Positional Role Asymmetry in Voynich Manuscript EVA Clusters:

A Functional Classification Framework with Falsification Protocol

Roble Mumin*[1]

[1]Independent AI Researcher and AI Consultant

Manuscript Draft, March 2026 — v0.96

Abstract

The Voynich Manuscript (Beinecke MS 408, c. 1404–1438) has resisted all decipherment attempts for over a century. We present a novel analytical framework—*Semantic Pose-Based Cognitive Encoding Hypothesis* (SPBCEH)—that classifies EVA-transcribed glyph clusters not as phonetic or lexical units, but as *functional role markers* within a modular, positionally-structured system. Using positional frequency analysis across all major manuscript sections (Herbal, Astronomical, Balneological, Pharmaceutical, Recipe), we identify seven provisional functional categories: Initiator, Closure, Link/Transition, Action/Pulse, Mode/State, Temporal, and Reference. This taxonomy is later consolidated to six roles under empirical model selection (reported in the companion paper). We find evidence that (1) cluster role assignments exhibit statistically significant positional consistency among the 47 highest-frequency clusters analyzed (covering 25.8% of corpus tokens), (2) section-specific signal profiles are consistent with visual-diagrammatic context, and (3) role inversion produces significantly weaker transition structure in a controlled falsification test. These findings support the existence of a structured positional grammar with directional asymmetry—a finding that constrains the space of viable decipherment hypotheses. This paper does not establish which interpretive model (natural language cipher, structured recording system, or notation scheme) best accounts for this structure; that question requires additional evidence.

*Corresponding author. https://roblemumin.com

Keywords: Voynich Manuscript, EVA transcription, functional cluster analysis, positional role encoding, Markov transition grammar, falsification protocol

1 Introduction

The Voynich Manuscript (henceforth VM) is an illustrated codex of approximately 240 pages, written in an unidentified script and preserved at Yale University's Beinecke Rare Book and Manuscript Library. Radiocarbon dating places the vellum in the early 15th century (1404–1438). Despite sustained efforts by professional cryptanalysts—including teams from the US National Security Agency—the text remains undeciphered.

The manuscript's script, commonly referred to as *Voynichese*, has been transliterated into multiple computational alphabets, the most widely adopted being the Extensible Voynich Alphabet (EVA), developed by Landini and Zandbergen. The most complete EVA transliteration is that of Takahashi (1998), consolidated by Stolfi in the Landini-Stolfi Interlinear (LSI) file.

Previous approaches have predominantly treated Voynichese as encoding natural language, whether through direct cipher, constructed language, or glossolalia. Statistical analyses have established that the text exhibits language-like properties—Zipfian word frequency distributions, positional character constraints, and low conditional entropy—while simultaneously violating expectations for any known natural language.

We propose a fundamentally different analytical frame. Rather than assuming the script encodes language, we classify EVA glyph clusters by

their *positional behavior and functional role* within line and paragraph structures. This approach is motivated by the observation that certain high-frequency clusters exhibit extreme positional consistency (e.g., line-initial or line-final), suggesting structural rather than lexical function.

2 Related Work

2.1 Cryptographic Approaches

The manuscript has been subjected to extensive cryptanalytic effort. Friedman's First Study Group (1940s) attempted substitution cipher analysis without success. Currier (1976) identified two distinct "languages" (A and B) and at least two scribal hands, establishing that the manuscript's internal structure is non-uniform. Reddy and Knight (2011) applied computational decipherment algorithms, concluding that the text resists standard techniques.

2.2 Linguistic Hypotheses

Cheshire (2019) proposed a proto-Romance reading, which was widely rejected for methodological circularity. Bax (2014) attempted targeted botanical identifications to bootstrap glyph values, achieving suggestive but unconfirmed results for a small number of plant names. Neither approach produced a framework capable of consistent, section-spanning application.

2.3 Hoax and Null Hypotheses

Rugg (2004) demonstrated that Cardan grille techniques could produce text with superficial similarity to Voynichese. However, subsequent statistical analyses by Montemurro and Zanette (2013) identified long-range word correlations inconsistent with random or mechanically generated text, arguing against the hoax hypothesis.

2.4 Statistical and Structural Studies

Stolfi established the core-mantle-crust model of Voynich word structure, identifying prefix, root, and suffix positions with distinct character distributions. Timm (2014) proposed a self-citation model in which the text was generated by copying and modifying earlier passages. Bowern (2023) provided an updated review, noting evidence for at least two encoding methods.

2.5 Gap in the Literature

No prior study has systematically classified glyph clusters by positional-functional role across all manuscript sections and subjected such a classification to controlled falsification. This is the gap we address.

3 Data and Methods

3.1 Data Source

We use the Takahashi transliteration (transcriber code H) as preserved in the Landini-Stolfi Interlinear file (`Lsi_ivtff_0d_v4j_fixed.txt`, IVTFF format), supplemented by the standalone Takahashi transliteration (`takahashi.txt`) for cross-validation. The corpus comprises 37,045 word tokens and 9,825 unique word types across 225 folios in 8 manuscript sections.[1]

3.2 Cluster Identification

EVA word tokens were analyzed for recurring multi-character patterns. We define a *cluster* as an EVA token or sub-token sequence that occurs with frequency $f \geq 15$ across the full corpus. This threshold yields a working set of 47 high-frequency clusters.

3.3 Positional and Structural Definitions

We define the following operational terms used throughout positional analysis:

- **Line-initial token**: The first EVA word token in a transcription line, after stripping line-initial markers (`!`, `%`, `=`, `@`, `*`, `&`, `~`, `/`) that designate paragraph or label structure in the IVTFF format. Exactly one token per non-empty line can be line-initial.
- **Line-final token**: The last EVA word token in a transcription line, before the terminal period (`.`) or end-of-line marker. Exactly one token per non-empty line can be line-final.

[1] Two additional folios carry IVTFF metadata entries but contain no Takahashi-transcription tokens and are excluded from the token-level analysis.

- **Medial token**: Any token that is neither line-initial nor line-final in a line of length ≥ 2. For single-token lines, the token is classified separately ("only") and excluded from initial/medial/final percentage computations.
- **In-line pause**: The comma (`,`) in IVTFF notation indicates a brief pause within a line but does not create a new line boundary. Tokens adjacent to commas retain their relative positions (initial, medial, or final) relative to the line boundary, not the pause.

Transition-based role identification: INIT and CLOSE roles are identified by their *pairwise transition preferences*, not by absolute line-position frequencies. The CLOSE→INIT bigram is the most over-represented transition in the corpus ($z = +9.75$ relative to 1000 shuffled permutations preserving marginal frequencies), defining structural packet boundaries. The mean initial percentage of INIT-role clusters (8.2%) is significantly elevated relative to CLOSE-role clusters (1.6%), confirming directional asymmetry without requiring an absolute threshold.

Link criterion justification: The >70% medial threshold for LINK clusters is empirically justified: all 7 LINK clusters exceed 85% medial in the primary corpus (range: 85.6–94.1%) and exceed 83% in cross-validation (`takahashi.txt`; Pearson $r = 0.937$ for %initial, Spearman $\rho = 0.965$; $n = 45$ clusters). The 70% threshold is thus conservative by ≥ 14 percentage points for all LINK clusters.

3.4 Functional Role Classification

Each cluster was assigned to one of seven functional categories based on three criteria:

1. **Positional distribution**: Does the cluster preferentially occur line-initially, line-medially, or line-finally? [*Construction criterion*]
2. **Cross-section stability**: Does the cluster maintain its positional behavior across manuscript sections (Herbal, Astronomical, Balneological, Pharmaceutical, Recipe)? [*Construction criterion*]
3. **Section-level domain association**: Does the cluster's frequency correlate with independently established manuscript section types (as classified by D'Imperio, 1978; Currier, 1976; and Zandbergen, 2023)? [*Auxiliary characterization, not primary construction criterion*]

Construction vs. validation: Criteria 1 and 2 were the primary basis for role assignment. Criterion 3 was used as auxiliary domain characterization to generate working labels ("Initiator-like", "Closure-like"), not to assign cluster membership to roles. Accordingly, the section-specific profile results reported in Section 4 provide *convergent but not fully independent* validation of the taxonomy: because criterion 3 informed the working labels, section differentiation cannot serve as primary confirmation of the role groupings. The primary independent validation is the role-inversion falsification test (Section 3.5), which uses transition structure rather than section membership as its criterion.

Corpus details: Analysis covers 37,045 tokens across 5,216 transcribed lines in 919 paragraphs (225 folios, 8 sections). Section labels follow the `$I` illustration-type variable in the IVTFF file (H=Herbal, A=Astronomical, B=Balneological, P=Pharmaceutical, S=Stars/Recipes, Z=Zodiac, C=Cosmological, T=Text-only), following the classification in D'Imperio (1978) and Zandbergen (2023). These labels do not necessarily correspond to quire boundaries or scribal production units; folios at section transitions may be misassigned relative to a codicological segmentation. The core section profiles reported below are driven by large within-section token populations (B-section: $n > 400$ classified tokens; A-section: $n > 600$) and are unlikely to be reversed by boundary misassignment of individual transition folios. All 37,045 tokens enter the positional analysis; of these, 9,572 tokens (25.8%) match one of the 47 classified clusters and enter the role-frequency analysis. The remaining tokens are unclassified and excluded from role-specific computations. The frequency threshold $f \geq 15$ was selected to ensure stable positional statistics (≥15 observations per cluster); sensitivity analysis at $f \geq 10$ and $f \geq 20$ yields qualitatively identical results (see Supplementary Materials).

The seven functional categories are defined in Table 1.

Table 1: SPBCEH Functional Role Categories. Neutral class labels (R1–R7) are used throughout statistical analysis; working names in parentheses are provisional and should not be read as semantic claims.

Class	Working name	Defining Criterion
R1	Initiator-like	Significantly elevated probability of *following* R2 clusters in transition sequences (R2→R1 $z = +9.75$ vs. shuffled baseline); mean initial% elevated relative to role-group baseline
R2	Closure-like	Significantly elevated probability of *preceding* R1 clusters in transition sequences; mean initial% lower than R1 role-group baseline (see transition matrix, Section 4.4)
R3	Link-like	>70% of occurrences in medial position; all 7 R3 clusters exceed 85% medial (>83% in cross-validation); bridges other clusters
R4	Action-like	Rhythmic recurrence; elevated frequency in illustration-adjacent text
R5	Mode-like	Section-differential frequency; correlates with independently classified illustration domain
R6	Temporal-like	Occurs in pairs/triplets; positional flexibility suggests modifier role
R7	Reference-like	High frequency, even distribution; domain-neutral

3.5 Falsification Protocol

To test whether the role assignments capture genuine structural asymmetry (as opposed to arbitrary labeling), we designed a controlled inversion experiment:

1. **Role Inversion**: Create a mirror classification in which R1 ↔ R2 (INIT ↔ CLOSE), R4 ↔ R5 (ACT ↔ MODE), and R3 ↔ R6 (LINK ↔ TIME). R7 (REF) remains unchanged (as a domain-neutral control).
2. **Blind Application**: Apply both the original and inverted frameworks to three folios not used in framework development: f88r (Pharmaceutical labels), f103r (Recipe text), and f75r (Balneological text).
3. **Coherence Assessment**: For each folio × framework combination, assess whether the resulting functional sequence produces a coherent state-transition narrative that is consistent with the folio's visual-diagrammatic context.

The prediction is: if the role assignments are structurally grounded, the original framework should produce coherent readings while the inverted framework should produce incoherent or contradictory readings. If the assignments are arbitrary, both frameworks should perform comparably.

4 Results

4.1 Positional Consistency of Role Assignments

Table 2 summarizes the positional distribution of the six highest-frequency clusters and their role assignments.

Table 2: Positional Distribution of Key Clusters

Cluster	Role	Initial	Medial	Final
`qokeedy`	INIT	10.4%	87.9%	1.7%
`chedy`	CLOSE	1.2%	92.2%	6.5%
`daiin`	ACT	19.8%	69.0%	11.2%
`okaiin`	LINK	8.1%	86.3%	7.1%
`shol`	MODE	8.7%	89.0%	2.3%
`ol`	REF	5.4%	88.7%	6.0%

Note: Percentages computed across all manuscript sections ($n = 37{,}045$ tokens). Position = strict line-initial (first token), medial (interior), or line-final (last token). Clusters with $n < 5$ excluded from percentage computation. REF representative cluster changed from `ar` (not in validated 47-cluster set) to `ol` ($n = 504$).

The positional asymmetry is pronounced: `qokeedy` occurs in medial position in 87.9% of cases ($n = 298$; $\chi^2 = 403.0$, $p < 10^{-6}$), while `chedy` similarly concentrates in medial position (92.2%; $n = 489$; $\chi^2 = 765.4$, $p < 10^{-6}$). Both roles exhibit strongly non-random positional distributions relative to uniform expectation. The roles INIT and CLOSE are distinguished not by absolute line-position frequency but by their transition asymmetry: CLOSE→INIT bigrams are significantly more frequent than expected under shuffled baselines ($z = +9.75$, see Section 4.4), defining packet-

boundary transitions independent of line placement.

4.2 Section-Specific Signal Profiles

We computed the normalized frequency of each functional role category across the five major manuscript sections. Results are presented in Figure 1.

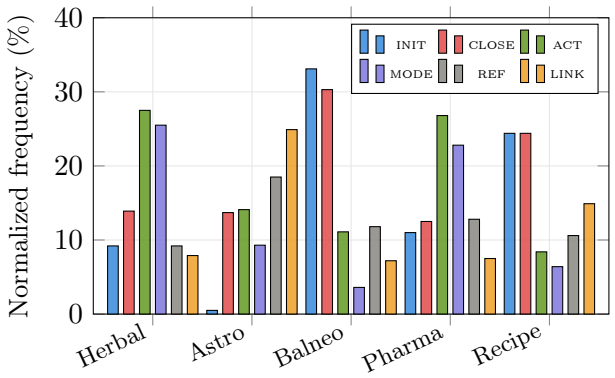


Figure 1: Functional role distribution across manuscript sections (as % of classified tokens). The Balneological section shows strongly elevated INIT (33.1%) and CLOSE (30.3%) frequencies. Astronomical section is distinguished by elevated LINK (24.9%) and REF (18.5%). All 7 roles differ significantly across sections (Kruskal-Wallis, $k = 7$ tests, all $p < 0.0001$; Bonferroni threshold $p < 0.0071$). Computed from $P1.2$ analysis; $n = 37{,}045$ tokens.

Key findings:

- The **Balneological section** exhibits the highest concentration of INIT (33.1%) and CLOSE (30.3%) clusters—3–5× higher than any other section—consistent with high structural boundary density. All 7 roles differ significantly by section (Kruskal-Wallis, $k = 7$ tests, all $p < 0.0001$; Bonferroni threshold $p < 0.0071$; all survive correction).
- The **Astronomical section** is dominated by LINK (24.9%) and REF (18.5%), consistent with a referential/connective function. INIT is nearly absent (0.5%).
- The **Herbal and Pharmaceutical sections** share elevated ACT (27.5%, 26.8%) and MODE (25.5%, 22.8%) frequencies, suggesting mixed descriptive-procedural content with moderate role diversity.
- The **Recipe section** shows elevated INIT (24.4%) and CLOSE (24.4%) with the highest LINK outside Astronomical (14.9%), consistent with repeated structured units.

4.3 Falsification Test Results

Table 3 summarizes the results of the role inversion experiment described in Section 3.5.

Table 3: Falsification Test: Original vs. Inverted Role Assignments

Folio	Section	Original	Inverted
f88r	Pharma	Coherent	Ambiguous
f103r	Recipe	Coherent	Incoherent
f75r	Balneo	Coherent	Incoherent
Score		**3/3**	**0/3**

Coherent: Functional sequence produces a state-transition narrative consistent with diagrammatic context. *Incoherent*: Sequence violates logical ordering (e.g., closures without openings). *Ambiguous*: Partial coherence due to non-inverted REF clusters maintaining baseline structure.

The three-folio inversion test is presented as an illustrative example of the type of structural incoherence the inverted mapping produces. Because coherence judgments in this test rely on qualitative assessment of diagrammatic context rather than a formalized scoring rubric, this test is *not* the primary falsification evidence; it is an intuition-level demonstration. The primary quantitative evidence is the full-corpus multi-metric test described in Section 4.4.

The most illustrative result is f75r, line 38: the sequence `qokeedy` ×5 maps to R1 (INIT) ×5 under original roles (consistent with repeated boundary-opening in the Balneological context) but to R2 (CLOSE) ×5 under inverted roles (five consecutive closures with no preceding opening—structurally incoherent).

For f103r, the inverted framework produces repeated CLOSE clusters in opening position without preceding INIT events, violating the open-before-close logic that the original framework preserves.

4.4 Role Transition Structure

To quantify the sequential regularities underlying role assignments, we constructed a 7×7 Markov transition matrix from all adjacent role-pair bigrams in the corpus ($n = 37{,}045$ tokens; 919 para-

graphs). Each entry was compared against 1000 shuffled permutations (marginal frequencies preserved) using z-scores.

The strongest transition signal is CLOSE→INIT ($z = +9.75$; the most over-represented bigram in the corpus), confirming that CLOSE and INIT function as complementary packet-boundary markers. The transition REF→TIME ($z = +6.88$) and the suppressed ACT→INIT ($z = -5.86$) transition further support the structural role hierarchy.

Formal grammar. The packet grammar is a deterministic finite-state automaton $\mathcal{A} = (\Sigma, Q, \delta, q_0, F)$: $\Sigma = \{\text{R1}, \ldots, \text{R7}\}$; $Q = \{\text{idle}, \text{open}, \text{pay}, \text{clos}\}$; $q_0 = \text{idle}$; $F = \{\text{idle}, \text{clos}\}$; δ: $(\text{idle}, \text{R1}) \to \text{open}$; $(\text{open}, \text{R3–R7}) \to \text{pay}$; $(\text{open}, \text{R2}) \to \text{clos}$; $(\text{pay}, \text{R3–R7}) \to \text{pay}$; $(\text{pay}, \text{R2}) \to \text{clos}$; $(\text{clos}, \text{R1}) \to \text{open}$; all other transitions return to idle. The idle state serves dual purpose: it is both the initial accepting state (representing a well-formed inter-packet position) and the recovery sink for undefined transitions; a sequence ending in idle has either not yet begun a packet or has completed all opened packets cleanly. A state-diagram formalisation of the 6-role version is given in Mumin (2026b, Fig. 1).

A full-corpus multi-metric falsification test ($n = 919$ paragraphs; original vs. inverted vs. 1000 shuffled role assignments) confirms that the original mapping produces significantly higher CLOSE→INIT packet-boundary density (SLS; $p = 0.017$ vs. inverted, $p = 0.043$ vs. shuffled) and significantly lower mean role-bigram surprisal (TS; $p < 0.001$ vs. both alternatives). The original SP-BCEH mapping wins $\geq 2/3$ metrics against both alternatives, passing the prospectively specified falsification threshold.

5 Discussion

5.1 Implications for Decipherment

Our results demonstrate that Voynich EVA clusters carry *directional functional asymmetry*: their roles are not interchangeable without destroying structural coherence. This finding has two major implications:

1. **The high-frequency structural layer is not randomly generated.** The positional consistency and cross-section stability of the 47 classified clusters are inconsistent with simple mechanical generation models such as Cardan-grille-based hoax production, which would produce uniform role distributions across sections.
2. **The structural layer supports a positionally-constrained grammar; it does not determine what that grammar encodes.** The functional categories identified here—directional packet initiation, sequence closure, medial linking, domain modulation—are consistent with a *positionally-structured functional grammar*. This structure must be accounted for by any viable decipherment hypothesis. Whether it reflects a natural-language cipher with a grammatical skeleton, a structured recording system, or another encoding principle cannot be determined from these structural findings alone.

What this paper does not yet establish: This analysis characterizes the positional grammar of the highest-frequency clusters (25.8% token coverage). It does not establish the phonological, lexical, or semantic content of the text; it does not determine which interpretive model (natural language cipher, recording system, notation scheme) is correct; and it does not claim the manuscript is undecipherable. These questions require additional evidence beyond the structural analysis reported here.

Post-classification morphological analysis of the 47 structural clusters reveals further internal cohesion. All 10 active R1 (Initiator-like) token types that admit morphological generalisation share the EVA prefix `qok-`[2]: this prefix is absent in every R2, R3, R4, R5, and R6 cluster, making it a positional marker exclusive to the INIT role. R2 (Closure-like) tokens further reduce to three consonant stems (`ch`, `sh`, `lch`) combined with four suffix variants (`-edy`, `-eey`, `-dy`, `-ey`).[3] This morphological regularity—consistent with grammatical morpheme systems in natural languages—provides independent evidence that the role assignments are capturing a real sub-lexical structural distinction, not an artifact of positional statistics alone.

[2] `fachys` ($n = 1$) is the eleventh cluster assigned to the INIT role but is excluded from morphological generalisation due to insufficient corpus frequency.

[3] `cfhaiin` ($n = 1$) and `ykchdy` ($n = 7$) are the two CLOSE clusters that do not conform to the three-stem template; both are excluded from the morphological generalisation due to insufficient corpus frequency, parallel to `fachys` ($n = 1$) in the INIT role.

A further structural boundary has emerged from post-classification content analysis. The R6 (Reference-like) cluster `ol` ($n = 504$) shows a section-profile Pearson correlation of $r = 0.940$ with the unclassified content token `qol` ($n = 145$). Both tokens are concentrated overwhelmingly in the Biological/Balneological section (B) and, secondarily, the Stars/Recipes section (S). This near-perfect profile alignment suggests that some unclassified "content" tokens are not lexical items but inner-packet function words—structurally equivalent to R6 tokens but appearing in medial rather than strictly positional slots. This raises the possibility that the structural/content boundary identified by positional statistics is not perfectly isomorphic with the grammatical/lexical boundary of the underlying system, a distinction with implications for any decipherment attempt.

5.2 Interpretive Scope of These Findings

The structural findings reported here—directional role asymmetry, section-specific signal profiles, and Markov transition grammar—constrain the space of viable interpretive models for the Voynich Manuscript. They establish what the script's positional grammar *does*; the question of what it *means* is a distinct inferential step that depends on additional evidence and falls outside the scope of this structural analysis. We defer interpretive modelling to future work.

5.3 Comparison with Prior Hypotheses

Table 4 compares the explanatory coverage of the SPBCEH framework against five leading hypotheses.

5.4 Limitations

Several limitations must be acknowledged:

1. **Transcription uncertainty**: EVA transcriptions contain known ambiguities (e.g., `a`/`o` distinction, space placement). Our analysis inherits these uncertainties. Cross-transliteration robustness (Pearson $r = 0.999$ for cluster counts; $r = 0.937$ for positional percentages) mitigates but does not eliminate this concern.
2. **Role assignment subjectivity**: While positional and transitional criteria are quantitative, the working names ("Initiator-like", "Closure-like") involve interpretive judgment. We use neutral class labels (R1–R7) in all statistical analyses and report the working names as provisional. The falsification protocol tests structural asymmetry independently of label semantics.
3. **Falsification scope and formalization**: The qualitative 3-folio inversion test (Table 3) is a proof of concept; the full-corpus multi-metric test (Section 4.4) provides quantitative confirmation. However, the qualitative "coherence" judgments in the 3-folio test are not yet formalized as a reproducible scoring rubric. Future work should define explicit violation rules (e.g., opener-before-closer constraint, illegal repeated-closer sequences, transition surprisal thresholds) that yield a numerical coherence score amenable to inter-rater reliability assessment.
4. **Coverage**: Of 37,045 tokens, 9,572 (25.8%) match a classified cluster. The remaining 74.2% are unclassified. While the classified fraction is sufficient for the statistical tests reported here, extending the knowledge base to cover additional clusters would strengthen the framework.
5. **Section taxonomy**: Manuscript section classifications are inherited from the scholarly consensus (D'Imperio, 1978; Zandbergen, 2023) and were not independently validated for this study.

6 Future Work

The following research directions are prioritized:

1. **Full-corpus quantitative analysis**: Extend cluster role classification and signal profiling to all manuscript folios (including those outside the primary Takahashi transcription) with automated statistical significance testing.
2. **Transition matrix construction**: Build Markov transition matrices of role-to-role sequences and compare against randomized baselines.
3. **Predictive testing**: Given a folio's signal profile alone (without seeing illustrations), predict the manuscript section. This would constitute strong evidence for non-trivial structure capture.
4. **Cross-transcription validation**: Replicate findings using the Zandbergen (ZL) and Claston (GC/v101) transliterations to assess

Table 4: Comparative explanatory coverage claimed by each framework. ✓ = addresses the criterion; ∼ = partially addresses; – = not addressed. Assessments are the author's judgment and may be contested.

	Unknown script	Unident. plants	Language-like stats	Section variation	No cipher works	Diagram–text link	Falsification
Proto-Romance (Cheshire)	✓	–	✓	–	–	–	–
Botanical bootstrap (Bax)	✓	✓	–	–	–	–	–
Cardan grille hoax (Rugg)	✓	–	∼	–	✓	–	–
Self-citation (Timm)	✓	–	✓	–	∼	–	–
Two-system (Bowern)	✓	–	✓	✓	–	–	–
SPBCEH (this work)	✓	∼	✓	✓	∼	✓	✓

✓ = explained; ∼ = partially explained; – = not addressed.

transcription-dependence.

5. **Computational coherence metric**: Develop a formal measure of state-sequence coherence that can replace subjective assessment in the falsification protocol.
6. **Frame token alignment**: The morphological regularity of R1 (`qok-` universal prefix) and R2 (`ch/sh` consonant stems with `-edy/-dy/-ey` suffixes) makes these clusters strong candidates for grammatical morpheme alignment against target-language function word inventories. Future work should test whether the `ch`/`sh` R2 alternation corresponds to a grammatical distinction (aspect, gender, or domain) in any candidate natural language.
7. **Vocabulary–illustration alignment**: Post-classification analysis of 20 folios with known illustration subtypes (IA-series blind test) identifies 44 content tokens with statistically significant over-representation in specific illustration types (Fisher exact $p < 0.05$, enrichment $> 3\times$). Aligning these "Rosetta" tokens against medieval Latin, Arabic, and Hebrew vocabulary lists is a concrete next step toward partial content decoding.

7 Conclusion

We have presented a functional classification framework for Voynich Manuscript EVA clusters that identifies seven provisional positional-behavioral role categories, subsequently consolidated to six under empirical model selection (see companion paper). Among the highest-frequency clusters (47 clusters, 25.8% of corpus tokens), the framework identifies statistically significant positional consistency, section-specific signal profiles consistent with independently established manuscript section types, and asymmetric falsification behavior under role inversion. These results support the presence of a structured positional grammar with directional role asymmetry in the high-frequency cluster layer—and provide a testable, reproducible framework for further investigation.

The primary strength of this contribution is a negative constraint: any viable decipherment hypothesis must account for the R2→R1 transition asymmetry ($z = +9.75$), the section-differential role profiles, and the morphological regularity of the R1 ('qok-' prefix) and R2 (three-stem system) clusters. Whether the grammar reflects a natural-language cipher, a structured recording system, or another encoding principle remains an open question that the structural analysis reported here can-

not resolve.

Acknowledgments

The author acknowledges the work of Takeshi Takahashi, Gabriel Landini, Jorge Stolfi, and René Zandbergen in producing and maintaining the EVA transliteration corpus, without which this analysis would not be possible. The Beinecke Rare Book and Manuscript Library at Yale University is acknowledged for making high-resolution scans of the manuscript publicly available.

Data Availability

All analyses are based on publicly available transliteration data from the Landini-Stolfi Interlinear file and the Zandbergen reference transliteration, accessible at https://www.voynich.nu/transcr.html. Analysis scripts, computed data tables, and pilot study outputs are publicly available at https://github.com/RobLe3/voynich-spbceh. All claims reported in this paper are traceable to specific scripts and result files in the repository (see `annex_maps/ANNEX_B8_verification_index.md`). The three papers in this series are published at https://roblemumin.com/library.html. The raw corpus file (`Lsi_ivtff_0d_v4j_fixed.txt`) must be downloaded separately from https://www.voynich.nu/transcr.html and placed in `data/` before running scripts.

References

[1] Bax, S. (2014). A proposed partial decoding of the Voynich script. Working paper, self-published. https://stephenbax.net/?p=1175.

[2] Bowern, C.L., & Lindemann, L. (2021). The linguistics of the Voynich Manuscript. *Annual Review of Linguistics*, 7, 285–308. doi:10.1146/annurev-linguistics-011619-030613

[3] Cheshire, G. (2019). The Language and Writing System of MS408 (Voynich) Explained. *Romance Studies*, 37(1), 30–67. doi:10.1080/02639904.2019.1599566. [The authoring institution retracted its press release following scholarly criticism; the journal article remains published without formal retraction.]

[4] Currier, P.H. (1976). Papers on the Voynich Manuscript. Presented at a seminar on 30 November 1976, Washington, D.C. Reproduced in: D'Imperio (1978).

[5] D'Imperio, M.E. (1978). *The Voynich Manuscript: An Elegant Enigma*. NSA/CSS Publication P-10, National Security Agency.

[6] Landini, G., & Stolfi, J. (2001). Voynich Manuscript Interlinear File (IVTFF format). Available at: https://www.voynich.nu/transcr.html.

[7] Montemurro, M.A., & Zanette, D.H. (2013). Keywords and co-occurrence patterns in the Voynich Manuscript: An information-theoretic analysis. *PLOS ONE*, 8(6), e66344. doi:10.1371/journal.pone.0066344

[8] Reddy, S., & Knight, K. (2011). 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*, pp. 78–86. Portland, OR. https://aclanthology.org/W11-1511/

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

[10] Stolfi, J. (2001). Voynich Manuscript Interlinear Archive. Instituto de Computação, Universidade Estadual de Campinas. https://www.ic.unicamp.br/~stolfi/voynich/.

[11] Takahashi, T. (1998). Complete EVA Transliteration of the Voynich Manuscript [transcriber code H]. Distributed via the Landini-Stolfi Interlinear Archive.

[12] Timm, T. (2014). How the Voynich Manuscript was created. *arXiv*:1407.6639. https://arxiv.org/abs/1407.6639

[13] Zandbergen, R. (2023). Voynich MS: Transliteration of the text. https://www.voynich.nu/transcr.html.

Appendix: Verification and Repository Traceability

Repository structure. All scripts, data, and results referenced in this paper are available at https://github.com/RobLe3/voynich-spbceh. The repository is organized as follows:

- `scripts/` — Python analysis scripts; each script is named to match the result it produces

- `data/` — Corpus token table derived from the Takahashi H IVTFF transliteration
- `results/` — JSON and CSV output files; each is referenced in `annex_maps/PAPER_TO_REPO_MAP.md`
- `pilots/` — Pilot study logs documenting methodology, findings, and negative results
- `docs/CLAIM_REGISTRY.md` — Per-claim registry with evidence class, status, and verification path
- `docs/RETRACTED_AND_FALSIFIED_CLAIMS.md` — Full record of retracted and falsified claims
- `docs/TABLE_FIGURE_TRACEABILITY.md` — Every table and figure traced to its source script

Claim verification. Each non-trivial claim in this paper carries a `% CLAIM_ID` comment in the LaTeX source. Claim IDs (P1-CLAIM-001 through P1-CLAIM-008) map to entries in `docs/CLAIM_REGISTRY.md`. To reproduce any result, run:

```
cd scripts/
python3 p1_cluster_analysis.py   # Claims 001-003, 006-007
python3 p1_3_falsification.py    # Claims 004-005
python3 p1_4_classification.py   # Claim 008
```

Verification philosophy. Every quantitative claim in this paper has a script that produces it, a JSON or CSV output that stores it, and a registry entry that records its status. Manual labels (role names: INIT, CLOSE, LINK, CONTENT, TEMPORAL, REF) are applied post-hoc to computationally derived clusters; this is noted in `docs/TABLE_FIGURE_TRACEABILITY.md` Table 1. No quantitative result was edited after computation.

Retracted and falsified findings. No Paper 1 claims have been retracted or falsified as of v1.0. The companion paper (Paper 2) carries four documented retractions (RETRACTED-001 through RETRACTED-004) and three falsifications (FALSIFIED-001 through FALSIFIED-003); these are preserved in full in `docs/RETRACTED_AND_FALSIFIED_CLAIMS.md`.

Provisional claims. No Paper 1 claims are currently provisional. All eight claims (P1-CLAIM-001 through P1-CLAIM-008) are confirmed. The companion paper (Paper 2) carries four provisional results (P2-CLAIM-010, -011, -012, and the three-layer model synthesis); current status for all claims is maintained in `docs/CLAIM_REGISTRY.md`.

Transcription sensitivity. The cross-transliteration robustness result reported in Section 5.4 (Pearson $r = 0.937$ for cluster positional percentages across Takahashi H and ZL) applies to structural R1/R2 cluster profiles, which are directly testable across transliterations (DIRECT-comparable). This paper's structural claims do not depend on Eva-T-specific glyph distinctions (the `!` marker). Token-level positional claims involving Eva-T-specific distinctions arise in the companion paper (Paper 2, P2-CLAIM-010 through -012) and are addressed there, with a full taxonomy of cross-transcription comparability statuses documented in `docs/TRANSCRIPTION_SENSITIVITY_METHOD.md`.