Building a Semantic Search Platform for Exploring Historical Chinese Corpora

Micah Kitsunai (National Taiwan University, Graduate Institute of Linguistics, Taipei, Taiwan)
Deborah Watty (National Taiwan University, Graduate Institute of Linguistics, Taipei, Taiwan)
Shu-Kai Hsieh (National Taiwan University, Graduate Institute of Linguistics, Taipei, Taiwan)

Abstract：This work introduces a historical corpus of the Chinese language spanning approximately 3,000 years and proposes a new corpus search system utilizing word embedding techniques and large language models (LLMs). The system adopts a hybrid search method that combines traditional keyword search with vector-based search based on semantic relationships. This approach enables searches for semantically similar words and visualizations of semantic change, which were challenging with conventional corpus search methods. Additionally, based on the collected corpus data, we implemented a feature to visualize changes in word meanings across specific periods and media types. This interface allows for a multifaceted analysis of language evolution, demonstrating a more effective analytical approach than traditional methods.

1. Introduction

The Chinese language has a history spanning over 3,000 years, with characters that have retained their form and continued use into the present. Such a longstanding writing system, with a continuous and recognizable script, is unparalleled.

Throughout these extensive historical periods, the meanings of words and grammatical structures have shifted with each era. These shifts have been the focus of diachronic linguistic research.

Traditionally, linguistic research in areas such as vocabulary and grammar was largely conducted through intuitive analysis and traditional textual study. However, since the 1970s, advancements in computer technology have made it possible to process large volumes of text data, leading to the rapid development of corpus linguistics. Today, data-driven approaches that quantitatively analyze semantic change and language usage trends are increasingly becoming the norm.

While diachronic corpora are used in studies of lexical semantic change, the time spans they typically cover are limited to roughly 10 to 100 years. This work, however, employs a corpus of Chinese texts spanning approximately 3,000 years, enriched with word embedding vector data, to uncover long-term trends in semantic change and the influence of historical and social factors from a novel perspective.

This corpus aims to provide users with an environment where they can easily search, analyze, and visualize patterns of semantic change across different historical periods, without requiring advanced technical skills.

This work proposes a method that combines keyword search, the standard approach in corpus search systems, with vector-based search, which enables meaning-based retrieval. This method offers the precision of keyword search along with the flexibility of vector search, holding potential to yield new insights across a wide range of fields, including linguistic research. Additionally, by leveraging large language models (LLMs), the study aims to enhance search and analysis accuracy and broaden applications, with the effectiveness of this approach evaluated through practical use cases.

2. Previous Research

2.1 Chinese Corpora

Notable examples of Chinese corpora include the

Academia Sinica Ancient Chinese Corpus[*1] and the Academia Sinica Balanced Corpus of Modern Chinese [*2]. However, to our knowledge, there is currently no Chinese corpus or search system that spans both ancient and modern periods and is openly accessible as a search tool.

In studies investigating the diachronic shifts in word meanings, such as those by [1] and [9], only self-collected data for specific words were utilized.

2.2 Word Embeddings

Word embedding techniques, which represent word meanings as numerical vectors, are widely used. Static embeddings, such as Word2vec [6], GloVe [8], and fastText [8], have played a significant role as foundational models for lexical semantic representation. However, these methods assign a fixed vector to each word, which poses a challenge in capturing context-dependent variations in meaning. For instance, polysemous or ambiguous words are assigned the same vector regardless of context, making static embeddings unsuitable for context-sensitive semantic representation.

To address this challenge, embedding methods that consider context have been developed, using self-attention mechanisms as represented by BERT[3]. This allows even the same word to be assigned different vectors depending on the context. As a result, the meanings of polysemous or ambiguous words can be accurately captured based on surrounding context. In Chinese, models such as macBERT[*3] and GuwenBERT[*4] (which is pre-trained on classical texts) provide embeddings specialized for specific language contexts.

Additionally, as a more comprehensive approach, [4] proposed dynamic embeddings, which incorporate not only contextual embeddings but also temporal and social information, broadening the scope of the model.

2.3 Semantic Change

Advancements in embedding techniques have made it easier to analyze semantic change, a phenomenon in which words acquire new meanings or undergo transformations in meaning over time. Semantic change encompasses various patterns, including extension (broadening of meaning), narrowing (restriction of meaning), and metaphorization (acquisition of metaphorical meanings) [2]. Approaches using large language models (LLMs), such as BERT[3], have been reported to generate high-precision embeddings for capturing semantic change over time. These models enhance accuracy compared to traditional methods, owing to their size and advanced capability to understand diverse contexts [7].

3. Methodology for Constructing the Chinese Historical Corpus

We developed a corpus encompassing Chinese text data from ancient to modern times.

3.1 Corpus Data

To construct the Chinese historical corpus for this work, text data spanning from ancient to modern periods was collected from three sources: CTEXT, CBETA, and PTT—as detailed in the following sections. The collected data covers an extensive range of over 2,500 years, encompassing various types of texts, from social media posts to classical literature, ensuring comprehensive coverage of text types.

For details regarding token counts and other information for each data source, refer to Table 1.

The data is organized into seven distinct periods, from Pre-Qin era to the Republic of China period. We used the dynasty divisions of CTEXT as a common categorization across the corpus, rather than the more detailed breakdown of CBETA, as shown in the dynastic distribution of texts in Figure 1.

3.1.1 CTEXT

The Chinese Text Project (CTEXT)[*5] is an open-source platform providing texts of classical Chinese literature. We selected CTEXT as a data source due to its extensive coverage of diverse fields, from Confucianism and Daoism to medicine and excavated documents, spanning from the pre-Qin period to the Republic of China. Text data was retrieved via the CTEXT API and indexed at the character level to enable later search and analysis.

3.1.2 CBETA

The Chinese Buddhist Electronic Texts Association (CBETA)[*6] is a text collection encompassing Buddhist literature from the Han dynasty to modern times.

[*1] https://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh

[*2] https://lingcorpus.iis.sinica.edu.tw/modern/

[*3] `https://github.com/ymcui/MacBERT/blob/master/README_EN.md`

[*4] `https://github.com/ethan-yt/guwenbert/blob/main/README_EN.md`

[*5] https://ctext.org/

[*6] https://www.cbeta.org/

Table 1 Data by Media Source

Media Source	Chinese Type	Tokens Collected	Percentage of Total
CTEXT	Ancient Chinese	571,490,912 tokens	49.9%
CBETA	Ancient Chinese	209,601,026 tokens	13.4%
PTT	Modern Chinese	778,447,272 tokens	36.6%
Total		1,559,539,210 tokens	

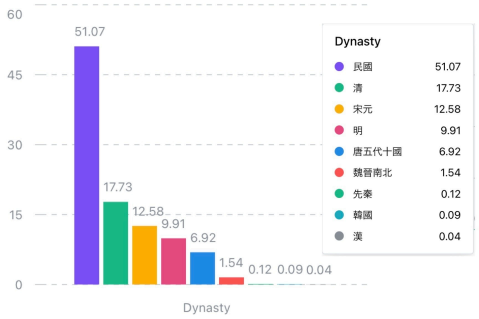


Figure 1 Distribution by Dynasty

To supplement CTEXT and obtain additional ancient data, CBETA was selected for its extensive temporal coverage. Text data and metadata were obtained by web scraping.

3.1.3 PTT

PTT[*7] is the largest electronic bulletin board in Taiwan. PTT was selected as a source of modern Chinese data due to the ease of categorizing posts by type using the “boards” on the platform, as well as its developer-friendly access for data retrieval. Given the vast amount of data available, we used data from five specific boards covering the years 2006 to 2024. For further details, see appendix Data was collected via Python-based web scraping, followed by tokenization and morphological analysis to prepare for subsequent analysis.

3.2 Indexing

In this work, the corpus data was converted into XML format and then indexed using BlackLab[*8], an open-source tool specialized for building language corpora.

Each entry is enriched with metadata such as dynasty, media type, author, and title, which can be utilized during corpus searches.

4. Creation of Word Embeddings

To incorporate semantic search into the corpus system, we applied embedding processing to a portion of the corpus data.

Embeddings were generated at the word level. While sentence-level embeddings are suitable for tasks like summarization, word-level vector data were deemed more appropriate for enhancing the precision of semantic search and for analysis in corpus linguistics.

4.1 Word Embedding Models

For generating context-aware embeddings for the corpus data, as mentioned in Chapter 2.2, we used GuwenBERT and macBERT. The GuwenBERT-base model employed is based on a RoBERTa model pre-trained on classical Chinese texts, while macBERT primarily targets modern Chinese.

Following the specifications of BERT, up to 512 tokens before and after each target word were used as input, generating a 768-dimensional word vector for each word, thus creating context-aware embeddings. For each period, we generated embeddings for an average of over 1,000,000 tokens, totaling 6,154,900 tokens across all periods.

Using BlackLab’ s API, index data was sequentially retrieved, vectorized, and stored in a vector database along with select metadata for linkage purposes. Large language models (LLMs) have facilitated efficient and highly accurate word embeddings. This approach significantly reduces the processing costs and adjustments that were previously necessary.

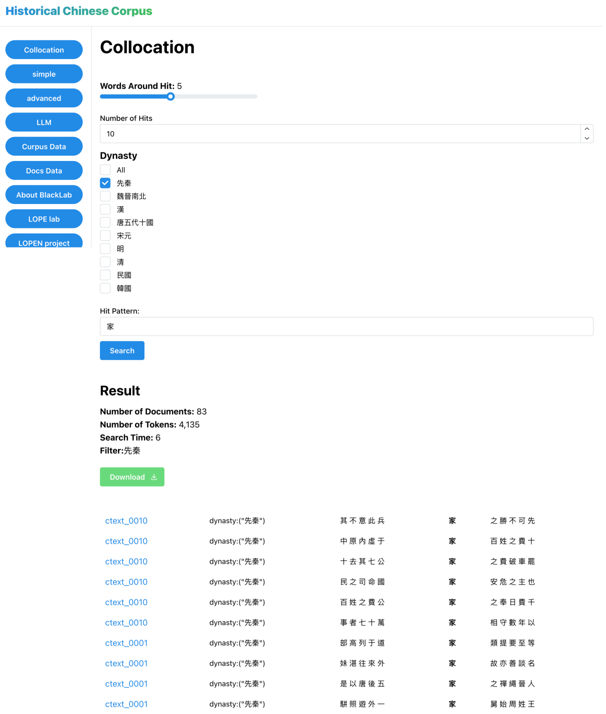
4.2 Vector Database

For the vector database, we used PostgreSQL along with its extension, pgvector[*9], to store the vector data.

Since BlackLab is optimized for storing and searching linguistic data, we implemented a separate vector database for storing and retrieving vector data, integrating it with BlackLab’s data.

The vector database is designed to store only essential information: the target word, the vectors generated by each of the two models, and the ID and position number of the sentence in which the word appears (from BlackLab). The connection with BlackLab is established through the sentence ID and position number,

[*7] https://term.ptt.cc/
[*8] https://inl.github.io/BlackLab/
[*9] https://github.com/pgvector/pgvector



Collocation

Words Around Hit: 5

Number of Hits: 10

Dynasty

☐ All
☑ 先秦
☐ 魏晉南北
☐ 漢
☐ 唐五代十國
☐ 宋元
☐ 明
☐ 清
☐ 民國
☐ 韓國

Hit Pattern: 家

Result

Number of Documents: 83
Number of Tokens: 4,135
Search Time: 6
Filter:先秦

ctext_0010	dynasty:("先秦")	其不意此兵	家	之勝不可先
ctext_0010	dynasty:("先秦")	中原內虛于	家	百姓之費十
ctext_0010	dynasty:("先秦")	十去其七公	家	之費破車罷
ctext_0010	dynasty:("先秦")	民之司命國	家	安危之主也
ctext_0010	dynasty:("先秦")	百姓之費公	家	之奉日費千
ctext_0010	dynasty:("先秦")	事者七十萬	家	相守數年以
ctext_0001	dynasty:("先秦")	部高列于道	家	類提要至等
ctext_0001	dynasty:("先秦")	妹湛往來外	家	故亦善談名
ctext_0001	dynasty:("先秦")	是以唐後五	家	之禪繩晉人
ctext_0001	dynasty:("先秦")	駢照遊外一	家	舅始周姓王

Figure 2 Corpus Interface

allowing retrieval of relevant context or additional information from BlackLab when needed.

This approach, where only the minimum information necessary for vector search is stored in the vector database and all other data is managed by BlackLab, enables efficient data partitioning and management.

5. Methodology for Developing the Corpus Search System

5.1 Corpus Search Interface

To provide an interactive corpus search interface for users, we developed a system using TypeScript and Express for the backend and React for the frontend. The backend centralizes the management of BlackLab' s API and the vector database, handling search functions and data linkage.

This setup allows users to effortlessly perform keyword searches in natural language as well as vector searches, without needing to be aware of complex data processing.

5.2 Hybrid Search Method

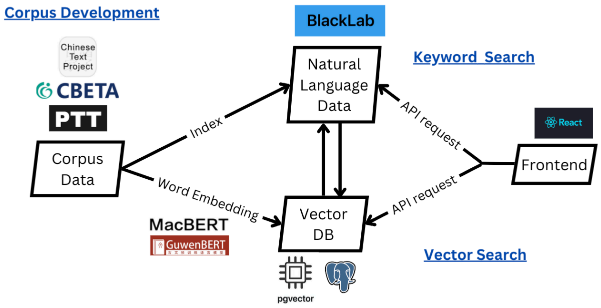


Figure 3 Pipeline of the Proposed Method

This work proposes a hybrid search method that combines BlackLab, optimized for traditional keyword searches in linguistic corpora, with vector search, which allows for consideration of semantic relationships (see Figure 3).

In conventional corpus searches, searches are primarily based on keyword or string matching, making it difficult to account for differences in word meaning or context. However, by integrating vector search, it becomes possible to retrieve semantically related words and concepts. The data indexed in BlackLab is embedded, inserted into the vector database along with metadata, and linked by document ID.

Our method combines BlackLab' s keyword search with vector search, enabling the discovery of a broader range of related information while using keyword accuracy to complement the ambiguity of vector search. This approach enables flexible and advanced searches beyond what conventional corpus search systems offer.

Furthermore, the vector data obtained through word embeddings is useful for analyzing diachronic and sociocultural semantic shifts within the corpus . For example, because the data stored in the vector database is linked to BlackLab by a common ID, searches can filter data by specific time periods or media types. This capability supports identifying trends in semantic change within particular historical or cultural contexts, contributing to search efficiency.

By combining the strengths of both semantic and keyword search, this hybrid method has the potential to contribute to new insights in linguistics and related fields.

6. Experiments

To validate the effectiveness of the proposed hybrid search method and compare it with conventional corpus search approaches, we conducted evaluations through specific case studies.

This case focuses on a feature that visualizes how the meanings of specific words have shifted over time.

As an example of a target word, we consider "家" (jia). In addition to the meaning of "house" (as a building), "家" can also refer to metaphorical meanings like "a place of origin" or "a small group sharing similar philosophies. These meanings are thought to vary in frequency across historical periods.

We developed a feature to visually display how these various senses of the target word have changed across different periods, leveraging vector search and BlackLab metadata.

Specifically, target words are extracted from the vector database, clustered using K-means, and plotted by cluster and historical period on a graph (see Figure 4). The horizontal axis represents time periods, while the vertical axis shows the proportion of occurrences for each cluster, allowing a visual understanding of the semantic evolution of the target word over time.

Furthermore, by clicking on a cluster point, users can view the context in which the word appears through integration with BlackLab (see Figure ??).

To determine the number of clusters and to extract example sentences for each clustered sense, we used the Chinese Wordnet (CWN)[5] for modern language. Example sentences from each cluster can be compared with those from the existing wordnet, allowing an LLM to assign a sense to each cluster.

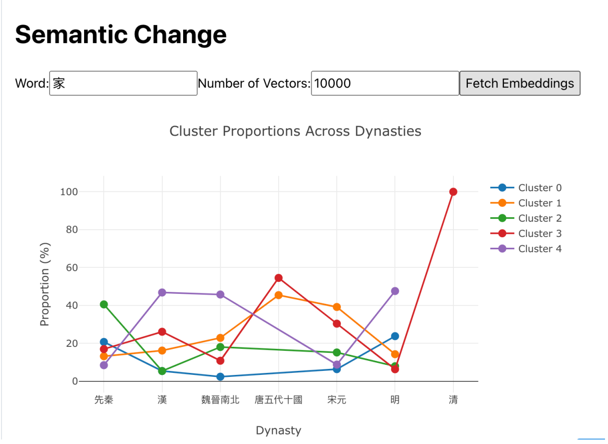
However, it should be noted that modern wordnets do not cover all meanings, especially for Classical Chinese. Additionally, word meanings can split or merge at specific points in time and may even disappear in certain eras, making it challenging to capture these changes accurately.

7. Discussion

7.1 Limitations

One limitation of this work is that much of the data used is biased toward classical texts, making it challenging to fully capture the linguistic features of ancient Chinese using resources aimed at modern language. Additionally, evaluating whether the embeddings accurately capture only the intended meanings remains a challenge.

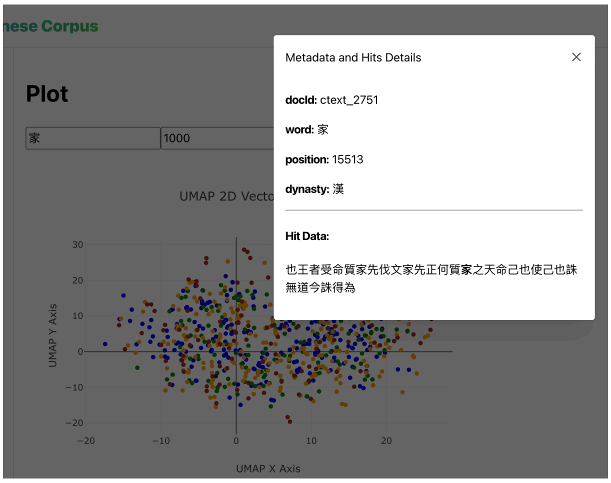
Furthermore, there is a noticeable bias in the current corpus data. For example, the PTT data used for modern Chinese primarily represents Taiwanese Chinese, while CBETA consists mainly of Buddhist texts. Consequently, the overall data may exhibit regional and thematic biases.



Semantic Change

Word: 家 Number of Vectors: 10000

Figure 4 Graph of Semantic Change



Metadata and Hits Details

docId: ctext_2751

word: 家

position: 15513

dynasty: 漢

Hit Data:

也王者受命質家先伐文家先正何質家之天命己也使己也誅無道今誅得為

Figure 5 Example of Contextual Search Results for Clustering of " 家"

7.2 Future Directions

Future work includes exploring more accurate word embedding models. In particular, we plan to investigate whether using GPT-based models instead of BERT-based ones improves embedding precision and semantic capture capabilities. Furthermore, by leveraging the natural language processing abilities of LLMs, we will explore using "agents" that perform information retrieval and task automation based on user instructions. This would allow for a streamlined approach, from input in natural language to output and visualization, aimed at advancing corpus linguistics.

The Chinese historical corpus and search system developed in this work are expected to contribute to corpus-driven linguistic research by not only enabling the analysis of predefined research questions but also facilitating the discovery of new research themes. Additionally, by making this system accessible to a broad user base beyond linguistic experts, we aim to provide a valuable resource that benefits a wide range of users beyond research.

8. Conclusion

In this work, we constructed a diachronic corpus spanning about 3,000 years of Chinese language and demonstrated the potential of applying word embedding techniques and large language models (LLMs) to envision the next generation of corpus search systems. We established a novel approach that integrates traditional keyword search, which corpus linguistics has predominantly relied upon, with semantic relation-based search methods.

References

[1] Chi, Y., Giunchiglia, F. and Xu, H.: Diachronic Semantic Tracking for Chinese Words and Morphemes over Centuries, *Electronics*, Vol. 13, No. 9, p. 1728 (2024).

[2] Closs Traugott, E.: On regularity in semantic change (1985).

[3] Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[4] Hofmann, V., Pierrehumbert, J. and Schütze, H.: Dynamic Contextualized Word Embeddings, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Zong, C., Xia, F., Li, W. and Navigli, R., eds.), Online, Association for Computational Linguistics, pp. 6970–6984 (online), DOI: 10.18653/v1/2021.acl-long.542 (2021).

[5] Huang, C.-R., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X. and Huang, S.-W.: Constructing chinese wordnet: Design principles and implementation, *Zhong-Guo-Yu-Wen*, Vol. 24, No. 2, pp. 169–186 (2010).

[6] Mikolov, T.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).

[7] Montariol, S.: Models of diachronic semantic change using word embeddings, PhD Thesis, Université Paris-Saclay (2021).

[8] Pennington, J., Socher, R. and Manning, C. D.: Glove: Global vectors for word representation, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543 (2014).

[9] Zhou, Y.: From 'deviate' to 'obedient' : A diachronic study on the semantic change of the auto-antonym guai in Chinese (2018).

Appendix

Table A·1 The Periodization of Dynasties Used in Our Corpus

Dynasty	**Period**	**Approximate Duration**
Pre-Qin	221 BCE - 206 BCE	15 years
Han	206 BCE - 220 CE	425 years
Wei, Jin, and Northern and Southern Dynasties	220 CE - 589 CE	370 years
Tang and Five Dynasties Ten Kingdoms	618 CE - 979 CE	365 years
Song and Yuan	960 CE - 1368 CE	410 years
Ming	1368 CE - 1644 CE	280 years
Qing	1636 CE - 1912 CE	280 years
Republic of China	1912 CE - Present	110 years

Table A·2 PTT Boards and Data Collection Periods

Board	**Description**	**Data Collection Period**
HatePolitics	Discussion on political issues, often critical in nature	2006-2010
Boy-Girl	Forum on relationships and dating advice	2006-2010
C_Chat	General chat board for comics, games, and pop culture	2007-2015
Gossiping	Informal discussions on various trending topics	2009-2010
Stock	Forum focused on stock market and investment discussions	2007-2015