ORTHOGRAPHIC MASS
A CIRCULAR-STATISTICS INVARIANT OF WRITTEN FORMS ACROSS WRITING SYSTEMS

BLAIZE ROUYEA AND COREY BOURGEOIS

Abstract. We define the *orthographic mass* of a written word as the magnitude of the first circular moment of the angular distribution of its UTF-8 nibbles. The construction is parameter-free and corpus-free: it reads only the bytes of a single written form. Our central claim is deliberately small and, we hope, durable. The quantity we define is not new. It is identical to the *mean resultant length* $\bar{R}$, the standard measure of concentration in directional statistics, and its complement $1 - \bar{R}$ is the circular variance. What is new is the object we point it at — the externalized written form under a fixed encoding — and what that instrument turns out to see.

Computed over 833,116 lemmas across fourteen languages and five writing systems, orthographic mass separates alphabetic scripts (high concentration) from logographic and abugida scripts (low concentration). Because mass is exactly $\bar{R}$, the relationship between script type and mass is a *consequence* of how each writing system populates the encoding, not a coincidence, and the value of a random byte string is governed by an exact law: $\mathbb{E}[m^2] = 1/k$ for a form of k nibbles. Reading mass as the first of a sequence of circular moments yields a typological *fingerprint*: the second moment resolves scripts the first collapses — most strikingly Arabic, which carries alphabetic-level mass yet a uniquely low second moment — and hierarchical clustering on the moment vector recovers the writing systems, with genealogical structure emerging within the Latin script as a correlate of shared orthographic statistics. We are explicit about scope: this measures written *form*, not meaning, not cognition, and not phylogeny.

Contents

Date: Draft: May 30, 2026.

1. Introduction

This began with a question that felt almost too simple to ask out loud: does a written word carry measurable structure in its *form alone* — in the shape of the marks, independent of what the word means?

We came to this from language and communication rather than from linear algebra, and our first instinct was to treat the question as one about meaning. An earlier version of this work was called *semantic mass*. That name was a mistake, and naming the mistake is the cleanest way to start. The quantity we had defined never touched meaning at any step. It is computed from the bytes of a word's written form and nothing else; two words with identical spelling have identical mass regardless of what they denote, and a word and a random string with the same byte statistics are, to this instrument, the same. Calling the result *semantic* claimed a bridge to meaning we had not built and did not need. The honest name for a number that depends only on the written form is *orthographic mass*,[1] and the rename is not cosmetic: it is the difference between a metaphor and a measurement.

Once we held the quantity at arm's length and asked plainly what it *was*, the rest of this paper followed. The answer turned out to be both humbling and freeing: orthographic mass is a quantity statisticians have studied for a century. That recognition is the spine of everything below. It lets us state exact properties instead of suggestive ones, and it lets the empirical regularities we observe inherit a body of theory rather than stand on their own assertion.

The paper proceeds in the order we actually walked it. Section 2 gives the construction. Section 3 establishes what the quantity is — its bounds, its identity with the mean resultant length, and the exact law governing random forms. Section 4 describes the data. Section 5 reports what the instrument sees across writing systems, including a generalization to higher circular moments that resolves distinctions the scalar misses. Section 6 is about scope: what orthographic mass is, and the several things it is not.

2. The Construction

Let w be a written word. Encode w to its UTF-8 byte sequence and expand each byte into its two hexadecimal digits, producing a *nibble sequence* $n_1, n_2, \ldots, n_k$ with each $n_t \in \{0, 1, \ldots, 15\}$. Each nibble is a discrete symbol, and we place the sixteen possible symbols at equal angles on the unit circle.

Definition 2.1 (Nibble embedding)**.** For a nibble value $j \in \{0, \ldots, 15\}$, define the angle $\theta(j) = 2\pi j/16$ and the unit vector $b_j = (\cos\theta(j), \sin\theta(j))$. A word w with nibbles $n_1, \ldots, n_k$ induces the angular sequence $\theta_t = \theta(n_t)$.

A word is therefore a finite walk over sixteen fixed directions. We summarize that walk by its *first circular moment* — the average of the unit vectors it visits.

Definition 2.2 (Orthographic mass and phase)**.** For a word w of k nibbles, let

$$c(w) = \frac{1}{k}\sum_{t=1}^{k} b_{n_t} = \Big(\tfrac{1}{k}\textstyle\sum_t \cos\theta_t,\ \tfrac{1}{k}\sum_t \sin\theta_t\Big).$$

The *orthographic mass* is the magnitude $m(w) = \|c(w)\|$, and the *phase* is the direction $\varphi(w) = \operatorname{atan2}\big(\sum_t \sin\theta_t,\ \sum_t \cos\theta_t\big)$.

[1] Throughout, *orthographic* is used in its linguistic sense — pertaining to orthography, the writing system of a language — and never in the sense of *orthographic projection* from descriptive geometry. The construction below is a trigonometric embedding onto a circle, not a parallel projection onto a plane.

Mass and phase are the polar coordinates of one vector, the resultant $c(w)$: mass is its length, phase its direction. No parameters are fitted, no corpus is consulted, no thresholds are chosen. The entire object is a deterministic function of the bytes of w.

We note, for readers tracing the construction to its broader setting, that this is the single-ring, feed-forward, read-only instance of the *orbience* framework [7], in which discrete state is read as geometry, phase, and magnitude on interacting rings. Orthographic mass uses one ring and reads it once. Nothing in this paper depends on that lineage; we record it only to place the construction.

3. What the Quantity Is

The value of stating the construction precisely is that its properties are now theorems rather than observations.

Proposition 3.1 (Bounds)**.** *For every word,* $0 \le m(w) \le 1$*. The upper bound follows from the triangle inequality,* $\|\frac{1}{k}\sum_t b_{n_t}\| \le \frac{1}{k}\sum_t \|b_{n_t}\| = 1$*, with equality iff all nibbles are identical; the lower bound is immediate.*

The next statement is the one that reorganizes the paper.

Proposition 3.2 (Identity with the mean resultant length)**.** *Orthographic mass is exactly the* mean resultant length $\bar{R}$ *of the angular sample* $\{\theta_t\}$*, the standard concentration statistic of directional statistics. Equivalently,* $1 - m(w)$ *is the sample* circular variance *of* $\{\theta_t\}$*.*

This is immediate once the previous definition is read in the right coordinates: $c(w)$ is the mean resultant vector of the angles θ_t, and its magnitude is by definition $\bar{R}$ [1, 2]. We verified the identity directly on 8,000 English forms, computing mass through the original implementation and through the textbook resultant-length formula independently; the two agreed to 5.5×10^{-16}, i.e. to floating-point identity.

The consequence is that orthographic mass is not an isolated coinage but an instance of a quantity with a mature theory. Concentration near 1 means the angles cluster in one direction; concentration near 0 means they are diffuse. The circular variance, the von Mises concentration parameter, and the standard tests of angular uniformity all attach to it without modification. In particular, the behavior of a structureless form is governed by an exact law.

Theorem 3.3 (Null law)**.** *If the nibbles* $n_1, \dots, n_k$ *are independent and uniform on* $\{0, \dots, 15\}$*, then* $\mathbb{E}[m(w)^2] = 1/k$*. Consequently* $\mathbb{E}[m(w)]$ *decays on the order of* $1/\sqrt{k}$*, and the mass of a random form tends to zero as its length grows.*

Proof. The sixteen embedding vectors are equally spaced, so $\sum_j b_j = 0$ and a uniform nibble has $\mathbb{E}[b] = 0$ with $\mathbb{E}\|b\|^2 = 1$. Writing $c = \frac{1}{k}\sum_t b_{n_t}$ with independent terms,

$$\mathbb{E}\|c\|^2 = \frac{1}{k^2}\sum_t \mathbb{E}\|b_{n_t}\|^2 + \frac{1}{k^2}\sum_{s\neq t} \mathbb{E}[b_{n_s} \cdot b_{n_t}] = \frac{1}{k^2}\,(k \cdot 1) + 0 = \frac{1}{k},$$

the cross terms vanishing by independence and zero mean. □

Simulation confirms the law across lengths:

The null law is more than a sanity check. It converts the claim "real words are not random" from an assertion into a measurement against a known baseline, and it foreshadows the mechanism of Section 5: anything that makes a script's nibble directions more diffuse pushes its mass toward the random floor.

Finally, mass is the first term of a sequence we will need.

Definition 3.4 (Moment spectrum)**.** For $p = 1, 2, 3, \dots$, the p-th circular-moment magnitude of w is $m_p(w) = \|\frac{1}{k}\sum_t e^{\,ip\theta_t}\|$. Orthographic mass is m_1. Higher orders measure concentration at finer angular periods; m_2, for instance, responds to axial (period-π) structure to which m_1 is blind.

nibbles k	measured $\mathbb{E}[m^2]$	$1/k$
4	0.243	0.250
8	0.126	0.125
16	0.064	0.063
32	0.031	0.031
64	0.015	0.016
128	0.008	0.008

4. Data and Reproducibility

We compute orthographic mass over lemma lists derived from WordNet [3] and its multilingual extensions, spanning fourteen languages and five writing systems, for a total of 833,116 lemmas. The pipeline — embedding, moment computation, and the analyses below — is implemented in the open `orbihex` toolkit. Because the operator is per-word and parameter-free, every figure in this paper is reproducible from the lemma lists alone.

Two honesty conditions are worth stating before any result. First, *the encoding is part of the operator.* Orthographic mass is a property of the pair (script, UTF-8 encoding) [4]; a different encoding would induce a different angular sequence. We regard this not as a confound to be removed but as the precise sense in which the quantity is *orthographic*: it measures the externalized form as actually written and encoded. Second, the moment-spectrum analyses of Sections 5.4–5.5 were run on the thirteen languages whose lemma lists are reconstructable in the public release; one language (Spanish) appears in the aggregate mass results but not in the per-moment extension, and we mark this where relevant.

5. What the Instrument Sees

5.1. **Mass distributions by writing system.** The first observation is that mass is not a single number per language but a distribution per writing system, and the distributions separate.

Aggregated to per-language means, the ordering is stable (Table 1). Germanic and Romance alphabets occupy a high-mass band near 0.49–0.56; Arabic, an alphabetic abjad, sits in the same band at 0.49; the abugida (Thai) and the logographic systems (Mandarin, Japanese) fall well below.

Table 1. Average orthographic mass by language. Lexicon size is the lemma count.

Language	Family	Script	Lemmas	m_1
Dutch	Germanic	LTR alph.	42,091	0.559
English	Germanic	LTR alph.	140,003	0.552
Norwegian	Nordic	LTR alph.	4,183	0.548
French	Romance	LTR alph.	48,783	0.521
Italian	Romance	LTR alph.	40,482	0.519
Catalan	Romance	LTR alph.	64,022	0.513
Spanish	Romance	LTR alph.	86,107	0.502
Finnish	Finno-Ugric	LTR alph.	117,681	0.499
Arabic	Afro-Asiatic	RTL abjad	19,074	0.494
Icelandic	Nordic	LTR alph.	11,346	0.409
Mandarin	Sino-Tibetan	Logographic	60,893	0.351
Thai	Tai-Kadai	Abugida	62,709	0.339
Japanese	Japonic	Mixed logo.	90,948	0.236

5.2. **Why concentration varies: the angular footprint.** The mechanism behind Table 1 is visible directly in each script's angular footprint — the distribution of nibble directions whose first moment *is* the mass.

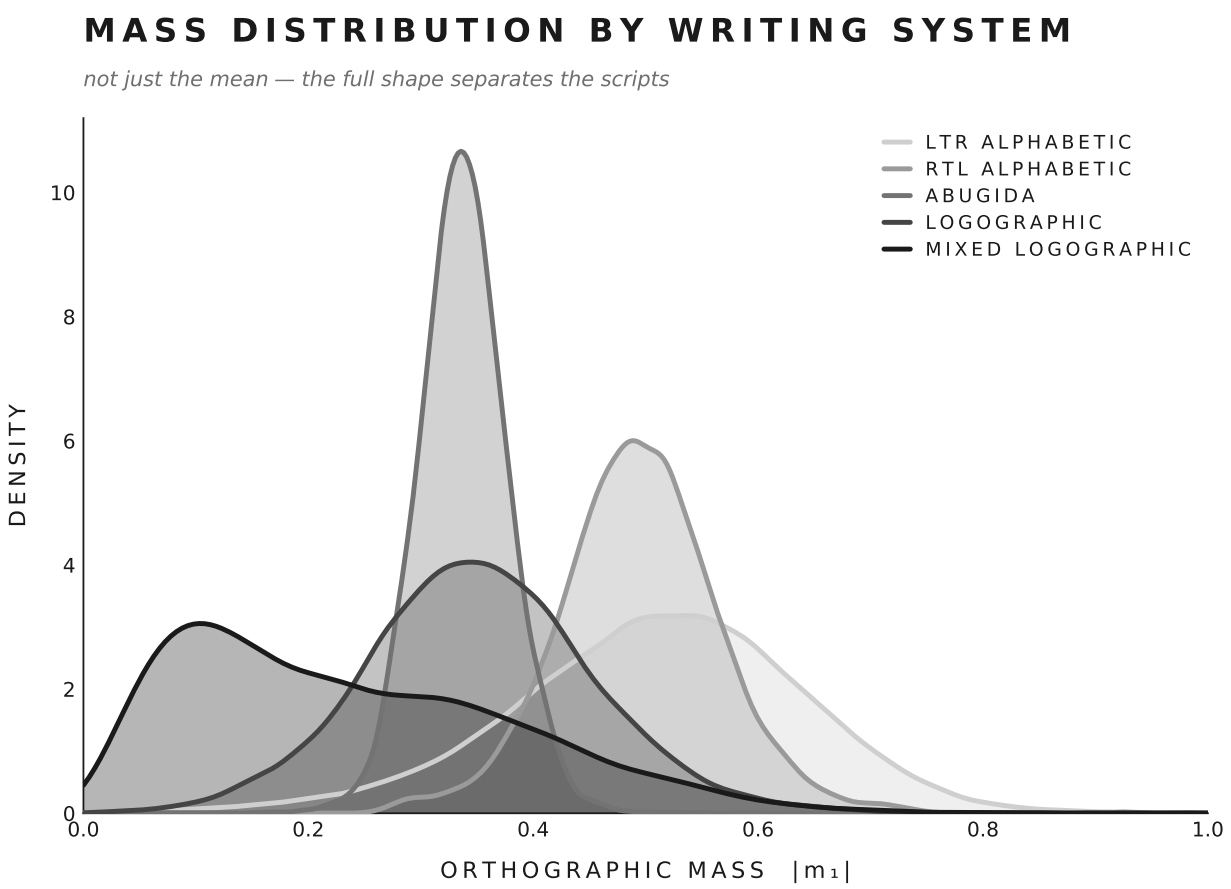

FIGURE 1. Distribution of orthographic mass by writing system, pooled across the languages of each class. Logographic and mixed-logographic forms concentrate at low mass; alphabetic forms at high mass; the abugida sits between. The separation is in the full shape, not only the mean.

A script's footprint is determined by which regions of the code space its characters occupy and therefore which nibble values its UTF-8 bytes produce. Latin lowercase letters live in a narrow byte range and yield a tight single lobe — high concentration, high mass. Scripts encoded in three-byte UTF-8 sequences over large code-point inventories (the CJK systems) spread their nibbles across many directions — diffuse footprints, low mass, drifting toward the random floor of Theorem 3.3. This is the writing-system effect stated mechanically: *mass is low precisely when the encoded form's directions are diffuse, and that diffuseness is a property of the script's place in the encoding.* It is not a statement about the complexity of the language, and it is certainly not a statement about its speakers.

The footprints also make the phase legible. The phase $\varphi(w)$ is the mean direction of a form's nibbles, and it cleanly separates Latin scripts, whose forms point into one angular region, from the non-Latin scripts, which sit in distinctly different regions of the circle by virtue of occupying different UTF-8 blocks.

5.3. **The complexity relationship as a corollary.** Because mass is the mean resultant length, the inverse relationship between script complexity and mass — heavier, more diffuse scripts carrying lower mass — is not a discovered correlation in need of explanation. It is a corollary of Proposition 3.2 and the footprint mechanism: a more diffuse angular distribution has, by definition, a shorter

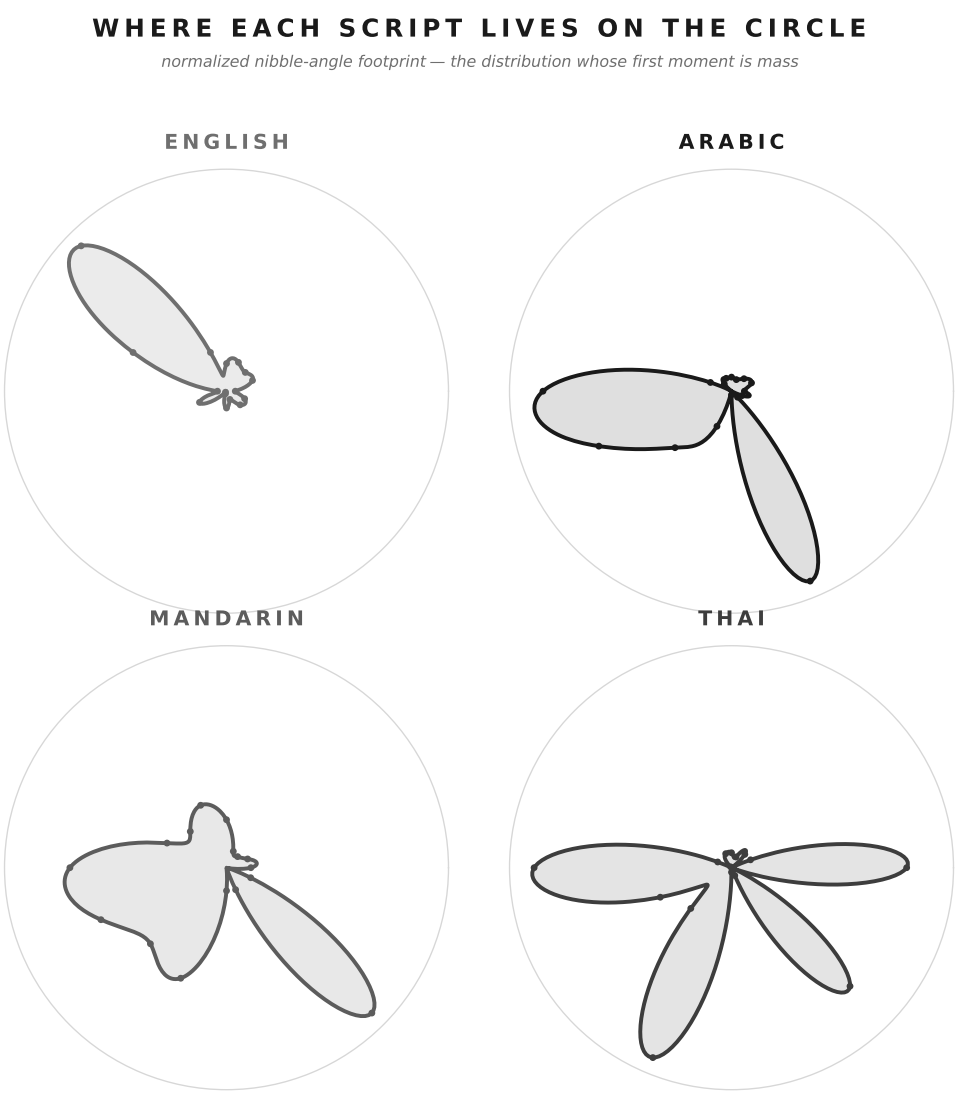

FIGURE 2. Normalized nibble-angle footprint for one script of each class. English concentrates in a single lobe (high mass). Arabic occupies two lobes roughly a quarter-turn apart. Mandarin and Thai spread their directions broadly. The footprint is the distribution; mass is the length of its resultant.

resultant. We therefore decline to fit a "complexity score," which would impose a subjective ordinal where the geometry already supplies the explanation.

5.4. **What mass alone misses: the second moment.** A single number is a single projection of a richer object, and reading only m_1 discards everything the higher moments carry. The clearest case is Arabic.

By mass alone, Arabic ($m_1 = 0.494$) is indistinguishable from a European alphabet — it sits inside the Latin band. But its second moment is $m_2 = 0.19$, while every Latin script has m_2 between 0.40 and 0.48. The second moment more than halves for Arabic while its first moment does not, and the footprint explains why: Arabic's two lobes lie roughly a quarter-turn apart, so under angle doubling they move to opposite sides of the circle and cancel. The first moment survives; the second does not.

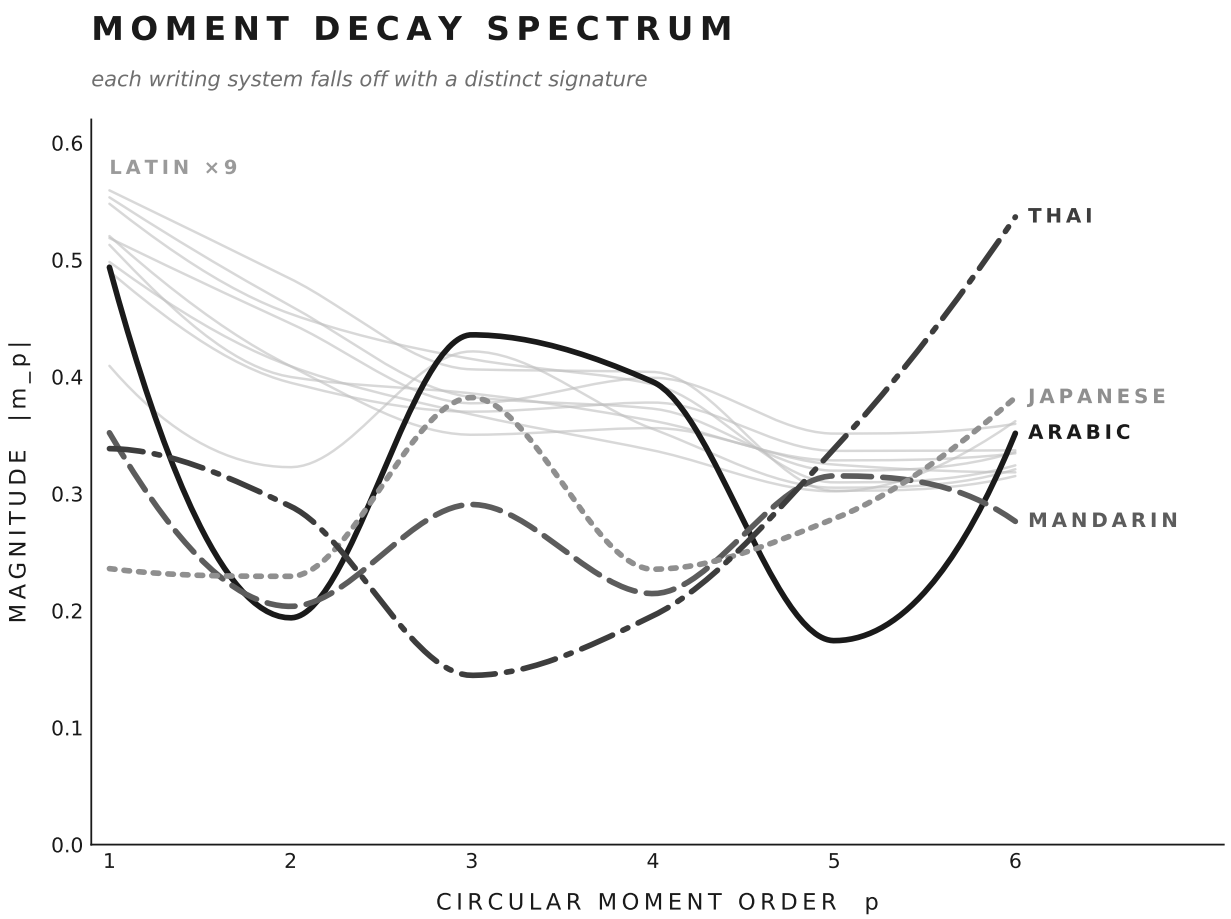

FIGURE 3. The moment decay spectrum $|m_p|$ for $p = 1, \ldots, 6$. The nine Latin alphabets fall in a tight band (grey); the four non-Latin scripts (lines) each fall off with a distinct signature. Arabic oscillates between even and odd orders; Thai collapses then rises; the curve itself is a fingerprint.

Across the scripts, m_1 and m_2 correlate at $r = 0.77$ — related, because both are concentration measures, but far from redundant. The information lives in the *residual*, where the second moment departs from what the first predicts. That residual is where Arabic and Japanese announce themselves.

5.5. **Typological clustering.** Treating the moment vector $(m_1, \ldots, m_6)$ as a per-script fingerprint, we cluster the languages and compare the result to clustering on mass alone.

The moment fingerprint recovers the four writing systems cleanly — Latin, Arabic, CJK, and Thai fall into their own clusters, with Arabic isolating from the Latin band where mass alone had buried it. Within the Latin clade, the Germanic core groups together and the Romance languages associate, which we read as a genealogical *correlate*: related languages share orthographic conventions — letter frequencies, diacritics, common digraphs — and those shared statistics survive into the moment geometry. We are careful here. The signal is orthographic, not phylogenetic; the geometry does not know the language tree, it reads the spelling that related languages happen to share, and the recovery is partial — Icelandic, with its þ/ð characters, and Finnish, the lone Uralic representative, sit apart from the families they nominally belong to.

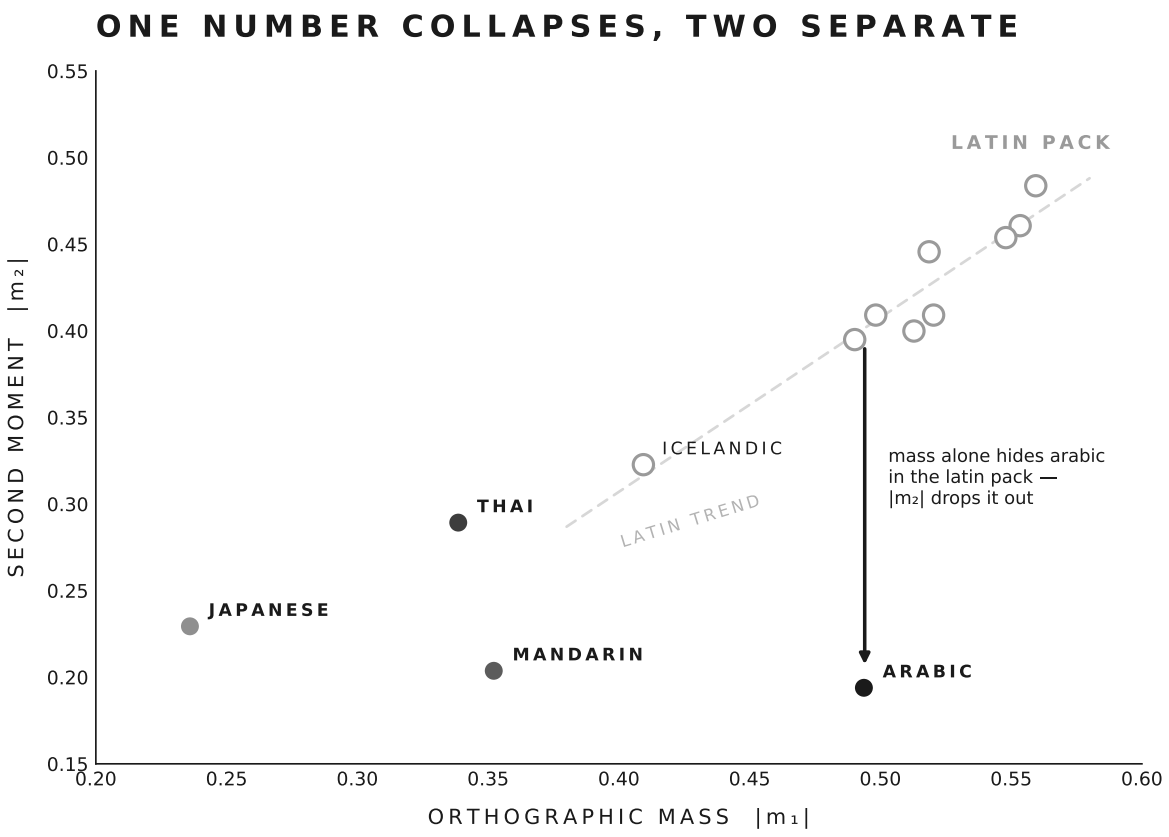

FIGURE 4. Orthographic mass against the second moment. Mass alone places Arabic in the Latin pack; the second moment drops it out. The arrow marks the separation a single number cannot make.

The contrast with the scalar is sharp. Clustering on mass alone commits cross-script errors — pairing Portuguese with Arabic, Finnish with Arabic, Icelandic with Mandarin — because those forms happen to share a mass value. The moment fingerprint makes none of these errors.

6. Scope: What This Is, and Is Not

We have tried to make only claims the construction supports, and the discipline is easier to keep if we state the boundary explicitly.

Orthographic mass is a classical circular-concentration statistic — the mean resultant length — computed over the UTF-8-encoded written form of a word. Its properties are exact (Propositions 3.1 and 3.2 and Theorem 3.3); its empirical regularities across writing systems are consequences of how scripts populate the encoding; and its generalization to higher moments is a typological fingerprint with demonstrated discriminative power.

Orthographic mass is not a measure of meaning. It does not touch semantics at any step, and the earlier name claimed otherwise. It is *not* a measure of cognition or of the complexity of a language or its speakers; the writing-system effect is a fact about encodings, not minds. It is *not* a phylogenetic marker; the family structure it recovers within a script is a correlate of shared orthography, not a reconstruction of descent. And it is *not* a new statistic. This last point is the one we most want to stand behind: the instrument is borrowed, fully and deliberately, from directional statistics. The contribution is the application of a proven invariant to a new object and the reading of its

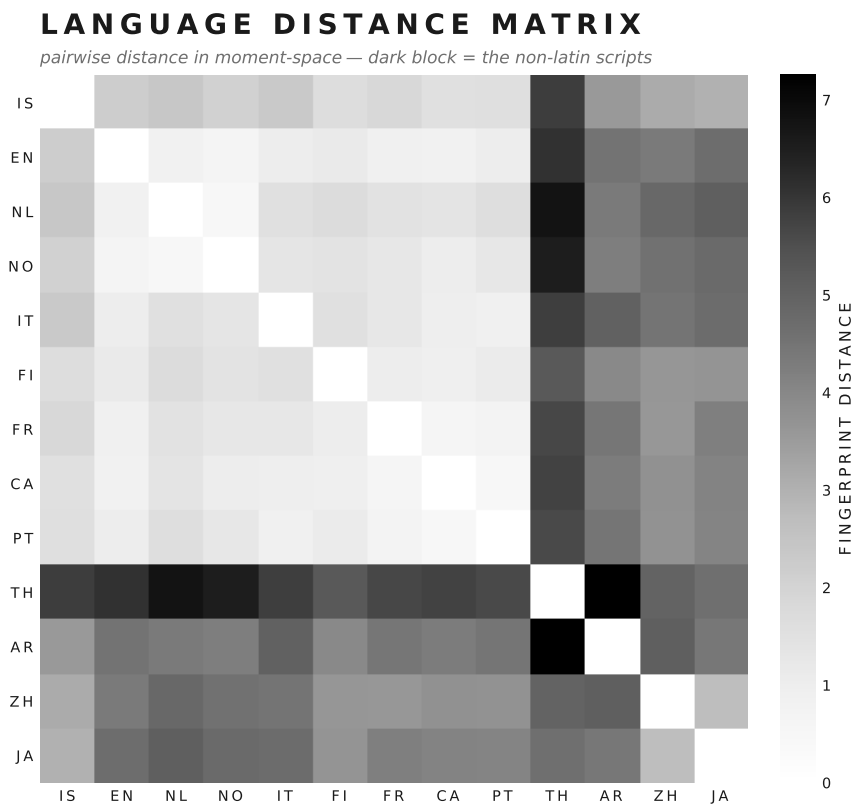

FIGURE 5. Pairwise distance between languages in moment space. The light block is the Latin alphabets (mutually close); the dark cross is the non-Latin scripts, far from the Latin block and, in the Thai–Arabic case, from one another.

moment spectrum as a script typology. We think that is a sturdier place to make a contribution than inventing a quantity no one has reason to trust.

Two limitations bound the empirics. The encoding dependence above means results are stated for UTF-8 and would differ under another encoding. And the Thai lexicon shows an anomalously long mean nibble length, consistent with unsegmented strings — Thai is written without spaces — so its higher-moment values should be read with caution pending a clean segmentation.

The natural relatives of this work are the literatures on directional statistics, where the mean resultant length and the circular moments live [1, 2], and on orthographic depth and processing [5], where the structure of writing systems is studied directly. Orthographic mass offers those literatures a parameter-free, corpus-free number with an exact null and a clean geometric reading.

7. Generalization and Future Work

The moment spectrum reframes the contribution. A written form is summarized not by one number but by the magnitude sequence $(m_1, m_2, m_3, \dots)$ — the first recovering script concentration, the higher orders resolving structure the first collapses. Mass is simply the first coordinate of a fingerprint.

One extension we have deliberately *not* taken in this paper, but which the operator invites, is the lift from words to utterances. The same first-moment construction applies to any nibble sequence, including a concatenated turn of speech or a speaker's running lexicon. A measure of whether two interlocutors' form-geometry converges over an interaction would be a geometric reading of *lexical entrainment* — a question squarely in interpersonal communication. We flag it as the obvious sequel

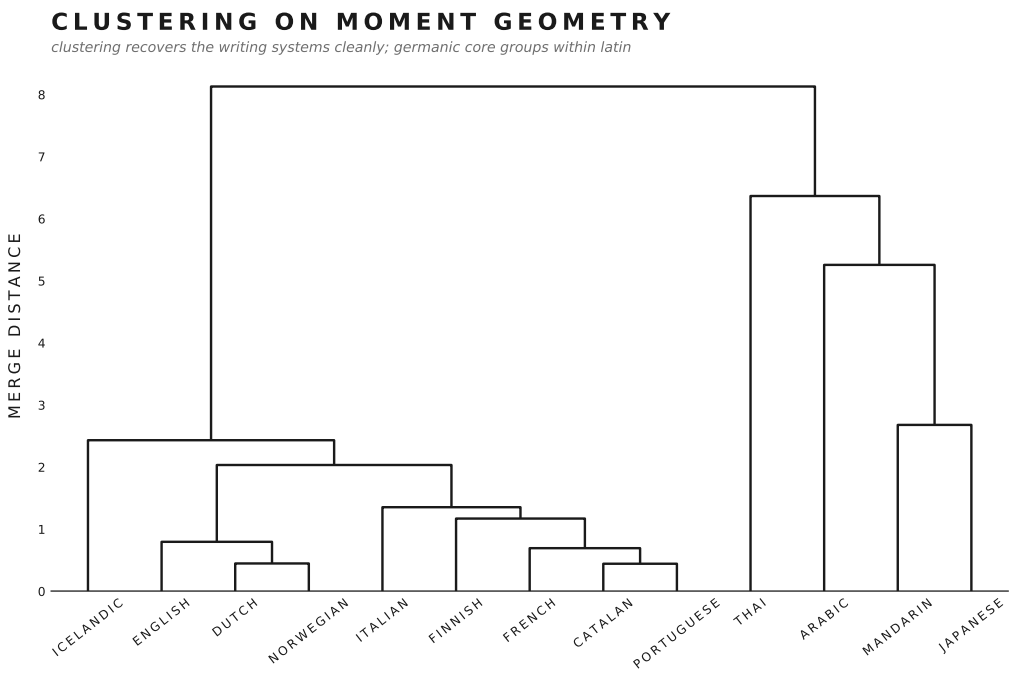

FIGURE 6. Hierarchical clustering on the moment fingerprint. The primary split separates the writing systems; within the Latin clade, the Germanic core (English, Dutch, Norwegian) groups together.

because it requires interactional data the present lexicon study does not contain, and we would rather name the open door than walk through it on this paper's evidence.

8. Conclusion

We set out to ask whether a written form carries measurable structure in its shape alone, and found that it does — but that the right way to say so is smaller and more durable than where we started. The structure is concentration, the measure is the mean resultant length, and the regularities across writing systems follow from how each script lives in the encoding. Naming the quantity correctly — orthographic, not semantic; a borrowed invariant, not a new one — costs a measure of novelty and buys, in exchange, the right to state exactly what is true. The novelty we keep is the one that survives scrutiny: a classical instrument, pointed at the written form, that recovers the typology of writing systems and extends to a moment spectrum that sees more than its first term.

References

[1] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 2000.
[2] N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1993.
[3] C. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
[4] The Unicode Consortium. *The Unicode Standard* (UTF-8 transformation format).
[5] R. Frost and L. Katz. Orthographic depth and its relation to reading and the structure of writing systems.
[6] B. Rouyea and C. Bourgeois. Prime patterns in arithmetic wheel-lanes. 2025.
[7] C. Bourgeois and B. Rouyea. orbience: A plastic multi-ring attractor brain. 2025.