Capturing Ancient Chinese Sense Induction with Automatic Pipelines

[1]Guan-Yu Tseng, [2]Chunki Lim, [3]Chih-Han Lin, [4]Tung-Le Pan,
[5]Yu-Chieh Wang, [6]Lang-Ching Yeh, [7]Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University
{[1]R14142007, [2]R14142001, [3]R13142001, [4]R12142006, [5]R14142006,
[6]R14142008, [7]shukaihsieh}@ntu.edu.tw

Abstract

While the study of diachronic semantic change has advanced alongside recent computational developments, structured lexical resources that reflect semantic evolution remain scarce for many languages, including Ancient Chinese. By systematizing the diachronic transformations within the Chinese Text Project (ctext, a large corpus of Ancient Chinese), we aim to bridge the gap between traditional philological inquiry and contemporary computational linguistics. This study proposes a pipeline that extracts contextualized embeddings from GujiBERT-fan, a language model pre-trained on pre-modern Chinese, and applies dynamic hierarchical clustering to identify distinct senses across historical periods. The pipeline operates at two levels: a global clustering that aggregates data across all periods to capture the full semantic space, and local clustering within each dynasty to reveal period-specific usage patterns. We test the pipeline with a pilot study on the character 手 (*shǒu*, "hand") across eight dynastic periods, covering over 185,000 occurrences. The results show that the pipeline can capture the diachronic shift from concrete to abstract senses, demonstrating its potential as a scalable method for mapping semantic evolution in historical languages.

Keywords: Chinese NLP, Semantic Change, Historical Linguistics

1. Introduction

Lexico-semantic change is fundamental to language evolution. Historically, the diachronic analysis of meaning was a time-consuming and labor-intensive task. Scholarly inquiry lacked a comprehensive overview of the field, as researchers were largely constrained to qualitative case studies involving only one or a small selection of lexemes at a time.

This changed with the introduction of computational methods. Scholars pioneered the application of distributional semantics to large-scale corpora (Hamilton et al., 2016; Schlechtweg et al., 2019), enabling time-aware semantic change studies for languages including English, German, Italian (Basile et al., 2020), and Latin (Perrone et al., 2021). These advancements help researchers track semantic change across thousands of words simultaneously.

However, Ancient Chinese-a language spanning over three millennia with a remarkably rich and continuous literary tradition—has received comparatively little attention in the area. Despite its potential to contribute profound insights into universal patterns of semantic evolution, the language has not yet been subjected to the same level of systematic, large-scale computational scrutiny, unlike Latin, English, or German.

To address this gap, this study proposes a computational framework to automatically map the Diachronic Semantic Landscape of the Chinese lexicon. The primary objective is not merely to rediscover established dictionary definitions, but to visualize the shifting densities and fuzzy boundaries of semantic fields across historical epochs. By leveraging dense vector representations, this research aims to quantify the topography of meaning, revealing how concepts expand, contract, and merge throughout Chinese history.

A key methodological challenge in applying lexical semantic change methods to Ancient Chinese is the severe imbalance between the sparse textual record and the voluminous late imperial corpora. To address this, the proposed pipeline adopts a two-level architecture combining Global and Local clustering: rather than choosing between joint and incremental approaches, each of which faces distinct limitations in this setting, the framework performs independent dynasty-specific clustering while maintaining a cross-period reference space, designed to handle the diachronic data sparsity inherent to historical Chinese corpora.

2. Literature Review

2.1. Lexical Semantic Change and Word Sense Induction

The computational detection of Lexical Semantic Change (LSC) has evolved considerably over the past decade. Early approaches relied on aligning static word embeddings across time pe-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 163–176.

riods (Hamilton et al., 2016), while more recent work leverages contextualized representations from pre-trained language models, clustering token-level embeddings to identify distinct word senses. This clustering-based approach is closely related to Word Sense Induction (WSI), which aims to discover the senses of a polysemous word directly from corpus data, without relying on a predefined sense inventory. In the diachronic setting, WSI extends to identifying how a word's set of senses changes over time. The SemEval-2020 Task 1 (Schlechtweg et al., 2020) introduced the first shared task for unsupervised LSC detection, providing manually annotated datasets for English, German, Latin, and Swedish. Many top-performing systems in that evaluation relied on clustering contextualized embeddings —the same general strategy we adopt in this study.

2.2. Joint vs. Incremental Methodologies

Traditional batch or "joint" clustering processes a longitudinal corpus as a single, static entity. While this approach maximizes the data available for each cluster, it faces significant challenges with "contextual saturation" (Periti and Montanelli, 2024), where the overwhelming volume of data from later periods can obscure subtle semantic nuances in earlier, sparser eras.

Conversely, incremental approaches like "What is Done is Done" (WiDiD) treat Semantic Shift Detection (SSD) as an organic, evolutionary process (Periti et al., 2025). By processing data chronologically and using prototypes from preceding periods, incremental models aim to capture "temporal transactions" in meaning. This avoids the risk of anachronistically projecting modern senses onto ancient texts.

2.3. Theoretical Limitations and Structural Recovery

Despite the practical appeal of incremental methods for large-scale corpora, theoretical analysis suggests inherent limitations. Ackerman and Dasgupta provided a formal proof that the incremental setting is "strictly weaker" than the batch model (Ackerman and Dasgupta, 2014). Specifically, certain "nice" cluster structures that are easily detectable when viewing the data in its entirety may be impossible for an incremental method to recover. They suggest, however, that these limitations can be mitigated by allowing for "extra clusters" or hierarchical refinements.

2.4. The Ancient Chinese Context

Computational processing of Ancient Chinese has gained increasing attention in recent years. The EvaHan shared task series, co-located with LT4HALA since 2022, has organized evaluations on word segmentation, POS tagging, and other fundamental NLP tasks for Ancient Chinese (Li et al., 2022). Pre-trained language models tailored to Classical Chinese, such as GujiBERT-fan (Wang et al., 2023), have also been developed. However, computational research on the diachronic semantics of Ancient Chinese —particularly sense induction over time —remains largely unexplored.

Applying LSC methods to this language presents a specific challenge. The literary record of the Pre-Qin period (before 221 BCE) is extremely limited compared to later dynasties. Under an incremental framework, initial cluster prototypes would be derived from this sparse early data, and any inaccuracy would propagate forward through all subsequent periods. Joint clustering avoids this error propagation by pooling all data together, but at the cost of allowing the massive late-imperial corpora (Ming and Qing) to dominate the semantic space, potentially flattening the very evolutionary trajectory the analysis seeks to reveal. Our study takes a different path. Rather than choosing between joint and incremental clustering, we perform independent clustering within each dynasty (Local Clustering) while also providing a cross-period aggregation (Global Clustering). Each dynasty's clustering is self-contained and not dependent on earlier prototypes, avoiding the error propagation of incremental methods. Meanwhile, the global view preserves rare or emergent senses that might be lost in any single period. Combined with a dynamic hierarchical sub-clustering step, this approach aims to trace the semantic evolution of the target lexeme at multiple levels of granularity.

3. Data and Methodology

The historical data are retrieved from *Chinese Text Project* (Sturgeon, 2011), an extensive corpus of Classical Chinese covering literary, philosophical, medical, poetic, and historiographic registers. It provides broad temporal and domain coverage, with bibliographic metadata supporting reliable periodization.

Our proposed methodology consists of a six-stage pipeline: **(1) Contextual Embedding, (2) OCR Cleaning, (3) Stratified Sampling, (4) Over-clustering and Merging, (5) Hierarchical Sub-clustering, (6) Human-in-the-Loop Semantic Annotation**.

To evaluate the feasibility of the proposed pipeline, we conducted a focused pilot analysis on the lexical item **手** (*shǒu*, "hand"), selected for its high token frequency, stable orthography across periods, and rich semantic diversity from physical to metaphorical uses.

3.1. Data Processing and Contextual Embeddings

We extracted all occurrences of *shǒu* from all the accessible texts in the *Chinese Text Project* database. The data was split into eight periods by major historical event such as dynasty changes: **先秦** (*xiān qín* 'pre-Qin', before 221 BCE), **秦漢** (*qín hàn* 'Qin-Han', 221 BCE–220 CE), **魏晉** (*wèi jìn* 'Wei-Jin', 221–509 CE), **隋唐** (*suí táng* 'Sui-Tang', 589–960 CE), **宋元** (*sòng yuán* 'Song Yuan', 960–1368 CE) **明** (*míng* 'Ming', 1368–1644 CE), **清** (*qīng* 'Qing', 1644–1912 CE) and **民國** (*mín guó* 'Republican', after 1912 CE). This sampling strategy covers all major dynastic periods and reflects a range of genres in each dynastic period for comprehensive, diachronic analysis.

After retrieving the raw data from the *Chinese Text Project*, we segmented the texts into individual sentences using standard terminal punctuation marks (i.e., "!", "?", "。", ";").

To capture the specific historical context, each complete sentence containing the target lexeme was processed through GujiBERT-fan (Wang et al., 2023), a transformer modeled after BERT (Devlin et al., 2018) and pretrained on the **四庫全書** *Sì kù quán shū*, an extensive collection of ancient Chinese text. From the resulting sequence outputs, we specifically isolated the contextualized vector representation corresponding to the character **手** (*shǒu*). These target word vectors were then stored with metadata (e.g., dynasty, work title) and original source sentences for subsequent clustering analysis.

Data Cleaning: To ensure data quality, instances containing obvious OCR (Optical Character Recognition) errors, such as abnormally long sentences, excessively repeated characters, and unidentifiable lexical items, were filtered out to yield a cleaner dataset.

Stratified Sampling: A stratified sampling method is implemented to prevent any single text or genre from dominating the semantic distribution. Within each historical period, the extraction was capped at a maximum of 200 instances of **手** (*shǒu*) per individual document. This threshold effectively mitigates the domination of voluminous historical records, ensuring that the resulting semantic landscape reflects a diverse and representative cross-section of language use for each era.

3.2. Clustering and Annotation

Over-clustering and Merging: Within each dynasty, the contextual embeddings of *shǒu* were grouped using **k-means** clustering to identify distinct senses. To capture fine-grained nuances, based on preliminary experiments, we adopted an over-clustering strategy by setting the initial number of clusters (K) to 20. Subsequently, an automated merging step was applied: clusters with a centroid cosine similarity exceeding the threshold of 0.95 were merged to become the same group. Each finalized cluster approximates a distinct contextualized usage pattern or lexical sense.

The choice of $K = 20$ was validated through a one-factor-at-a-time sensitivity analysis, in which we tested $K \in \{10, 15, 20, 25, 30\}$, *merge_ratio* $\in \{0.90, 0.92, 0.95, 0.97\}$, and *split_ratio* $\in \{0.1, 0.3, 0.5\}$ for both global and local clustering, while holding the remaining parameters fixed at their baseline values. Although $K = 25$ yielded a marginally higher global Silhouette Score (0.054 vs. 0.049, $\Delta = +0.005$), its final cluster count after merging (16) was nearly identical to that of $K = 20$ (17), indicating that the additional initial clusters were redundant. $K = 20$ was therefore selected as the setting that achieves comparable clustering quality while avoiding unnecessary over-fragmentation.

Table 1: Hyperparameter configurations for Global and Local clustering.

Parameter	Global	Local
K (initial clusters)	20	20
merge_ratio	0.95	0.95
split_ratio	0.1	0.3
sub-K	6 (fixed)	dynamic

Hierarchical Sub-clustering: Due to the semantic continuity of natural language, the merging phase often produces overly large, coarse-grained clusters. To extract the nuanced senses of *shǒu*, we incorporated a dynamic hierarchical sub-clustering mechanism. This process is governed by two parameters[1]: a *split-ratio* acts as the threshold to trigger further division (e.g., a ratio of 0.3 means any cluster containing over 30% of the

[1]For example, if the Ming dynasty corpus contains 10,000 occurrences of **手**, setting the *split-ratio* to 0.3 means any merged cluster exceeding 3,000 instances will trigger the sub-clustering pipeline. Once triggered, the algorithm evaluates the cluster's exact size to assign a proportional *sub-*K. A cluster of 4,000 instances might be automatically partitioned into 3 sub-groups, whereas a massive cluster of 20,000 instances in another dynasty would be assigned a higher *sub-*K (e.g., 6). This dynamic allocation prevents both under-clustering and over-fragmentation.

dynasty's total data will be sub-clustered). Rather than using a fixed number, the value of *sub-*K is dynamically determined based on the absolute data volume of the target cluster, ensuring that the semantic granularity scales proportionally with the size of the data.

Human-in-the-Loop Semantic Annotation — Seven Major Categories for shǒu: While the automated pipeline efficiently discovers latent semantic structures, assigning meaningful linguistic interpretations to these mathematical clusters requires domain expertise. We adopted a human-in-the-loop approach to annotate the final output. To enhance efficiency, two large language models, Google Gemini Pro and Anthropic Claude Sonnet (both accessed in February 2026), were employed as independent pre-annotators. By examining the representative "core sentences" located closest to the centroid of each cluster, two LLMs mapped the fine-grained clusters into seven macro-categories. Discrepancies between the two models were flagged for human review, with human experts in Chinese linguistics making the final determination.

This coding scheme reflects the diachronic semantic extension of *shǒu*: (1) **Body_Medical** (the literal physical organ, limbs, and traditional medical concepts like meridians and symptoms); (2) **Physical_Action** (concrete body movements, holding, spatial relations, and tool usage); (3) **Social_Interaction** (interpersonal contact, social signaling, and metaphorical relationships); (4) **Text_Culture** (writing, calligraphy, official edicts, and divination); (5) **Power_Skill** (abstract agency, control, subordinates, and titles for experts/masters); (6) **Grammar_Suffix** (lexicalized occupational suffixes); and (7) **NOISE** (OCR errors and unidentifiable strings). This human-validated categorization ensures that our data-driven clusters are robustly grounded in usage-based linguistics. The complete hierarchical coding scheme, detailing the macro-categories, subcategories, and their corresponding core examples, is presented in Table 2.

4. Results and Analysis

In this section, we present the clustering results of the contextual embeddings for 手 (*shǒu*). We first present a unified diachronic projection to present the overall clustering (Global Clustering, Section 4.1), followed by an examination of the dynasty-specific semantic landscapes to capture the semantic distributions within each specific era (Local Clustering, Section 4.2).

All visualizations in this study follow the same encoding scheme. In the subsequent scatter plots, each data point represents a single occurrence of the character 手, encoded as a 768-dimensional GujiBERT-fan contextual embedding. While the clustering operates in this high-dimensional space, the coordinates are projected onto a 2D plane via Principal Component Analysis (PCA) strictly for visualization purposes.

Data points are color-coded to represent the six overarching macro-categories (along with a grey NOISE category) defined in our semantic annotation scheme. Text boxes indicate the exact centroids of the dynamically generated, fine-grained clusters. To contextualize the semantic shifts quantitatively and mitigate biases from absolute data volume, each visualization panel reports three key metrics: the absolute sample size (n) in the title, the relative proportion of each macro-category in the top-right corner, and the Silhouette Score (Rousseeuw, 1987) in the bottom-left/right to validate the structural cohesion of the clustering.

4.1. Analysis of Global Clustering

The global clustering analysis provides a diachronic overview of the semantic landscape of 手 (*shǒu*) by aggregating data from all eight historical periods. This unified projection (Figure 1) shows the primary senses that persist across all periods. The macro-categories identified—including **Body_Medical**, **Physical_Action**, **Power_Skill**, **Social_Interaction**, **Text_Culture**, and **Grammar_Suffix**—form the foundational semantic infrastructure of the character. For instance, the clustering isolates core anatomical references and medical concepts within the **Body_Medical** sector, while simultaneously capturing abstract extensions in the **Power_Skill** and **Social_Interaction** domains.

However, because the global view is an aggregation of 185,792 occurrences, it may be susceptible to corpus imbalance, particularly the dominance of late imperial texts (Ming and Qing). To discern the "mainstream" semantic senses of 手 within a specific time frame and to avoid being overshadowed by high-volume eras, it is necessary to refer to the dynasty-specific distributions (Local Clustering) to examine how these semantic weights shifted over time.

4.2. Analysis of Local Clustering

By examining the dynasty-specific landscapes (Figure 2), we can observe the fine-grained evolution of the character's semantic priorities. Each panel reveals the dominant usage types of 手 unique to its respective era. In the early stages (Pre-Qin), the semantic gravity is heavily concentrated in the concrete domains of **Body_Medical** and **Physical_Action**, reflecting a period where 手

Table 2: The Hierarchical Semantic Coding Scheme for the Character **手** (*shǒu*)

Macro Category	Sub-categories (Senses)	Typical Examples (Hanzi + *Pinyin* + Translation)
Body_Medical	Literal_Hand, Limbs, Anatomy	右手 *yòu shǒu* 'right hand'; 手足 *shǒu zú* 'hands and feet'
	Symptom (Medical state)	手足厥冷 *shǒu zú jué lěng* 'coldness in hands and feet'
	Meridian (TCM concept)	手太陰 *shǒu tài yīn* 'Hand-Taiyin (Lung Meridian)'
Physical_Action	Holding, Combat_Hold	手執 *shǒu zhí* 'to hold in hand'
	Spatial, State	手中 *shǒu zhōng* 'in the hands'
	Gesture, Take_Action, Instrumental	下手 *xià shǒu* 'to take action/strike'; 以手 *yǐ shǒu* 'using hands'
Social_Interaction	Connection, Contact_Sign	攜手 *xié shǒu* 'hand in hand'; 拍手 *pāi shǒu* 'to clap hands'
	Metaphor (Relationships)	猶人手足 *yóu rén shǒu zú* 'like hands and feet (metaphor for brotherhood)'
Text_Culture	Letter_Edict, Writing, Artifact	手詔 *shǒu zhào* 'imperial edict written by the emperor'
	Divination, Etymology	艮手 *gèn shǒu* 'Gen represents the hand (divination)'
	Literary (Poetic expressions)	飛來我手 *fēi lái wǒ shǒu* 'flies into my hand'
Power_Skill	Agency, Control, Possession	出其手 *chū qí shǒu* 'directed by someone'; 權在手 *quán zài shǒu* 'power in hand'
	Subordinate	手下 *shǒu xià* 'subordinates'
	Expertise, Capability	妙手 *miào shǒu* 'highly skilled person / master'
Grammar_Suffix	Occupation_Suffix	水手 *shuǐ shǒu* 'sailor'; 弓箭手 *gōng jiàn shǒu* 'bowman'
NOISE	OCR_Error	差手靈蓋 *chā shǒu líng gài* (Meaningless OCR noise)

was primarily utilized for literal bodily functions and direct physical movements.

Significant diachronic variations emerge during the medieval and late imperial periods. From the Sui–Tang through the Song–Yuan dynasties, we observe a "semantic migration" toward the upper-left quadrants of the vector space. During these eras, the proportions of **Social_Interaction** and **Power_Skill** clusters expand significantly, indicating that the character's role in social connections and expressions of agency was becoming mainstream. By the Ming and Qing periods, the emergence of the **Grammar_Suffix** cluster signifies the final stage of this evolution, where the character became structurally stabilized as an occupational marker. These local shifts highlight that while the global senses remain relatively consistent, the "mainstream" usage of the word has undergone a transformation from the physical to the abstract.

5. Discussions

The Global and Local clustering approaches serve different analytical purposes. The Global Cluster concatenates data across all millennia, providing a macro-historical landscape of the target lexeme. The global space helps preserve rare senses that would be lost in dynasty-specific clustering. For instance, an abstract sense of *shǒu* (e.g., representing power or agency) might be exceedingly rare in the Pre-Qin period. In a dynasty-specific Local Cluster, such sparse occurrences would likely be absorbed into dominant physical action clusters or entirely discarded as statistical noise. However, in the Global space, the explosion of these abstract usages in later eras (e.g., Ming and Qing) constructs a massive, well-defined semantic territory. The sparse Pre-Qin instances can thus find their "home" within this global distribution, allow-

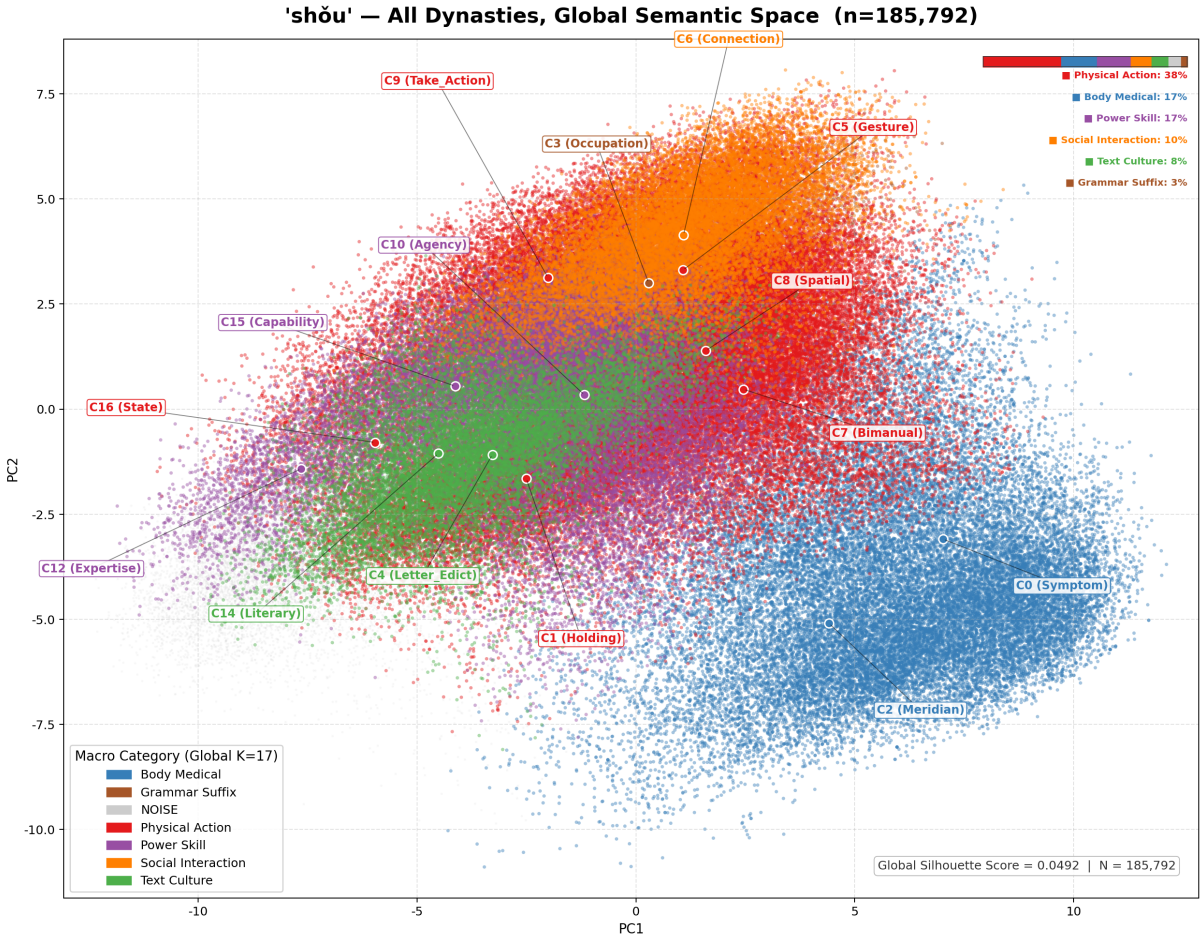

Figure 1: **Unified diachronic semantic space of 手 ($n = 185,792$)**
This plot maps all historical periods onto an absolute PCA coordinate system. Initialized with $K = 20$ and parameterized with a *merge-ratio* of 0.95 and *split-ratio* of 0.1 (*sub-K*=6), the pipeline dynamically yielded 17 semantic clusters (C0–C16).

ing us to trace the earliest seeds of semantic shift. Conversely, Local Clusters are indispensable for providing time-specific snapshots, demonstrating the mainstream semantic priorities and contextual boundaries unique to each individual dynasty.

This approach differs from traditional sense inventory in that it represents meaning as a continuous, multi-dimensional space rather than a discrete list. The character's meaning is not a collection of static definitions, but a dynamic integration of semantic fields such as social and medical fields, grammatical role such as occupational suffix, and prototypical senses.

The clustering output identifies overarching semantic fields (e.g., the medical and social domains), clear syntactic transformations (e.g., the emergence of grammatical suffixes), sense prototypes where usage is highly conventionalized. By moving away from the rigid constraints of a 1D sense list, our approach respects the "semantic continuum" inherent in natural language, where boundaries between capability, agency, and physical action are often fluid and overlapping.

6. Conclusion

In this work, we present a computational framework to automatically map the Diachronic Semantic Landscape of the Chinese lexicon. By integrating contextualized embeddings from a specialized Classical Chinese language model (Wang et al., 2023) with a dynamic, unsupervised clustering pipeline, we demonstrate a methodology that can reveal semantic structure which static dictionary entries do not capture.

The implications of this mapping extend well beyond computational linguistics. The framework can help digital humanities scholars to trace how the senses of key terms shift over time, complementing traditional philological method.

The methodology established in this study offers a scalable framework that could potentially be extended to other languages with extensive diachronic corpora, such as Sanskrit or Ancient

Figure 2: **Dynasty-specific (Local) semantic landscapes of 手**

To capture the semantic distributions within each specific era, the clustering pipeline was initialized with $K = 20$ for every period. We applied a *merge-ratio* of 0.95 and a *split-ratio* of 0.3, with the *sub-*K dynamically determined by the absolute sample size (n) of the respective era. Visual encodings (colors, centroids, and metrics) follow the principles detailed in Section 4.

Greek. By demonstrating the feasibility of unsupervised semantic mapping in a historical context, this research contributes to the ongoing development of dynamic historical lexicons. Ultimately, we hope this pipeline offers a practical, reproducible method for studying lexical change in Chinese and other historical languages.

Furthermore, as the current framework was evaluated through a pilot study of a single high-frequency lexeme (*shǒu*), future work will expand the dynamic clustering analysis to a broader set of target characters. As demonstrated in the generalizability analysis presented in Appendix A, the pipeline produces coherent semantic clusters for two additional lexemes — 水 (*shuǐ*, "water") and 道 (*dào*, "way/road/principle") — suggesting that the approach is not limited to concrete body-part nouns. Extending the analysis to a wider range of lexical types, including abstract concepts and function words, remains an important direction for future work.

7. Limitations

While the integration of GujiBERT-fan and dynamic hierarchical clustering provides a macro-level perspective on the semantic evolution of 手 (*shǒu*), two primary methodological limitations should be acknowledged.

First, regarding the granularity of semantic representation, contextualized embeddings are highly sensitive to broader usage-type distributions and semantic fields (e.g., medical contexts or grammatical functions) rather than fine-grained, dictionary-level senses (Periti and Tahmasebi, 2024). Furthermore, applying a hard-clustering algorithm (K-means) inherently imposes discrete mathematical boundaries upon lexical semantics, which naturally operates on a continuous spectrum. While the visual overlaps in our PCA projections help illustrate transitional "bridging contexts," the generated clusters should be interpreted as semantic prototypes rather than mutually exclusive lexicographic categories.

Second, a persistent challenge in historical corpus linguistics is diachronic data sparsity, particularly the severe corpus imbalance between early eras (e.g., Pre-Qin) and late imperial periods. This sparsity reflects two compounding factors: the historically limited survival of early texts, and the greater susceptibility of archaic orthography to OCR errors, which results in higher data attrition during the cleaning stage. Although our methodology employs stratified sampling to mitigate volume imbalances, it cannot fully compensate for these constraints. Consequently, while the observed spatial expansion from concrete to abstract domains provides strong correlational evidence for metaphorical extension, establishing direct diachronic causality requires future research to supplement this distant-reading approach with qualitative close-reading of historical texts, alongside cor-

pora with denser coverage of underrepresented dynastic periods and improved OCR pipelines.

8. Bibliographical References

Margareta Ackerman and Sanjoy Dasgupta. 2014. Incremental clustering: The case for extra clusters. *arXiv preprint arXiv:1406.6398*.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. *DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task*, pages 411–419.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang Qu, and Dongbo Wang. 2022. The first international Ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 135–140, Marseille, France. European Language Resources Association.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–36.

Francesco Periti, Sergio Picascia, Stefano Montanelli, Alfio Ferrara, and Nina Tahmasebi. 2025. Studying word meaning evolution through incremental semantic shift detection. *Language Resources and Evaluation*, 59:1363–1399.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2021. Lexical semantic change for ancient greek and latin. *CoRR*, abs/2101.09069.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

9. Language Resource References

Donald Sturgeon. 2011. Chinese text project: a dynamic digital library of premodern chinese. `https://ctext.org/`.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, and Xiyu Wang. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts.

A. Appendix: Generalizability Analysis on 水 (shuǐ) and 道 (dào)

To evaluate the generalizability of the proposed pipeline beyond the pilot lexeme 手 (shǒu), we applied the same methodology to two additional characters: 水 (shuǐ, "water") and 道 (dào, "way/road/principle"). These characters were selected to represent distinct semantic profiles — 水 as a high-frequency concrete noun with rich metaphorical extensions, and 道 as a polysemous term spanning concrete, philosophical, administrative, and grammaticalized senses.

A.1. 水 (shuǐ, 'water')

The global clustering (Figure 3) analysis of 水 ($n = 313,184$) yielded 12 semantic clusters organized into eight macro-categories: Geography, Physical Usage, Literary, Geomancy, Cosmology, Body Medical, Military, and NOISE. Geography (24%) and Physical Usage (21%) dominate the semantic space, reflecting the character's stable concrete core across all dynasties. The Literary category (19%) captures the extensive use of 水 as a poetic image, while Geomancy (12%) reflects its centrality in fengshui traditions.

The dynasty-specific local clustering (Figure 4) reveals a clear diachronic trajectory. In the Pre-Qin period ($n = 1,722$), the semantic space is dominated by Geography and Cosmology, consistent with the cosmological significance of water in early Chinese thought. From the Song-Yuan period onward, the Literary category expands substantially, reflecting the flourishing of water imagery in classical poetry.

A.2. 道 (dào, 'way/road/principle')

The global clustering (Figure 5) of 道 ($n = 331,910$) yielded 22 semantic clusters across nine macro-categories: Philosophy, Speech Act, Geography, Administration, Religion, Onomastics, Cognition, Medical, and NOISE. Philosophy dominates at 41%, reflecting the centrality of 道 as a philosophical concept. Most notably, Speech Act (15%) forms a visually distinct cluster in the embedding space, spatially separated from the philosophical senses, providing strong evidence for a grammaticalization pathway from lexical verb to quotative marker.

The local clustering (Figure 6) results reveal a particularly rich diachronic trajectory. In the Pre-Qin period ($n = 3,226$), the semantic space is almost entirely dominated by Philosophy, with no Speech Act clusters present. The Speech Act category becomes prominent from the Song-Yuan period and expands significantly through the Ming and Qing dynasties. Notably, the Republican period achieves the highest Silhouette Score across all dynasties (0.1718), suggesting that the semantic categories of 道 have reached their highest degree of differentiation in the modern era.

A.3. Cross-lexeme Comparison

Taken together, the results for 水 and 道 demonstrate that the pipeline generalizes effectively across lexemes with distinct semantic profiles. While 手 exhibits a trajectory from concrete bodily meaning toward abstract and grammaticalized senses, 水 shows stable concrete dominance with expanding literary and cosmological extensions, and 道 displays the most complex trajectory, spanning philosophical elaboration, religious extension, administrative specialization, and grammaticalization into a quotative marker. These contrasting patterns confirm that the pipeline is sensitive to the unique semantic histories of individual lexemes while maintaining a consistent methodological framework across all three cases.

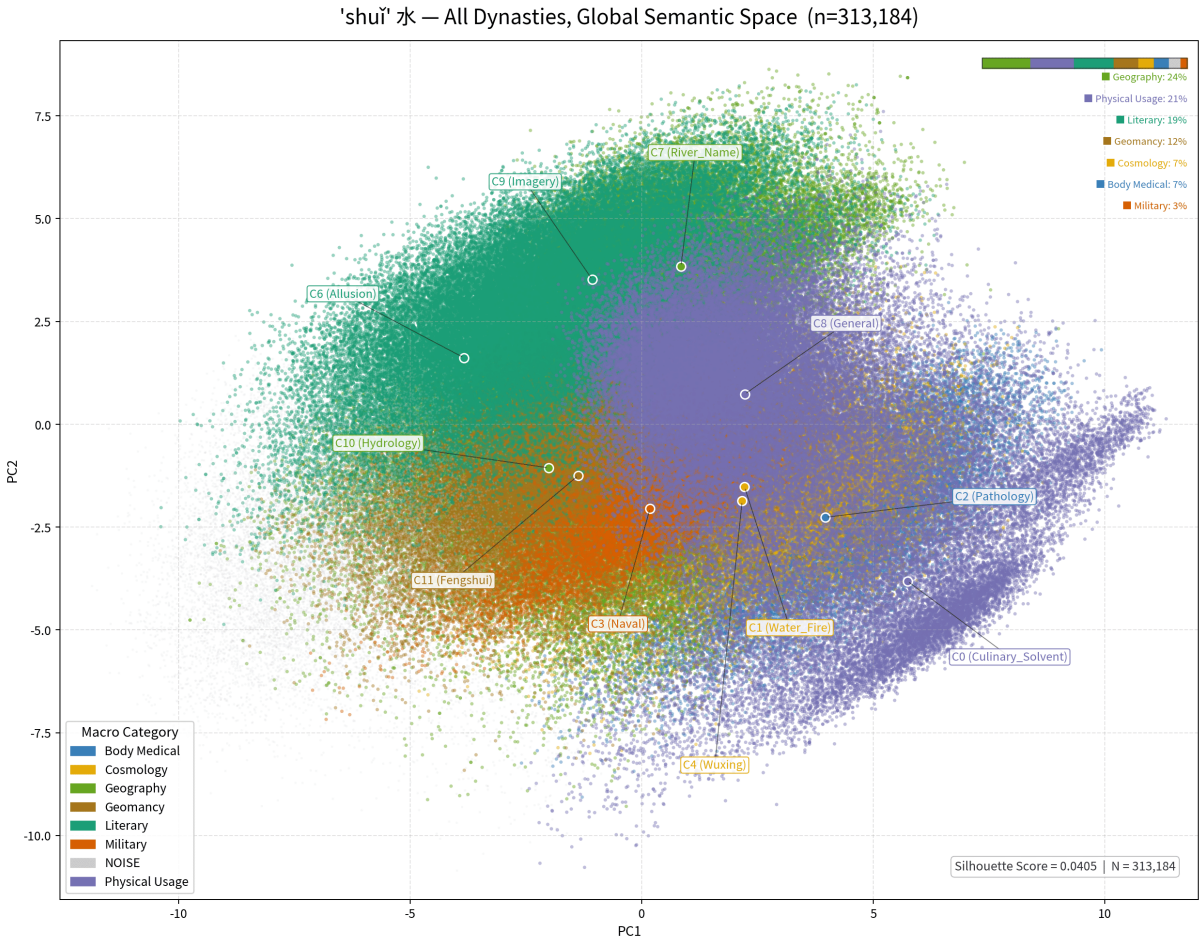

Figure 3: **Unified diachronic semantic space of 水 ($n = 313,184$).** This plot maps all historical periods onto an absolute PCA coordinate system. Initialized with $K = 20$ and parameterized with a *merge-ratio* of 0.95 and *split-ratio* of 0.1 (*sub-K*=6), the pipeline dynamically yielded 12 semantic clusters (C0–C11).

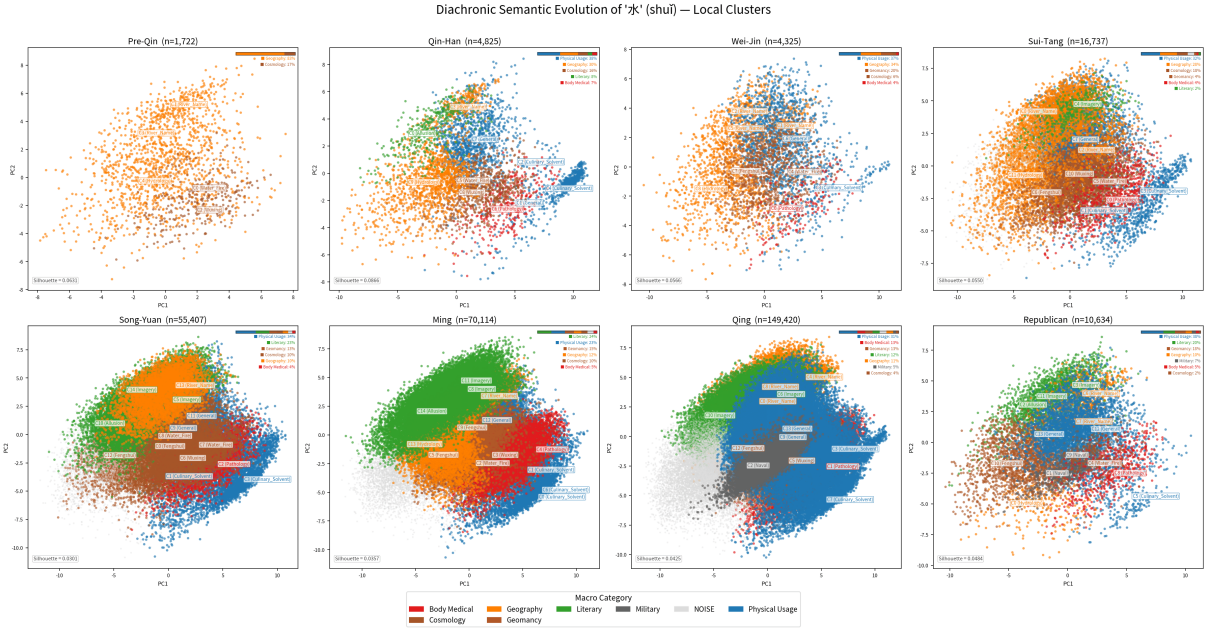

Figure 4: **Dynasty-specific (Local) semantic landscapes of 水**. To capture the semantic distributions within each specific era, the clustering pipeline was initialized with $K = 20$ for every period. We applied a *merge-ratio* of 0.95 and a *split-ratio* of 0.3, with the *sub-*K dynamically determined by the absolute sample size (n) of the respective era.

Table 3: The Hierarchical Semantic Coding Scheme for the Character **水** (*shuǐ*)

Macro Category	**Sub-categories (Senses)**	**Typical Examples (Hanzi + *Pinyin* + Translation)**
Geography	River Name	濼水 *Luòshuǐ* 'Luo River'; 濮水 *Púshuǐ* 'Pu River'
	Hydrology	措置水之術 *cuòzhì shuǐ zhī shù* 'techniques for water management'
Physical_Usage	Culinary Solvent	水煎熟 *shuǐ jiān shú* 'boiled in water'
	General	清水 *qīng shuǐ* 'clear water'
Cosmology	Wuxing (Five Elements)	水乘火 *shuǐ chéng huǒ* 'water overcomes fire'
	Water-Fire Pair	水火之數 *shuǐ huǒ zhī shù* 'the numerology of water and fire'
Body_Medical	Pathology (TCM)	利水 *lì shuǐ* 'to promote diuresis (TCM)'
Military	Naval	水師提督 *shuǐ shī tídū* 'naval commander'
Literary	Allusion	易水 *Yìshuǐ* 'Yi River (allusion to Jing Ke)'
	Imagery	遠水縈紆而來 *yuǎn shuǐ yíng yū* 'distant water winding toward us'
Geomancy	Fengshui	水口關欄 *shuǐkǒu guānlán* 'water mouth closure (feng shui)'
NOISE	OCR Error	近蒙專使至虔遠致時服寢衣之魏尋附居布謝必連 (Meaningless OCR noise)

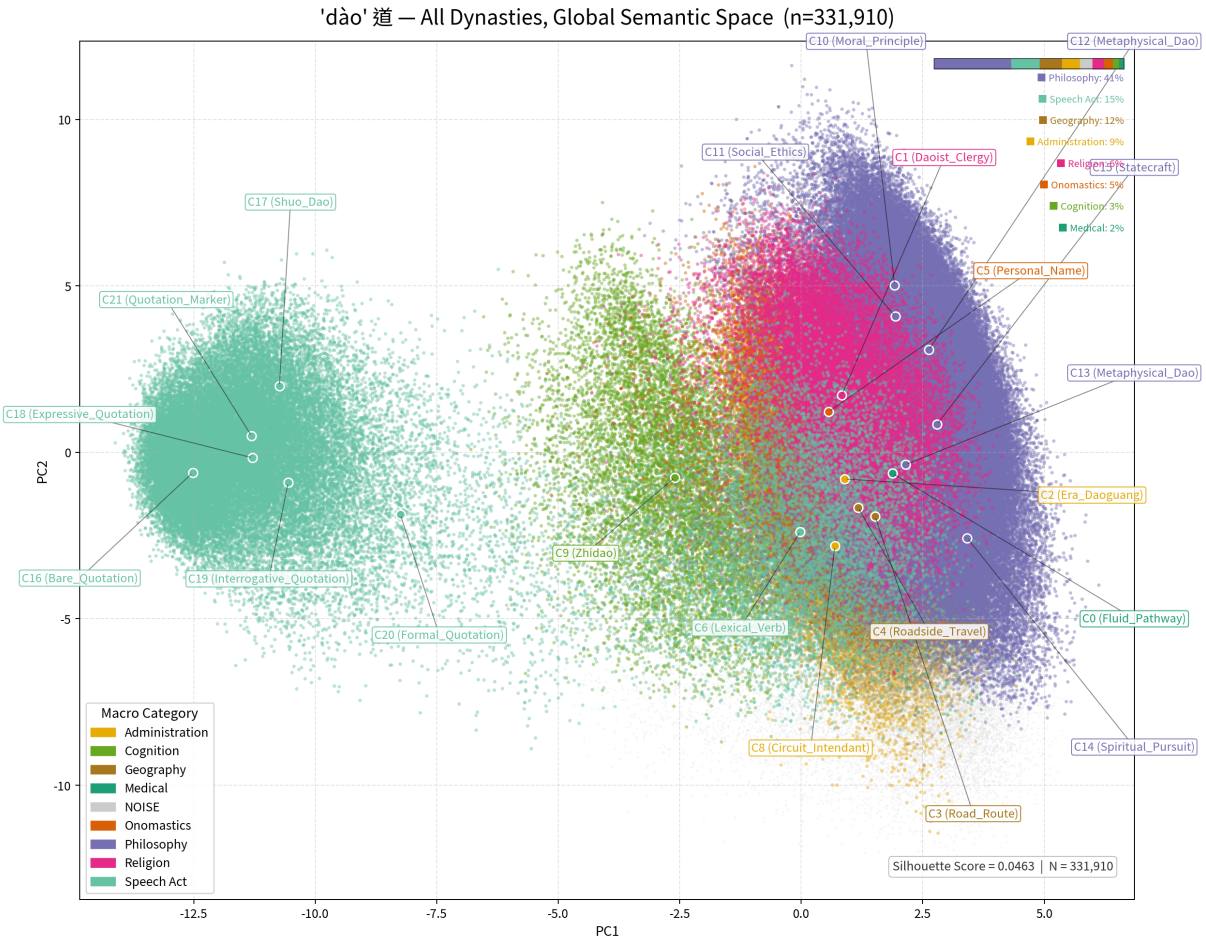

Figure 5: **Unified diachronic semantic space of 道 ($n = 331,910$).** This plot maps all historical periods onto an absolute PCA coordinate system. Initialized with $K = 20$ and parameterized with a *merge-ratio* of 0.95 and *split-ratio* of 0.1 (*sub-K*=6), the pipeline dynamically yielded 22 semantic clusters (C0–C21).

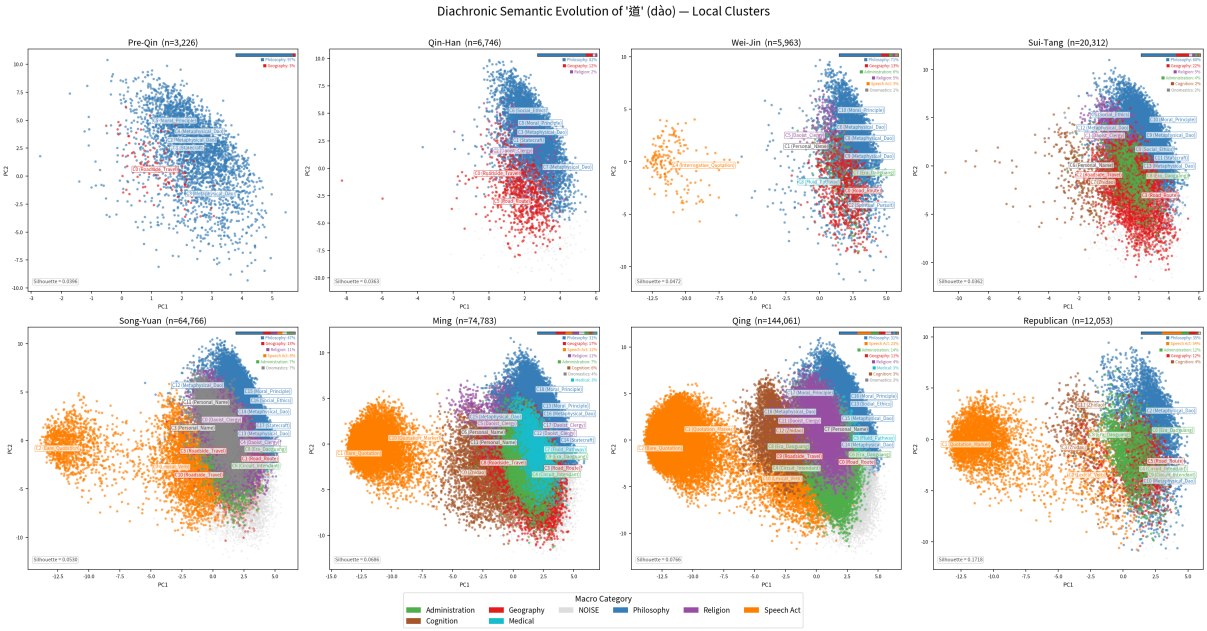

Figure 6: **Dynasty-specific (Local) semantic landscapes of 道**. To capture the semantic distributions within each specific era, the clustering pipeline was initialized with $K = 20$ for every period. We applied a *merge-ratio* of 0.95 and a *split-ratio* of 0.3, with the *sub-*K dynamically determined by the absolute sample size (n) of the respective era.

Table 4: The Hierarchical Semantic Coding Scheme for the Character **道** (*dào*)

Macro Category	**Sub-categories (Senses)**	**Typical Examples (Hanzi + *Pinyin* + Translation)**
Philosophy	Moral Principle, Political Ethics	知生之道 *zhī shēng zhī dào* 'the way of sustaining life'
	Metaphysical Dao	道之精妙 *dào zhī jīng miào* 'the subtlety of the Dao'
	Spiritual Pursuit, Statecraft	求道志 *qiú dào zhì* 'aspiration to seek the Dao'
Speech_Act	Lexical Verb	必能道之 *bì néng dào zhī* 'must be able to speak of it'
	Bare Quotation, Shuo Dao	說道：「弟之苦衷實難告人」 *shuō dào* 'said: …'
	Expressive, Interrogative, Formal	大笑道 *dà xiào dào* 'laughed and said'; 問道 *wèn dào* 'asked'
Geography	Road Route	不敢循驛道 *bù gǎn xún yì dào* 'dared not follow the post road'
	Roadside Travel	伏在道側蓬蒿之內 *fú zài dào cè* 'hid beside the road'
Religion	Daoist Clergy	道士復歸 *dào shì fù guī* 'the Taoist returned'
	Personal Name	杜遵道 *Dù Zūndào* (personal name)
Administration	Circuit Intendant	分巡福寧道 *fēn xún Fúníng dào* 'Circuit Intendant of Funing'
	Era Name (Daoguang)	道光中黃氏後人 *Dàoguāng zhōng* 'during the Daoguang reign'
Onomastics	Personal Name	蜀人吳師道為漢州太守 *Wú Shīdào wéi Hànzhōu tàishǒu* 'Wu Shidao served as Prefect of Hanzhou'
Cognition	Zhidao (Know)	知道今日饑荒苦 *zhī dào jīnrì jīhuāng kǔ* 'knows today's suffering'
Medical	Fluid Pathway (TCM)	通調水道 *tōng tiáo shuǐ dào* 'to regulate the water pathway (TCM)'
NOISE	OCR Error	栗難曉氣通南者謝班中絕景 (Meaningless OCR noise)