

Structured Knowledge Accumulation: Real-Time Discovery of the Universal Language Manifold

Bouarfa Mahi
Quantiota
Email: info@quantiota.org

December 3, 2025

Abstract

Human languages differ in vocabulary, syntax, and surface structure, yet they all share an underlying dynamical principle: meaning emerges through the progressive reduction of uncertainty. This paper proposes that the universal structure behind all languages can be revealed through *Structured Knowledge Accumulation* (SKA), a real-time, entropy-based learning framework. While large language models learn static correlations across tokens, SKA follows the physical process that creates meaning by enforcing the *law of entropic least action*, a variational principle governing the flow of information in time. To use a universal and modality-independent input, we take spoken words—the raw acoustic stream that comes out of our mouths—as the fundamental signal driving the real-time learning process. By analyzing this continuous sound flow, SKA reconstructs a latent *language manifold*—a geometry of knowledge states whose evolution is independent of the spoken or written language. This framework reveals English, French, Arabic, Mandarin, and all other languages to be different coordinate projections of the same entropy-evolving informational structure. This framework unifies linguistics, cognition, and real-time learning under a single physical law, offering a new foundation for understanding meaning, translation, and intelligence beyond classical symbolic or neural approaches.

1 Introduction

Human languages differ widely in vocabulary, syntax, phonology, and cultural development, yet they all express meaning through the same physical channel: a continuous acoustic stream whose structure evolves in time. This paper proposes that the universality of language can be understood not through symbolic rules or innate grammatical templates, but through a physical principle governing the flow of information. Within the Structured Knowledge Accumulation (SKA) framework, meaning emerges from the progressive reduction of entropy under a forward-only variational law—the *entropic least action*. When applied to raw speech, this principle reconstructs a latent geometry of knowledge states: the **Universal Language Manifold (ULM)**.

Unlike classical linguistic theories that posit Universal Grammar, or modern self-supervised speech models that learn statistical structure from large audio corpora, the ULM does not rely on symbolic abstractions or learned discrete units. Instead, it derives linguistic universality from the physics of sound production and real-time cognitive processing. The acoustic waveform carries formants, harmonics, prosody, and coarticulatory patterns that constrain how meaning can be encoded; SKA organizes these constraints into a dynamic Riemannian manifold whose curvature reflects the structure of uncertainty reduction. Different spoken languages correspond to different coordinate projections of this same manifold.

This formulation explains linguistic diversity without abandoning universality: English, Arabic, Mandarin, and thousands of others become distinct coordinate systems defined over a shared

entropy-evolving geometry. Translation between languages thus becomes a coordinate transformation rather than the alignment of symbolic tokens. Because the manifold is reconstructed from real-time acoustic trajectories, the framework naturally integrates linguistic theory, the physics of speech, and information geometry into a single physical model of language.

To validate this hypothesis empirically, we rely on a multilingual audio corpus with identical semantic content across languages. Spoken Bible recordings in over 2,000 languages enable direct comparison of acoustic entropy trajectories under constant meaning, allowing the SKA framework to isolate invariant manifold structure across linguistic families. This makes the Universal Language Manifold a falsifiable scientific theory rather than a symbolic speculation.

2 Literature

The search for a **Universal Language Manifold** connects four traditionally separate research domains: linguistic theory, self-supervised learning from raw speech, the physics of acoustic signals, and the SKA framework’s entropy-based information geometry. Below we review the relevant work across these domains and show how they converge toward a unified physical foundation for linguistic universality.

2.1 Linguistics and the Search for Universal Structure

The generative tradition initiated by Chomsky argues that the diversity of human languages is governed by a set of shared underlying principles known as *Universal Grammar*. Classic formulations [24, 25] propose an innate, biologically specified architecture that constrains the space of possible grammars. More recent work explores mathematical and algebraic formulations of linguistic universals, such as the algebraic structure of Merge and its invariants [26]. Crucially, Universal Grammar has remained biologically specified and modality-agnostic in principle, yet has never been derived from the physical signal itself. The present framework closes this gap: the algebraic invariants of Merge and the constraints on possible grammars emerge as geometric consequences of entropic least action on the acoustic manifold, without requiring any innate symbolic module. Parallel research in historical linguistics documents that writing systems emerged only around 3300–3200 BC [1, 3, 14], while spoken language predates writing by tens of thousands of years. These findings support the view that linguistic structure originates in the acoustic stream and cognitive-perceptual constraints rather than cultural inscription systems.

2.2 Self-Supervised Learning from Raw Acoustic Streams

Modern self-supervised speech models demonstrate that linguistic structure can be learned directly from the raw audio waveform. Systems such as wav2vec 2.0 [15], HuBERT [16], and RawNet [17] bypass human-designed features (e.g., MFCCs, phonemes), mapping the acoustic waveform into structured latent spaces that encode phonetic, articulatory, and prosodic information. Research on *Acoustic Unit Discovery* shows that meaningful subword units can be extracted without any labels [18], and unsupervised word segmentation from discretized acoustic units [19] demonstrates that lexical structure can emerge from the continuous signal alone. Recent advancements in scaling to multilingual settings, such as Whisper [22] and MMS [23], extend SSL from raw audio to diverse languages, showing emergent structure in multimodal and whispered speech without labels, further supporting the discovery of universal representations across linguistic diversity. These findings empirically support the SKA hypothesis that the acoustic stream contains sufficient structure for real-time discovery of a universal manifold.

2.3 Acoustic Physics and Information Geometry

Spoken language is not merely a symbolic sequence but a continuous physical signal governed by the laws of acoustics, resonance, and nonlinear dynamics. The patterns formed by formants, harmonics, coarticulation, and prosody embed structural information that constrains how meaning can be encoded. This connects naturally to information geometry, where the evolution of knowledge states is modeled as trajectories on curved manifolds shaped by entropy and variational principles [29]. The SKA framework uses the *law of entropic least action* to define how information flows in time, aligning linguistic universality with physical invariants rather than genetically pre-specified rules.

2.4 Integration with the SKA Framework

The Structured Knowledge Accumulation (SKA) framework proposes that meaning emerges from the continuous reduction of entropy under a forward-only learning dynamic. When applied to the raw acoustic stream, SKA reconstructs a latent *language manifold*: a Riemannian geometry of knowledge states whose curvature is invariant across all spoken languages. Within this formulation:

- linguistic structure arises from physical constraints in the acoustic-entropy field;
- languages correspond to different coordinate projections of the same manifold;
- no symbolic priors, syntactic templates, or phonetic categories need to be built in;
- the universality of language is explained by geometry, not innate grammatical rules.
- translation between any pair of languages becomes a coordinate transformation on the manifold rather than alignment of discrete symbolic trees.

This positions SKA as a unifying framework that bridges generative linguistics, modern speech self-supervision, and information-theoretic physics, offering a physical basis for linguistic universals grounded in the raw acoustic signal.

3 Dataset

The empirical validation of the Universal Language Manifold (ULM) relies on a comprehensive multilingual audio corpus that enables direct comparison of linguistic projections while maintaining semantic constancy. This project utilizes spoken Bible recordings from Faith Comes By Hearing (FCBH), a repository providing complete audio Bibles in over 2,000 languages [30]. This dataset is uniquely suited to the Structured Knowledge Accumulation (SKA) framework for several key reasons, ensuring robust testing of the ULM hypothesis.

- **Semantic Constancy Across Languages:** The dataset features the same textual content translated across all languages, enabling precise comparison of how diverse linguistic systems project identical semantic structures onto the manifold. This constancy isolates coordinate variations, allowing SKA to reconstruct invariant geometric features without confounding factors from content differences.
- **Continuous, Natural Speech Streams:** The recordings consist of uninterrupted, full-text narrations, ideal for extracting real-time acoustic entropy trajectories. This format supports the law of entropic least action by providing raw, modality-independent signals for manifold reconstruction, as required by SKA’s forward-only processing.

- **Global Linguistic Coverage:** Encompassing nearly every major language family—including Indo-European, Sino-Tibetan, Afro-Asiatic, Niger-Congo, and many rare or endangered languages—the corpus offers an unparalleled empirical sampling. This breadth facilitates statistical validation of the manifold’s universality, with coverage representing over 90% of the world’s spoken languages [30].
- **Neutral, Uncharted Reading Style:** The audio employs a consistent, non-dramatized narration style, preserving the pure acoustic structure (e.g., formants, harmonics, and prosody) essential for entropy analysis. This minimizes extraneous variations, ensuring high-fidelity input for geometric computations.
- **Multi-Scale Temporal Alignment:** Fine-grained verse-level timestamps can be automatically generated using forced alignment techniques based on Dynamic Time Warping (DTW) with Mel-frequency cepstral coefficients (MFCC). This enables controlled experiments on entropy trajectories at multiple temporal scales—from phoneme to word to verse to chapter—while maintaining semantic consistency across languages. The alignment process synthesizes text fragments using text-to-speech engines and applies the Sakoe-Chiba Band DTW algorithm to map synthesized audio onto real recordings with millisecond precision, providing natural segmentation points for manifold analysis at varying levels of linguistic granularity.

The following sections describe how SKA processes these acoustic streams to reconstruct the universal manifold geometry and validate coordinate invariance across language families.

4 Methods

5 Results

6 Discussion

7 Conclusion

References

- [1] Schmandt-Besserat, Denise. *How Writing Came About*. University of Texas Press, Austin, 1996. (Authoritative account: dates the emergence of writing in Mesopotamia to ca. 3300–3200 BC.)
- [2] Schmandt-Besserat, Denise. *Before Writing, Volume I: From Counting to Cuneiform*. University of Texas Press, Austin, 1992. (Explains precursor token systems and the transition to proto-cuneiform writing.)
- [3] Englund, Robert K. “Proto-Cuneiform Account-Books and Journals.” In Michael Hudson and Cornelia Wunsch (eds.), *Creating Economic Order: Record-Keeping, Standardization, and the Development of Accounting in the Ancient Near East*, CDL Press, Bethesda, MD, 2004, pp. 23–46. (Detailed analysis of Uruk proto-writing ca. 3300 BC.)
- [4] Cooper, Jerrold S. “Babylonian Beginnings: The Origin of the Cuneiform Writing System in Comparative Perspective.” In Stephen D. Houston (ed.), *The First Writing: Script Invention as History and Process*, Cambridge University Press, Cambridge, 2004, pp. 71–99. (High-level comparative analysis; confirms earliest writing in southern Mesopotamia.)

- [5] Glassner, Jean-Jacques. *The Invention of Cuneiform: Writing in Sumer*. Translated by Zainab Bahrani and Marc Van De Mieroop. Johns Hopkins University Press, Baltimore, 2003. (Leading academic treatment of the emergence of writing ca. 34th century BC.)
- [6] Nissen, Hans J., Peter Damerow, and Robert K. Englund. *Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East*. Translated by Paul Larsen. University of Chicago Press, Chicago, 1993. (Seminal reference: dates proto-cuneiform to between 3400–3000 BC.)
- [7] Cooper, Jerrold S. “Writing in Early Mesopotamia.” In Daniel C. Snell (ed.), *A Companion to the Ancient Near East*, Blackwell Publishing, Malden, MA / Oxford, 2008, pp. 213–231. (Comprehensive chronological synthesis of early writing systems.)
- [8] Houston, Stephen D. (ed.). *The First Writing: Script Invention as History and Process*. Cambridge University Press, Cambridge, 2004. (Comparative academic volume on independent origins of script technologies.)
- [9] Fischer, Steven Roger. *A History of Writing*. Reaktion Books, London, 2004. (Global scholarly treatment; places earliest writing at ca. 3300 BC.)
- [10] Woods, Christopher (ed.). *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. Oriental Institute Museum Publications 32. Oriental Institute of the University of Chicago, Chicago, 2010. (Excellent archaeological catalogue on the earliest scripts.)
- [11] Woods, Christopher. “The Earliest Mesopotamian Writing.” In Christopher Woods, Geoff Emberling, and Emily Teeter (eds.), *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*, University of Chicago Press, Chicago, 2015, pp. 33–50. (Updated and authoritative overview; dates proto-cuneiform to ca. 3350 BC.)
- [12] Englund, Robert K. “The Proto-Cuneiform Texts from the Erlenmeyer Collection: A Quantitative Overview.” In Pierre Lombard and Mahfouz al-Khalaf (eds.), *From Sumer to Babylon: The Erlenmeyer Collection of Ancient Near Eastern Art*, Kuwait National Museum, 2015, pp. 45–62. (Updated corpus analysis refining chronological estimates.)
- [13] Glassner, Jean-Jacques. *L’invention de l’écriture cunéiforme*. Éditions du Seuil, Paris, 2020. (Modern synthesis arguing for invention during the 34th century BC.)
- [14] Matthews, Roger, et al. “Seals and Signs: Tracing the Origins of Writing in Ancient South-West Asia.” *Antiquity*, 98(400), 2024, pp. 1–18. (Recent peer-reviewed article connecting seal imagery to proto-writing formation.)
- [15] Baevski, A., Schneider, S., & Auli, M. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020.
- [16] Hsu, W.-N., Zhang, Y., Glass, J., et al. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. 2021.
- [17] Jung, J., Heo, H., Park, J., et al. “RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification.” INTERSPEECH 2019.
- [18] Feng, S., Xu, S., & Li, H. “Unsupervised Acoustic Unit Discovery by Leveraging a Language-Independent Subword Discriminative Feature Representation.” INTERSPEECH 2021.
- [19] Jansen, A., Hore, A., & Bacchiani, M. “Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings.” ACL Anthology, 2022.

- [20] Pascual, S., Rabaoui, S., & Dauphin, Y. *PASE: Problem Agnostic Speech Encoder*. 2019.
- [21] Liu, A., Hsu, W.-N., Chang, J., et al. *Generative Spoken Language Modeling from Raw Audio*. 2021.
- [22] Zhang, Y., et al. *Robust Speech Recognition via Large-Scale Weak Supervision* (Whisper). arXiv:2212.04356, 2023.
- [23] Pratap, V., et al. *Scaling Speech Technology to 1,000+ Languages* (MMS). arXiv:2305.13516, 2023.
- [24] Chomsky, Noam. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, 1965.
- [25] Jackendoff, Ray. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford, 2002.
- [26] Berwick, Robert C., Chomsky, Noam, & Marcolli, Matilde. *Mathematical Structure of Syntactic Merge: An Algebraic Model for Generative Linguistics*. MIT Press, 2023.
- [27] Haspelmath, Martin. “Human Linguisticity and the Building Blocks of Languages.” *Frontiers in Psychology*, 10:3056, 2020.
- [28] Pinker, Steven. *The Stuff of Thought: Language as a Window into Human Nature*. Viking, New York, 2007.
- [29] B. Mahi, *Structured Knowledge Accumulation: Geodesic Learning Paths and Architecture Discovery in Riemannian Neural Fields*, arXiv preprint, 2025 (to appear).
- [30] Faith Comes By Hearing GitHub Repository. Available at: <https://github.com/faithcomesbyhearing>.