A Formal Grammar of the Voynich Manuscript Notation System

Complete generative specification derived from computational reverse engineering of 37,465 tokens

---

1. Alphabet and Slot Inventories

Let every token *w* in the manuscript be a concatenation of four slots: ***w* = p · g · c · s**, where $p \in$ P (prefix), $g \in$ G (gallows), $c \in$ C (core), $s \in$ S (suffix). This decomposition has zero reconstruction errors across all 7,598 types and 37,465 tokens.

Slot	Symbol	|·E|	ements
Prefix	P	8	∅, o, y, d, s, ch, sh, qo
Gallows	G	9	∅, k, t, p, f, ckh, cth, cph, cfh
Core	C	2,001	Open set (1–10 chars); ∅ = empty core
Suffix	S	33	7 families: Y, N, L, R, BARE, M, OTHER

2. Information-Theoretic Decomposition (Chain Rule)

The token entropy decomposes exactly under the chain rule with no residual:

$$H(w) = H(p) + H(g|p) + H(c|p,g) + H(s|p,g,c) = 2.788 + 1.374 + 3.622 + 2.527 = 10.311 \text{ bits}$$

Slot	H (bits)	% of H(w)	Interpretation
Prefix p	2.788	27.0%	Positional/functional class marker
Gallows g|p	1.374	13.3%	Independent encoding axis
Core c|p,g	3.622	35.1%	Content payload (opaque)
Suffix s|p,g,c	2.527	24.5%	Functional modifier
Total	**10.311**	**100.0%**	**Δ = 0.000 (exact)**

3. Slot Coupling Constraints

Slots are not independent. The joint distribution P(*p, g, c, s*) is constrained by:

$$\text{(i) Prefix} \times \text{Gallows: Cramér's V} = 0.266 \text{ (moderate coupling)}$$

$$\text{(ii) Core} \rightarrow \text{Suffix: MI(s; c|p,g)} = 0.551 \text{ bits, driven by core-final character}$$

$$\text{(iii) Prefix} \rightarrow \text{Section: MI(section; p)} = 0.154 \text{ bits (positional encoding)}$$

$$\text{(iv) Core} \rightarrow \text{Section: Section explains 8\% of core entropy}$$

The core is a single complex slot with internal character-level dependencies (MI(body, coda) = 1.310 bits) but no decomposable independent sub-slots. An optional onset modifier {ch, sh} covers 18.6% of full-core tokens with near-zero independent information (MI = 0.022 bits).

4. Two Registers

The manuscript operates in two registers under the same 4-slot grammar, distinguished by line length:

	Naming Register (labels)	Descriptive Register (body)
Trigger	Line length ≤ 2 words	Line length > 2 words
Coverage	983 lines (19%), 1,170 tokens	4,179 lines (81%), 36,295 tokens
P(prefix = o)	0.493	0.214
P(core = ∅)	0.272	0.536
H(core) share	46.5% of H(w)	34.5% of H(w)
H(suffix) share	12.7% of H(w)	24.8% of H(w)
Type uniqueness	86.6% (1-word), 98.9% (2-word)	—
Cross-folio overlap	2.0%	—
Function	Individual item names	Descriptive/procedural entries

5. Generative Line Grammar (Body Text)

$$\text{DOCUMENT} \rightarrow \text{PARAGRAPH}^{+}$$

$$\text{PARAGRAPH} \rightarrow \text{FIRST\_LINE CONTINUATION}^{*}$$

```
FIRST_LINE → PARA_MARKER MIDDLE* CLOSER
CONTINUATION → OPENER MIDDLE* CLOSER
```

Zone	P(p)	P(g ≠ ∅)	P(c = ∅)	Suffix bias
PARA_MARKER	p = ∅ (79%)	88%	13%	Y (35%), R (24%)
OPENER	d (15%), ∅ (25%), y (14%)	46%	39%	Y (31%), R (20%), N (20%)
MIDDLE	o (22%), ch (18%), qo (15%)	42%	57%	Y (41%), N (16%), L (16%)
CLOSER	∅ (26%), o (23%), d (17%)	33%	45%	M (15%), Y (39%), N (15%)

M-suffix is a line-boundary marker: 14.9% at final position vs 1.8% at penultimate (Z = 57.8). EC/FC alternation within lines follows a Bernoulli process with *p*(EC) = 0.54 (observed mean run 2.14 vs expected 2.17).

6. Transition Grammar

Adjacent tokens within a line are weakly coupled: MI(p_{n+1}; s_n) = 0.095 bits (3.5% of H(p)). The dominant transitions are:

```
Y-suffix → qo-prefix (1.80× expected) R-suffix → qo-prefix (0.33× depleted)
N-suffix → ch/sh-prefix (1.40×) BARE-suffix → ∅-prefix (1.85×)
```

Position within line is a stronger predictor of token identity than the preceding word: H(*w*|position) reduces H(*w*) by 11.7% vs 10.4% for sequential MI. Jointly, position + section explain 26.8% of token entropy.

7. Core Internal Constraints

The core is NOT decomposable into independent sub-slots. Three internal regularities exist within the single slot:

Regularity	Strength	Coverage
Last character → suffix coupling	MI = 0.551 bits (Z = 103.7)	100% of full-core tokens
ch/sh onset modifier (bench)	91.8% pedal recycling	18.6% of full-core tokens
First character → section	MI = 0.154 bits	100% of full-core tokens

8. Completeness and Limits of This Grammar

What this grammar specifies: The complete structural architecture of every well-formed VMS token and its placement within lines, paragraphs, and sections. The entropy budget is exact; no information remains structurally unaccounted for.

What this grammar does not specify: The *content* of the core slot. C contains 2,001 opaque types carrying 35.1% of all token information. Without an external key, these remain uninterpretable symbols. The suffix families (Y, N, L, R, BARE, M) have functional roles that are characterised positionally and distributionally but not semantically.

Classification: This system is not a natural language (rigid 4-slot structure, mid-word entropy peak, Bernoulli EC/FC alternation), not a simple cipher (section-specific vocabulary, folio coherence at 1.76×, label uniqueness), and not meaningless (genuine information discontinuity at word boundaries, page-specific content). It is a **structured notation system** consistent with formulaic reference works (pharmacopoeiae, herbals, astrological tables) from the 15th-century Northern Italian tradition.

Derived from computational analysis of the Voynich Manuscript (Beinecke MS 408) using the ZLZI/EVA transcription system. All statistics computed from enriched_records.pkl (37,465 tokens, 7,598 types, 5,162 lines, 226 folios, 9 sections). Segmentation rules: p70_rules_canonical.json (210 rules).