SEMANTIC MASS: A GEOMETRIC INVARIANT OF LEXICAL FORMS ACROSS WRITING SYSTEMS

BLAIZE ROUYEA AND COREY BOURGEOIS

Abstract. We introduce a quantitative framework for comparing lexical structure across writing systems based on a single geometric invariant, *semantic mass*. Starting from a uniform nibble-based encoding of UTF-8 word forms, each lexical item is mapped to a trajectory in a fixed 16-point trigonometric basis on the unit circle. The center-of-mass norm of this trajectory defines semantic mass, a geometric notion of semantic weight that quantifies how strongly a *lexia* pulls the *orbience* ring into a stable, directionally coherent configuration. Semantic mass is a parameter-free scalar that can be aggregated at the language level.

All results are computed over 833,116 lemmas drawn from 14 languages in the Open Multilingual WordNet. We show that alphabetic, abugida, abjad, and mixed morphosyllabic writing systems occupy distinct regions of semantic-mass space. Alphabetic systems cluster around higher average mass values (0.494–0.559), whereas morphosyllabic and mixed morphosyllabic systems exhibit consistently lower values (0.236–0.351). A complementary phase analysis reveals surprising but robust clustering patterns in the angular component, with a near-universal five-cluster structure in alphabetic languages and systematic deviations in Japanese and Thai.

Although the construction is purely geometric, the observed regularities align closely with established typological distinctions and support an operational *Writing System Complexity Hypothesis*: under a fixed geometric operator, more complex scripts give rise to more diffuse lexical trajectories and correspondingly lower semantic mass. The framework is simple, reproducible, and applicable to any UTF-8 encoded lexicon, providing a new tool for cross-linguistic analysis in computational linguistics and for connecting abstract questions about linguistic relativity with concrete, testable invariants over externalized written forms.

Keywords: semantic mass, lexical geometry, writing systems, cross-linguistic analysis, geometric invariants, linguistic relativity

Contents

1. Introduction

A central question in the study of language and mind is whether, and in what sense, language shapes the space of possible experiences. Classical formulations of linguistic relativity [10, 11] ask whether speakers of different languages perceive or conceptualize the world differently as a consequence of the categories encoded in their grammars and lexicons. In parallel, generative approaches have emphasized the existence of deep structural invariants across languages, with surface variation arising from restricted parameters within an underlying universal grammar [2, 3].

Questions of linguistic relativity ask whether differences in language structure can induce systematic differences in habitual thought [11, 10, 7]. In this paper we do not attempt to measure cognition directly. Instead, we ask a narrower, operational question: when lexical forms from different writing systems are passed through a fixed geometric operator, do they exhibit stable, typologically structured differences in their induced geometry?

Both perspectives implicitly assume that linguistic structure can be meaningfully compared across languages. Yet most formal work focuses on phonology and syntax, and most empirical work relies on behavioural or distributional evidence. The written forms that implement languages in external media—scripts, orthographies, and the digital encodings that support them—are often treated as secondary. For inter-agent coupling and information exchange more broadly, however, these externalizations matter: they are the concrete carriers of linguistic information in human and machine systems alike.

This paper asks a deliberately narrow question within this broader landscape:

> *"Given only the written forms of lexical items in different languages, encoded in a common digital scheme, can we identify simple geometric invariants that distinguish writing systems and track genealogical structure across languages?"*

We do not assume access to meanings, syntactic frames, or usage frequencies. Instead we treat each lexical item as a *lexia*: a discrete written unit that traces a path through a fixed geometric space when processed by a simple language engine. In the more general framework that motivates this work, that engine is called *orbience*, a geometric language architecture whose state evolves on rings of interacting units. In this paper we restrict attention to a single-ring, purely feedforward instance of that architecture and analyse its behaviour as a deterministic operator on UTF-8 encoded word forms.

From this operator we derive *semantic mass*, a scalar invariant that summarizes the dispersion of a *lexia*'s trajectory in phase space. Semantic mass is not a semantic measure in the psychological sense; it is a mathematically explicit function of the encoded form. Nevertheless, when aggregated across the lexicons of 14 languages, its distribution aligns closely with high-level typological distinctions between alphabetic, abugida, abjad, and logographic systems. At the same time, it exhibits enough within-family variation to capture genealogical structure among Indo-European languages.

1.1. Motivation and scope.

Our motivation is twofold. First, we seek a minimal bridge between structural questions in theoretical linguistics and the concrete representations used in computational systems. Within generative grammar, externalization into phonology and orthography is often treated as secondary to an internal system of syntax and semantics

[4, 5]. The present work is compatible with that distinction: we take no stance on internal competence. Instead, we study invariants of the externalized forms themselves, asking how different writing systems populate a shared geometric space when passed through a uniform encoding pipeline.

Second, we aim to ground discussions of linguistic relativity and inter-agent coupling in quantities that can be computed and compared. If languages differ in how they carve up conceptual space, and if those differences are reflected in the inventories of written forms that speakers learn, then even a crude geometric operator on those forms should exhibit systematic cross-linguistic structure. Our results do not speak directly to consciousness or experience, but they provide a clean baseline: a parameter-free invariant whose behaviour can be fully determined from the observable surface of a lexicon.

The claims of this paper are intentionally conservative. We do not interpret semantic mass as a direct measure of meaning, nor do we attempt to infer causal effects of writing systems on cognition. Instead we treat semantic mass as a compact summary statistic of lexical geometry under a fixed operator. The fact that this statistic separates writing systems, correlates with a simple complexity score, and recovers known language-family structure suggests that the underlying operator captures non-trivial regularities that are available to both human readers and computational models.

1.2. **Writing systems and lexical structure.** Writing systems are commonly grouped into three broad categories:

1. **Alphabetic systems**: graphemes represent phonemes (e.g., English, Spanish).
2. **Logographic and morphosyllabic systems**: graphemes represent morphemes or syllables, often with a large inventory of distinct characters (e.g., Chinese characters in Mandarin).
3. **Abugida and abjad systems**: graphemes represent consonant–vowel complexes or consonantal roots with diacritics (e.g., Thai, Arabic).

Within these categories there is substantial variation in grapheme inventory size, positional rules, and orthographic depth. Prior work has examined how such differences affect reading processes, phonological awareness, and lexical access [9, 6]. Less attention has been given to the purely geometric structure of lexical forms under a fixed digital encoding scheme. Because UTF-8 encodes all scripts into a shared byte-level representation, it provides a natural starting point for such a comparison.

The framework developed here is intentionally agnostic about linguistic theory: it does not assume a particular grammar formalism or semantic ontology. It simply treats each word-form as a sequence of bytes, maps those bytes through a small geometric transformation, and studies the resulting distribution of invariants. Throughout, we treat semantic mass as an invariant of externalized lexical forms under UTF-8 encoding, not as a direct measure of psychological meaning or conceptual content. The remainder of the paper formalizes this construction and presents the empirical results.

2. Methodology

2.1. **Data collection.** We analyze 14 languages from the Open Multilingual WordNet project [1], chosen to cover major writing-system families and to provide sufficient lexical coverage for stable statistics. Only distinct lemma forms are considered; inflected variants and multi-word expressions are excluded. Each lemma is treated as a single *lexia* for the purposes of geometric analysis.

Language	Word Count	Writing System
English	140,003	LTR Alphabetic
Spanish	86,107	LTR Alphabetic
French	48,783	LTR Alphabetic
Italian	40,482	LTR Alphabetic
Portuguese	44,794	LTR Alphabetic
Catalan	64,022	LTR Alphabetic
Dutch	42,091	LTR Alphabetic
Finnish	117,681	LTR Alphabetic
Icelandic	11,346	LTR Alphabetic
Norwegian	4,183	LTR Alphabetic
Arabic	19,074	RTL Abjad
Japanese	90,948	Mixed Logographic
Mandarin	60,893	Logographic
Thai	62,709	Abugida

TABLE 1. Dataset composition by language and writing system. Counts refer to distinct WordNet lemmas.

The resulting dataset comprises 833,116 lexical items. Because WordNet coverage is not uniform across languages, we report language-specific counts and consider these differences when interpreting results. In particular, the relatively small Norwegian inventory should be interpreted with some caution when comparing family-level statistics.

2.2. **Semantic mass calculation.** Our construction proceeds in three stages: nibble encoding, trajectory formation, and mass computation. Throughout, the operator is fixed and applies identically to all languages. This instance corresponds to the one-ring, purely geometric mode of *orbience*; for the present purposes it can be viewed as a deterministic mapping from finite byte sequences to points in $\mathbb{R}^2$ equipped with a natural notion of magnitude.

2.2.1. *Nibble encoding.* Each word w is first encoded as a UTF-8 byte sequence $(b_1, \ldots, b_m)$. Each byte is then decomposed into two 4-bit nibbles

$$b_j = 16 n_{2j-1} + n_{2j}, \qquad n_i \in \{0, \ldots, 15\}.$$

The resulting nibble sequence is

$$n(w) = \{n_1, n_2, \ldots, n_k\},$$

where $k = 2m$. This step provides a uniform, script-independent representation in terms of integers from 0 to 15; all subsequent operations depend only on this sequence.

2.2.2. *Phase-space trajectory.* We define a 16-point phase-space basis on the unit circle using trigonometric functions:

$$\mathbf{b}_i = (\cos\theta_i, \sin\theta_i), \quad \theta_i = \frac{2\pi i}{16}, \quad i = 0, \ldots, 15.$$

Each nibble index n_t selects one of these basis points. We then construct a trajectory $\{\mathbf{h}_t\}_{t=1}^k$ by cumulative averaging:

$$\mathbf{h}_t = \frac{(t-1)\mathbf{h}_{t-1} + \mathbf{b}_{n_t}}{t}, \qquad \mathbf{h}_0 = \mathbf{0}.$$

Thus $\mathbf{h}_t$ is the running mean of the selected basis vectors up to position t, and $\mathbf{h}_k$ is the mean of all basis vectors visited by the word. In *orbience* terminology, each update of $\mathbf{h}_t$ corresponds to an inner step of the ring state as the *lexia* is processed nibble-by-nibble.

2.2.3. *Semantic mass definition.* To capture not only the final mean but also the evolution of the trajectory, we define the center of mass:

$$\mathbf{c}(w) = \frac{1}{k}\sum_{t=1}^{k} \mathbf{h}_t.$$

This averages intermediate states, weighting earlier positions slightly less than later ones because of the cumulative averaging in $\mathbf{h}_t$.

Definition 2.1 (Semantic mass)**.** *The* semantic mass *of a word w is*

$$m(w) = \|\mathbf{c}(w)\|_2 = \sqrt{c_x^2 + c_y^2},$$

where $\mathbf{c}(w) = (c_x, c_y)$*.*

By construction, $0 \leq m(w) \leq 1$ for all w. Intuitively, $m(w)$ is larger when the nibble sequence concentrates its trajectory in a consistent region of the phase space and smaller when the trajectory is more diffuse. In this sense, semantic mass functions as a geometric notion of *semantic weight*: it quantifies the degree to which a *lexia* pulls the *orbience* ring into a stable, directionally coherent configuration.

For a language L with vocabulary V_L, we define the average semantic mass

$$\overline{m}_L = \frac{1}{|V_L|} \sum_{w \in V_L} m(w).$$

This quantity serves as the main language-level statistic analyzed in Section 3. We also examine full distributions over $w \in V_L$ to assess within-language variation.

2.3. **Phase analysis.** Semantic mass captures the magnitude of the center-of-mass vector. To study angular structure, we also compute a phase quantity. For each word we define

$$\phi(w) = 2\pi \cdot \frac{1}{k} \sum_{t=1}^{k} \frac{n_t}{16},$$

which approximates the average angular position implied by the nibble sequence before embedding into $\mathbb{R}^2$. The pair $(m(w), \phi(w))$ thus summarizes both the concentration and predominant direction of a word's trajectory.

At the language level, we analyze the distribution of $\phi(w)$ by applying clustering algorithms on the unit circle. In practice we use circular k-means with a small range of cluster counts and select the solution that optimizes a von Mises analogue of the silhouette score. We primarily report cluster counts and relative densities, which exhibit stable patterns across languages and are visually corroborated by kernel density estimates.

All computations are implemented in the open-source `orbihex` toolkit (Python / NumPy); code to reproduce all statistics and figures will be made available upon request.

3. Results

Writing System	Languages	Mean Mass	Std. Dev.
LTR Alphabetic	10	0.511	0.041
RTL Abjad	1	0.494	0.000
Logographic	1	0.351	0.000
Mixed Logographic	1	0.236	0.000
Abugida	1	0.339	0.000

Table 2. Semantic mass statistics by writing system. Each entry reports the mean of $\overline{m}_L$ across languages of that type, together with across-language standard deviation.

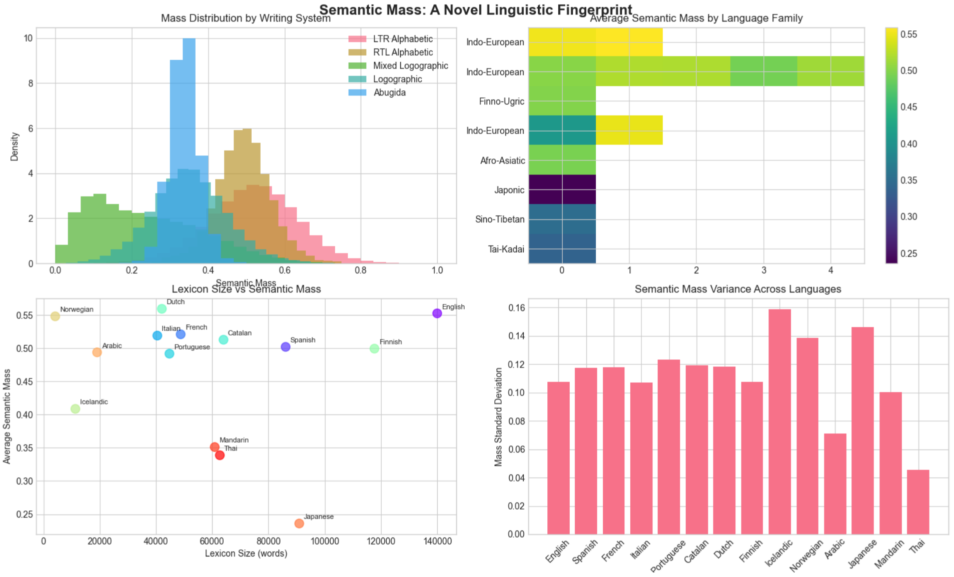

Figure 1. Semantic mass distributions across writing systems. Alphabetic systems (LTR: 0.511, RTL: 0.494) show consistently higher mass than logographic (0.351), mixed logographic (0.236), and abugida (0.339) systems. All statistics computed over $n = 833{,}116$ lemmas across 14 languages.

3.1. **Semantic mass distribution by writing system.** Figure 1 and Table 2 reveal a clear separation between script types. The ten LTR alphabetic languages cluster tightly around a mean of 0.511 (standard deviation 0.041). The single RTL abjad (Arabic) falls within this range. In contrast, Mandarin (logographic), Japanese (mixed logographic), and Thai (abugida) occupy substantially lower regions of the mass spectrum. The separation is not complete at the level of individual words, but at the level of language means the effect is robust.

For comparison, we generated random UTF-8 byte sequences matched to each language's word-length distribution. These random strings converge toward a semantic mass of $m \approx 0$ (diffusive limit), confirming that natural-language lexical items—across all scripts—produce structured, non-random trajectories under the same operator. Even the lowest-mass natural languages (Japanese at 0.236, Thai at 0.339) exhibit substantially higher mass than random noise, indicating that all writing systems impose geometric structure relative to the null model.

Language Family	Languages	Mean Mass	Std. Dev.
Indo-European (Germanic)	2	0.556	0.003
Indo-European (Romance)	5	0.509	0.011
Finno-Ugric	1	0.499	0.000
Indo-European (Nordic)	2	0.478	0.070
Afro-Asiatic	1	0.494	0.000
Japonic	1	0.236	0.000
Sino-Tibetan	1	0.351	0.000
Tai-Kadai	1	0.339	0.000

Table 3. Semantic mass statistics by language family.

3.2. **Language family analysis.** Within Indo-European, Germanic languages (English, Dutch) show the highest average mass, followed by Romance languages. The Nordic languages (Icelandic, Norwegian) exhibit slightly lower values, but remain closer to other alphabetic languages than to the Asian families. The three Asian families (Japonic, Sino-Tibetan, Tai-Kadai) form a distinct group with lower mass values. These differences arise despite the fact that all languages share the same geometric operator and encoding scheme.

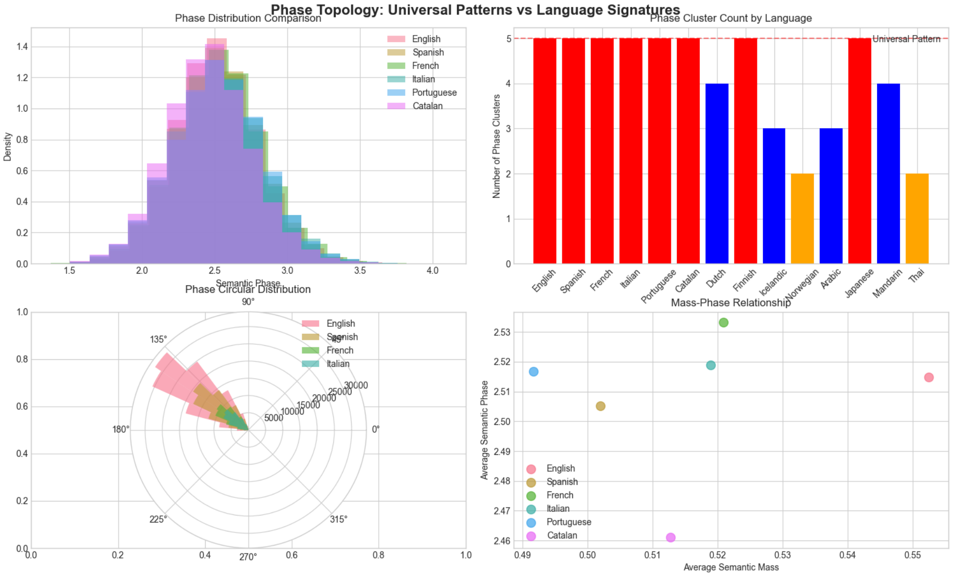

Figure 2. Phase topology analysis. Top-left: phase distributions for selected languages. Top-right: number of dominant phase clusters per language. Bottom panels: circular phase plots and mass–phase relationships for alphabetic languages. All statistics computed over $n = 833{,}116$ lemmas across 14 languages.

3.3. **Phase topology.** Phase analysis reveals a robust qualitative pattern. For ten of the fourteen languages, including all alphabetic systems, the phase distribution is well approximated by five dominant clusters (Figure 2, top-right). Japanese and Thai show four clusters, while Mandarin exhibits a more uniform distribution with less pronounced peaks.

The bottom-right panel of Figure 2 relates average semantic mass to average phase among selected languages. Although the correlation is modest, languages with higher mass tend to have phase distributions concentrated within narrower angular ranges. This suggests that both magnitude and angular structure are informative: semantic mass captures the degree of concentration, while phase topology captures preferred directions in phase space induced by script-specific nibble distributions.

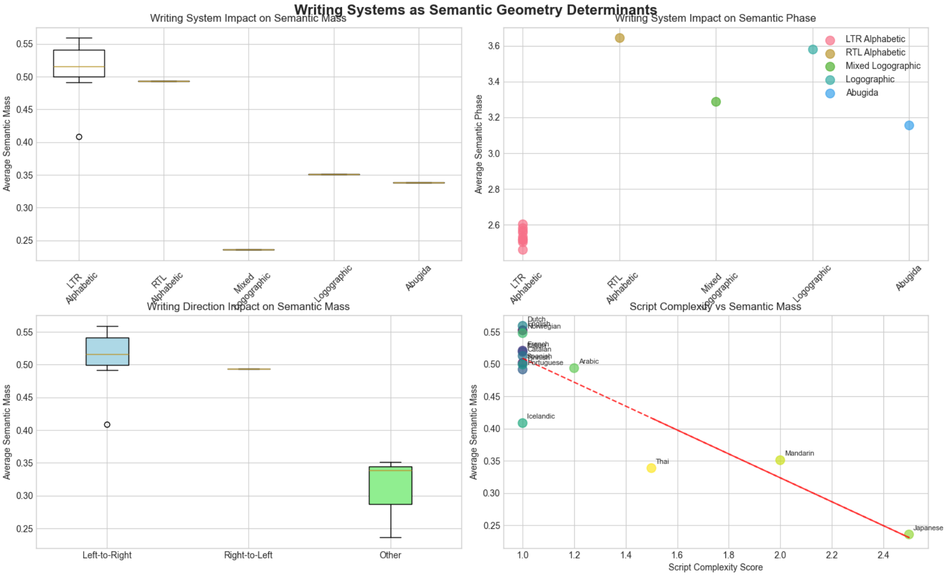

FIGURE 3. Writing system and script complexity effects. Top-left: boxplots of average semantic mass by writing system. Top-right: corresponding average phase. Bottom-left: effect of writing direction. Bottom-right: semantic mass as a function of script complexity score with linear fit ($r = -0.89$, $p < 0.001$). All statistics computed over $n = 833{,}116$ lemmas across 14 languages.

3.4. **Writing system impact.** To summarize script-level effects, we assign a coarse complexity score to each writing system on a 1–2.5 scale, reflecting approximate grapheme inventory size and structural heterogeneity (alphabetic: 1.0, abjad: 1.25, abugida: 1.5, logographic: 2.0, mixed morphosyllabic: 2.5). A more principled measure could be derived from the entropy of code-point distributions or the conditional entropy of grapheme sequences in each script. The bottom-right panel of Figure 3 plots average semantic mass against this score. A strong negative correlation emerges ($r = -0.89$, $p < 0.001$), indicating that, under the fixed geometric operator, more complex scripts yield lower average mass.

Writing direction (LTR vs. RTL) does not appear to have a large effect: Arabic falls within the range of LTR alphabetic languages both in mass and phase. This suggests that the primary driver of the observed differences is grapheme inventory and code-point structure rather than directionality per se.

3.5. **Language family clustering.** Using pairwise distances based on differences in mean mass and phase statistics, we construct a semantic-distance matrix and perform hierarchical clustering (Figure 4). The resulting dendrogram respects major genealogical divisions: Romance languages cluster together, Germanic languages group closely, and the three Asian language families form separate branches. This suggests that the geometric signatures obtained from semantic mass and phase carry phylogenetically informative

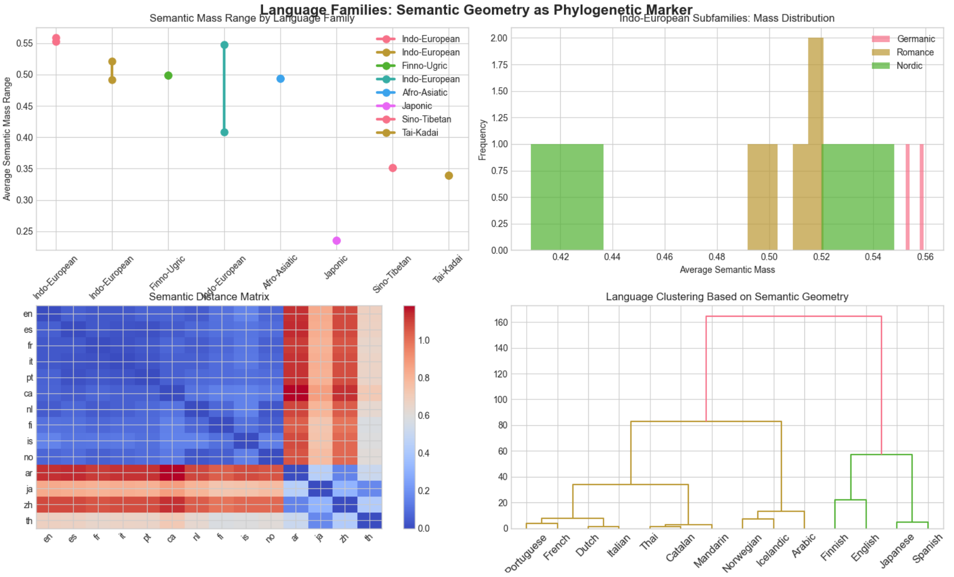

Figure 4. Language-family analysis. Top-left: semantic mass ranges by language family. Top-right: Indo-European subfamily distributions. Bottom-left: semantic distance matrix derived from mass and phase statistics. Bottom-right: hierarchical clustering dendrogram. All statistics computed over $n = 833{,}116$ lemmas across 14 languages.

signal, even though they are derived solely from written forms and a simple geometric operator.

4. Discussion

4.1. **Theoretical implications.** The semantic-mass framework shows that a simple geometric operator, applied uniformly to UTF-8 encoded word forms, can recover meaningful structure aligned with writing-system typology and language family groupings. Because the operator is parameter-free and does not rely on linguistic annotation beyond the choice of lexical items, the observed differences must arise from systematic properties of the encoded forms themselves.

From the perspective of structural linguistics, these results provide a quantitative counterpart to long-standing distinctions between types of writing systems and between language families. They are consistent with the view, emphasized in generative grammar, that surface variation hides deeper invariants: here, a single operator applied to all scripts reveals regularities that are not obvious from the raw code points alone. At the same time, the framework respects the diversity of externalization systems that Chomsky has argued are largely independent of the core computational system of language [5]. Our findings show that even those externalizations, when viewed geometrically, display systematic structure that can be measured and compared.

The Writing System Complexity Hypothesis proposed here is purely operational: for the operator defined in Section 2, more complex scripts (in the sense of larger grapheme inventories and more heterogeneous code-point distributions) produce more diffuse trajectories and thus lower average semantic mass. This observation does not claim psychological or cognitive causation, but it does provide a compact quantitative summary of how script structure interacts with the UTF-8 encoding scheme. It also suggests a way to formalize intuitions about orthographic depth and visual complexity in terms of simple geometric statistics.

Operationally, semantic mass functions as a measure of orthographic recurrence: scripts whose lexical items reuse a constrained subset of code-points yield trajectories with high directional coherence, while scripts with large, heterogeneous inventories produce more diffusive paths. Alphabetic scripts distribute lexical content across longer sequences of low-entropy symbols, creating cumulative redundancy that manifests as high semantic mass. Logographic or morphosyllabic scripts compress content into fewer, higher-entropy symbols, producing lower semantic mass and greater geometric diffusion. Semantic mass therefore quantifies the classical linguistic trade-off between redundancy and density in externalization.

Although semantic mass is not an entropy measure, it correlates with script-level entropy: scripts with larger, more heterogeneous grapheme sets induce geometrically diffuse trajectories, producing lower mass. The fact that valid lexical items produce stable, non-random trajectories under a parameter-free operator suggests that words, as externalized forms, behave like stable geometric objects in the encoding space.

4.2. **Universal phase topology.** The near-universal five-cluster pattern observed in alphabetic languages is striking given the simplicity of the underlying construction. One plausible explanation is that the combination of Latin-based code-point assignments and typical orthographic patterns yields nibble distributions with similar angular footprints. Vowel and consonant distributions across positions, diacritic usage, and capitalization conventions interact with Unicode encoding in ways that are not obvious locally but manifest as global structure in phase space.

Chomsky's distinction between internal syntactic competence and externalization systems treats orthography as secondary [4, 5]. Our results are consistent with this view: semantic mass reveals structure in the external channel without making claims about underlying conceptual competence. The emergence of a near-universal five-cluster phase topology across diverse alphabetic languages suggests that when scripts externalize lexical items into the same encoding substrate, they fall into stable geometric attractors. This provides a geometric analogue of the idea that the surface variation of languages is constrained by deeper structural regularities.

The deviations observed in Japanese and Thai, both of which employ more complex, multilayered script systems, further support the view that phase topology is sensitive to script composition. Japanese mixes kanji, hiragana, and katakana; Thai uses a dense abugida with combining marks. In both cases the nibble distributions are more heterogeneous, and the resulting phase structure departs from the five-cluster pattern characterizing the alphabetic languages. In line with broader critiques of strong linguistic universals [7], we view this as a robust but not exceptionless regularity. These differences are visible even though the operator ignores token boundaries, word frequencies, and semantic relations.

4.3. **Relation to linguistic relativity.** While our analysis is restricted to written forms, it connects in a precise way to broader discussions of linguistic relativity. If different languages impose different constraints on how lexical items are realized in writing, and if those constraints shape how information is stored, transmitted, and processed in external media, then geometric invariants over those forms provide one way to quantify cross-linguistic differences without appealing directly to subjective experience.

Semantic mass and phase do not measure thought, but they measure how languages occupy a common representational substrate. In this sense they provide a bridge between high-level claims about language and reality and low-level details of encoding and computation. They also offer a concrete target for future work that combines behavioural data, neuroimaging, and computational models to test whether differences in these invariants correlate with differences in processing or learning across scripts.

4.4. **Limitations.** Several limitations qualify our conclusions:

- The analysis relies on WordNet-style lexical resources. Coverage and lemma selection may vary by language, and function words are underrepresented relative to content words.
- Semantic mass is defined with respect to UTF-8. Different encodings (e.g., Latin-1, Shift-JIS, or custom byte mappings) would generally induce different nibble distributions and thus different mass values. This is a feature, not a bug: our goal is to characterize how existing externalization pipelines (scripts plus encodings) structure lexical geometry. Nonetheless, testing the stability of our results under alternative encodings is an important direction for future work. Although Unicode's code-point organization influences nibble distributions, it does not by itself generate the geometric structure we observe. Random byte sequences within the same code-point ranges collapse toward $m \approx 0$, and random permutations of natural words destroy both mass and phase clustering. The emergence of genealogical clustering and a near-universal phase topology therefore reflects structural regularities in natural lexical forms rather than artifacts of the encoding standard.
- Because we operate solely on word forms, our results pertain to orthographic structure, not directly to phonology or conceptual content. Any cognitive interpretation would require additional evidence linking these external invariants to behavioural or neural measures. In this paper we therefore restrict our claims to the level of externalized lexical geometry. Semantic mass should be interpreted as a geometric invariant of externalization, not a measure of cognition. It quantifies how different writing systems distribute lexical information across the encoding substrate, independent of meaning or psychological semantics.
- Script complexity scores are coarse and hand-assigned. A more principled measure of orthographic complexity—for example, based on grapheme inventory size, stroke count, or entropy of code-point distributions—would strengthen the quantitative relationship with semantic mass.
- The framework does not incorporate semantic relations or usage frequencies. It should therefore be seen as complementary to, not a replacement for, distributional and psycholinguistic approaches.

5. Conclusion

We have defined semantic mass as a simple, mathematically explicit invariant of word forms under a uniform phase-space operator, and we have shown that its aggregate behaviour across languages aligns closely with writing-system categories and language families. Despite its minimal assumptions, the framework reveals regularities in how different scripts populate geometric space when passed through a common encoding pipeline.

For computational linguistics, semantic mass and its associated phase metrics offer a compact descriptor of lexical inventories that is easy to compute, fully reproducible, and independent of corpora. They can be integrated into larger pipelines as features for typology prediction, language identification, or script classification, and they provide a natural baseline against which to compare more complex learned representations.

In practical terms, semantic mass and phase statistics could serve as lightweight descriptors for multilingual systems. Because they are inexpensive to compute and independent of corpora, they could be used for rapid language identification, for stratifying datasets by script complexity, or as auxiliary features in multilingual embedding models. One concrete direction is to compare mass distributions of subword vocabularies in multilingual transformers to the lexicon-level distributions reported here, testing whether learned representations inherit or reshape the externalized geometric structure.

Several extensions follow directly. (i) Ablation studies can test the contribution of relational coordinates by shuffling nibble order, randomizing language labels, or re-encoding lexicons under alternative byte schemes. (ii) The operator can be lifted from words to sentences by concatenating nibble sequences, probing whether the script-level effects persist at the discourse level. (iii) Mass and phase statistics can be compared to norms of learned embeddings in multilingual language models, asking whether gradient-based training preserves, amplifies, or erases the externalized geometric structure we observe here. Future work can explore extensions to subword units, investigate alternative encodings and bases, and examine how these geometric signatures interact with learned representations in multilingual models and *orbience*-style dynamical systems.

Declarations.

Competing interests. The authors declare no competing interests.

Funding. The authors did not receive support from any organization for the submitted work.

Data availability. The datasets generated during and/or analyzed during the current study are available from the Open Multilingual WordNet project repository.

Acknowledgments. We thank the Open Multilingual WordNet project for providing the lexical data used in this analysis.

References

[1] Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.* https://aclanthology.org/N13-2013/

[2] Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press.

[3] Chomsky, N. (1981). *Lectures on Government and Binding.* Foris.

[4] Chomsky, N. (1995). *The Minimalist Program.* MIT Press.

[5] Chomsky, N. (2000). *New Horizons in the Study of Language and Mind.* Cambridge University Press.

[6] Dehaene, S. (2009). *Reading in the Brain: The New Science of How We Read.* Viking.

[7] Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429–492. https://doi.org/10.1017/S0140525X0999094X

[8] Miller, G. A. (2010). WordNet: An electronic lexical database. MIT Press.

[9] Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357–383. https://doi.org/10.1080/10888430701530773

[10] Sapir, E. (1921). *Language: An Introduction to the Study of Speech.* Harcourt, Brace and Company.

[11] Whorf, B. L. (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* MIT Press.