The Irreversibility Depth of Natural Language

Richard Hoekstra

April 2026

cs.CL (Computational Linguistics)

Abstract

We introduce the *irreversibility depth* D* of an information source, defined via the Helmholtz-Hodge decomposition of the empirical Markov transition field on de Bruijn context graphs. At context depth D, the byte-level transition structure decomposes uniquely into an exact (reversible) component and a harmonic (irreversible) component. The harmonic fraction f(D) decays as a stretched exponential with exponent k = 1.91, and D* marks the extrapolated depth at which the harmonic fraction vanishes. We measure D* across 49 natural languages, 7 programming languages, synthetic music, and genomic DNA. For English Wikipedia, D* = 7.5 bytes; a formality ladder spans from MIDI music (D* ~ 4) through programming languages (D* ~ 7) to literary prose (D* ~ 10) and E. coli DNA (D* ~ 165 bases). A cross-linguistic atlas of 49 languages reveals three principal components explaining 83% of variance in 8 topological invariants. An n-gram asymmetry catalogue shows that function-word boundaries carry the strongest cycle currents, with a scale crossover from morphological (D = 4) to collocational patterns (D = 7). We establish an anti-dispersion law: strong cycles die first under depth filtration (survivors mean |asym| = 5.96 vs. non-survivors 6.10). GPT-2 generated text matches real English at D <= 3 but is anomalously rigid at D = 5 — a structural difference invisible to perplexity. We confirm time-reversal invariance of the harmonic energy (0.01%–1.3% deviation) and define the emergent irreversibility Phi_Hodge, which is positive for similar sources (+0.014) and negative for dissimilar sources (-0.036). Three falsification results bound the framework. D* classifies information sources by their time-asymmetry structure without domain-specific knowledge and provides a new diagnostic for language model quality.

---

1. Introduction

1.1 Motivation

Every information source has a time arrow. Reading “the cat sat on the” forward, “the” after “on” is highly predictable; reading backward, “on” after “the” is far less constrained. This directional asymmetry is fundamental to how language encodes meaning, yet it has no standard measurement. Perplexity captures predictability but not directionality. Entropy rate is symmetric by construction in the stationary limit. We lack a quantity that answers: *at what scale does the time arrow of a source vanish?*

We propose such a quantity: the irreversibility depth D*, derived from the Helmholtz-Hodge decomposition of the Markov transition field on de Bruijn context graphs. The construction requires no parser, no grammar, no domain knowledge — only a byte stream and a sparse linear solve.

1.2 The Hodge decomposition on graphs

The Helmholtz-Hodge decomposition is a classical result in discrete differential geometry. Any scalar function on the edges of a graph (a 1-cochain) decomposes uniquely into an exact component (the gradient of a vertex potential) and a harmonic component (in the kernel of the graph Laplacian). On a 1-complex (graph without higher-dimensional faces), the co-exact component vanishes and the decomposition is bipartite: exact + harmonic.

The exact component represents reversible structure — edge values that can be expressed as potential differences between vertices. The harmonic component represents irreversible structure — cycle currents that flow preferentially in one direction and cannot be reduced to any vertex potential.

1.3 Our construction

Given a byte stream and a fixed context depth D, we construct the order-D de Bruijn graph whose vertices are observed D-grams and whose edges are empirical transitions. The Markov transition field A(c -> c') = log P_emp(c'|c) is a natural 1-cochain that encodes the full statistical structure of the source at order D. Decomposing A into its exact and harmonic components measures how much of the transition structure is reversible (a vertex gradient) versus irreversible (a cycle current).

We define the harmonic fraction f(D) and show that it decays with D as a stretched exponential. The irreversibility depth D* is the extrapolated depth where f -> 0. Our central finding is that D* is a robust, informative invariant of information sources.

1.4 Related work

The Hodge decomposition on graphs was formalized for data analysis by Jiang et al. (2011) in the context of ranking from pairwise comparisons. Schaub et al. (2020) applied it to edge flows on simplicial complexes. The de Bruijn graph is standard in genomic assembly (Compeau et al., 2011) and has been used for formal language analysis (Shallit, 2009), but its Hodge structure has not been studied.

Time-reversal asymmetry in Markov chains connects to non-equilibrium statistical mechanics (Jiang et al., 2004) and the cycle decomposition of stationary distributions (Kalpazidou, 2006). Our contribution is a practical, byte-level measurement of this asymmetry across context scales and source types.

1.5 Contributions

1. The irreversibility depth D* as a new invariant of information sources.
2. A stretched-exponential decay law f(D) = exp(-(D/D_0)^k) with k ~ 2 for natural language.
3. A formality ladder spanning from D*(music) ~ 4 to D*(DNA) ~ 165.
4. A 49-language cross-linguistic atlas with PCA on 8 topological invariants.
5. An anti-dispersion law: strong cycles die first under depth filtration.
6. Detection of structural differences between GPT-2 and natural text invisible to perplexity.
7. The emergent irreversibility Phi_Hodge as a measure of source coupling.
8. Three falsification results bounding the scope of the framework.

---

2. Method

2.1 The Markov transition field

Given a corpus of N bytes over alphabet Sigma (|Sigma| = 256 for text, 4 for DNA), fix a context depth D. The order-D de Bruijn graph G_D = (V, E) has:

- Vertices V = {c in Sigmaˆ D : count(c) >= tau} (observed D-grams above a frequency threshold tau)
- Directed edges E = {(c, c') : c' = c[1:] + b for some b in Sigma, count(c -> c') > 0}

The Markov transition field is the scalar 1-cochain:

```
A(c -> c') = log P_emp(c'|c) = log [count(c -> c') / count(c)]
```

This field is not a priori exact. It is not the coboundary of any vertex function because the empirical Markov chain is generically irreversible.

2.2 Hodge decomposition

The oriented incidence matrix d_0 in Rˆ{|E| x |V|} has entries (d_0)_{e,v} = +1 if v = head(e), -1 if v = tail(e). The weighted graph Laplacian is L = d_0ˆT W d_0 where W = diag(w_e) with edge weights w_e = count(e).

The exact component is A_exact = d_0 phi where phi solves the weighted normal equation:

```
L phi = d_0^T W A
```

We add a Tikhonov ridge (10ˆ{-8} * I) to handle the constant nullspace. The harmonic component is the residual:

```
A_harmonic = A - A_exact
```

The harmonic fraction is the ratio of weighted energies:

```
f(D) = sum_e [w_e * A_harmonic(e)^2] / sum_e [w_e * A(e)^2]
```

Orthogonality between the exact and harmonic components is verified numerically at every depth: the cross-term <A_exact, A_harmonic>_w is less than 10ˆ{-9} of ||A||ˆ2.

2.3 The irreversibility depth

We fit f(D) to a stretched exponential:

```
f(D) = exp(-(D/D_0)^k)
```

and define the irreversibility depth as:

```
D* = D_0 * (-ln epsilon)^{1/k}
```

for threshold epsilon = 0.01. We also report the simpler linear extrapolation D*_lin where the linear fit of f(D) crosses zero; the two estimates agree within 10%.

2.4 The n-gram asymmetry

For each (D+1)-gram s with frequency above threshold, the time-reversal asymmetry is:

```
asym(s) = log P(s[D] | s[0:D]) - log P(s[0] | s[1:D+1])
        = log [count(s[1:D+1]) / count(s[0:D])]
```

This measures local irreversibility: how much more predictable the last byte is from its prefix than the first byte is from its suffix.

2.5 Emergent irreversibility

Given two sources A and B, interleave them at chunk size C to produce a coupled source AB. Define:

```
Phi_Hodge = f(AB) - max(f(A), f(B))
```

Positive Phi indicates emergent cycle currents at chunk boundaries (constructive interference); negative Phi indicates destructive interference.

---

3. Results

3.1 D-sweep on English Wikipedia

On enwik8 (first 2 MB, threshold tau = 8):

D		V			E
2	3,189	34,136	30,948	0.813	4e-12
3	13,642	75,881	62,240	0.689	3e-11
4	27,997	95,255	67,259	0.528	1e-10
5	38,645	88,232	49,588	0.367	4e-10

At D = 2, the source is 81% harmonic — overwhelmingly irreversible. By D = 5, only 37% remains harmonic. The Hodge orthogonality is numerically clean at every depth.

3.2 Stretched exponential fit

Four candidate models were fitted to the f(D) decay:

Model	Form	Parameters	Quality
Linear	1.124 - 0.150 * D	zero at D ~ 7.50	–
Exponential	A * exp(-D/D_0)	A = 1.45, D_0 = 3.77	poor (predicts f > 1 at D = 2)
Power law	C * D^{-beta}	C = 1.57, beta = 0.84	poor
Stretched exp	exp(-(D/D_0)^k)	**k = 1.91, D_0 = 4.96**	**best**

The stretched exponential with k = 1.91 ~ 2 fits best, implying a Gaussian-like distribution of dependency scales. Both the stretched exponential and linear fits agree on D* ~ 7.5 bytes as the irreversibility depth of English. The pure exponential is qualitatively wrong (super-exponential decay rules it out). The power law cannot match the concavity.

The irreversibility depth **D* = 7.5 bytes** means that beyond approximately 8 bytes of context, the Markov chain over English text becomes effectively reversible — the time arrow vanishes.

3.3 The formality ladder: 7 corpora

The same measurement on seven text and code corpora (2 MB each, tau = 4):

Source	Type	D=2	D=3	D=5	D*
Pratchett	literary fiction	0.808	0.694	0.486	**9.5**
enwik8	encyclopedia	0.812	0.693	0.410	**8.0**
Legal	legal text	0.796	0.667	0.392	**7.9**
Python	dynamic code	0.778	0.592	0.310	**7.0**
Ruby	dynamic code	0.779	0.599	0.304	**6.8**
Zig	static code	0.755	0.558	0.282	**6.6**
Lean	proof code	0.748	0.552	0.232	**6.3**

Three clusters emerge:

- **Natural language** (D* ~ 8–10): literary fiction, encyclopedia, legal text. The most irreversible — the richest cycle-current structure.
- **Dynamic code** (D* ~ 7): Python, Ruby. Intermediate.
- **Formal languages** (D* ~ 6–7): Zig, Lean. The most reversible — the most constrained syntax.

The ranking is consistent at every depth and matches linguistic intuition about "formality" — but is derived purely from byte-level topology, without any parser or grammar.

Within natural language, the spread is notable: literary fiction (D* = 9.5) requires nearly 10 bytes of context before its time arrow vanishes, while legal English (D* = 7.9) is more formulaic. The 1.6-byte gap between Pratchett and legal text within the same language is driven entirely by genre and register, not by the grammar of English.

A sliding-window analysis (199 windows of 50 KB across 5 MB of enwik8 at D = 3) confirms intra-corpus variation: f_harm ranges from 0.477 (structured biographical lists) to 0.631 (dense encyclopedic prose), a span of 0.15 exceeding 6 standard deviations.

3.4 Cross-linguistic atlas: 49 languages

We extend the measurement to 49 natural languages using Wikipedia text (200 KB per language, D = 1..5). To correct for UTF-8 encoding differences, we compute D*_chars = D*_bytes / bpc, where bpc is the bytes-per-character ratio. Eight invariants are extracted per language: D*_chars, kappa (stretched-exponential shape), lambda_2 ratio, mu_1 (Fiedler vector asymmetry), gradient entropy, Fiedler entropy, concentration ratio, and bpc.

The D*_chars ranking:

Cluster	D*_chars	Languages
Latin-script European	2.3–2.9	en, fr, de, nl, sv, no, da, it, es, pt, ro, pl, cs, hr, la

Cluster	D*_chars	Languages
Cyrillic	~1.2	ru, uk, bg
Greek / Hebrew	1.2–1.3	el, he
Abugida (South Asian)	0.6–1.0	hi, bn, ta, te, my
CJK	0.8–1.0	zh, ja, ko
Thai	~0.8	th

PCA on 49 languages x 8 invariants:

PC	Variance explained	Dominant invariants
PC1	55.7%	D*_chars, kappa, bpc (encoding + morphological type)
PC2	15.1%	lambda_2 ratio, Fiedler entropy (spectral structure)
PC3	12.0%	Fiedler entropy, lambda_2 ratio, mu_1 (partition asymmetry)
Total	**82.8%**	

Three components explain 82.8% of variance. PC1 is dominated by the script encoding effect (multibyte vs. single-byte), separating Latin-script languages from CJK and Brahmic scripts. PC2 and PC3 are encoding-independent and measure genuine structural differences: spectral regularity of the context graph (PC2) and asymmetry of its Fiedler partition (PC3).

Language families cluster together in invariant space: Romance and Germanic languages decay at similar rates, Slavic languages faster, CJK fastest. The f(D) curves do not cross between families — the ranking is consistent at every depth.

3.5 Music and DNA: the complete ladder

MIDI music (364 KB synthetic, six genres):

D	All genres	Chorale	Jazz	Minimalist
2	0.612	0.671	0.599	0.367
5	0.164	0.034	0.073	0.115

D*(music) ~ 4–5. Chorales are nearly palindromic (f = 0.034 at D = 5): functional harmony works in both directions. Jazz retains more irreversibility (0.073) from chromatic substitutions. Minimalist music (Reich-style phasing) is the most irreversible at D = 5 (0.115) because phase drift is unidirectional.

E. coli K-12 genome (4.64 MB, alphabet {A,C,G,T}):

D	E. coli f(D)	Shuffled f(D)
2	0.993	0.9999
5	0.971	0.9998
7	0.962	0.9991

The harmonic fraction barely decays: 0.993 to 0.962 over D = 2..7. Linear extrapolation gives D*(E. coli) ~ 165 bases ~ 55 codons, but this estimate is fragile: the decay from 0.993 to 0.962 (3.1 percentage points over 5 depth steps) is small enough that the extrapolation depends heavily on the functional form assumed. The stretched exponential fit is poorly constrained because f is near 1.0 at all measured depths. The D* ~ 165 figure should be read as order-of-magnitude (D* » 10), not as a precise estimate. The shuffled control gives D* ~ 7,564 — the real genome is approximately 3% more reversible than random, reflecting codon bias and complementary base pairing.

After alphabet-size normalization against an i.i.d. baseline, English is *more* reversible than DNA at all depths. The raw comparison (DNA f = 0.98 vs. text f = 0.69) is an alphabet-size artifact: with |Sigma| = 4, the de Bruijn graph is near-complete at accessible depths, maximizing cycle density regardless of source structure.

The complete formality ladder:

```
D* ~     4     MIDI music
D* ~     6     Lean proofs
D* ~     7     Zig / Ruby / Python
D* ~     8     Wikipedia / legal English
D* ~    10     literary fiction (Pratchett)
D* ~   165     E. coli genome
```

A factor of ~40 from music to DNA, derived without domain knowledge.

3.6 n-gram asymmetry catalogue

The top asymmetric 5-grams in enwik8 (5 MB, tau >= 32) are exclusively function-word boundaries:

5-gram		asym
[space]of[space]3	6.58	function word
-[space]the	6.47	function word
[pipe]the	6.46	function word
[space]the[space]j	6.38	function word
2[space]and	6.25	function word

The top ~50 entries are all variants of "the", "of", "and", "to" at word boundaries. No SVO patterns, no semantic structure — the time arrow of English lives in function words.

Scale crossover as D increases:

D	Dominant pattern class	Examples
4	Function-word boundaries	`the`, `of`, `and`, `to`
6	Markup + word fragments	`&quot;`, `of th[e]`
7	Collocations + URLs	`of the X`, `in the X`, `http://`

This is a mechanical decomposition of English by topological asymmetry alone, recovering the hierarchy morpheme -> word -> phrase without linguistic input.

3.7 Byte-class analysis

Classifying 5-grams by the character class of their middle byte reveals a scale crossover:

Class		asym	at D=2
Markup	**2.92**	**1.59**	0.91
Newline	2.74	1.61	1.04
Space	1.92	1.40	**1.19**
Lowercase	1.79	1.40	1.10
Digit	1.25	1.07	**0.43**

Markup dominates asymmetry at D = 2–3 (HTML tags are strongly directional: `<ref>` predicts what follows, `</ref>` does not predict what preceded). Space takes over at D = 6 (function-word boundaries persist longer). Digits are the most reversible class at every depth — number sequences are nearly palindromic.

The scale crossover is the byte-class fingerprint of the f(D) decay: the drop from 0.81 (D = 2) to 0.37 (D = 5) is driven by markup asymmetry resolving first (D = 4–5), leaving function-word boundary asymmetry (D = 6+), which itself vanishes by D ~ 8. Two waves of irreversibility, two byte classes.

3.8 Anti-dispersion

Tracking the top-30 asymmetric 5-grams (D = 4) into the D = 7 catalogue: only 6/30 (20%) survive. The survivors have *lower* mean amplitude (|asym| = 5.96) than the non-survivors (|asym| = 6.10).

This is suggestive of anti-dispersion: strong cycles die first, inverting the soliton intuition where amplitude protects against dispersion. In language, amplitude and locality appear positively correlated: the most frequent patterns are the most local and resolve first with context.

Caveat. This observation is based on a sample of 30 n-grams tracked across two depth levels. The amplitude difference (6.10 vs 5.96) is small. Larger-scale tracking (hundreds of n-grams, multiple corpora) is needed to establish anti-dispersion as a robust law rather than a suggestive pattern.

This tentatively yields a taxonomy:

	Short-lived	Long-lived
Strong	function words, morphology	*rare (anti-dispersion if confirmed)*
Weak	noise	collocations, constructions

If confirmed at scale, the "grammar" of a source in the Hodge sense would be not the loud 80% (function words, articles, agreement markers) but the quiet 20% that survives all context depths.

3.9 GPT-2 crossover

GPT-2 generated text (HuggingFace wiki-intro dataset, 2 MB) compared to enwik8:

Source	D=2	D=3	D=5
enwik8	0.813	0.689	**0.367**
GPT-2 small	0.794	0.661	0.438
GPT-2 medium	0.801	0.671	0.446
GPT-2 large	0.803	0.672	0.438

The f(D) curves cross at D ~ 4. At D <= 3, GPT-2 text is slightly *less* irreversible than real English (1–2% lower — smoother local statistics). At D = 5, GPT-2 is 7–8% *more* irreversible than real English. GPT-2 text over-commits to local patterns, creating artificial cycle currents at the 5-gram scale that real English does not have.

Model size closes the short-range gap monotonically (small -> medium -> large at D = 2–3) but not the long-range one: at D = 5, all three model sizes are tied at f ~ 0.44, well above enwik8's 0.37. The medium-range rigidity is architectural, not a scaling problem.

The Hodge decomposition detects a structural difference between AI-generated and human text that is invisible to perplexity.

3.10 Emergent irreversibility (Phi_Hodge)

Interleaving two sources at chunk size C (D = 3, tau = 4, 100 KB corpora):

Pairing	C	f(A)	f(B)	f(AB)	Phi_Hodge
enwik8 / enwik8	64	0.641	0.599	0.654	**+0.014**
enwik8 / enwik8	4096	0.641	0.599	0.643	+0.002
enwik8 / Ruby	256	0.641	0.444	0.616	**-0.025**
enwik8 / Zig	256	0.641	0.449	0.605	**-0.036**

Similar sources create constructive cycle interference (Phi > 0): interleaving two halves of Wikipedia produces *more* irreversibility than either half alone, because the chunk boundaries create new cycle currents. The effect scales as ~1/C (more frequent switching = more boundary currents).

Dissimilar sources create destructive interference (Phi < 0): text + code interleaving *reduces* irreversibility. The two sources have largely disjoint context sets, so boundary transitions contribute to the exact (gradient) component rather than the harmonic (cycle) component.

3.11 Time-reversal invariance

The harmonic energy ||A_harmonic||^2 is invariant under byte-reversal of the corpus:

D	E_harm (fwd)	E_harm (rev)	**Rel. diff**	E_exact (fwd)	E_exact (rev)	Rel. diff
2	10,045,448	10,044,651	**0.01%**	2,324,409	2,290,944	1.4%
3	4,886,878	4,873,123	**0.28%**	2,272,781	1,992,447	12.3%
4	1,851,459	1,839,549	**0.64%**	1,794,504	1,592,686	11.2%
5	534,804	527,920	**1.29%**	1,213,326	1,065,076	12.2%

The harmonic energy is near-invariant (0.01%–1.3%) while the exact energy changes substantially (1.4%–12.3%). The harmonic component measures the *structure* of cycle currents (invariant under reversal); the exact component measures the *potential* (direction-dependent). The n-gram asymmetry signs reverse exactly (sum = 0.000000 to floating-point precision).

3.12 Three falsification results

Three predictions that would follow from a naive reading of the Hodge framework fail empirically, bounding the scope of the theory.

3.12.1 The concentration ratio C(D) < 1. If harmonic energy were uniformly distributed across the cycle space, f_harm would equal b_1/|E|. Empirically, f_harm < b_1/|E| at every depth. The concentration ratio C(D) = f_harm / (b_1/|E|) decays from 0.90 (D = 2) to 0.65 (D = 5) on the 2 MB enwik8 subsample. Only 65–90% of the topological cycle space carries energetically active currents; the rest is topologically present but energetically silent.

3.12.2 The de Bruijn graphs are not Ramanujan. For the enwik8 D = 2 graph (mean degree ~ 21.4, q ~ 20.4), the measured second eigenvalue lambda_2 = 7.10 exceeds the Ramanujan bound 2*sqrt(20.4) = 4.27 by 66% (ratio = 1.66). The de Bruijn graphs of natural language do not have optimal spectral expansion. Spectral gap arguments that assume Ramanujan-like mixing do not apply.

3.12.3 Two definitions of D* disagree. The energetic D* (where f(D) extrapolates to zero) is approximately 7.5, while the topological D* (where b_1/|E| extrapolates to zero) is approximately 9.4. The gap reflects a concentration effect: energetic irreversibility dies faster than topological irreversibility because energy vacates cycles before they topologically vanish.

---

4. Discussion

4.1 D* as the Reynolds number of information sources

The irreversibility depth D* marks the transition from a turbulent regime (high cycle currents, strong directional preferences) to a laminar regime (nearly reversible, gradient-dominated). The analogy to the Reynolds number in fluid dynamics is more than metaphorical. The Hodge decomposition of the edge field is the discrete analogue of the Helmholtz decomposition of a velocity field into curl-free and divergence-free components. The harmonic fraction f(D) measures the fraction of the "flow" that carries vorticity. D* is where the vorticity vanishes.

The formality ladder then has a physical reading: literary prose is "turbulent" (many competing cycle currents, high D*), proof code is "laminar" (few cycles, low D*), and DNA is "extremely turbulent" (near-complete cycle saturation, very high D*).

4.2 The stationarity-reversibility separation

The stationarity scale W* ~ 262,144 bytes (the optimal sliding-window size for enwik8) and the irreversibility depth D* ~ 7.5 bytes differ by five orders of magnitude. They measure different properties: W* is the mixing time (how long until the source forgets its initial state), D* is the reversibility depth (how much context until the time arrow vanishes).

A source can be highly non-stationary (long memory, slow mixing) yet nearly reversible at modest context depth. English is exactly this: its time arrow lives in the first 8 bytes, but its topic structure persists over 256 KB. The two scales decouple completely.

4.3 GPT-2 as a Hodge diagnostic

The crossover at D ~ 4 suggests the Hodge profile f(D) as a new diagnostic for language model quality. A model that matches the source's f(D) at all depths has reproduced the full cycle-current structure. The fact that GPT-2 is more irreversible than real English at D = 5 — creating artificial medium-range rigidity — indicates that autoregressive transformers commit too strongly to local patterns. The text they produce is statistically smooth (low perplexity) but topologically rigid (excess cycle currents).

Critically, model scaling does not fix this: GPT-2 small, medium, and large all converge to f ~ 0.44 at D = 5, above enwik8's 0.37. The rigidity is architectural. Future work should test whether larger models (GPT-3, GPT-4) or different architectures (state-space models, diffusion language models) close the D = 5 gap.

4.4 Anti-dispersion and the structure of grammar

The anti-dispersion law inverts the soliton intuition from nonlinear wave theory. In language, amplitude and locality are positively correlated: the most frequent patterns are the most local. This means the "loudest" grammatical signals — articles before nouns, subject-verb agreement, function-word placement — are the first to be absorbed into the potential as context grows. The "quietest" signals — collocations, constructions, idioms — form the persistent core.

This reframes what "grammar" means topologically. The persistent 20% of cycle currents that survive all context depths are not the surface rules easily articulated by linguists but the long-range structural regularities that hold language together over distances beyond D*. These are precisely the patterns that language models find most difficult, and that the Hodge decomposition isolates.

4.5 The concentration ratio and the limits of topology

The falsification results bound the framework in a useful way. The harmonic fraction is not determined by graph topology alone — it depends on the energy distribution over cycles, which is source-specific. Sources with identical de Bruijn graph structure could have different harmonic fractions. The concentration ratio C(D) itself appears to be a meaningful invariant: natural language maintains higher concentration (0.75–0.91) than code (0.75–0.88), suggesting that natural language distributes its irreversibility more uniformly across cycles.

The non-Ramanujan character of the de Bruijn graphs explains why C(D) < 1: bottlenecks and clusters in the graph prevent Markov currents from accessing the full cycle space. The measured deviation (C ~ 0.85 on average) quantifies "topological waste" — the fraction of cycle space that exists but carries no current.

---

5. Conclusion

The Hodge decomposition of the Markov transition field on de Bruijn graphs provides a general-purpose measurement of the time-asymmetry structure of information sources. Three new invariants

— the irreversibility depth D*, the stretched-exponential shape parameter k, and the anti-dispersion law — classify sources without domain-specific knowledge. The formality ladder from music through code through text to DNA emerges from pure byte-level topology.

The practical implications are twofold.

First, the Hodge profile f(D) is a diagnostic for language model quality that detects structural differences invisible to perplexity. GPT-2's anomalous rigidity at D = 5, invisible to standard evaluation metrics, is immediately visible in the harmonic fraction. As language models are deployed in high-stakes settings, structural diagnostics that go beyond perplexity will become essential.

Second, the n-gram asymmetry catalogue and byte-class analysis provide a new lens on linguistic structure. The scale crossover from function-word morphology (D = 4) to collocations (D = 7), the anti-dispersion law (strong cycles die first), and the byte-class fingerprint of the f(D) decay are empirical facts about natural language that emerge without any linguistic input. They suggest that the irreversibility depth is a natural coordinate for computational typology.

The 49-language atlas is a first step. With larger corpora and higher context depths, D* and its associated invariants could provide a typological classification complementary to existing morphological and syntactic universals — one grounded not in human-defined categories but in the topology of the byte stream.

All code and data are available at the accompanying repository.

References

- Compeau, P.E.C., Pevzner, P.A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11), 987–991.
- Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. (2011). Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1), 203–244.
- Jiang, D.-Q., Qian, M., and Qian, M.-P. (2004). *Mathematical Theory of Nonequilibrium Steady States*. Springer.
- Kalpazidou, S.L. (2006). *Cycle Representations of Markov Processes*. Springer.
- Schaub, M.T., Benson, A.R., Horn, P., Lippner, G., and Jadbabaie, A. (2020). Random walks on simplicial complexes and the normalized Hodge 1-Laplacian. *SIAM Review*, 62(2), 353–391.
- Shallit, J. (2009). *A Second Course in Formal Languages and Automata Theory*. Cambridge University Press.

Appendix A: Compression validation

As a practical validation that D* captures compression-relevant structure, we built a PPM-C byte-level compressor guided by the Hodge decomposition. Starting from a Hodge-motivated parent+delta baseline (3.68 bpb), three components — exclusion, PPM-C escape, and online adaptation — reduce the rate to 2.16 bpb on enwik8 (1M bytes), a 41% reduction competitive with bzip2. Twenty controlled experiments identified seven improvements and thirteen honest negatives;

the best single intervention was skip-gram context blending (-0.089 bpb). All results were verified with a Lean 4 formalization proving the compressor's roundtrip property with zero sorry's, yielding K(enwik8_prefix) <= 2.16 * |x| + O(1) as a machine-verified Kolmogorov complexity upper bound.

The most robust negative finding: every attempt to replace PPM-C's escape formula u/(t+u) fails. The escape mechanism is a local optimum. One interpretation is that the escape operator has nilpotent structure (applying escape twice yields nothing, analogous to epsilonˆ2 = 0), but this analogy has not been formalized and should be treated as suggestive rather than proved. What is established empirically is that improvements live outside the escape mechanism — modifying which contexts exist and how distributions are blended, without touching the core formula.

Appendix B: Harmonic energy as a generalization predictor

We tested five predictions from the geometric framework underlying the Hodge decomposition (Hoekstra 2026). Four were falsified: inter-worker holonomy K(c) does not beat KL divergence, the scaling exponent is alpha = 0.148 (not the predicted 1/3), curvature-gated mixtures underperform the trie alone, and trajectory-based shock loops do not cluster on article boundaries.

One prediction survived in modified form: per-vertex harmonic energy from the Hodge decomposition outperforms holonomy as a generalization predictor (+0.27 vs. +0.21 partial Spearman at 1M bytes), though it remains weaker than conditional entropy (+0.57) and KL divergence (+0.39). The ranking H > D > H_harm > K is stable across scales. The geometric signal is real but not dominant — a useful bound on the scope of the framework.

Appendix C: Notation summary

Symbol	Definition
D	Context depth (bytes)
G_D	Order-D de Bruijn graph
A(c -> c')	Markov transition field: log P_emp(c'|c)
d_0	Oriented incidence matrix
L	Weighted graph Laplacian: d_0ˆT W d_0
phi	Vertex potential solving L phi = d_0ˆT W A
A_exact	Exact component: d_0 phi
A_harmonic	Harmonic component: A - A_exact
f(D)	Harmonic fraction:
D*	Irreversibility depth: D where f -> 0
k	Stretched-exponential shape parameter
D_0	Stretched-exponential scale parameter
b_1	First Betti number:
C(D)	Concentration ratio: f_harm / (b_1/|E|)
asym(s)	Time-reversal asymmetry of n-gram s
Phi_Hodge	Emergent irreversibility of coupled sources