Tracing the Evolution of Language Through Symbolic Trees and Graphs

1. Introduction

The evolution of language encompasses the transformation of human expression across time, culture, and technological advancement. Written symbols—whether etched into stone or typed into code editors—are not static. They adapt, merge, and diverge, forming a living record of human communication and cognition. This research proposes a framework for tracing the evolution of such symbols and their grammatical structures through computational models, specifically hierarchical trees and directed acyclic graphs (DAGs).

This inquiry begins with pictoglyphic systems such as Chinese Hanzi, Japanese Kanji, Korean Hanja, and Southeast Asian derivatives like Chữ Nôm and Khmer. These systems are chosen for their rich symbolic history, internal visual logic, and cross-cultural diffusion. By decomposing characters into visual components (e.g., radicals and strokes), we aim to reconstruct a probable tree of symbolic ancestry. The complexity of character evolution—where symbols often recombine or adapt meanings—necessitates extending this tree into a DAG to more accurately represent the real-world processes of language change.

Beyond symbol analysis, we expand our scope to grammatical evolution. Languages evolve not only in how they look but in how they function. Thus, this research introduces methods for modeling grammar trees (e.g., abstract syntax trees or ASTs) and extending them into functional DAGs. These representations allow us to visualize and analyze changes in syntax, logic, and structure across time or paradigm shifts (e.g., procedural to object-oriented programming).

To support this effort, we leverage large language models (LLMs) as generative and analytical tools. LLMs trained on vast multilingual and multigrammatical corpora can infer relationships, identify latent structural similarities, and assist in generating hypotheses about language ancestry. Their integration into this research offers a scalable way to synthesize data, normalize cross-language inconsistencies, and propose evolutionary pathways.

A key methodological step involves comparing character sets used in basic literacy education (e.g., grade school curricula) with those found in full literary dictionaries and university-level materials. This allows for the modeling of character acquisition, complexity growth, and usage frequency over time. Overlaying these findings with spoken language patterns enables the incorporation of phonetic and contextual variations. Since the juxtaposition of characters in written form affects pronunciation and meaning, context analysis becomes essential in determining accurate lineage and function.

Ultimately, this framework aspires to provide a visually navigable, data-driven method for understanding linguistic and grammatical evolution. By quantifying transformation and structure at each step, we also propose a model for estimating the relative maturity or evolutionary stability of a language or grammar. While initially applied to human language, this framework is designed to be extensible to machine grammars such as C, Java, and Python—revealing deeper patterns in how humans and machines shape their modes of expression.

2. Background and Related Work

2.1 Evolution of Writing Systems

The study of writing system evolution has long focused on the transition from pictographic representations to logographic, syllabic, and alphabetic forms. East Asian scripts—such as Chinese Hanzi and its derivatives—provide an unusually clear record of symbolic transformation, with roots in oracle bone script evolving over millennia through clerical and regular styles. Scholars have examined the use of radicals as semantic classifiers and the phonetic components embedded in compound characters, forming the basis of modern character dictionaries and ontologies. Existing character decomposition datasets, such as the Chinese Ideographic Description Sequences (IDS), offer structured ways to parse and represent these relationships. However, most of this work remains manually curated and lacks a scalable graph-based model capable of generalizing across scripts and historical contexts.

2.2 Graph Theory in Linguistics

Graphs have been used in linguistics to model syntactic structure, morphological relationships, and lexical networks. Syntactic trees are a standard representation in computational linguistics, often used in parsing algorithms and machine translation systems. DAGs, in particular, offer a way to represent reentrant structures and shared subcomponents, as seen in semantic role labeling and graph-based dependency parsing. Recent research has also applied phylogenetic trees and DAGs to trace the lineage of natural languages, particularly in Indo-European studies. However, these models are rarely extended to visual or symbolic evolution, nor are they routinely combined with grammar evolution in a unified framework.

2.3 Large Language Models and Generative AI

The advent of large language models (LLMs) has significantly expanded the capacity to analyze linguistic data at scale. Models such as GPT, LLaMA, and PaLM, trained on multilingual corpora, have demonstrated an emergent ability to reason about syntactic structure, infer semantic equivalence, and even generate novel linguistic forms. These models can aid in tasks such as symbol-to-meaning mapping, cross-lingual analogy generation, and the alignment of grammar rules across languages. Furthermore, their generative capabilities make them suitable for hypothesizing plausible but undocumented evolutionary paths, offering a complementary tool to historical linguistic analysis.

3. Methodology

3.1 Character Set Stratification

We begin by stratifying the writing system into educational and literary layers. Character sets from grade-school literacy curricula serve as the foundational tier, representing core vocabulary and primary symbolic forms. These are contrasted with comprehensive literary dictionaries and academic corpora that contain rare, archaic, or domain-specific characters. By analyzing these layers independently and jointly, we capture growth trajectories in symbolic complexity and frequency of usage. This stratification aids in identifying stable roots versus divergent branches in the language tree.

3.2 Decomposition and Tree Construction

Characters are decomposed into sub-symbolic components using visual parsing and existing decomposition databases (e.g., IDS). These components form the nodes of a symbolic ancestry tree, with hierarchical relationships determined by form similarity, semantic lineage, and phonetic inheritance. This tree forms the core structure of our visual model.

3.3 Transition to Directed Acyclic Graphs

To account for recombination and shared ancestry, the character tree is converted into a directed acyclic graph (DAG). Nodes may have multiple parents or children, reflecting convergent evolution, shared radicals, or multi-role phonetic components. This model more accurately reflects historical processes of borrowing, adaptation, and cross-linguistic influence.

3.4 Spoken Language and Context Overlay

To model the dynamic interplay between written and spoken language, phonetic overlays are applied to character sequences. Characters that shift in pronunciation depending on juxtaposition (tone sandhi, compound-induced shifts, etc.) are flagged, and context-dependent phonological patterns are extracted. This phonetic DAG complements the symbolic DAG and introduces a temporal and contextual axis to the graph.

3.5 LLM-Aided Inference and Normalization

LLMs are used throughout the methodology to augment manual and rule-based methods. Tasks include:

- Semantic clustering of characters based on usage context.

- Suggesting likely historical precursors and descendants of given characters.

- Aligning grammar structures across human and programming languages.

- Detecting patterns in phonetic shifts across corpora.

This multi-pronged approach enables both macro and micro analyses of linguistic evolution, while maintaining a scalable, repeatable computational process.

4. Applications

4.1 Language Lineage and Ancestry Mapping

By visualizing symbolic evolution as a DAG, we can map the lineage of characters and their transformations over time. This has direct applications in historical linguistics and comparative philology, enabling researchers to explore converging paths, identify semantic borrowing, and clarify uncertain ancestries in character systems.

4.2 Maturity Modeling of Languages and Grammars

The DAG structure enables quantification of structural complexity, reuse, and change across nodes. These properties form the basis of a proposed maturity model, estimating how stable or dynamic a language is over time. Applied to both human and programming languages, this model offers a diagnostic view of linguistic growth, stabilization, and obsolescence.

4.3 Complexity and Contextual Dynamics

Coupling DAG analysis with contextual overlays—such as phonetic shifts in spoken language or syntactic rules in programming—reveals the deeper complexity embedded in language. This allows the identification of ambiguous regions, structural inflection points, or regions of high entropy. Such insights can inform language education, natural language processing (NLP), and automated refactoring tools.

4.4 Implications for Security and Analytics

Understanding the structure and evolution of machine grammars has implications for cybersecurity. Languages with high structural ambiguity or inconsistent rule evolution may be more prone to exploitation. DAG-based models can support static analysis tools, vulnerability detection, and grammar hardening. Similarly, in data analytics, the ability to trace code or query language lineage can enhance reproducibility, maintenance, and governance.

Together, these applications illustrate the transformative potential of applying linguistic and symbolic evolution models to a broad array of domains—linking language heritage, computational structure, and modern analytical needs.