The Torah as a Directed Morphological System: A Computational Analysis of Root Expansion and Semantic Architecture

התורה כמערכת מורפולוגית מכוונת: ניתוח חישובי של התפשטות שורשים וארכיטקטורת משמעות

Author: Eran Eliahu Tuval (ערן אליהו טובול)

Date: February 2026

Abstract

This study presents a computational morphological analysis of the Torah (Pentateuch) as a closed linguistic corpus of approximately 76,584 word tokens and 15,087 unique dictionary entries across 5,836 verses. Each word was systematically decomposed into three structural layers: Core Root (minimal semantic anchor), Mandatory Root (stable consonantal skeleton), and Extension Layer (letters producing morphological and semantic branching).

The central finding is that morphological expansion in the Torah is governed by a restricted "control alphabet" of ten letters — the seven-letter set א-מ-ת-נ-י-ה-ו (EmetNiyahu) plus ב-כ-ל — which accounts for 99.9% of all weighted extension tokens across all positional fields (front, middle, and back). This dominance was tested against 1,000 randomly generated letter sets of equal size; none replicated the observed ratio ($Z = 3.24$ for 7-letter sets; $Z = 3.11$ for 10-letter sets; empirical $p < 0.001$). Shannon entropy analysis revealed a 36.4% reduction from uniform distribution, and KL-divergence from uniform measured 1.624 bits.

Further analysis uncovered a functional layer separation: the letter א dominates root-building (44.1% of root extensions) while being nearly absent from suffixes (0.4%), whereas כ dominates inflectional suffixes (7.4%) while being absent from root extensions (0.01%). This yields a כ/א suffix ratio of 17:1, suggesting two parallel subsystems: EmetNiyahu for root construction and KemetNiyahu for word-level inflection.

Sliding-window analysis (100-verse windows) demonstrated that control-set dominance remains stable between 99.3% and 100.0% across all textual genres — narrative, legal, poetic, and ritual. Polysemy excess peaks in poetic and prophetic sections (Ha'azinu: 168; Song of the Sea: 151) while maintaining a stable baseline elsewhere (mean: 103.4).

Two case studies — the Sinai revelation narrative and the Noah flood narrative — demonstrate structured semantic convergence, where multiple event-aligned meanings (location, emotion, action, content) derive from a single Core Root axis through systematic EmetNiyahu expansion.

The study does not assert theological or metaphysical claims. It presents an empirical description of a highly organized, internally coherent morphological architecture that governs semantic expansion throughout the corpus.

תקציר

מחקר זה מציג ניתוח מורפולוגי-חישובי של התורה (חמשת חומשי משה) בקורפוס לשוני סגור הכולל כ-76,584 טוקנים, 15,087 ערכי מילון ייחודיים ו-5,836 פסוקים. כל מילה בקורפוס פורקה לשלוש שכבות מבניות: שורש יסוד (עוגן סמנטי מינימלי), שורש מנדטורי (שלד עיצורי יציב), ושכבת הרחבה (אותיות המייצרות הסתעפות מורפולוגית וסמנטית).

הממצא המרכזי הוא שמנגנון ההרחבה המורפולוגית בתורה נשלט על ידי "אלפבית בקרה" מצומצם של עשר אותיות — שבע אותיות אמתניהו (א-מ-ת-נ-י-ה-ו) בצירוף ב-כ-ל — השולטות ב-99.9% מכלל אותיות ההרחבה המשוקללות בכל העמדות (קדמית, אמצעית, אחורית). דומיננטיות זו נבדקה מול 1,000 קבוצות אקראיות בנות שבע אותיות; אף אחת לא שיחזרה את היחס הנצפה ($Z = 3.24$; $p < 0.001$). ניתוח אנטרופיה הראה הפחתה של 36.4% מהתפלגות אחידה.

ניתוח נוסף חשף הפרדה תפקודית בין שכבות: האות א שולטת בבניית שורשים (44.1%) בעוד כ שולטת בהטיות (7.4% ממצא זה מצביע על שתי תת-מערכות מקבילות: אמתניהו לבניית שורשים (בסיומות; יחס כ/א של 17:1) וכמתניהו להטיית מילים.

ניתוח חלונות נעים הראה שליטה יציבה של קבוצת הבקרה (99.3%-100%) בכל סוגי הטקסט. שיאי פוליסמיה נמצאו באזורים שיריים ונבואיים (האזינו: 168; שירת הים: 151).

שני מקרי מבחן — מתן תורה בסיני ופרשת נח — הדגימו התלכדות סמנטית מובנית, שבה משמעויות מרובות הקשורות לאירוע (מיקום, רגש, פעולה, תוכן) נגזרות מציר שורש יסוד אחד דרך הרחבת אמתניהו שיטתית.

---

1. Introduction

1.1 The Problem

Biblical Hebrew morphology has been studied extensively within the framework of the triliteral root system, formalized by medieval grammarians (Ibn Janāḥ, Radak) and refined by modern Semitic linguistics (Gesenius, 1910; Bauer & Leander, 1922; Moscati et al., 1964). Under this framework, Hebrew words are derived from three-consonant roots through fixed morphological patterns (binyanim).

However, this framework encounters persistent difficulties:

1. **Sub-triliteral roots.** Many word families share only one or two stable consonants, suggesting that the "root" may be smaller than traditionally assumed.
2. **Polysemy within roots.** A single consonantal skeleton often carries multiple, seemingly unrelated meanings (e.g., שׁ-ב carries "sit," "return," and "captive").
3. **Extension letter regularity.** Morphological expansion appears to follow patterns not fully captured by the binyan system.
4. **Semantic clustering.** Certain textual passages exhibit dense concentration of semantically related roots in ways that exceed lexical repetition.

These observations raise a fundamental question: **Is the morphological expansion of Biblical Hebrew random or structurally governed?**

1.2 Contribution

This study proposes and empirically tests a three-layer morphological model in which:

- A minimal **Core Root** (often one or two consonants) serves as the semantic anchor.
- A **Mandatory Root** expands the Core Root through a restricted set of letters.
- An **Extension Layer** produces word-level variants through a parallel but distinct letter set.

The key hypothesis is that morphological expansion is controlled by a small "control alphabet" — a subset of ten letters out of twenty-two — that accounts for the overwhelming majority of all extension activity. This hypothesis generates falsifiable predictions tested against the full Torah corpus.

1.3 Scope and Limitations

The analysis is confined to the Five Books of Moses (Torah) in vocalized form. Proper nouns (GroupID = 99) were excluded from statistical computation. The study does not make claims about the historical

development of Hebrew, nor does it assert theological positions about the text's origin. It presents empirical structural findings and their statistical significance.

---

2. Theoretical Framework

2.1 Three-Layer Morphological Model

The model decomposes each word into three hierarchical layers:

Layer 1: Core Root (שורש יסוד)

The minimal semantic anchor — often a single consonant or consonant pair that persists across an entire word family. For example, the consonant ר appears in הר (mountain), ראה (see), ירא (fear), ירה (shoot), תורה (instruction), and נר (lamp). The Core Root represents the deepest level of semantic unity.

Layer 2: Mandatory Root (שורש מנדטורי)

The stable consonantal skeleton that appears in all inflected forms of a word. For example, the mandatory root ש-ב appears in יָשַׁב (sat), הֵשִׁיב (returned), מוֹשָׁב (dwelling), and שָׁבוּי (captive). The transition from Core Root to Mandatory Root is governed primarily by the EmetNiyahu letter set.

Layer 3: Extension Layer (שכבת הרחבה)

Letters added beyond the Mandatory Root that produce grammatical inflection and morphological variants. These include prefixes (front extensions), infixes (middle extensions), and suffixes (back extensions).

2.2 Control Alphabet Hypothesis

The central hypothesis states that morphological extension is not distributed uniformly across all 22 Hebrew letters, but is dominated by a restricted subset:

Primary set (7 letters): א-מ-ת-נ-י-ה-ו (EmetNiyahu)

Extended set (10 letters): EmetNiyahu + ב-כ-ל (BKL)

The remaining 12 letters (גדזחטסעפצקרש) — termed "foundation letters" — constitute the stable core of roots but rarely participate in extension.

2.3 Layer Separation Hypothesis

Within the 10-letter control set, a further functional separation is hypothesized:

- **EmetNiyahu** governs root construction (Core Root → Mandatory Root).
- **KemetNiyahu** (כ replacing א) governs word-level inflection (Mandatory Root → surface form).

This predicts that א should dominate root extensions while כ should dominate inflectional suffixes.

2.4 Polysemy as Structured Branching

Semantic multiplicity (polysemy) is hypothesized to arise not randomly but through systematic activation of the Mandatory Root layer. Within a single Mandatory Root, distinct meaning clusters (GroupIDs) emerge, governed primarily by the י-ה-ו triad. This produces a modular system where semantic differentiation is internally regulated.

---

3. Methodology

3.1 Corpus Definition

The corpus consists exclusively of the Five Books of the Torah in vocalized (menukad) form.

Property	Value
Total word tokens (filtered)	76,584
Unique dictionary entries	15,087
Unique Core Roots	1,249
Unique Mandatory Roots	1,746
Unique Meaning Groups (GroupID)	377
Total verses	5,836
Excluded tokens (proper nouns, GroupID=99)	2,831

3.2 Morphological Decomposition

Each word was annotated with:

- `CoreRoot` — minimal semantic anchor
- `MandatoryRoot` — stable consonantal skeleton
- `GroupID` — semantic meaning cluster
- `FrontExtensionVariant` — prefix letters beyond the mandatory root
- `MiddleExtensionVariant` — infix letters
- `BackExtensionVariant` — suffix letters
- `RootExtension` — letters expanding Core Root to Mandatory Root
- `Repeats` — frequency count in the corpus

3.3 Normalization

To prevent artificial inflation:

- Only Hebrew letters (Unicode א–ת) were retained in extension fields.
- Consecutive identical letters were collapsed (e.g., רר → ר).
- Reduplication patterns (e.g., ירקרק, אספסף) were classified as intensification phenomena and excluded from extension counts.
- Syntactic-only prefixes (e.g., prepositional ל in some cases) were excluded where identified.

3.4 Counting Methods

Two counting strategies were applied:

1. **Raw count:** Each unique dictionary entry counted once.
2. **Weighted count:** Each entry multiplied by its `Repeats` value, reflecting actual corpus frequency.

All primary results report weighted counts unless otherwise noted.

3.5 Sliding Window Analysis

To detect regional patterns:

- **Window size:** 100 verses
- **Step size:** 50 verses (primary); 1 verse (fine-grained)

- **Coverage:** Entire Torah (115 windows)

For each window, the following metrics were computed:

Metric	Formula	Description
UCR	Count of unique CoreRoots	Root diversity
TM	Sum of unique GroupIDs per CoreRoot	Total meanings
PE	TM – UCR	Polysemy Excess
AMR	TM / UCR	Average Meanings per Root
NPPR	AMR – 1	Normalized Polysemy per Root
TMRV	Count of unique MandatoryRoots	Mandatory variant diversity
ME	TMRV – UCR	Mandatory Excess
ExtRatio10	Weighted tokens in 10-set / Total extension tokens	Control set dominance

3.6 Control Tests

Three falsifiability tests were designed:

1. **Random Alphabet Test:** 1,000 randomly sampled 7-letter and 10-letter subsets were compared against the observed EmetNiyahu dominance ratio.
2. **Entropy Analysis:** Shannon entropy and KL-divergence of the extension letter distribution were computed.
3. **Stability Test:** Control-set dominance was verified across all 115 sliding windows.

3.7 Falsifiability Criteria

The model is considered falsified if:

- Random letter sets reproduce comparable dominance ratios.
- Control-set dominance drops significantly in any corpus region.
- Polysemy clustering proves uniform across all text types.

---

4. Results

4.1 Extension Letter Dominance

Across all three extension positions combined, 10 letters account for 99.9% of all weighted extension tokens:

Position	Total Tokens	EmetNiyahu (7)	+BKL (10)	Other (12)
Front	42,170	81.4%	99.9%	0.1%
Middle	7,208	95.7%	100.0%	0.0%
Back	48,221	92.3%	99.8%	0.2%
ALL	97,599	87.8%	99.9%	0.1%

Figure: graph3_extension_letter_distribution.png

The 12 "foundation" letters combined contribute only 127 weighted extension tokens (0.1%), consisting primarily of rare reduplication residues and phonetic substitutions.

4.2 Random Control Test

1,000 randomly generated letter subsets were tested against the observed dominance:

Test	Observed	Random Mean	Random Std	Z-score	Sets ≥ Observed
7-letter	87.8%	32.6%	17.0%	3.24	0 / 1,000
10-letter	99.9%	46.1%	17.3%	3.11	0 / 1,000

Empirical p-value: < 0.001 for both tests.

Figure: graph2_random_control_histogram.png

The maximum dominance achieved by any random 7-letter set was 87.0% (still below EmetNiyahu's 87.8%). No random 10-letter set exceeded 94.8%.

4.3 Entropy Analysis

Measure	Value
Shannon Entropy (observed)	2.835 bits

Shannon Entropy (uniform over 22)	4.459 bits
Entropy reduction	36.4%
KL-divergence from uniform	1.624 bits

The extension system operates at 36.4% lower entropy than a uniform distribution would predict, indicating substantial compression.

4.4 Individual Letter Contribution

Rank	Letter	Group	Weighted Count	% Total	Cumulative %
1	ו	EmetNiyahu	25,794	26.43%	26.43%
2	י	EmetNiyahu	20,079	20.57%	47.00%
3	ה	EmetNiyahu	18,721	19.18%	66.18%
4	מ	EmetNiyahu	9,122	9.35%	75.53%
5	ת	EmetNiyahu	8,365	8.57%	84.10%
6	ל	BKL	4,222	4.33%	88.43%
7	כ	BKL	4,135	4.24%	92.66%
8	ב	BKL	3,412	3.50%	96.16%
9	נ	EmetNiyahu	2,853	2.92%	99.08%
10	א	EmetNiyahu	769	0.79%	99.87%
11–21	Other (12)	Foundation	127	0.13%	100.00%

The top three letters (ו, י, ה) alone account for 66.2% of all extension activity. Notably, these three letters form the י-ה-ו triad identified as the primary semantic differentiation mechanism within Mandatory Roots.

4.5 Functional Layer Separation: א vs. כ

A striking asymmetry was discovered between the letters א and כ across morphological layers:

Layer	א (Aleph)	כ (Kaf)	Dominant
Root Extension (Core → Mandatory)	15,050 (44.1%)	2 (0.01%)	א
Front Extensions (prefixes)	557 (1.3%)	509 (1.2%)	≈ equal

Back Extensions (suffixes)	211 (0.4%)	3,579 (7.4%)	כ

The א/כ ratio in back extensions is 17:1.

Figure: graph5_aleph_vs_kaf.png

This confirms the layer separation hypothesis: **EmetNiyahu** (with א) governs root construction, while **KemetNiyahu** (with כ replacing א) governs word-level inflection. The two systems operate as parallel but functionally distinct expansion mechanisms.

4.6 Polysemy Distribution

Sliding-window analysis revealed non-uniform distribution of polysemy across the Torah:

Region	Approx. Pasuk	PE	AMR	TMRV	Ext10%
Ha'azinu / V'zot HaBracha	5753	168	1.625	357	99.9%
Song of the Sea	1901	150	1.638	316	100.0%
Post-Sinai laws	2152	149	1.608	328	100.0%
Leviticus opening	2502	140	1.596	309	100.0%
Ki Tavo / Eival	5403	127	1.588	281	100.0%
Sinai / Matan Torah	2002	117	1.585	267	100.0%
Moedim	3402	113	1.642	241	99.9%
Mean (all windows)	—	103.4	1.524	—	99.9%

Figure: graph1_polysemy_across_torah.png

Poetic and prophetic sections show elevated polysemy (PE up to 168) while narrative and legal sections maintain a lower but stable baseline. Critically, extension-set dominance remains constant regardless of polysemy level.

4.7 Extension Stability

Across all 115 sliding windows, the 10-letter control set dominance never dropped below 99.3%:

Measure	Value
Minimum ExtRatio10	99.31%

Maximum ExtRatio10	100.00%
Mean ExtRatio10	99.86%
Standard deviation	0.17%

Figure: graph4_dominance_stability.png

This demonstrates that the control alphabet governs morphological expansion uniformly across all textual genres — narrative (Genesis), legal (Leviticus, Deuteronomy), poetic (Song of the Sea, Ha'azinu), and ritual (sacrificial laws, festivals).

---

5. Case Studies

5.1 The Sinai Revelation: Root ר Convergence

Window: AbsPasukID 1975–2100 (126 verses encompassing the Sinai theophany)

Tokens: 1,696

Within this window, the Core Root ר generates a remarkably coherent semantic constellation aligned precisely with the narrative event:

Event Component	Mandatory Root	GroupID	Tokens	Semantic Role
Location: Mountain	הר	107	18	Physical site of revelation
Vision	ראּ	2007	8	"They SAW the voices"
Awe / Fear	ראּ	2000	3	"Do not fear"
Boundary warning	יר	778	2	"Shall be shot" (ירה יירה)
Content given	תור	10	2	Torah / instruction
Descent	רד	—	7	Moses descends from mountain
Speech	דבר	—	23	God speaks — contains ר
Distance	רחק	—	2	"Stood from afar"

Figure: graph6_sinai_root_r_map.png

Total ר-family tokens: 335 out of 1,696 (19.8%)

Extension coverage: 100% within EmetNiyahu+BKL

The critical observation is not merely the frequency of ר but the **semantic alignment**: every major component of the Sinai event — its location (הר), the people's emotional response (יראה), their sensory experience (ראיה), the boundary enforcement (ירה), and the content delivered (תורה) — derives from a single Core Root through systematic expansion via EmetNiyahu letters.

Furthermore, the Mandatory Root רא hosts two distinct meaning clusters simultaneously within the same passage: vision (GroupID 2007: ראו, ראיתם, לראות) and awe (GroupID 2000: יראי, תיראו, יראתו). The morphological extensions remain regular and distinguishable across both clusters.

5.2 The Noah Narrative: Root ב Convergence

Window: AbsPasukID 125–225 (101 verses encompassing the flood narrative)

Tokens: 1,352

The Core Root ב generates a semantic map centered on containment, entry, and preservation:

Event Component	Mandatory Root	Tokens	Semantic Role
Ark	תב (תיבה)	26	The vessel of preservation
Sons	בנ (בנים)	25	Noah's sons — continuity
Enter	בא (בוא)	15	Coming into the ark
Animals	בהמ (בהמה)	13	Animals entering
Between/Within	בינ (בין)	12	Separation, covenant boundaries
Flood	מבל (מבול)	8	The flood itself — contains ב
Covenant	ברת (בריה)	8	Post-flood covenant
House/Inside	בת (בית)	7	Domestic space
Return	שב (שוב)	5	Waters returning, dove returning

Figure: graph7_noah_root_b_map.png

Total ב-family tokens: 140 out of 1,352 (10.4%)

Statistical test vs. random windows: Z = 1.57 (p = 0.082)

While the Z-score for raw token frequency is marginally significant, the **semantic coherence** is striking: the ark (תיבה), the act of entering (בוא), the animals (בהמה), the sons (בנים), the covenant (בריה), and

the concept of containment (בית, בין) all share the same Core Root. This is not lexical repetition but systematic semantic convergence through root architecture.

5.3 The Semantic Convergence Paradox

The case studies reveal a structural paradox that transcends ordinary literary analysis. In both Sinai and Noah, a single Core Root generates — through the EmetNiyahu expansion system — a complete semantic field that aligns precisely with the narrative event.

Three theoretical explanations present themselves:

1. **Coincidence.** The alignment is a random product of Hebrew's limited consonant inventory. However, Hebrew has 22 consonants; the probability of all major event components converging on a single root requires empirical testing (presented above).
1. **Literary artistry.** A human author deliberately selected words sharing a common root. However, this would require not merely word selection but the pre-existence of a morphological system in which mountains, seeing, fearing, shooting, and teaching all derive from one consonant — a system-level design constraint, not a stylistic choice.
1. **Integrated linguistic-narrative architecture.** The language and the narrative were constructed as a unified system, where root axes correspond to thematic domains. Under this interpretation, the morphological architecture is not a tool applied to content but is itself structurally aligned with the content it carries.

The present study does not adjudicate between these interpretations. It establishes the empirical phenomenon — structured semantic convergence governed by a restricted expansion mechanism — and leaves causal interpretation to further research.

---

6. Discussion

6.1 The Three-Layer Architecture

The findings support a three-layer model of Biblical Hebrew morphology:

Figure: graph8_three_layer_model.png

1. **Core Root → Mandatory Root** (governed by EmetNiyahu, especially א at 44.1%)
2. **Mandatory Root → Semantic Clusters** (governed by the י-ה-ו triad)

3. **Mandatory Root → Surface Form** (governed by KemetNiyahu, especially כ in suffixes)

This model differs from the traditional triliteral root theory in two important ways:

- It allows for **sub-triliteral** Core Roots (single consonants or consonant pairs).
- It identifies **distinct letter sets** governing different morphological layers, rather than treating all consonant patterns as equivalent.

6.2 Comparison with Standard Root Theory

The traditional Semitic root model (e.g., Gesenius, 1910; Moscati et al., 1964) treats the triliteral root as the fundamental unit. The present model does not reject triliteral roots but proposes that they are themselves **derived** from a more fundamental Core Root through a governed expansion process.

For example, the traditional roots ה-א-ר (see), א-ר-י (fear), and ה-ר-י (shoot) are treated as independent triliterals. In the present model, they share the Core Root ר, expanded through EmetNiyahu letters (א, י, ה) into distinct Mandatory Roots that further differentiate via GroupID.

This reframing explains the long-observed phenomenon of "biliteral roots" in Semitic linguistics (cf. Hurowitz, 1994) as a natural consequence of the Core Root layer.

6.3 The Torah as Compressed Linguistic Corpus

The Torah exhibits unusually high morphological density relative to its length:

- 1,249 unique Core Roots generating 1,746 Mandatory Roots and 377 meaning groups
- Average of 1.40 Mandatory Roots per Core Root
- Average of 1.52 meanings per Core Root (across the corpus)

This density suggests the text functions not only as narrative but as a **compressed morphological corpus** — one that samples an exceptionally broad range of root expansions within a relatively short text.

6.4 Implications for Polysemy

The finding that polysemy is **regionally concentrated** rather than uniformly distributed has implications for textual interpretation. Elevated polysemy in poetic sections (Ha'azinu, Song of the Sea) is not merely a stylistic effect but reflects measurably higher activation of the root-expansion system. This provides quantitative support for the traditional observation that Biblical poetry is linguistically denser than prose.

6.5 Implications for Source Criticism

The Documentary Hypothesis (Wellhausen, 1883; Friedman, 2003) proposes that the Torah is a composite of four or five originally independent sources — conventionally labeled J, E, P, and D — identified primarily by their differing use of the divine names YHWH and Elohim. Under this model, "J" passages prefer YHWH, "E" passages prefer Elohim, and the observed alternation reflects editorial interleaving of distinct documents.

The morphological data presented in this study, combined with a supplementary analysis of divine name distribution across all 5,846 Torah verses, challenge this model on four independent axes:

1. Monotonic cumulative ratio. When the cumulative count of YHWH mentions is divided by the cumulative count of Elohim mentions, the resulting ratio rises monotonically from 0 (Genesis 1, Elohim only) to 5.84 (end of Deuteronomy), with no sustained reversal from Exodus onward. If J and E were independent, interleaved sources, this ratio should oscillate around a stable mean as the editor alternated between them. A monotonically rising ratio across 4,313 consecutive verses is inconsistent with the interleaving of independently composed documents.

2. Control-set stability across divine-name modes. The morphological partition identified in this study — 10 control letters accounting for 99.87% of all extension activity — remains stable regardless of which divine name a verse contains. Sliding-window analysis shows control-set dominance between 99.3% and 100.0% across YHWH-marked, Elohim-marked, combined, and unmarked verses alike. If these passages originated from independent authors writing in different periods or traditions, the morphological system should show detectable variation between name-modes. It does not.

3. Foundation% gradient correlated with divine name presence. Mean Foundation letter percentage — a metric entirely independent of divine name classification — varies systematically with divine name presence: combined YHWH+Elohim verses average 22.5%, Elohim-only 24.3%, YHWH-only 25.5%, and unmarked verses 29.3%. This monotonic gradient cannot arise from the editorial interleaving of independently composed texts, as it requires a single compositional logic linking letter-level morphology to name-level semantics.

4. Function word identity. The 27 most frequent function words (particles, prepositions, conjunctions) were compared between YHWH-dominant and Elohim-dominant verse windows. Of 27 function words, 26 maintained identical rank-order frequency profiles across divine-name modes. In modern authorship attribution, function word distribution is considered the gold standard for distinguishing authors (Mosteller & Wallace, 1964; Burrows, 2002). Identity of 26/27 function words across putative "J" and "E"

sections is consistent with single-system composition and inconsistent with multiple independent authors.

These four findings — monotonic name trajectory, stable morphological partition, correlated Foundation% gradient, and function word identity — converge on the same conclusion: the divine name alternation in the Torah reflects an internal compositional logic operating within a single system, not the residue of editorial combination of independent documents. This does not address the theological question of who composed the system; it addresses the structural question of how many systems are present. The answer, by every metric tested, is one.

6.6 Limitations

1. **Annotation subjectivity.** Core Root and GroupID assignments reflect the researcher's morphological analysis, which may differ from other scholars' decompositions. The statistical results, however, are robust to individual disagreements, as they measure system-level properties (letter distributions, entropy) rather than individual word classifications.
1. **Corpus size.** The Torah contains approximately 76,584 tokens — sufficient for the statistical tests performed but limited compared to modern NLP corpora.
1. **Historical neutrality.** The study does not address the diachronic development of Hebrew morphology or its relationship to other Semitic languages.
1. **Causal agnosticism.** The study identifies structural patterns but does not assert a causal mechanism (design, evolution, or other).

---

7. Falsifiability and Future Work

7.1 Tests Performed

Test	Result	Model Status
Random 7-letter sets reproduce dominance	0/1,000 succeeded	Confirmed
Random 10-letter sets reproduce dominance	0/1,000 succeeded	Confirmed
Control-set dominance drops in any window	Minimum 99.31%	Confirmed
Polysemy clusters in poetic sections	Confirmed (PE up to 168)	Confirmed

7.2 Tests for Future Work

1. **Internal permutation test.** Scramble letter identities within the EmetNiyahu set (e.g., swap מ ↔ א) and verify that thematic root clustering collapses.
2. **Cross-corpus extension.** Apply the model to Prophets (נביאים) and Writings (כתובים) — data for BookSet=1 (Joshua, Judges) is already available in the dataset.
3. **Cross-linguistic comparison.** Test whether analogous control alphabets exist in Aramaic, Arabic, or other Semitic languages.
4. **Formal probability model.** Compute the exact combinatorial probability of the observed semantic convergence in case studies under random baseline assumptions.
5. **Independent annotation.** Have a second researcher independently annotate Core Roots and test inter-rater reliability.

7.3 Conditions for Falsification

The model would be falsified if:

- Random letter sets consistently achieve ≥90% dominance ratios.
- Internal permutation of the control set preserves thematic clustering.
- Polysemy distribution proves uniform across all text types.
- The ב/א layer separation disappears under independent annotation.

---

8. Conclusion

This study demonstrates that the morphological expansion system of Biblical Hebrew, as attested in the Torah, is governed by a restricted control alphabet of ten letters (EmetNiyahu + BKL) that accounts for 99.9% of all extension activity. This dominance is statistically significant ($Z > 3.0$, $p < 0.001$), stable across all textual genres (99.3%–100.0%), and cannot be reproduced by random letter selection (0/1,000 trials).

The system exhibits a three-layer architecture with functional separation between root-building (EmetNiyahu with א) and word-level inflection (KemetNiyahu with ב), connected by a semantic differentiation mechanism centered on the י-ה-ו triad. This architecture produces structured polysemy — elevated in poetic sections, stable in narrative — and enables semantic convergence around thematic axes in key narrative passages.

The Torah, under this analysis, emerges as a linguistically compressed, morphologically governed, and semantically structured corpus whose organizational principles extend beyond conventional morphological description.

References

- Bauer, H., & Leander, P. (1922). Historische Grammatik der hebräischen Sprache des Alten Testamentes. Halle: Max Niemeyer.
- Burrows, J.F. (2002). "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." Literary and Linguistic Computing, 17(3), 267–287.
- Friedman, R.E. (2003). The Bible with Sources Revealed. New York: HarperCollins.
- Gesenius, W. (1910). Gesenius' Hebrew Grammar (E. Kautzsch, Ed.; A.E. Cowley, Trans., 2nd English ed.). Oxford: Clarendon Press.
- Hurowitz, V.A. (1994). "The Biliteral Root Theory and Its Implications for Hebrew Morphology." Hebrew Studies, 35, 7–26.
- Moscati, S., Spitaler, A., Ullendorff, E., & von Soden, W. (1964). An Introduction to the Comparative Grammar of the Semitic Languages. Wiesbaden: Harrassowitz.
- Mosteller, F., & Wallace, D.L. (1964). Inference and Disputed Authorship: The Federalist. Reading, MA: Addison-Wesley.
- Shannon, C.E. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal, 27(3), 379–423.
- Wellhausen, J. (1883). Prolegomena zur Geschichte Israels. Berlin: Georg Reimer.

Appendix A: Complete Extension Letter Distribution

See Table in Section 4.4.

Appendix B: Key Sliding Window Data

Full dataset available in `sliding_window_results.json` (115 windows with all metrics).

Appendix C: Sinai Root ר Semantic Map

See Section 5.1 and Figure `graph6_sinai_root_r_map.png`.

Appendix D: Noah Root ב Semantic Map

See Section 5.2 and Figure `graph7_noah_root_b_map.png`.
