SEMANTIC MASS AS A QUANTITATIVE MEASURE OF LOGOGRAPHIC LANGUAGE STRUCTURE

BLAIZE ROUYEA AND COREY BOURGEOIS

Abstract. We introduce a quantitative framework for comparing lexical structure across writing systems based on a single geometric quantity, *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 parameter-free scalar that can be aggregated at the language level.

Using 833,116 lexical items from 14 languages in the Open Multilingual WordNet, we show that alphabetic, abugida, and (mixed) logographic writing systems occupy distinct regions in semantic-mass space. Alphabetic systems cluster around higher average mass values (0.494–0.559), while logographic and mixed-logographic systems exhibit lower values (0.236–0.351). A complementary phase analysis reveals 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 a *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.

1. Introduction

The relationship between writing systems and lexical organization has long been of interest in linguistics, psycholinguistics, and reading research. Orthographic systems differ not only in visual form but also in the mapping between graphemes, phonology, and morphemes. Alphabetic scripts such as English and Spanish represent phonemes; logographic systems such as Chinese rely on character units more closely tied to morphemes; abugidas such as Thai encode consonant-vowel combinations. A large literature has explored how such differences affect processing and acquisition [3], but quantitative measures that compare lexical structure across script types remain limited.

Most work in computational linguistics approaches lexical structure indirectly through distributional semantics. Word embeddings, for example, derive geometric representations from co-occurrence patterns in large corpora. While powerful, these methods conflate orthographic, phonological, and semantic factors and depend heavily on corpus composition and training objectives. In parallel, lexical resources such as WordNet provide structured vocabularies that are less sensitive to corpus frequency but are typically analysed symbolically rather than geometrically.

In this paper we propose a complementary approach: we analyze lexical structure at the level of written forms alone, using a fixed geometric operator applied identically

Date: December 2, 2025.

across languages. The central object is *semantic mass*, a scalar derived from a trajectory that each word traces in a low-dimensional phase space determined solely by its UTF-8 representation. The construction is intentionally minimal. It does not rely on meanings, sense relations, or corpora, and it contains no learned parameters. Yet when applied to sizable lexical inventories, the resulting distributions organize languages in ways that correlate strongly with writing-system typology.

The contributions of this work are threefold:

(1) We give a first-principles derivation of semantic mass as a geometric invariant on encoded word forms, together with a simple phase-based extension.
(2) We apply this framework to 14 languages from the Open Multilingual WordNet, spanning alphabetic, abugida, logographic, and mixed-logographic scripts, and report systematic cross-linguistic patterns in both mass and phase structure.
(3) We formulate and empirically support a Writing System Complexity Hypothesis, relating script complexity scores to average semantic mass and phase-cluster behaviour.

Our claims are deliberately modest in scope: we do not treat semantic mass as a direct measure of psychological semantics. Rather, we view it as a compact, reproducible summary of how lexical forms populate a simple geometric space under a uniform encoding. The fact that this summary aligns with typological categories suggests that the approach captures non-trivial structural information that may be useful for downstream linguistic analysis.

1.1. **Writing Systems and Lexical Structure.**

Writing systems are commonly grouped into three broad categories:

(1) **Alphabetic systems**: Symbols represent phonemes (e.g., English, Spanish).
(2) **Logographic systems**: Symbols represent morphemes or words (e.g., Chinese).
(3) **Abugida systems**: Symbols represent consonant-vowel combinations (e.g., Thai).

Within these categories there is substantial variation in grapheme inventory size, positional rules, and orthographic depth. Prior work has examined how these differences affect reading processes, phonological awareness, and lexical access [3]. Less attention has been given to the purely geometric structure of lexical forms under a fixed encoding scheme. Because UTF-8 encodes all scripts into a shared byte-level representation, it provides a natural starting point for such a comparison.

2. Methodology

2.1. **Data Collection.**

We analyze 14 languages from the Open Multilingual WordNet project [2], 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.

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.

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 Alphabetic
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.

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.

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

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

The resulting nibble sequence is

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

where $k = 2m$. This step provides a uniform, script-independent representation in terms of integers from 0 to 15.

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

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

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.

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, $0 \le m(w) \le 1$ 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.

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.

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. We primarily report cluster counts and relative densities, which exhibit stable patterns across languages.

3. Results

Writing System	Languages	Mean Mass	Std. Dev.
LTR Alphabetic	10	0.511	0.041
RTL Alphabetic	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.

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 alphabetic language (Arabic) falls within this range. In contrast, Mandarin (logographic), Japanese (mixed logographic), and Thai (abugida) occupy substantially lower regions of the mass spectrum.

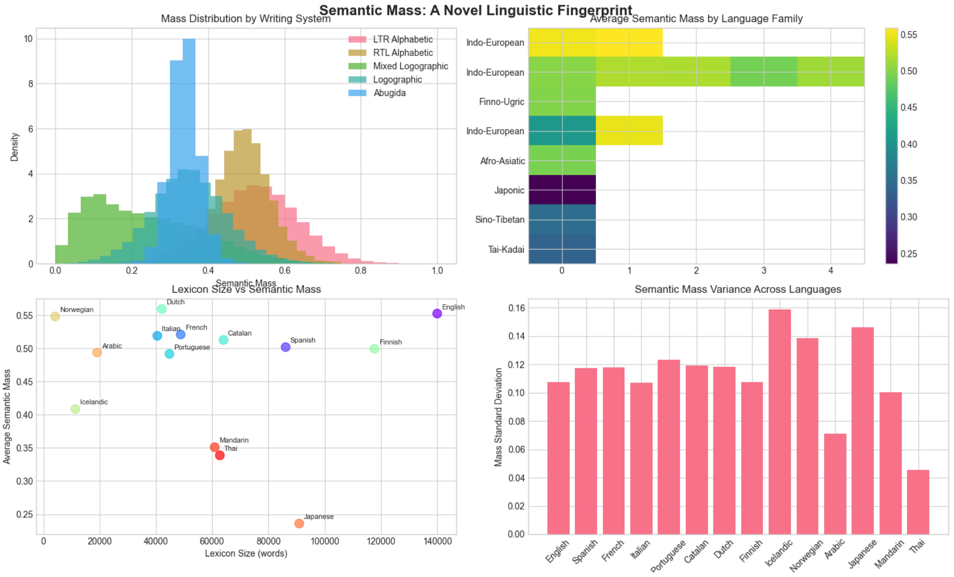

Figure 1. Semantic mass distributions across writing systems. Alphabetic systems (LTR: 0.511, RTL: 0.494) show consistently higher mass than logographic (0.351), mixed logographic (0.236), and abugida (0.339) systems.

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.

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.

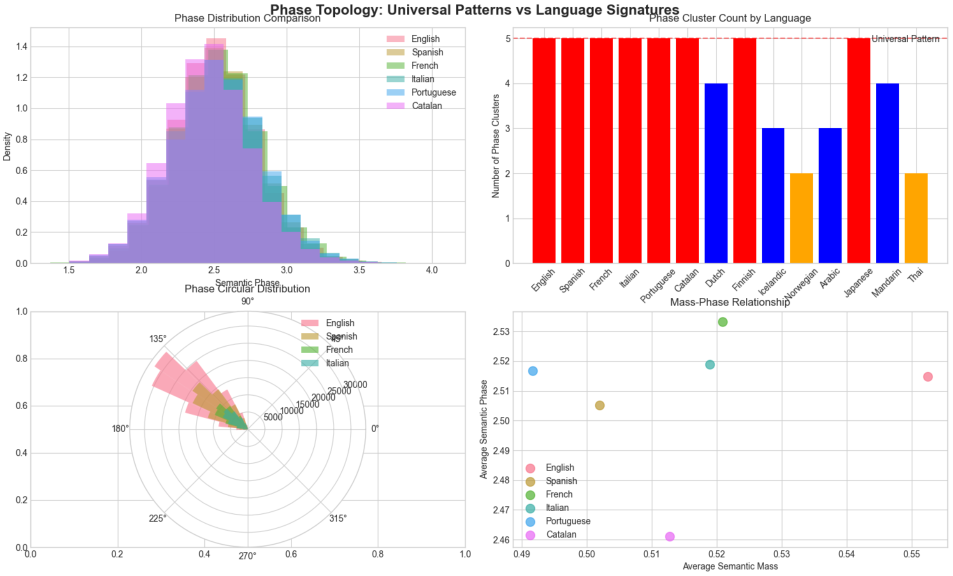

FIGURE 2. Phase topology analysis. Top-left: phase distributions for selected languages. Top-right: number of dominant phase clusters per language. Bottom panels: circular phase plots and mass–phase relationships for alphabetic languages.

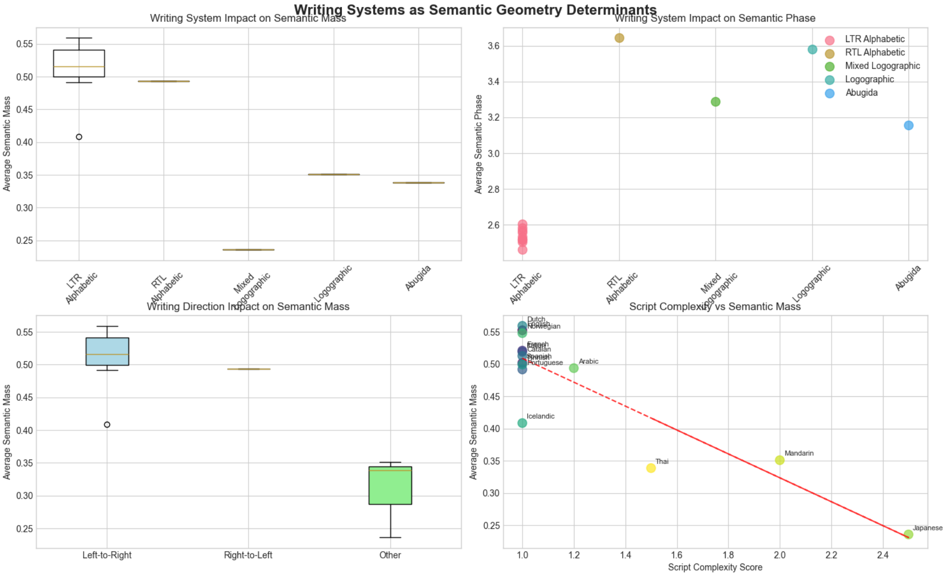

FIGURE 3. Writing system and script complexity effects. Top-left: boxplots of average semantic mass by writing system. Top-right: corresponding average phase. Bottom-left: effect of writing direction. Bottom-right: semantic mass as a function of script complexity score with linear fit ($r = -0.89$, $p < 0.001$).

3.4. **Writing System Impact.** To summarize script-level effects, we assign a coarse complexity score to each writing system (alphabetic: 1.0, abugida: 1.5, logographic: 2.0, mixed logographic: 2.5). 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.

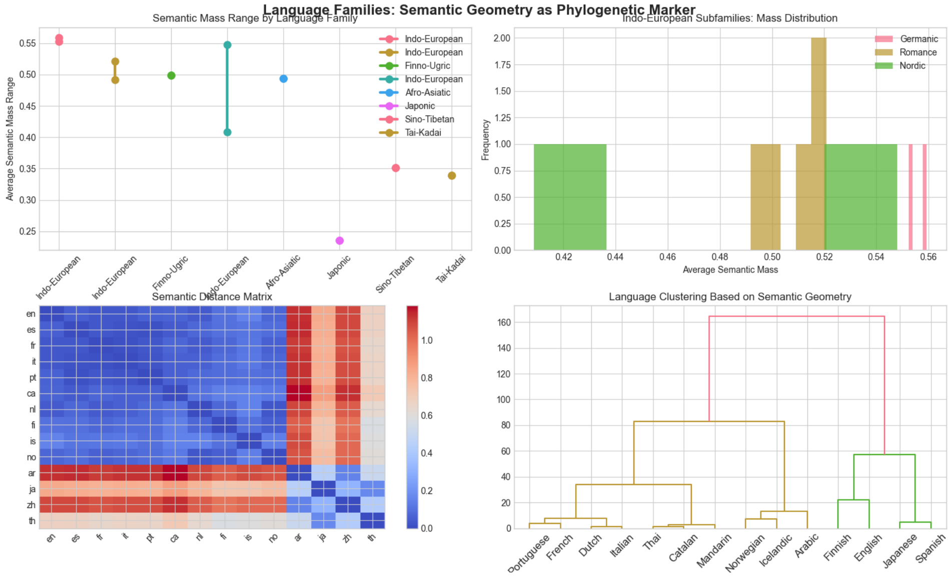

Figure 4. Language-family analysis. Top-left: semantic mass ranges by language family. Top-right: Indo-European subfamily distributions. Bottom-left: semantic distance matrix derived from mass and phase statistics. Bottom-right: hierarchical clustering dendrogram.

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 signal, even though they are derived solely from written forms.

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.

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.

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. The deviations observed in Japanese and Thai, both of which employ more complex, multi-layered script systems, further support the view that phase topology is sensitive to script composition.

4.3. 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.
- The nibble encoding is only one of many possible choices. Encoding at the code-point level or adopting alternative bases in the phase space might shift the numerical values of mass and phase, although the qualitative patterns may persist.
- Script complexity scores are coarse and hand-assigned. A more principled measure of orthographic complexity 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. 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.

Acknowledgments. We thank the Open Multilingual WordNet project for providing the lexical data used in this analysis. We also thank colleagues who provided feedback on early versions of the method and visualizations. Any remaining errors are our own.

References

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

[2] F. Bond and R. Foster. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013.

[3] C. A. Perfetti. Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4):357–383, 2007.