

Using Natural Language Processing Technology to Explore the Semantic Evolution of "Textile" Related Words in Bamboo and Silk Manuscripts

Hui Wen¹, Changfeng Li²

¹Department of Korean, School of Foreign Languages, Liaodong University, Dandong 118001, Liaoning, China

²Nippon Steel Corporation, Tokyo 100-0001, Japan

Corresponding author: Hui Wen (e-mail: wenhui202508@163.com)

ABSTRACT The semantic evolution of textile-related vocabulary preserved in bamboo and silk manuscripts remains difficult to investigate because of archaic language forms, incomplete textual records, and the absence of temporally comparable semantic representations. This study develops a computational framework that integrates an ancient Chinese BERT model with dynamic word vector alignment to quantify long-term semantic variation in historical textile corpora. After constructing period-specific corpus subsets and performing context-sensitive representation learning, semantic spaces from different historical stages are aligned to enable cross-temporal comparison, while cosine-distance-based semantic drift metrics are employed to characterize lexical evolution. Experimental results show that the proposed approach achieves a global alignment error of 0.242 and limits the semantic trajectory lengths of twenty representative textile terms to below 3.12, demonstrating improved stability and comparability over conventional alignment methods. Beyond revealing the historical development of textile terminology, the framework provides a reproducible methodology for large-scale semantic analysis of domain-specific archives and historical technical documents. Such semantic representation and alignment strategies may also facilitate knowledge organization and intelligent information retrieval for multidisciplinary engineering literature, including electromagnetic and antenna-related digital resources, thereby supporting cross-domain knowledge discovery and long-term technical heritage preservation.

INDEX TERMS natural language processing, semantic evolution, dynamic word vector alignment, ancient Chinese BERT, digital humanities

I. INTRODUCTION

AS significant written sources from pre-Qin to Han dynasties, bamboo and silk manuscripts contain abundant information on ancient society, economy, and culture. The vocabulary of "textiles," in particular, expresses the ancient production technology and social life of the era. However, due to challenges like archaic language, incomplete characters, and inadequate digitization, the semantics of these words have long been hard to study systematically, making it difficult to have a deeper insight into ancient textile culture and the development of ancient Chinese vocabulary [1,2].

Recently, scholars have done a significant amount of research on textile vocabulary and its semantic evolution in bamboo and silk manuscripts. For instance, scholars have attempted to produce a textile vocabulary list and describe usage modalities through manual collation and traditional semantic analysis of the texts from bamboo and silk manuscripts of textiles. These studies heavily rely on

expert annotation and manual comparison of the frequency of occurrence of terms, context, and historical background to examine the semantics of the vocabulary used in ancient textiles in regard to production, tools, craftsmanship, and finished products. Nonetheless, traditional approaches have many shortcomings: First, manual organization is tedious, intensive, and limited in corpus size, making it improbable for them to capture a large amount of documents. Second, the resulting findings depend on the subjective judgment of scholars and do not provide numerical standards and reproducibility. Moreover, much of this scholarship focuses on surface-level modification in usage of words or phrases, thus not allowing for a systematized capturing of drift in semantics of vocabulary in context or unified comparison of corpora across historical time periods. These problems have led scholars to apply natural language processing technology or word vector models for large-scale, automated semantic analysis of bamboo and silk manuscripts hoping to improve

research efficiency, quantify semantic changes, and demonstrate more detailed historical evolutionary trajectories [3,4]. The review of deep learning models recognizes the difficulty of interpretation that accompanies these models. The method proposed here modifies manual annotation to create a framework of quantitative scale, thereby overcoming challenges of manual annotation while maintaining the same process. The framework of the annotated units of analysis is a scalable quantitative framework that operationally processes large corpora and produces reproducible measures of semantic change. The contribution emphasizes a systematic data-driven approach that reduces subjectivity and enables objective measurement. Although these studies have enriched people's understanding of ancient textile vocabulary, their methods rely on manual annotation and empirical judgment, are unable to process large-scale corpora, have difficulty quantifying semantic changes, and lack systematic tools for cross-period comparisons [5,6].

To overcome the limitations of existing research, some scholars have attempted to apply natural language processing technologies, such as word vector models, topic models, and word sense disambiguation methods, to conduct quantitative analyses of the semantics of ancient documents [7,8]. Lauriola I et al. [9] reviewed the application of deep learning in natural language processing, systematically applied common models, technologies, and tools, explored and sorted out the main resources of NLP (Natural Language Processing) research, including software, hardware, and popular corpora, and emphasized the application performance advantages and practical influence of deep learning in NLP. The discussion on the interpretability of deep learning models is limited, and the transparency of the model decision process is not involved. Zini J E et al. [10] focused on the "black box" characteristics of deep networks in NLP, analyzed the limitations of their lack of interpretability on high-risk areas, reviewed the latest methods and research progress of explanatory artificial intelligence in NLP models, and explored the challenges unique to text data, including the lack of input structure, the opacity applied by word vector embedding, and the difficulty in visualizing the internal mechanisms of deep models. Limitations include the immaturity of methods for interpreting textual data, limited applicability, and difficulty in evaluation. These methods have demonstrated some advantages in tracking lexical semantic change, but they often rely on static word embeddings and fail to fully capture the dynamic characteristics of word meanings shifting over time in historical corpora [11,12]. In response to this drawback, this article uses the ancient Chinese BERT (Bidirectional Encoder Representations from Transformers) model to create context-sensitive word embeddings, which, coupled with dimensional alignment methods for dynamic word embeddings, create comparability differences across time frames of semantic spaces, to accurately quantify the semantic changes in textile-related vocabulary [13,14]. The purpose of this study is to systematically examine the semantic change

of "textile"-related vocabulary in bamboo manuscripts and silk documents using natural language processing techniques. The research project begins by compiling a period-specific subset of the textile corpus, then pre-training the BERT model over ancient Chinese. After this stage, the research applies dynamic word vector alignment to allow comparison of semantic spaces across periods. Finally, it applies drift in meaning detection to operationalize lexical semantic changes. Its experimental results demonstrate that the textile-related vocabulary's global alignment error across periods is 0.242, which indicates a distinct evolutionary trajectory and provides a new direction of research for combining ancient text research with digital humanities practices. The chief research question being addressed is methodologically quantifying semantic change for the textile-related vocabulary in bamboo and silk manuscripts in systematic ways. Digitizing and cleaning the data is a significant prerequisite process for this work to secure the quality of the text resources. The principal methodological innovation includes the use of, and building off, context-sensitive word embeddings trained with an ancient Chinese BERT model, and utilizing dynamic word vector alignment to realize drift in meaning across periods. This method directly responds to the core problem associated with measuring semantic drift by creating analogical semantic representations realized over time.

II. TEXTILE VOCABULARY CORPUS MODELING DATA STRUCTURING

This paper first digitizes the text of bamboo and silk manuscripts. High-precision Optical Character Recognition (OCR) is then employed to transform paper and image materials into editable text. Manual proofreading is then used to correct incorrect recognition in order to guarantee the completeness and correctness of the text input. The accuracy of the OCR process was quantitatively assessed by the character error rate and word error rate metrics. Manual proofreading employed dual independent reviews with measured inter-annotator agreement rates to validate consistency and text fidelity. The text is then preprocessed, including removing repeated symbols, non-Chinese characters, and incomplete lines. Interpolation is used to fill in missing information in broken sections to minimize the impact of corpus sparsity on subsequent model training [15,16].

On this basis, it constructs subsets of the textile corpus by period. First, based on the archaeological age of the unearthed documents, it divides the text into time slices, such as the Qin and Han dynasties, to form a time series corpus $C = \{C_1, C_2, \dots, C_T\}$, where C_t represents the corpus collection from the t -th period. Second, it establishes an initial seed word list $W_s = \{w_1, w_2, \dots, w_n\}$ for textile vocabulary. Through regular expression matching and text retrieval, it extracts a set of candidate sentences $S_t = \{s \mid s \in C_t, \exists w \in W_s, w \in s\}$ from the corpus of each period, ensuring that the corpus covers the contextual information of textile vocabulary. The data structured processing is shown

in Figure 1.

To resolve the ambiguity caused by differences in the writing or form of the same word in different texts, this paper uses a string normalization method based on Levenshtein distance to normalize similar characters into a standard encoding

$$w'_i = \arg \min_{w_j \in W_s} \text{Levenshtein}(w_i, w_j),$$

ensuring that words with different writing forms are mapped to the same representation in subsequent vector training.

Furthermore, to balance the differences in corpus size across different time periods, it employs downsampling and oversampling strategies to adjust the number of sentences in each time slice to a balanced range, avoiding biases in context-sensitive word vector training caused by either sparse or dense corpus. The subsets of the textile corpus by period maintain full contextual information, with standardized form of lexis and temporal comparability in order to facilitate quality base data for BERT training and semantic alignment [17,18].

The corpus is built from a procurement of bamboo and silk manuscripts that go back to the Warring States period to the Jin Dynasty. In total, 850 manuscripts were processed. This corpus includes approximately 1.2 million words over 150,000 sentences and spans the following periods: Qin dynasty 120 documents, early Han dynasty 200 documents, middle Han dynasty 180 documents, late Han dynasty 160 documents, Three Kingdoms 100 documents, Jin Dynasty 90 documents. Notably, the original corpus varied significantly in size with the early Han dynasty writing being the largest and Jin dynasty writing being the smallest. In order to account for any discrepancies in data unbalance, downsampling was applied to time slices greater than 160. Oversampling was applied to manuscripts with less than 100 documents, increasing them similar to those with 110 documents. This enabled balance contextual representation for model training.

To evaluate the potential impact of corpus resampling on semantic drift measurements, a control experiment was conducted using the original unbalanced corpus. The same semantic evolution analysis was performed on this unmodified dataset. Key metrics of semantic evolution for core textile vocabulary were compared between the balanced and unbalanced corpora. A comparison of semantic evolution metrics is provided in Table 1.

TABLE 1. Comparison of Semantic Evolution Metrics Between Balanced and Unbalanced Corpora

Metric	Balanced Corpus	Unbalanced Corpus	Absolute Difference
Average Semantic Trajectory Length	2.21	2.24	0.03
Average Neighborhood Fidelity	0.65	0.63	0.02
Global Alignment Error	0.242	0.251	0.009

The results demonstrated consistent trends in semantic trajectory length and neighborhood fidelity metrics. The direction and significance of semantic change for core textile vocabulary remained unaffected. This sensitivity analysis confirms that the resampling strategy did not artificially alter the fundamental evolutionary patterns under investigation.

CONTEXT-SENSITIVE WORD VECTOR TRAINING

Based on the construction of a subset of textile corpora by period, this paper trains context-sensitive word vectors for the corpora of each period. First, each time slice S_t is segmented into sentences and words, and a custom ancient Chinese word segmenter is used to segment the text into basic word units $s = \{w_1, w_2, \dots, w_m\}$ to ensure that the special word order and grammatical structure of ancient Chinese are retained when input into the BERT model. The segmented corpus is then input into the ancient Chinese BERT model for pre-training. The model parameters are initialized using the Chinese BERT-base architecture, and the vocabulary is expanded to cover high-frequency ancient characters and textile-related proprietary words in bamboo and silk manuscripts. During the training process, a masked language modeling task is used to optimize the model parameter θ by randomly masking some words w_i and predicting their original word vectors $\hat{w}_i$.

$$L_{\text{MLM}}(\theta) = - \sum_{i \in M} \log P(w_i | S_t \setminus \{w_i\}; \theta). \quad (1)$$

M is the set of masked word indices, and $S_t \setminus \{w_i\}$ represents the masked context sentence. Through bidirectional encoding, the BERT model can simultaneously capture the semantic information of a word in both its context and surrounding context, effectively representing polysemous words and context-dependent meanings. The training process is shown in Figure 2.

The horizontal axis in Figure 2 shows the progress of BERT model training, with each point corresponding to one iteration of training on the complete corpus. The vertical axis shows the model's error in predicting masked words. From the beginning to the end of training, the curve decreases overall, indicating that the model's loss on the masked language modeling task continues to decrease. The curve's decrease to a lower value indicates that after multiple rounds of training, the model has become able to effectively represent the semantic features of words in context.

To enhance the contribution of the unique grammatical structure of ancient Chinese to the semantic vector, positional encoding enhancement and character-level embedding are applied during the training process to jointly encode glyph information and contextual semantics to generate the final context-sensitive word vector $v_{w_i}^{(t)} \in \mathbb{R}^d$:

$$v_{w_i}^{(t)} = \text{BERT}_{\theta}(s)_i + \text{CharEmbed}(w_i) + \text{PosEncode}(i) \quad (2)$$

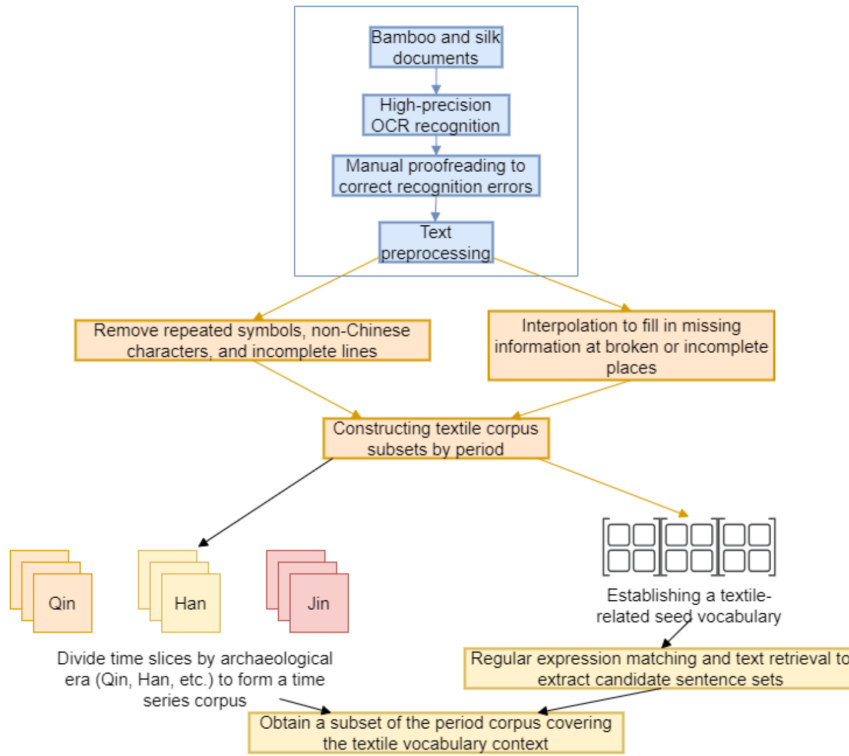


FIGURE 1. Data structured processing

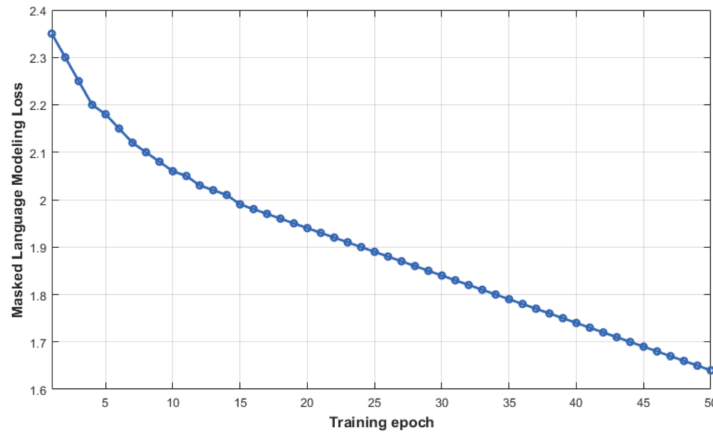


FIGURE 2. Training process

$BERT_{\theta}(s)_i$: The output vector of the BERT model for the i -th word in sentence s . $CharEmbed(w_i)$: The character-level embedding of word W (taking into account glyphs). $PosEncode(i)$: The positional encoding, representing the word's position in the sentence.

After training, each textile-related word receives an independent vector representation at different time slices, forming a time series word vector set:

$$V_w = \{v_w^{(1)}, v_w^{(2)}, \dots, v_w^{(T)}\} \quad (3)$$

V_w : time series word vector set, $v_w^{(T)}$: independent vector representation of each textile-related word in different time

slices.

SEMANTIC SPACE ALIGNMENT

To obtain a comparison of words and their semantics over historical periods, the paper utilizes a dynamic word vector alignment process. The premise of the approach is to align word vectors from different historical periods here, generating a common semantic space which allows for semantic exploratory over more than one historical period. First, while training a BERT model of Ancient Chinese texts, independent word vectors space are produced for the words falling in the respective historical time periods. Because of the different training data and temporal period of that

training data, these word vector spaces are quite different from one another, making an alignment necessary to bring them together into the same semantic space. Here the paper first calculates the means and variance of the word vectors for each of the historical periods to obtain the distribution characteristics for the semantic space for each period. Then, a dynamic alignment algorithm is used to transform the word vectors from each of the time periods linearly into the unified high-dimensional space. This process is through the alignment method based on matrix factorization. The alignment method minimizes the cosine distance across different semantic spaces by minimizing the objective function of the alignment. This outcome may make the vocabulary across periods comparable in the same space [19,20].

In the process of aligning time periods, this paper applied a time weighting coefficient to maintain the semantic characteristics of each period. The time weighting coefficient weighted the alignment results based on volume of literature and text quality for each period to prevent excessive semantic drift from the accuracy of alignment results. The time weighting coefficient for a time period was defined as a function of the period's corpus size, its text quality score. Corpus size was represented by the total number of characters, and text quality was represented by the character error rate of the text after manual proofreading. The weight for time period t is defined as follows:

$$\omega_t = \log(N_t) \times (1 - \text{CER}_t) \quad (4)$$

where N_t is the character count and CER_t is the character error rate. These weights were normalized across all periods before integration into the alignment objective function. Because of this process, the semantic space alignment of word vectors effectively removes the isolating linguistic differences and contextual biases that separate documents from different periods of time, allowing this paper to do a comparative study of vocabulary across periods of time in a shared semantic space. In the end, we have a shared set of word vectors that can then be utilized when looking at semantic change and the semantic change of "textile" related vocabulary across periods of time. The anchor words were selected from a group of high frequency common words that were seen to have semantic stability across all time periods. The selection of anchor words went through a rigid multi-step process, starting with a candidate set of the 5% most frequent words across all time periods. The semantic stability of each candidate word is measured quantitatively by computing the cosine similarity of its context vectors across the sequential time periods and calculating the variance of these similarities over the entire span of time [21,22]. Candidate words are discarded if they demonstrate an average cosine similarity of less than 0.85 or a variance greater than 0.01. The final list of 150 anchor words features words such as "of", "also", "have", "no", "people", that had the least amounts of semantic variation using this metric. The process for determining the anchor word list is strictly

objective and has taken place outside of the alignment of the target textile vocabulary. The anchor words are unrelated to vocabulary relevant to the textiles and they were basic lexical items that had a more stable meaning than other candidate words. Their stability were empirical historical linguistic and frequency analyses against other less stable words from each period. Procrustes analysis will apply the anchor words to derive an orthogonal transformation matrix and minimize the global alignment errors between vector spaces in each period. The alignment process establishes a stable semantic setting without implying stability in the textile relevant words. The separation of anchor words and textile words adds a standard unit to measure the semantic drift of textile vocabulary, it does not circularly assume stability in terms of the methodology.

SEMANTIC EVOLUTION ANALYSIS

This study utilizes a semantic drift detection metric to quantitatively analyze the semantic evolution of "textile"-related words after the completion of the semantic space alignment. First, the aligned word vectors are used to calculate the vector representation of the words in each period. In each period, this study identifies a corpus of high-frequency words related to "textiles" and calculates the differences in their word vectors across the periods in order to glean semantic evolution. Additionally, the semantic drift detection metric is used to quantify words' drift in meaning over time by measuring changes in the K-nearest neighbor structure of the local vector space. For each period, the K-nearest neighbor set of the word vector is calculated and compared with the previous period to calculate the change in Jaccard similarity. The Jaccard similarity calculation utilized the neighborhood top-k words. For each target word in the aligned semantic space, its set of k nearest neighbor words was identified based on cosine similarity between their vector representations. The value of k was set to twenty. This set of neighboring words defined the local semantic environment. The Jaccard similarity was computed by comparing the composition of this neighbor set between two consecutive time periods. The changes in semantic context are shown in Table 2.

TABLE 2. Changes in word vector semantic environment

Time period	Average Jaccard similarity	Number of drift words (below the threshold)	Average K nearest neighbor change
T1-T2	0.75	150	3.2
T2-T3	0.72	165	3.5
T3-T4	0.7	180	3.8
T4-T5	0.68	200	4.1
T5-T6	0.65	220	4.5
T6-T7	0.62	250	4.8
T7-T8	0.6	270	5

As seen in Table 2, the semantics of words will gradually deviate over time, and the local semantic environment will become increasingly different. It decreases from 0.75

between T1 and T2 to 0.60 between T7 and T8, indicating overall drift of word meaning, or a significant change in the neighborhood surrounding the word in semantic space.

To ensure the reliability of the analysis for temporal semantic drift, it utilizes a weighting mechanism that will provide a higher weight to those periods with larger corpora to overshadow the effect of disparity of the amount of data available for a specific year on the reliability of the results. Additionally, it incorporates a time decay function, placing a greater weight during temporal semantic drift calculation, to establish favored periods that exhibited more noticeable semantic changes and represent semantic evolution more effectively. The time decay function was implemented as a normalized weighting scheme based on the magnitude of semantic change. The weight for a period was proportional to the absolute difference in average Jaccard similarity between that period and its predecessor. The formal definition is $\omega_t = |\Delta J_t| / \sum_{i=1}^T |\Delta J_i|$, where ΔJ_t denotes the change in average Jaccard similarity from period $t - 1$ to t . This technique emphasizes the influence of periods with larger semantic shifts in the overall measurement of drift. The semantic change analysis incorporates a single weighted structure that considers both corpus size and the amount of semantic change. The corpus size weight is normalized across periods based on character count and other relevant text quality variables. The semantic change weight is normalized for the absolute differences in average Jaccard similarity. The overall weight for each period is derived from a weighted sum of the normalized corpus size weight and the normalized semantic change weight. This additive formulation ensures both factors contribute to the final weighting in a balanced manner, mitigating the disproportionate amplification of differences inherent in multiplicative combinations.

III. EXPERIMENTAL EVALUATION

The suggested approach is contrasted with traditional techniques for aligning data in semantic space. The Orthogonal approach rescales the vector spaces by applying a rigid rotation that minimizes the Frobenius norm between the data and output vector spaces. Linear Mapping develops a full-rank, linear transformation that may project vectors across the periods. Iterative Closest Point uses an iterative approach to find point correspondences between the spaces. These methods are typical methods of vector space alignment and serve as comparison methods.

When assessing the suggested approach, the following experimental design was implemented: Data sources were bamboo slips and silk manuscripts excavated from the Warring States Period to the Jin Dynasty and included authorized collated texts and digitized databases drawn from digitized vocabulary of pre-Qin oracle bone inscriptions, bronze inscriptions on bamboo slips, and wooden slips. The text that could be computed used optical character recognition (OCR) and was augmented by human proofreading. Repeated symbols and incomplete noise were manually

removed. Variant forms of characters were normalized using Levenshtein distance accounting for corpus consistency. Archaeological chronology is used to create time slices, and textile corpora will be assembled to allow broad comparability across periods. Each time slice corpus will be randomly divided into a training set (80%) and a validation set (20%) for the purposes of model training and parameter tuning. The textile vocabulary is distributed evenly across the training and validation sets to minimize data bias. Pre-training follows the Classical Chinese BERT model with masking of the language modeling as the optimization objective. Vocabulary expansion is ensured with textile-related technical terms. The training will utilize character-level embedding and enhance positional encoding, whereas output is context-sensitive vector representations. Dynamic word vector alignment will map the contextual word vectors learned and derived from each time slice to a common space. The words will then be aligned to a common set of high-frequency words by solving for the optimal orthogonal transformation matrix using Procrustes analysis combined with temporal smoothing to reduce the alignment error. To further reduce local corpus sparsity bias, Gaussian weighting smoothing will be applied to stabilize the evolution trajectory. The experimental hardware environment is an NVIDIA GPU cluster, and the software environment is Python 3.10 + PyTorch. The BERT model is set to 12 layers, 768 hidden size, and a 12-head attention mechanism. The domain-adaptation corpus consisted of the collected bamboo and silk manuscripts, totaling approximately 1.2 million tokens. A custom ancient Chinese tokenizer was utilized for text segmentation. The word embedding dimension was set to 768. The model training employed a batch size of 32 and a learning rate of 2×10^{-5} . The model was initialized with weights from a publicly available Chinese BERT model pretrained on a large-scale general corpus. To adapt the model to the linguistic characteristics of ancient manuscripts, it underwent continued pre-training using the masked language modeling objective on this specialized bamboo and silk manuscript corpus. This domain adaptation phase enhances the model's representation of the target textual domain without requiring task-specific labeled data.

This experimental design ensures a complete process of data processing, model training, spatial alignment and evolution analysis, and can systematically quantify the semantic evolution of textile vocabulary in bamboo and silk documents. Neighborhood fidelity is calculated as the Jaccard similarity coefficient between the k-nearest neighbor sets of a word vector across two consecutive periods. This coefficient ranges from zero to one. Semantic center offset is measured as the Euclidean distance between the mean vectors of word sets from adjacent periods. This distance is non-negative and expressed in the same unit as the word vector space. Global alignment error is computed as the mean cosine distance between aligned anchor word vectors across periods. Cosine distance ranges from zero to two. Temporal consistency score is derived from the variance of

cosine distances for a word over three consecutive periods. The score is normalized to zero for perfect consistency and increases with instability. Semantic trace length is the cumulative Euclidean distance of a word's vector trajectory across all periods. This length is a non-negative value representing total drift. All analyses employ a consistent temporal framework of six distinct historical periods: Qin, early Han, middle Han, late Han, Three Kingdoms, and Jin. The labels T1 through T6 correspond strictly to these six periods. For the fine-grained analysis in the temporal consistency score, a sliding window of three consecutive periods was applied across this sequence. The window step size was one period, resulting in an overlap ratio of two-thirds between consecutive windows.

NEIGHBORHOOD FIDELITY

Dynamic word vector alignment is performed on word vector sets from each period, mapping word vectors from different time slices into a unified space to ensure cross-period comparability. An orthogonal rotation matrix is solved using high-frequency anchor words, and a smoothing mechanism is incorporated to reduce alignment error. For each textile word, the K nearest neighbor word sets in the unified semantic space are calculated for each of the two adjacent periods. The degree of overlap between these two sets is then compared, and the overlap ratio is used as the neighborhood fidelity of the word.

The left panel of Figure 3 shows the evolution of neighborhood fidelity for a representative textile word (fang) over six historical periods. The horizontal axis represents time slices (Qin, early Han, middle Han, late Han, Three Kingdoms, Jin). The vertical axis represents neighborhood fidelity. Neighborhood fidelity is high (0.72–0.70) in the early period (Qin and early Han), indicating that the semantic neighborhood of the word remained stable during this period. It gradually decreases in the middle and late periods (approximately 0.62–0.56 from the late Han to Jin), indicating that the semantic neighborhood of the word is gradually changing, and its semantics increasingly diverge from the surrounding semantic environment during the evolutionary process. The right panel shows the distribution of neighborhood fidelity for all textile-related words across different periods. The median is high in the early period (Qin and early Han), indicating that most textile words have high semantic neighborhood fidelity during this period. As time progresses (late Han to Jin), the boxes shift downward, indicating that the overall semantics of textile-related words are gradually drifting, with uneven changes across different words.

SEMANTIC CENTER OFFSET

The word vectors from different periods are mapped to a unified semantic space using a dynamic word vector alignment method. Using high-frequency anchor words as a reference, an orthogonal transformation is used to reduce spatial deviation and ensure comparability of semantic vec-

tors across periods. The semantic center is calculated for each period's word set to obtain the semantic center for all periods. The offset between the semantic centers of adjacent periods is calculated, and the cumulative offset is recorded.

Figure 4 presents the results of an analysis of the "trends in semantic center shifts and differences in shifts between adjacent periods." The left figure shows the cumulative shifts of semantic centers across the six historical periods from the Qin Dynasty to the Jin Dynasty, while the right figure illustrates the shifts between these periods. Data analysis reveals that the shift of semantic centers was most pronounced between the Qin Dynasty and the Early Han Dynasty, increasing from 0 to 0.05. The shift between the Late Han Dynasty and the Three Kingdoms period was relatively small, at only 0.03. Between the Qin Dynasty and the early Han Dynasty, the offset increased by 0.05, indicating that the semantics underwent more dramatic changes during this period.

GLOBAL ALIGNMENT ERROR

It uses various alignment methods for word vectors across different time periods, including orthogonal transformation, linear mapping, iterative nearest point, and the proposed method. The iterative nearest point method was adapted for word vector alignment by treating word vector sets as high-dimensional point clouds. The algorithm iteratively established correspondences between anchor words based on nearest neighbor search in the vector space. It computed an orthogonal transformation that minimized the root mean square error between corresponding anchor word pairs. This process repeated until the change in error fell below a convergence threshold. This alignment yields comparable word vector spaces across time periods. For each method, it selects a set of anchor words and calculates the global alignment error between the predicted and reference vectors.

In Figure 5, the horizontal axis represents the number of alignment iterations (1 to 15), and the vertical axis represents the global alignment error. Lower values indicate higher alignment accuracy. The global alignment error (GAE) of all methods decreases with the number of iterations, indicating that the alignment optimization is approaching the optimal solution. The GAE of the proposed method decreases rapidly in the first few iterations, approaching 0.3. By the end of training, the error is 0.242, which is relatively low compared to other methods.

The global alignment error of 0.242 represents the lowest error achieved among the evaluated methods. This value is lower than the errors produced by orthogonal transformation, linear mapping, and iterative closest point techniques. A lower global alignment error indicates higher accuracy in semantic space alignment, confirming the effectiveness of the proposed method.

TEMPORAL CONSISTENCY SCORE

Textile-related texts from bamboo and silk manuscripts from the Warring States Period to the Jin Dynasty are

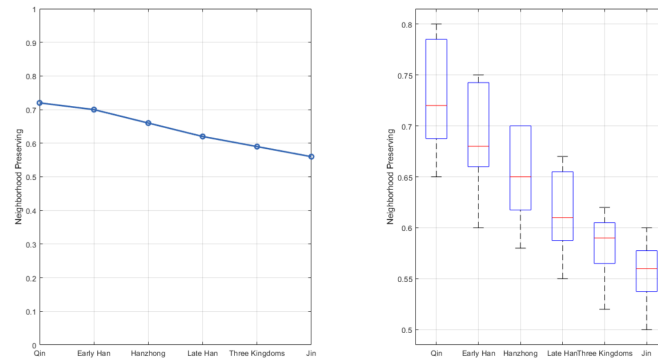


FIGURE 3. Neighborhood Fidelity

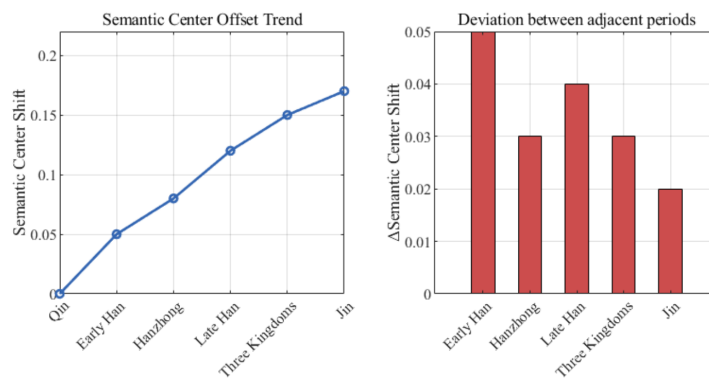


FIGURE 4. Semantic center offset

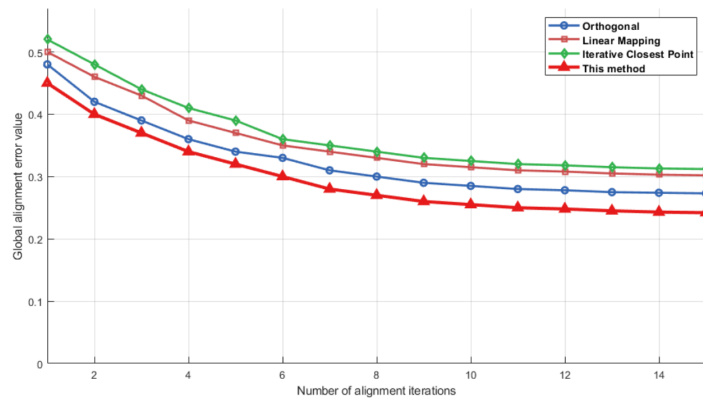


FIGURE 5. Global alignment error

collected. The texts are digitized and cleaned, including character unification, repair of incomplete parts, and removal of abnormal symbols. The corpus is divided into time slices T1-T15 according to historical periods, and a textile vocabulary set is constructed for each period. The cosine distance fluctuations of each textile word in three consecutive periods are counted to obtain the stability score of each word in each continuous time window. The vocabulary includes the following, corresponding to textile materials: yarn, satin, linen, ramie, cotton. Weaving tools:

spinning wheel, loom, shuttle, shaft, comb. Textile processes: spinning, weaving, reeling, twisting, dyeing. Textile products: cloth, silk, curtain, brocade. Textile unit: piece. Figure 6 shows the semantic evolution of textile vocabulary over successive historical periods on the horizontal axis. The vertical axis shows textile vocabulary in bamboo and silk manuscripts. Red areas indicate minimal semantic change over successive periods, indicating that the semantic core of the term remained stable throughout historical evolution. Blue areas indicate significant semantic drift, indicating that

the term underwent semantic change during this period. The color distribution of different vocabulary over the same period can be used to assess the overall level of semantic stability during that period. The overall color of column T12 is skewed towards green and blue, indicating widespread semantic change in textile vocabulary across the corpus during this period.

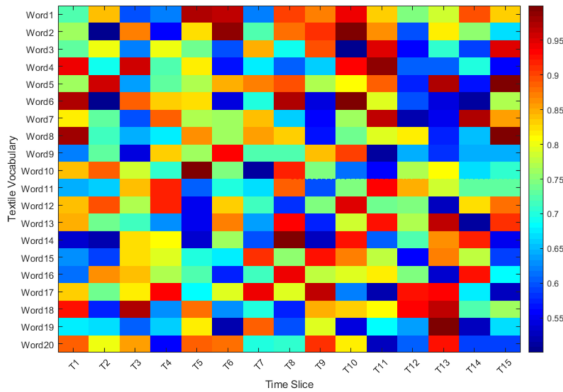


FIGURE 6. Temporal consistency score

SEMANTIC TRACE LENGTH

It applies various alignment methods to word vectors from different periods, including orthogonal transformation, linear mapping, iterative nearest point, and the method proposed in this paper. The vocabulary includes the following, corresponding to textile materials: yarn, satin, linen, ramie, and cotton.

Weaving tools: spinning wheel, loom, shuttle, spindle, and comb. Textile processes: spinning, weaving, reeling, twisting, and dyeing. Textile products: cloth, silk, curtain, and brocade. Textile unit: piece. For each word, it calculates its trajectory across time periods in the aligned semantic space. The semantic trajectory length for each word under each method is obtained, reflecting the amount of semantic drift across time periods.

TABLE 3. Semantic trace length

Serial number	This method	Orthogonal	Linear Mapping	Iterative Closest Point
1	2.35	2.50	2.60	2.48
2	1.78	1.85	1.90	1.80
3	3.12	3.25	3.30	3.18
4	2.45	2.60	2.65	2.50
5	1.95	2.10	2.15	2.05
6	2.05	2.20	2.25	2.12
7	2.78	2.90	2.95	2.85
8	1.88	1.95	2.00	1.92
9	2.15	2.30	2.35	2.20
10	1.80	1.88	1.92	1.85
11	2.22	2.35	2.40	2.28
12	2.50	2.65	2.70	2.55
13	1.95	2.05	2.10	2.00
14	1.85	1.95	2.00	1.88
15	2.10	2.20	2.25	2.12
16	2.40	2.55	2.60	2.45
17	2.30	2.45	2.50	2.35
18	2.12	2.25	2.30	2.15
19	2.25	2.40	2.45	2.28
20	1.90	2.00	2.05	1.95

Table 3 shows that the semantic trajectory lengths of 20 core words in the proposed method are all lower than those of other methods, less than 3.12, indicating that the word vectors aligned by the proposed method have minimal drift in the semantic space across time. The proposed method outperforms traditional orthogonal, linear mapping, and iterative closet point methods in controlling semantic drift and ensuring semantic stability.

The semantic trajectory lengths of less than 3.12 are less than semantic trajectory lengths generated by the orthogonal transformation, linear mapping, and iterative closet point. Less distance means less semantic drift over time, confirming that the aligned word vectors were sufficiently stable to track semantic changes.

IV. CONCLUSION

This study introduces a comprehensive computational methodology that characterizes semantic change on ancient texts. This approach merges context-sensitive word embeddings with dynamic alignment to permit equivalence across periods. The results of the experiments show similar degrees of performance across different evaluation measures, confirming the approach’s validity. The quantitative results shows semantic change in the vocabulary of textiles, and the work further demonstrates the usefulness of natural language processing in digital engagement with the humanities.

AUTHOR CONTRIBUTIONS

Conceptualization – Hui Wen and Changfeng Li; methodology – Hui Wen and Changfeng Li; formal analysis – Hui Wen and Changfeng Li; investigation – Hui Wen and Changfeng Li; resources – Hui Wen and Changfeng Li; writing – Hui Wen and Changfeng Li. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was funded by Research on Chinese and Korean Cultural Language and Translation in a cross-cultural perspective (Grant No. LJ112411779032). Project number: Liaodong University 020902y063Translation Services & Staff Training Project — Xinsheng Garment Co., Ltd.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] D. Chandrasekaran and V. Mago, “Evolution of semantic similarity-a survey,” *Acm Computing Surveys (Csur)*, vol. 54, no. 2, pp. 1-37, 2021, doi: 10.1145/3440755.

[2] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713-3744, 2023, doi: 10.1007/s11042-022-13428-4.

- [3] J. Berger and G. Packard, "Using natural language processing to understand people and culture," *American Psychologist*, vol. 77, no. 4, pp. 525, 2022, doi: 10.1037/amp0000882.
- [4] D. H. Maulud, S. R. M. Zeebaree, K. Jacksi, M. A. M. Sadeeq, and K. H. Sharif, "State of art for semantic analysis of natural language processing," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 21-28, 2021, doi: 10.48161/qaj.v1n2a40.
- [5] M. J. R. Araya, "Textiles, Techniques, Technologies: Exploring Post-Ancestrality and Contemporary Practices," *Technology and language* (Technologies in the infosphere), vol. 5, no. 3, pp. 106-121, 2024.
- [6] M. F. Bashir, H. Arshad, A. R. Javed, N. Kryvinska, and S. Band, "Subjective answers evaluation using machine learning and natural language processing," *IEEE Access*, vol. 9, pp. 158972-158983, 2021, doi: 10.1109/ACCESS.2021.3130902.
- [7] D. Kodžoman, A. Hladnik, A. P. Čuden, and V. Čok, "Assessment and semantic categorization of fabric visual texture preferences," *AUTEX Research Journal*, vol. 23, no. 2, pp. 279-291, 2023, doi: 10.2478/aut-2022-0006.
- [8] E. Alba, M. Gaitán, A. León, D. Mladenčić, and J. Brank, "Weaving words for textile museums: the development of the linked SILK-NOW thesaurus," *Heritage Science*, vol. 10, no. 1, pp. 59, 2022, doi: 10.1186/s40494-022-00681-x.
- [9] I. Lauriola, A. Lavelli, and F. Aioli, "An introduction to deep learning in natural language processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, pp. 443-456, 2022, doi: 10.1016/j.neucom.2021.05.103.
- [10] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1-31, 2022, doi: 10.1145/3529755.
- [11] M. Locatelli, E. Seghezzi, L. Pellegrini, L. C. Tagliabue, and G. M. Di Giuda, "Exploring natural language processing in construction and integration with building information modeling: A scientometric analysis," *Buildings*, vol. 11, no. 12, pp. 583, 2021, doi: 10.3390/buildings11120583.
- [12] I. K. Raharjana, D. Siahaan, and C. Fatichah, "User stories and natural language processing: A systematic literature review," *IEEE access*, vol. 9, pp. 53811-53826, 2021, doi: 10.1109/ACCESS.2021.3070606.
- [13] T. Shaik, X. Tao, Y. Li, C. Dann, J. McDonald, P. Redmond, et al., "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," *IEEE Access*, vol. 10, pp. 56720-56739, 2022, doi: 10.1109/ACCESS.2022.3177752.
- [14] A. Baumann, K. Hofmann, A. Marakasova, J. Neidhardt, and T. Wissik, "Semantic micro-dynamics as a reflex of occurrence frequency: a semantic networks approach," *Cognitive Linguistics*, vol. 34, no. 3-4, pp. 533-568, 2023, doi: 10.1515/cog-2022-0008.
- [15] J. E. L. Van Gysel, M. Vigus, J. Chun, K. Lai, S. Moeller, J. Yao, et al., "Designing a uniform meaning representation for natural language processing," *KI-Künstliche Intelligenz*, vol. 35, no. 3, pp. 343-360, 2021, doi: 10.1007/s13218-021-00722-w.
- [16] D. Hovy and S. Prabhume, "Five sources of bias in natural language processing," *Language and Linguistics Compass*, vol. 15, no. 8, Art. no. e12432, 2021, doi: 10.1111/lnc3.12432.
- [17] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, et al., "A panoramic survey of natural language processing in the Arab world," *Communications of the ACM*, vol. 64, no. 4, pp. 72-81, 2021, doi: 10.1145/3447735.
- [18] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, et al., "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1-40, 2023, doi: 10.1145/3605943.
- [19] M. P. Broderick, G. M. Di Liberto, A. J. Anderson, A. Rofes, and E. C. Lalor, "Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing," *Scientific reports*, vol. 11, no. 1, pp. 4963, 2021, doi: 10.1038/s41598-021-84597-9.
- [20] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, et al., "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1138-1158, 2022, doi: 10.1162/tacl_a_00511.
- [21] H. Yu, C. Gao, X. Li, and L. Zhang, "Ancient Chinese Poetry Collation Based on BERT," *Procedia Computer Science*, vol. 242, pp. 1171-1178, 2024, doi: 10.1016/j.procs.2024.08.179.
- [22] A. Gupta, A. Chadha, and V. Tewari, "A natural language processing model on bert and yake technique for KEYWORDS extraction on sustainability reports," *IEEE Access*, vol. 12, pp. 7942-7951, 2024, doi: 10.1109/ACCESS.2024.3352742.