **Frequency alignment** (Phase 5–6): initial assignment of EVA characters to Hebrew consonants via ranked frequency matching, validated independently through Italian phonemic transliteration.[1]

2. **Allograph detection** (Phase 7–9): positional profile analysis identified functional equivalences:

   - `f` = `p` (both → lamed): cosine similarity 0.987, context overlap 0.67
   - `i` (standalone) = `d` (both → resh): `ii` → he, standalone `i` → resh

[1]The Hebrew and Italian optimization paths converged on 84.2% (16/19) of characters. A convergence control using a simplified steepest-ascent hill-climber (10 restarts) on shuffled text and random strings yields 0–12% convergence (0–2/17 characters agreeing with the published mapping), consistent with the chance expectation ($1/22 \approx 4.5\%$ per character). The 84.2% agreement between independent derivation paths greatly exceeds any control.

- `k/t` confirmed equivalent (cosine 0.999) but mapped to distinct letters (tav/tet)

3. **Digraph resolution** (Phase 9): `ch` identified as a single cipher unit mapping to kaf (cohesion $P(h|c) = 82.7\%$, +72 lexicon matches).

4. **Positional splitting** (Phase 9): two position-dependent rules recovered two additional Hebrew letters:

   - EVA `n` at Hebrew word-initial position → bet (elsewhere → dalet): +2,162 net lexicon matches
   - EVA `r/ii` at Hebrew word-initial position → samekh (elsewhere → he): +563 net matches

   These splits were selected from 36 explicitly tested hypotheses (6 anomalous position–letter pairs × 5 unmapped Hebrew letters, plus 6 allographic variants). The full search space comprises $19 \times 3 \times 5 = 285$ possible splits. Both adopted splits survive Bonferroni correction at $\alpha/285 = 1.75 \times 10^{-4}$: `n`→bet $z = 40.1$, $p_{\text{corrected}} < 10^{-100}$; `r`→samekh $z = 19.5$, $p_{\text{corrected}} < 10^{-80}$ (sign test on gained vs. lost matches).[2]

The resulting mapping covers 19 of the 22 Hebrew consonants. The three unmapped letters (zayin, tsade, qof) were exhaustively probed via positional splits, digraph candidates (`sh`), and constrained optimization; all strategies yielded non-significant gains. Table 1 presents the complete mapping.

Table 1: Complete EVA → Hebrew mapping (19/22 consonants). Direction: RTL. Preprocessing: `ch`→kaf, `q`-prefix stripped, `ii`→he, standalone `i`→resh. Post-decode: dalet at position 0 → bet; he at position 0 → samekh.

EVA	Hebrew	Name	Type	Notes
`a`	y	yod	convergent	
`c`	A	aleph	convergent	
`ch`	k	kaf	digraph	Phase 9 B3
`d`	r	resh	convergent	
`e`	p	pe	convergent	
`f`	l	lamed	allograph	= `p`
`g`	X	chet	convergent	
`h`	E	ayin	convergent	
`ii`	h	he	composite	
`i`	r	resh	composite	allograph of `d`
`k`	t	tav	convergent	
`l`	m	mem	convergent	
`m`	g	gimel	convergent	
`n`	d/b	dalet/bet	positional	bet at initial
`o`	w	vav	convergent	
`p`	l	lamed	convergent	
`q`	—	prefix	stripped	
`r`	h/s	he/samekh	positional	samekh at initial
`s`	n	nun	convergent	
`t`	J	tet	convergent	
`y`	S	shin	convergent	

The **reading direction** was confirmed as right-to-left via permutation testing (500 permutations per direction): RTL achieves 41.65% match rate vs LTR 33.14% (+8.51 pp, $z_{\text{direct}} = 22.97$).

[2]The sign-test z-scores treat each gained or lost match as an independent Bernoulli trial under H_0: $P(\text{gain}) = P(\text{loss}) = 0.5$. Bonferroni correction for the full 285-hypothesis search space leaves significance unchanged.

Both directions produce significant permutation z-scores (RTL $z = 4.17$, LTR $z = 2.44$), but RTL is strongly preferred across both Currier languages (A: +13.07 pp; B: +6.87 pp).

The **mapping optimality** was verified via constrained letter audit (Phase 10, 15): for each EVA character, all 22 Hebrew alternatives were tested while enforcing one-to-one constraints. With the curated lexicon (45,713 forms, excluding corpus-attested forms), 14 of 17 base letters are optimally assigned. The three non-optimal letters show marginal gains ($< 2\%$ token improvement) that fail permutation significance tests.

Cross-validation

To address the concern that the mapping was derived and validated on the same data, we performed two cross-validation experiments using the honest lexicon (45K forms).

Hand-based split. Hand 1 pages (6,832 tokens) serve as the training set; all other hands (30,193 tokens) form the test set. We ran the constrained per-letter audit independently on each subset. Full held-out derivation (re-running the hill-climbing optimizer on each subset) would require re-implementing the historical 10-group homophonic architecture; instead, we test whether the existing mapping is *independently optimal* on each data partition. Results: 15/17 letters optimal on training, 13/17 on test, with 13/17 independently optimal on both subsets. The four non-shared letters (`c`, `g`, `m`, `t`) show small gaps ($\leq$137 tokens). The permutation test ($n = 200$) on the held-out (non-Hand 1) subset yields $z = 5.72$ ($p = 0.005$).

Random 50/50 split. We randomly partitioned pages into two halves (10 repetitions, different random seeds). For each split we audited mapping optimality on the training half and computed a permutation z-score on the test half. All 10 splits are significant: mean $z = 5.85 \pm 0.32$ (range 5.36–6.36), mean optimality 13.5/17 (range 13–14), test match rate $25.8\% \pm 0.7\%$.

A mapping that achieves $z => 5$ on held-out data across all splits, with 13/17 letters independently optimal on disjoint subsets, cannot be dismissed as an artefact of overfitting to one corpus partition.

2.3 Hebrew Lexicon

A key methodological concern is lexicon size inflation. We constructed lexicons at three coverage tiers to control for this effect:

Table 2: Hebrew lexicon tiers. The "honest" tier (45K) excludes corpus-attested forms from Sefaria, which contribute high match rates for both real and random mappings.

Tier	Forms	Sources
Biblical (STEPBible)	5,475	Biblical Hebrew headwords
Curated ("honest")	45,713	STEPBible + Jastrow + Klein + curated
Full (corpus)	491,148	+ Sefaria corpus (freq$\geq$5)

The Sefaria corpus contributes 445,458 attested forms from a 250-million-token corpus of Hebrew texts. While these forms are genuine Hebrew, the sheer coverage means that random consonant strings of typical Voynich word lengths match at 26.8% ("monkey baseline"). We therefore report all key results at the *honest* tier and note full-lexicon results parenthetically.

3 Results

3.1 Match Rate and Lexicon Calibration

Table 3 presents the core result: match rates of the decoded corpus against each lexicon tier, compared to random-mapping and random-string baselines.

Several patterns emerge:

Table 3: Reality test: decoded Voynich text vs. baselines across lexicon tiers. "Random mapping" = 1000 random EVA→Hebrew permutations. "Random strings" = consonant strings matching observed length distribution, drawn with Hebrew character frequencies.

Lexicon tier	Real	Rnd. map	z	Rnd. str.	Real/Str.
STEPBible (6.5K)	17.2%	9.0%	3.6	∼5%	3.4×
Honest (45K)	25.6%	∼12%	∼4.0	∼11%	2.3×
Full (491K)	45.7%	29.9%	3.7	26.8%	1.7×

1. The z-score remains in the range 3.6–4.0 across all lexicon tiers, indicating a **fixed differential** of approximately 8 percentage points above random mappings. If the mapping were a correct cipher key, additional legitimate lexical forms should widen this differential; the plateau suggests partial or coincidental correspondence.

2. The real-to-random-string ratio *decreases* with lexicon size (3.4× → 1.7×), demonstrating that large lexicons inflate both sides. The honest tier provides the most informative signal.

3. Even the smallest lexicon (STEPBible, 5,475 biblical forms) yields $z = 3.6$, ruling out the possibility that the signal is purely an artifact of lexicon size.

3.2 Null Model Test

To determine whether the decoded text contains genuine Hebrew-like structure beyond simple character matching, we compared the real decoded corpus against 500 synthetic corpora generated by random character substitution preserving word lengths and overall Hebrew character frequencies (Table 4).

Table 4: Null model test (honest lexicon, 33,320 tokens, 500 iterations).

Test	Real	Synthetic	z	Verdict
Match rate	21.48%[3]	8.91%±0.13%	98.2	significant
Gloss entropy	5.56	10.16±0.04	121.1	significant
Top-5 concentration	33.3%	1.8%±0.2%	178.1	significant
Bigram plausibility	−2.970	−3.095±0.003	40.9	significant

The real decoded text matches the lexicon at 2.41× the synthetic rate. Crucially, the matched words show Zipfian concentration: the top 5 decoded types account for 33.3% of all matches (vs. 1.8% for synthetics), and the gloss entropy is far lower (5.56 vs. 10.16), indicating that the mapping produces a small set of high-frequency forms rather than uniformly distributed matches. Hebrew bigram transition probabilities are also significantly better fit ($z = 40.9$).

This establishes that the signal is **structurally embedded in character sequences**, not merely an artifact of character frequency overlap. However, it should be noted that the null model is deliberately minimal: it preserves only marginal character frequencies and word lengths, discarding all sequential structure (bigram, trigram, and positional patterns). Any systematically derived mapping would preserve the EVA text's sequential structure when projected onto the Hebrew alphabet, and would therefore outperform this baseline. The null model thus provides a *floor* (the decoded text is better than random character sequences), not a ceiling. The permutation tests in §3.3, which compare against random bijective mappings preserving all source-text structure, provide the more appropriate measure of statistical significance.

3.3 Permutation Tests

We tested the mapping against domain-specific vocabulary using 1,000 random EVA→Hebrew permutations per test:

Table 5: Permutation tests against domain vocabulary (1,000 permutations each).

Domain	Real matches	Mean random	z	p
Botanical (plant names)	6	1.38	3.2	0.017
Domain anchors (d≤1)	99	—	4.2	0.004
Zodiac vocabulary	2	0.44	0.9	0.071 (ns)
Semantic coherence (max consec.)	10	6.05	4.35	0.002
Semantic coherence (high lines)	1,445	253	14.54	0.001

Botanical and domain-anchor tests are significant after FDR correction; zodiac is not (only 1 exact match out of 3 expected). Semantic coherence — measured as the tendency for lexicon-matched words to cluster in consecutive positions — is highly significant, confirming that matches are **non-randomly distributed** within the text.

However, manual inspection of the highest-scoring passages reveals that the glosses do not form coherent sentences. A typical "100% semantic" passage in a herbal section reads: "six – die – poor – poor – back." The statistical clustering reflects repeated common forms, not semantic coherence.

3.4 Mapping Stability

The constrained letter audit (Table 6) shows that 14 of 17 base EVA characters are optimally assigned under the honest lexicon, and all three non-optimal letters have marginal gaps:

Table 6: Non-optimal letter assignments in constrained audit (honest lexicon, 45K forms). Only 3 of 17 letters show any improvement potential, all non-significant under permutation testing.

EVA	Current	Best alt.	Gap (tokens)	Gap (%)	Significance
`t`	tet	tav	+154	+0.9%	$z = 1.09$, $p = 0.133$ (ns)
`m`	gimel	tsade	+118	+1.2%	$z = 1.38$, $p = 0.094$ (ns)
`c`	aleph	he	+14	+0.1%	ns

The mapping is therefore **locally optimal**: no single-letter change produces a statistically significant improvement. The gimel→tsade swap was independently investigated with 1,000 permutations across both lexicon tiers and definitively rejected (honest $z = 1.38$; full $z = 0.22$).

3.5 Alternative Hypotheses

3.5.1 Naibbe Verbose Cipher

Greshko (2025) proposed that the Voynich text could be a verbose homophonic cipher. We tested this by simulating 200 Naibbe encryptions of Italian text, applying our Hebrew mapping to the output, and comparing match rates.

The Naibbe simulation produces 20.7% ± 1.6% match rate vs. real 40.3% ($z = 12.1$). Additionally, 8 of 9 diagnostic indicators — including index of coincidence (IC= 0.077, mono range 0.060–0.085), Gini coefficient (0.469, mono range 0.30–0.55), and conditional entropy ratio ($H_1/H_0 = 0.613$, mono range 0.50–0.75) — favor a monoalphabetic interpretation. Only the hapax ratio (0.71) falls in the homophonic range.

Verdict: the Naibbe verbose cipher hypothesis does not account for the observed Hebrew signal. The text's statistical properties are consistent with monoalphabetic substitution.

3.5.2 Judeo-Italian Substrate

We tested whether the Hebrew signal could arise from Italian text transliterated via documented Judeo-Italian conventions (tet for /t/, tsade for /c/ dolce, qof for /c/ dura, matres lectionis for vowels). After transliterating 60,738 Italian forms:

- Judeo-Italian strict match: 5.0% (1,674/33,412 tokens)
- Hebrew match: 40.2% (13,438 tokens)
- JI explains only 10.0% of Hebrew-matched types (101/1,009)
- Permutation test: $z = 4.59$, $p = 0.005$ — the JI signal is real

The Judeo-Italian signal is statistically significant but explains only a small fraction of the Hebrew correspondence. Notably, 46 word types (566 tokens) match Italian forms but *not* standard Hebrew, suggesting a possible Italian substrate. The three unmapped Hebrew letters (zayin, tsade, qof) are all used in JI transliteration conventions.

Verdict: Judeo-Italian is a plausible **partial** component but cannot be the primary explanation.

3.5.3 Judeo-Arabic Substrate

Given that Arabic and Hebrew share a large number of trilateral roots and that Judeo-Arabic (Arabic written in Hebrew characters) was widely used in medieval Jewish intellectual culture, we tested whether the decoded output could reflect Arabic rather than Hebrew. We transliterated 37,746 Arabic forms (AraMorph 1.2.1 stems excluding proper nouns, plus CAMeL Morph non-proper lemmas and roots) into Hebrew consonants via standard Judeo-Arabic orthographic conventions.

- Arabic match: 3.8% of types (296/7,752), 14.4% of tokens (4,801/33,412)
- 90% of Arabic matches (266/296 types) are also in the Hebrew lexicon — shared Semitic roots
- Only 30 types match Arabic but not Hebrew (491K); examples include *sryr* (*sarīr*, "bed," freq=103) and *Srwr* (*shurūr*, "malice")

The Arabic signal is weaker than Hebrew at every lexicon tier and is almost entirely explained by shared Semitic cognates rather than independent Arabic vocabulary.

Verdict: the decoded text is **not** Judeo-Arabic. The small Arabic-specific overlap does not add explanatory power beyond the Hebrew hypothesis.

3.5.4 Ladino (Judeo-Spanish)

Ladino — medieval Spanish written in Hebrew characters — was widely used by Sephardic communities in Italy after the 1492 expulsion. We transliterated 5,902 Ladino word forms to Hebrew consonants and tested against the decoded output: only 0.5% of types (42/7,752) and 3.0% of tokens (990/33,412) match. Of the 42 matches, 33 are also in the Hebrew lexicon (generic short forms like *syr* ← "ser"). Only 9 types are Ladino-specific, all at frequency ≤31 and semantically trivial.

Verdict: Ladino is **conclusively excluded**.

3.5.5 Currier Language Split

Both Currier languages produce significant permutation scores independently (A: $z = 4.02$; B: $z = 3.85$), confirming that the Hebrew signal is not concentrated in one "language." However, Language A achieves significantly higher match rates (full lexicon): 45.7% vs. 38.7% (+7.0 pp, $z = 11.64$, $p < 0.0001$). Cross-language testing confirms that Aramaic matches at only 0.2–0.4% in both languages, excluding a different Semitic substrate.

3.6 Scribal Variation

Following Davis (2020)'s identification of five scribal hands, we computed match rates per hand (Table 7).

Table 7: Match rates per scribal hand (honest lexicon, paragraph text only). Hands 3, 5, Y omitted (<30 pages).

Hand	Pages	Tokens	Honest %	Perm. z	Lang.
1	86	6,545	**28.8%**	3.79**	A
2	45	9,087	24.9%	3.64**	B
?	66	10,868	23.0%	4.24**	mixed
X	6	2,703	21.9%	3.89**	B
4	8	765	21.7%	n/a	**A**

The central finding is that the previously reported Currier A>B advantage (+7.0 pp) is entirely driven by **Hand 1**. Hand 4, which is also classified as Language A, achieves only 21.7% — *lower* than Hand 2 (Language B, 24.9%). In paragraph text overall, Hand 1 (28.8%) vs. Hand 2 (24.9%): +3.9 pp, $z = 3.12$, $p < 0.002$. A content-controlled comparison restricted to herbal pages (full lexicon) shows an even larger gap: Hand 1 48.6% vs. Hand 2 38.6% (+9.9 pp, $z = 8.18$, $p < 0.0001$).

A per-character frequency comparison (proportion z-test with Bonferroni correction for 19 characters, $\alpha_{\text{adj}} = 0.0026$) reveals the structural source of the gap. Hand 4 uses EVA `e` at roughly twice the rate of Hand 1 (11.1% vs. 5.3%), making it the single largest divergence. Because `e`→pe occurs disproportionately in non-matching contexts, this alone accounts for a substantial portion of the match-rate difference. Hand 4's small sample size (765 paragraph tokens) limits statistical power, with only 2–3 characters reaching Bonferroni significance. The full 19-character comparison is available in the supplementary data.

Hand 1 is the primary herbal scribe (86 pages, virtually all herbal section), responsible for 75% of Language A pages. The mapping appears to be **specifically tuned to Hand 1's orthographic conventions**, with diminishing performance on all other hands including the other Language A scribe.

3.7 Layout-Aware Analysis

The IVTFF transcription encodes three distinct text layout types: paragraph (continuous text, ∼8 words/line), label (captions under figures, 85% single-word), and circular (ring text around diagrams). Previous section-level analyses mixed these types.

Table 8: Match rates by layout type (honest lexicon).

Layout	Words	Decoded	Honest %	vs. Paragraph
Paragraph	33,684	32,852	**24.7%**	—
Circular	2,286	2,201	20.4%	$z = -4.48$***
Label	1,022	946	13.2%	$z = -8.10$***

Labels match at roughly **half** the paragraph rate (13.2% vs. 24.7%, $z = -8.10$, $p < 10^{-15}$). This is counter-intuitive if labels represent identifiable nouns (plant names, star names), which should be easier to match. The low label rate suggests that these words are proper names or technical terms absent from standard Hebrew lexicons.

Critically, the zodiac section (Z) contains **zero paragraph text** — its 1,322 words are entirely labels (367) and circular text (955). This explains its anomalously low match rate (12.2%) in section-level analyses. When restricted to paragraph text, section rates become substantially more uniform (18.5%–27.7%), with the zodiac dropping out entirely (Table 9).

Table 9: Section match rates: mixed vs. paragraph-only (honest lexicon). Zodiac (Z) has no paragraph text and is omitted from the paragraph column.

Sec.	Name	Mixed %	Para. only %	Δ	Para. words
C	cosmological	22.8	**27.7**	+4.9	1,500
P	pharmaceutical	24.6	26.9	+2.3	2,202
H	herbal	24.1	26.7	+2.6	10,328
B	balneological	23.4	25.4	+2.0	6,527
T	text	19.4	22.5	+3.1	1,531
S	astronomical	18.8	21.8	+3.0	10,505
A[4]	—	12.9	18.5	+5.6	259
Z	zodiac	12.2	—	—	0

3.8 Domain-Specific Lexicon Test

To test whether the decoded text exhibits domain specificity — a requirement for meaningful decipherment — we assembled a curated lexicon of 420 medieval Hebrew terms across four domains: botanical (148 terms from Shem Tov and Talmud), astronomical (76 from ibn Ezra), medical (96 from Asaph ha-Rofe and medieval translations), and balneological (64 from Mishnah Mikvaot and medieval sources). For each domain, we counted decoded matches per manuscript section and tested whether matches concentrate in the expected sections (e.g., botanical terms in herbal/pharmaceutical) via both chi-square and permutation tests (1,000 permutations of section labels).

Table 10: Domain lexicon concentration by manuscript section. Concentration ratio > 1 indicates above-baseline concentration in expected sections.

Domain	Expected	Matches	Conc.	Perm. z	p
Balneological	B	649	**2.08**	8.53	<0.001
Astronomical	S,Z,C	1,170	0.72	3.51	0.001
Medical	B,P	189	1.03	2.79	0.002
Botanical	H,P	372	1.15	−6.00	1.000 (ns)

Balneological vocabulary shows the strongest domain specificity: decoded water/bathing terms concentrate at 2.08× baseline in the balneological section ($z = 8.53$, $p < 0.001$). Astronomical and medical domains are also significant by permutation but with weak or dispersed concentration. Botanical terms, surprisingly, show *anti-concentration* ($z = -6.00$): herbal sections contain *fewer* botanical matches per word than other sections. This negative result is consistent with the section-entropy finding (§3.7) that the same high-frequency glosses dominate all sections, with no domain specialization in decoded vocabulary.

3.9 Scribal Error Correction

Approximately 75% of decoded word types do not match the honest lexicon. We tested whether single-character scribal errors — visually confusable EVA glyph substitutions — could account for part of this gap. Eleven confusion pairs were defined from EVA glyph shape analysis (a/o, e/i, n/r, c/e, d/s, k/t, f/p, l/r, m/n, i/n, c/h). For each of the 7,933 unmatched types, we generated all distance-1 visual variants (substituting one confusable glyph at a time), decoded each variant, and checked against the honest lexicon (45K forms).

This procedure recovered 835 types (10.5% of unmatched) corresponding to 8,544 tokens (31.1%). The most productive substitution was o→a (2,077 tokens), followed by a→o (1,108) and n→i (793). Recovery rates were uniform across scribal hands (9.6%–15.6%) and manuscript sections (9.3%–13.9%), with no hand or section standing out.

A permutation control measured whether the *specific* confusion pairs outperform arbitrary ones. We shuffled the confusion-pair labels 200 times, each time applying the same correction procedure. The result is a clear null: type $z = -0.07$ ($p = 0.53$), token $z = -0.03$ ($p = 0.51$). Shuffled pairs recover the same number of words as the real visual-confusion pairs. The recovery is therefore an artefact of lexicon density — any single-character substitution has a non-trivial chance of hitting the 45K-form lexicon — rather than evidence of actual scribal copying errors.

This null result has a constructive implication: the 75% unmatched words cannot be recovered by error correction against the current lexicon. The gap reflects **lexicon inadequacy** — medieval Hebrew technical vocabulary (botanical, astronomical, medical) is underrepresented in the biblical and talmudic sources that constitute the honest lexicon — rather than scribal noise. Progress toward readability requires a better-suited lexicon, not error correction.

3.10 Most Frequent Decoded Words

Table 11 shows the most frequent decoded Hebrew forms with known dictionary glosses. These words account for a disproportionate share of all matches, consistent with the Zipfian concentration observed in the null model test.

Table 11: Top 10 decoded Hebrew forms by token frequency (glossed subset, excluding Sefaria corpus forms).

Hebrew	Freq.	Gloss
bhyr	846	bright, brilliant (of light)
bhy	481	chaotic
mwk	383	be poor
sy	358	thee
Spk	345	to pour
bhyt	327	[variant of bhy]
bryt	325	covenant
syr	308	[noun, feminine]
my	257	who? whose?
Sr	243	[noun, masculine]

Notably, none of these words are domain-specific (botanical, astronomical, or medical). They are generic Hebrew roots — a pattern that persists across all manuscript sections, including the herbal pages where botanical terminology would be expected.

4 Discussion

4.1 What the Signal Is

The Hebrew cipher hypothesis produces a statistically genuine signal that survives multiple controls:

- Significant across three independent lexicon tiers ($z = 3.6$–4.4)
- Robust to permutation testing (botanical $p = 0.017$, anchors $p = 0.004$)
- Structurally embedded in character sequences beyond marginal frequency overlap
- Present in both Currier languages independently (A: $z = 4.02$, B: $z = 3.85$)
- Monoalphabetic (8/9 Naibbe diagnostics favor mono)
- Not explained by Aramaic (0.2–0.4% match), Italian (4.5% match), Judeo-Italian (5.0% match), Judeo-Arabic (3.8% types, 90% shared cognates), or Ladino (0.5% types)

The signal is quantitatively modest but reproducible: a fixed ∼8 percentage-point advantage over random mappings, corresponding to approximately 2,700 excess matched tokens (out of 33,000) beyond what chance would produce.

4.2 What the Signal Is Not

The decoded text does **not** read as Hebrew. Several lines of evidence confirm this:

1. **Incoherent glosses**: the highest-scoring semantically accessible passages produce word sequences like "bright – die – poor – poor – back" in herbal sections.
2. **No domain specialization**: the same generic words (bhyr, mwk, my) dominate all sections. An herbal text should show botanical concentration; a zodiac text should show astronomical terms.
3. **Labels worse than text**: figure captions, which should be identifiable nouns, match at only 13.2% — nearly half the paragraph rate.
4. **Scribe-specific signal**: the mapping works best for one of five scribes (Hand 1, 28.8%) and substantially less well for all others (21–25%), including the other Language A scribe (Hand 4, 21.7%).
5. **Match rate ceiling**: even with 491K lexical forms, only 45.7% of tokens match — compared to an expected >80% for correctly deciphered consonantal text matched against a comprehensive lexicon.

4.3 Interpretation

We consider four possible interpretations of the signal:

1. **Partially correct mapping**. Some letter assignments may be correct while others are not, producing an 8 pp excess above random. Under this interpretation, the mapping captures a subset of the true cipher relations, but the incorrect letters corrupt the decoded output. The local optimality of 14/17 letters argues against major errors, though marginal letters (tet/tav, gimel/tsade) remain ambiguous.

2. **Structural mimicry**. Hebrew consonantal writing and Voynich text may share phonotactic properties (word length distribution, character frequency profiles, bigram patterns) without a direct cipher relationship. The null model test argues against pure frequency matching, but subtler structural similarities could produce the observed signal.

3. **Shared substrate**. If the Voynich text encodes a Romance language through a Hebrew-like consonantal framework — as in Judeo-Italian writing conventions — the decoded output would resemble Hebrew morphologically without being Hebrew semantically. The 5% JI match rate and the 46 JI-only word types support this possibility.

4. **Non-linguistic structure**. Generative mechanisms such as Rugg's Cardan grille (Rugg, 2004) can produce text with natural-language statistics. If the Voynich text is generated rather than encoded, any monoalphabetic mapping would produce pseudo-meaningful output tuned to the generator's character frequencies. However, the grille hypothesis struggles to explain the significant permutation test results and the non-random spatial clustering of matches.

The data do not decisively distinguish between these interpretations. The concentration of signal in Hand 1 (interpretation 1 or 3), the lack of domain-specific vocabulary (against 1), and the Zipfian structure of matches (against 4) each constrain the space of viable explanations without resolving it.

4.4 Cross-Analysis with Independent Token Classification

An independent computational analysis of the Voynich text by DiPrima (DiPrima, 2026) classifies all EVA tokens into morphological components and functional categories using distributional methods without assuming linguistic content. His framework decomposes each token into prefix, middle, and suffix, assigns each middle to one of three *kernel operators* (K, H, E) based on positional and co-occurrence properties, and groups 479 token types into 49 instruction classes.

We decoded each of his 8,150 token types through our Hebrew mapping and tested whether his classifications predict our lexicon match rate. Table 12 summarizes the results using the honest lexicon (45K forms).

Table 12: Hebrew match rate (honest lexicon, types) by DiPrima's token classifications. All χ^2 tests are significant at $p < 10^{-6}$.

Axis	Category	Types	Matched	Rate
Kernel	K (energy)	2,111	156	7.4%
	H (transition)	1,207	48	4.0%
	E (stability)	1,552	28	1.8%
System	multi (A+B)	1,498	203	13.6%
	A only	2,164	97	4.5%
	B only	3,585	154	4.3%
Regime	Precision	1,088	119	10.9%
	High energy	777	36	4.6%
	Settling	1,699	28	1.7%

A potential confound is word length: short tokens match any lexicon at higher rates (63.6% at 2 chars, 7.7% at 5 chars, <1% beyond 7). To control for this, we computed match rates within fixed-length subsets. At length 4 (679 types): K matches at 33.3% vs E at 9.6% (3.5×). At length 6 (1,845 types): K at 3.2% vs E at 0.2% (16×). The kernel effect persists after length control.

Two findings merit attention. First, tokens appearing in both Currier systems ("multi") match at 13.6% — three times the rate of system-specific tokens (∼4.5%). The Hebrew signal resides in shared vocabulary, not system-specific material. Second, the kernel gradient (K > H > E) suggests that our mapping differentially captures specific functional categories of the text, regardless of whether those categories represent linguistic or non-linguistic structure.

An epistemological tension deserves acknowledgment: DiPrima's framework assumes the Voynich text represents non-linguistic control programs (distillation instructions), yet his distributional classifications successfully predict our Hebrew lexical match rates. This convergence admits two interpretations: (a) the linguistic and distributional structures are genuinely correlated — both methods detect the same underlying organization of the text; or (b) both systems detect shared statistical regularities (word-length distributions, positional character patterns) without either correctly identifying the generative mechanism. The cross-analysis demonstrates that *something real* is being captured by independent methods, but does not adjudicate between linguistic and non-linguistic interpretations of that signal.

4.5 Independent Validation by Hebrew Language Model

To assess the quality of our lexicon matches independently of dictionary coverage, we submitted all 1,098 glossed word types to DictaLM (Shmidman et al., 2024a), a Hebrew-specialized large language model trained on 200 billion tokens of Hebrew text (dicta-il/DictaLM-3.0-1.7B-Instruct). Each word was evaluated for plausibility as genuine Hebrew (valid / possible / invalid), approximate meaning, historical period, and whether our assigned gloss was accurate. Validation was conducted via the Featherless.ai API.

Table 13: DictaLM validation results for 1,098 glossed word types. Token counts reflect the 16,100 matched tokens (of 37,025 corpus tokens); token-weighted rates show the higher frequency of confirmed forms.

Category	Types	Type %	Tokens	Token %
Valid (genuine Hebrew)	211	19.4%	9,113	56.6%
Possible (plausible form)	763	70.1%	5,359	33.3%
Invalid	114	10.5%	1,628	10.1%

The key finding is a strong frequency asymmetry: although only 19.4% of matched *types* are rated valid, those 211 types account for 56.6% of all matched *tokens*, confirming that high-frequency forms are the genuine Hebrew signal. A permutation test restricted to the 211 DictaLM-confirmed forms yields the strongest result of this study: 21.43% match rate vs. 0.18% for random mappings, giving $z = 56.71$ and a real-to-random ratio of 118×. For comparison, the full 491K-form lexicon (minus 114 invalid forms) produces $z = 3.70$ with a ratio of only 1.9×, and the honest lexicon (cleaned of 49 invalid forms) produces $z = 46.00$ with a ratio of 78×. The signal is thus highly concentrated in approximately 200 high-frequency Hebrew roots whose validity is independently confirmed.

The 114 invalid types are predominantly low-frequency Sefaria-corpus entries with no dictionary gloss, consistent with our earlier finding that corpus-attested forms inflate match rates without adding semantic content. No letter mapping was positively refuted: the 3 mappings with fewer than 40 associated types (chet, aleph, dalet) lack statistical power to adjudicate, but none showed consistent invalidity.

The period classification is overwhelmingly **medieval** (899 of 974 valid + possible forms), with almost no biblical classifications. This suggests that if the decoded text is genuinely Hebrew, it is closer to medieval or rabbinic Hebrew than to the biblical register targeted by the STEPBible component of our lexicon — potentially explaining why biblical glosses frequently misidentify the intended meaning. Consistent with this, DictaLM independently reinterpreted

mwk (freq 383, our gloss "be poor") as "cotton/wool" in the rabbinic sense — more appropriate for herbal and pharmaceutical contexts — and identified syr as "pot/cauldron" rather than a generic feminine noun, a gloss consistent with a recipe or preparation genre. Mixed Aramaic-Hebrew forms were also flagged, characteristic of medieval Jewish texts rather than biblical Hebrew.

4.6 Syntactic Analysis with DictaBERT

To move beyond generative LLM assessment, we applied DictaBERT-parse (Shmidman et al., 2024b), a transformer-based Hebrew dependency parser trained on Universal Dependencies treebanks. Unlike DictaLM, this model produces structured syntactic analysis: POS tags, morphological features, and dependency trees. We computed a syntactic quality score (0–6) based on the presence of a root, verb, noun, function words, core arguments (subject/object), and dependency diversity.

We selected the 47 best consecutive 5-word windows in which all words are DictaLM-valid — by construction, the most favorable windows in the entire corpus. This deliberately optimistic selection means that the results below represent an *upper bound* on syntactic quality (Table 14).

Table 14: DictaBERT syntactic quality scores (0–6 scale) for three text types. Permutation z compares real word order against random reorderings of the same vocabulary.

Text type	Score	Notes
Real Hebrew	6.0/6	100% verbs, arguments, function words
Voynich phrases	2.74/6	23% verbs, 13% func. words, 2% arguments
Random consonants	2.5/6	baseline
Permutation test (z)		-0.51 (real order $\leq$ random)

Even under this optimistic selection, only 23% of Voynich phrases receive a VERB tag (vs. 100% for real Hebrew), 13% have function words, and only 2% (1 phrase) have core syntactic arguments.

A permutation test comparing the real consecutive phrases against random sequences drawn from the same word vocabulary yields $z = -0.51$: the actual word order scores *lower* than random orderings. This rules out the possibility that the decoded text contains latent syntactic structure not captured by the lexical analysis.

The parser treats most decoded tokens as NOUN chains in construct state (*smixut*), which is the Hebrew parser's default for sequences of unrelated nominals. The conclusion is unambiguous: **the signal is purely lexical; word ordering carries no detectable syntactic information.**

4.7 Limitations

- **Lexicon coverage**: the honest lexicon (45K forms) represents a specific historical stratum (biblical + talmudic + curated). Medieval Hebrew technical vocabulary — especially botanical, astronomical, and medical terms — is poorly represented. A scribal error correction experiment (§**??**) confirms that the 75% unmatched gap is not recoverable through single-character corrections; the shortfall is lexical, not scribal.

- **Single mapping**: we test one mapping derived through a specific optimization trajectory. The space of possible 19-letter monoalphabetic mappings is vast ($22!/3! \approx 10^{20}$); our mapping is locally optimal but may not be globally so.

- **Transcription uncertainty**: the EVA transcription, particularly the Takahashi layer,

contains uncertain readings (marked with `!` and `?` in the source). These affect approximately 3% of characters.

- **Morphological opacity**: Hebrew consonantal text requires morphological analysis for full comprehension. Our matching is purely lexical (exact string match against consonantal forms). A morphology-aware approach might recover additional signal — or additional noise.
- **DictaLM calibration**: the 1.7B-parameter model used for Hebrew validation has limited capacity. A blinded calibration experiment (100 known Hebrew forms, 100 random consonantal strings, 100 Voynich forms, submitted without origin labels) yielded 100% precision but only 2% recall (strict): the model recognized only 2/100 known Hebrew words and 0/100 random strings (FPR = 0%). Of 100 blinded Voynich forms, 3 were accepted (vs. 0 expected by chance). The extremely low recall means the model is too conservative to serve as a reliable standalone validator; the "possible" category (70.1% of Voynich forms in the main experiment) reflects this conservatism rather than genuine ambiguity. The $z = 56.71$ DictaLM-filtered score should be interpreted as a lower bound, since the model rejects most valid Hebrew.

5 Meta-Analysis: Comparison with Published Research

No systematic comparison of Voynich hypotheses using unified statistical criteria has been published to date. In this section we compute information-theoretic metrics that enable direct comparison with published results and construct a comprehensive evaluation of 15 prior studies against our data.

5.1 Character Entropy

Bowern and Lindemann (2021) report that the Voynich text has a conditional character entropy of order 2 (h_2) of approximately 2 bits — anomalously low for natural language, where $h_2 \approx 3$–4 bits. This has been cited as evidence against linguistic content. We computed h_2 for three representations of the text: raw EVA, our decoded Hebrew, and a reference Hebrew corpus sampled from the Sefaria 250-million-token corpus (Table 15).

Table 15: Character entropy at orders 0–2. Hebrew reference generated by frequency-weighted sampling from the Sefaria corpus (37,025 tokens). $h_k = H(X_n|X_{n-1}, \ldots, X_{n-k})$ in bits.

Metric	EVA	Decoded	Hebrew ref.
h_0 (entropy)	3.86	3.82	4.12
h_1 (cond. bigram)	2.37	2.70	3.98
h_2 (cond. trigram)	2.12	2.44	3.72
Alphabet size	22[5]	19	22
Characters	191,545	164,736	148,621

The mapping **increases** h_2 by +0.32 bits (from 2.12 to 2.44), closing approximately 20% of the gap between raw EVA and real Hebrew ($3.72 - 2.12 = 1.60$ bits). The direction is consistent with partial decipherment: a correct cipher key should decompress the entropy of the ciphertext toward natural-language values. However, the magnitude is modest — the decoded text remains much closer to raw EVA than to the Hebrew reference ($h_2 = 2.44$ vs. 3.72), and the 80% residual gap confirms that the mapping is, at best, a partial decipherment.

The h_1 shift is larger (+0.33 bits), consistent with the mapping's positional rules (initial bet/samekh splits) introducing context-dependent variation that reduces short-range predictabil-

ity. The near-constant h_0 (3.86 → 3.82) reflects the reduction from 22 EVA characters to 19 Hebrew consonants, approximately offset by the more uniform distribution of Hebrew letters.

5.2 Morphological Complexity (MATTR)

Lindemann (2022) found an anomalously high Moving Average Type-Token Ratio (MATTR, window 50) for the Voynich text, suggesting low morphological complexity. We computed MATTR for the three text representations:

Text	MATTR (window=50)
EVA	0.876
Decoded Hebrew	0.865
Hebrew reference	0.977

The decoded text shows marginally lower MATTR than raw EVA (−0.011), indicating a slight increase in morphological regularity from the letter-mapping process. However, the Hebrew reference has even higher MATTR (0.977), reflecting the rich type inventory of real Hebrew consonantal text. The Voynich-decoded text does not move toward the Hebrew reference, suggesting that while character-level entropy improves, word-level diversity does not.

5.3 Zipf Distribution

The decoded text produces a Zipf slope of −0.90, close to the theoretical −1.0 for natural language and consistent with Landini (2001)'s earlier analysis of raw EVA (slope −0.87). The Hebrew reference slope (−0.73) is shallower, reflecting the flatter frequency distribution of a large-vocabulary language sampled with replacement. The near-Zipfian distribution is consistent with linguistic structure but is not diagnostic, as various non-linguistic processes also produce Zipf-like distributions (Timm and Schinner, 2014).

5.4 Comparative Evaluation of Published Claims

Table 16 summarizes how our results compare with 15 published studies. We classify each comparison as **confirms** (our data independently supports their claim), **refutes** (our data contradicts their central claim), **contradicts** (mutual tension with our findings), or **neutral** (insufficient evidence to adjudicate).

Of the 15 comparisons, **8 confirm** our results (including two that we extend with new scribe-level and Currier-level data), **1 is refuted** by our evidence (Cheshire's Proto-Romance hypothesis, contradicted by the 5% Italian vs. 25% Hebrew match rate), **4 are in tension** with our findings (our data provide evidence against these claims but do not conclusively exclude them), **1 is contradicted**, and **1 is neutral**.

The entropy analysis (§5.1) provides the most informative new test. Bowern and Lindemann (2021)'s finding of anomalously low h_2 is often cited against the possibility of monoalphabetic cipher. Our result shows that the mapping partially shifts h_2 toward natural-language values: an increase of 0.32 bits, closing approximately 20% of the gap to the Hebrew reference (3.72 bits). This is consistent with the mapping capturing some cipher structure, but the modest magnitude (80% of the gap remains) means the evidence is suggestive rather than conclusive.

The principal tension is with Montemurro and Zanette (2013): their keyword clustering analysis predicts domain-specific vocabulary in different manuscript sections, but our decoded glosses show no such specialization. This contradiction has two possible explanations: (a) the clustering they detected operates at a sub-word level that our word-level decoding does not capture, or (b) the keywords they identify are structural artifacts (frequent words occurring in section-specific positions) rather than semantic content.

Table 16: Comparison with published Voynich research. See text for details.

Study	Claim	Our result	Verdict
Reddy and Knight (2011)	Hebrew abjad (perplexity)	19/22 consonants mapped	**C**
Bowern and Lindemann (2021)	$h_2 \approx 2$, not mono	h_2: 2.12 → 2.44 (decoded)	**C**
Kondrak and Hauer (2016)	Hebrew most probable	Match 17–25% honest	**C**
Cheshire (2019)	Proto-Romance	Italian 5% vs Hebrew 25%	**R**
Greshko (2025)	Naibbe cipher	$z = 12.1$, 8/9 mono	**T**
Rugg (2004)	Cardan grille (hoax)	Perm. $z = 3.6$–4.4	**T**
Schinner (2007)	Stochastic process	IC=0.077 (mono range)	**T**
Davis (2020)	5 scribal hands	Hand 1 drives signal	**C+**
Montemurro and Zanette (2013)	Domain keywords	Same glosses all sections	**X**
Landini (2001)	Zipf-like distribution	Slope −0.90	**C**
Currier (1976)	Two languages A/B	Both sig.; A=Hand 1	**C+**
Timm and Schinner (2014)	Non-linguistic mechanism	Perm. $z = 3.6$–4.4	**T**
Amancio et al. (2013)	Language-like statistics	IC, Zipf confirm	**C**
Lindemann (2022)	High MATTR	0.876 → 0.865 (decoded)	**N**
Stolfi (2005)	Prefix-root-suffix grammar	Matches Hebrew morphology	**C**

C = confirms, C+ = confirms and extends, R = refutes, T = in tension, X = contradicts, N = neutral.

6 Conclusion

We have conducted the most comprehensive statistical evaluation to date of the hypothesis that the Voynich Manuscript encodes Hebrew consonantal text. Our 19-letter monoalphabetic mapping produces a significant, reproducible signal ($z = 3.6$–4.4 across lexicon tiers) that survives permutation testing and is in tension with tested alternative explanations (homophonic cipher, Aramaic, Italian, Judeo-Italian, Judeo-Arabic, Ladino), though it does not conclusively exclude all non-Hebrew generative mechanisms.

The signal is real. The decipherment is not. The decoded text does not read as Hebrew and fails to produce coherent passages. A domain-specific lexicon test (§3.8) finds one significant result — balneological vocabulary concentrates at 2.08× baseline in the bathing section ($z = 8.53$) — but botanical terms show no specialization. The signal is concentrated in one scribal hand and in continuous text rather than figure labels.

We propose that the correspondence reflects either a partially correct mapping (some letters right, others not), a shared phonotactic substrate (possibly Judeo-Italian or another Hebrew-script Romance language), or structural properties of the Voynich text that happen to align with Hebrew consonantal morphology. Cross-analysis with an independent distributional classification of the same tokens (DiPrima, 2026) reveals that the mapping differentially captures specific functional categories of the text (kernel-K tokens match at 3.5× the rate of kernel-E tokens at matched word length), suggesting the signal is not uniformly distributed across the text's internal structure. Distinguishing between these possibilities requires the identification of additional cipher structure beyond monoalphabetic substitution.

Independent validation by DictaLM (§4.5) confirms that the mapping captures real Hebrew vocabulary: 19.4% of matched types are confirmed as genuine Hebrew words (token-weighted: 56.6%). A permutation test restricted to these 211 confirmed forms yields $z = 56.71$ with a real-to-random ratio of 118× — the strongest result in this study. The overwhelming me-

dieval/rabbinic period classification of confirmed words suggests that a medieval Hebrew lexicon would be more appropriate than the biblical tier we have primarily used. A scribal error correction experiment (§??) reinforces this conclusion: single-character visual corrections recover 10.5% of unmatched types, but a permutation control shows this recovery is no better than chance ($z \approx 0$), confirming that the unmatched gap reflects lexicon inadequacy rather than copyist noise.

Dependency parsing with DictaBERT further sharpens the diagnosis: the 47 best consecutive phrases (all words DictaLM-validated) score only 2.74/6 on syntactic quality, barely above random consonant strings (2.5/6) and far below real Hebrew (6.0/6). A permutation test confirms that the actual word order scores no better than random reorderings of the same vocabulary ($z = -0.51$). The signal is therefore purely lexical: individual words match Hebrew roots, but their sequential arrangement carries no syntactic information. This does not resolve the readability gap but narrows the diagnosis to a precise boundary: real vocabulary, random syntax.

The complete mapping, decoded corpus, and all statistical results are available in the project repository for independent verification.

Data Availability and Reproducibility

All source code, the EVA transcription, and the complete mapping are available at https://github.com/antenore/voynich-toolkit under the MIT license. The analysis pipeline is fully reproducible:

```
git clone https://github.com/antenore/voynich-toolkit.git
cd voynich-toolkit
pip install -e .
voynich --force full-decode        # decode corpus
voynich --force meta-analysis      # h2, MATTR, Zipf, literature table
voynich --force null-model-test    # null model (1-3 min)
voynich --force scribe-analysis    # per-hand match rates (3 min)
voynich --force naibbe-test        # Naibbe hypothesis (40s)
voynich --force layout-analysis    # label vs paragraph
voynich --force cross-analysis    # cross-analysis (requires epilectrik repo)
voynich --force dictalm-validate  # DictaLM validation (55 min, API key required)
```

Each command produces JSON (machine-readable), TXT (human-readable), and where applicable LaTeX table output in `output/stats/`. The SQLite database (`voynich.db`, regenerable via `python scripts/build_sqlite_db.py`) contains all intermediate results in queryable form. Hebrew lexicon data requires separate preparation (`voynich enrich-lexicon`) due to third-party licensing; instructions are provided in the repository. The EVA transcription file (`eva_data/LSI_ivtff_0d.txt`) is included in the repository.

Acknowledgments

This work used the EVA transcription by Takeshi Takahashi and the interlinear file maintained by the Voynich community. Hebrew lexicon data from STEPBible, Jastrow's Dictionary of the Talmud, the Klein Etymological Dictionary (via Sefaria API), and the Sefaria open-source corpus. Computational analysis was conducted with the assistance of Claude (Anthropic) for code development and statistical analysis.

References

C. Bowern and S. J. Lindemann. The linguistics of the Voynich manuscript. *Annual Review of Linguistics*, 7:285–308, 2021.

G. Cheshire. The language and writing system of MS408 (Voynich) explained. *Romance Studies*, 37(1):23–34, 2019.

G. Kondrak and B. Hauer. Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computational Linguistics*, 4:75–86, 2016.

S. J. Lindemann. Quantitative approaches to the Voynich manuscript. PhD thesis, Yale University, 2022.

S. Reddy and K. 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*, pages 78–86, 2011.

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

J. Stolfi. A quantitative study of the script of the Voynich manuscript. Technical report, Institute of Computing, University of Campinas, 2005.

D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr., and L. da F. Costa. Probing the statistical properties of unknown texts: application to the Voynich manuscript. *PLoS ONE*, 8(7):e67310, 2013.

J. DiPrima. Distributional analysis of Voynich manuscript token structure. GitHub repository, https://github.com/epilectrik/voynich, 2026.

P. Currier. Papers on the Voynich manuscript. In *New Research on the Voynich Manuscript: Proceedings of a Seminar*, Washington, DC, 1976.

L. Fagin Davis. How many scribes? A paleographic study of the Voynich manuscript. *Manuscript Studies*, 5(2):164–186, 2020.

M. Greshko. The Naibbe cipher: a verbose homophonic substitution for the Voynich manuscript. *Cryptologia*, 2025.

G. Landini. Evidence of linguistic structure in the Voynich manuscript using spectral analysis. *Cryptologia*, 25(4):275–295, 2001.

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

G. Rugg. An elegant hoax? A possible solution to the Voynich manuscript. *Cryptologia*, 28(1):31–46, 2004.

A. Timm and A. Schinner. A possible generating mechanism for the Voynich manuscript. *Cryptologia*, 38(4):311–328, 2014.

S. Shmidman, A. Gueta, et al. DictaLM: A large generative language model for Hebrew. *arXiv preprint arXiv:2407.07080*, 2024.

S. Shmidman, E. Gueta, et al. DictaBERT: A state-of-the-art Hebrew NLP suite. In *Findings of ACL 2024*, pages 4523–4538, 2024.