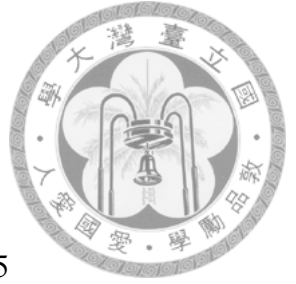


國立臺灣大學文學院語言學研究所  
碩士論文

Graduate Institute of Linguistics  
College of Liberal Arts

National Taiwan University  
Master Thesis



0.5

詞向量的語意變遷計算模型：  
以「家」為例

Modeling semantic change with word embeddings:  
a case study of *jiā*

陳蓓怡  
Pei-Yi Chen

指導教授：謝舒凱博士  
Advisor: Shu-Kai Hsieh, Ph.D.

April 2021  
中華民國 110 年 4 月



0.5

## 摘要

本研究欲從語料量化與計算的觀點切入詞彙語意變遷的語言現象。近年來，文字在網路上大量流傳，加上社會快速變遷，語意表達亦不斷變化。與此同時，歷史文本的電子化數量的增長，使我們得以從中分析、挖掘詞彙所蘊含的詞意，開展了更多與歷時語意相關的研究可能。

語言，將所思所想傳遞、紀錄，並在說話者使用語言時，不斷被重塑與流傳 (**blank1999new**)。從共時 (synchronic) 的角度來看，語意存在各種變異 (variation)，而在歷時 (diachronic) 的脈絡下，經過時間累積而則彰顯了各種的變遷。近年來的歷史詞彙語意研究，從詞意的改變、新舊字詞的興衰，探索其背後的運作機制與認知層面，已開始摸索出語意變遷 (semantic change) 的規律性 (regularities) (**blank1999new**)。語料庫作為語言使用的經驗素材，提供了我們從中觀察、歸納出可質化、量化的語言分析；而歷時語料庫更因應科技進步，結合了計算語言學界近年來的語言向量表徵、神經語言統計模型等新方式探求語意在時間洪流下的變動與趨勢。

然而在歷時語料中，有些詞彙並無明顯的詞頻變化，其多義行為亦造成研究者面對巨量資料時的困擾。本論文的目的，在於結合語料統計模型與計算語意學的表徵模型，探究漢語的語意變遷。從數位化的原始語料中，以共現 (co-occurrence) 分佈的趨勢發覺意義分布的異同，並從語境詞向量 (contextualized word embeddings) 將多義性 (polysemy) 的變動做形式表達。期待以量化的方式量測語意變遷的程度，並以質化分析輔證已知的例子，並發掘更多可能的例子與規律。我們以歷時語料庫 (中國哲學書電子計畫 (**sturgeon2019c**text)) 與現代漢語語料庫 (中研院漢語平衡語料庫 (**chen1996sinica**)) 為語料來源，建立歷時詞向量並搭配詞彙資料庫，並參考 **hamilton2016cultural** 的全域鄰近詞法，以搭配詞的相似度數值組成二階

向量 (second-order embedding)，提高語意表徵的精確度來比較各時代向量的方法，求其相關係數和語意變遷程度之間的關聯。並從詞彙的意義分布與互動，描繪出不同詞意的消長與變動。此外，本研究也同時採用以變異程度為基礎的近鄰群聚分析法 (Variability-based Neighbor Clustering, VNC) (gries2012variability)，此階層式的分群可勾勒出綜合性評估各觀察變項的影響下，漢語詞彙發展的時代區分。<sup>0.5</sup>

計算語意學與歷史語意學的整合研究可以使我們在經驗基礎上回溯驗證個別詞彙的意義變化，更進一步梳理整體的原理原則。詞彙反映人們對於新事物賦予新名的動機、社會概念的更迭也同時牽動詞彙之間的關聯。本研究的應用範圍更可擴及到詞彙與文化變遷的探索。

**關鍵詞：**語意變遷、歷時語意、向量表徵、階層式集群



0.5

# Abstract

This research aims to investigate the topic of historical semantic change from the perspective of quantitative and computational linguistics. With a rapid accumulation of texts in the digital era, attention is called upon a more temporal-aware interpretation of language use and meaning construction. Meanwhile, the digitalization of historical texts opens up more research opportunities to trace the diachronic development of words and meanings. Especially, semantic change motivated by linguistic features and factors can be explored in a data-driven approach.

Language is a means of communication through which ideas are conveyed, stored, and recorded, and in essence, constant change and evolution occurs as the speakers use the language with the passage of time (**blank1999new**).

The dynamics of meaning construction is embodied in the emergence and loss of senses, as well as the split and shifts, which contributes to the different distributions and interactions of words, reflects the regularities and adaptability of the language, and the cognition and culture operating behind (**blank1999new**). Synchronic variations can be dealt with through a diachronic lens. Corpus-based, data-driven approach enables an observation and derived generalizations of semantic change. Coupled with the advances in vector space models and statistical analysis, the changes in meaning are explored. Polysemy is a driving force of semantic change. Concepts and meanings are structured in words and language use, and how word-formation is realized in Chinese is addressed in the development of monosyllabic to disyllabic words, which not only allows us to explore the influence of homophony, the interaction between words, and the growth of disyllabic words and compounds. Seeing that historical textual data are in demand, computational semantics and statistical models resolves the dilemmas.

On top of that, it is possible that semantic change occurs not in observed frequency, but other distributional ways, making the encoded meanings distinctively different from

previous time periods. As distributed models like word embeddings are receiving much attention, historical semantic change is a research topic that should enter the discussions. In the field of corpus linguistics, such research method are based on co-occurrences of words in context, and the co-occurrence distribution represents the similarities and differences in meaning interactions.

0.5

The diachronic corpus consists of texts from the following sources: the Chinese Text Project (**sturgeon2019ctext**) and Academia Sinica Balanced Corpus of Modern Chinese for modern Chinese (**chen1996sinica**). By applying a quantitative inquiry into semantic change, we will measure the degrees of semantic change, support known change cases, and discover unknown ones, with the consultation of lexical databases. Firstly, the global measures proposed by **hamilton2016cultural** is adopted. Second-order embeddings comprised of similarity scores of keywords are formed to compare the meaning representations of different eras. The lower the correlation between two temporally-adjacent vectors, the higher the degrees of semantic change. Secondly, based on the distribution and interaction of a word's senses, the semantic trajectories of the word will be traced. Finally, this study will proceed with periodization analysis using the Variability-based Neighbor Clustering (VNC) method (**gries2012variability**). As a hierarchical clustering method, it is bottom-up, as opposite to the decisive clustering, a comprehensive evaluation of the influence of the selected linguistic factors in this study is implemented to explore how the development of meaning construction can be understood under different stages. In sum, this study explores the phenomenon of semantic change in retrospect to derive the semantic development in diachrony. The computational/statistical modeling of historical lexical semantic change will shed new light on how the language community describes and makes sense of the society that is also constantly changing.

**Keywords:** Semantic change, diachronic lexical semantics, distributed representations, hierarchical clustering

# Table of Contents

0.5



# List of Figures

0.5



# List of Tables

0.5








0.5

# Chapter 1

## Introduction

Language is constantly changing and evolving. The emergence of new senses, the demise of old ones, and the polysemous nature of linguistic expressions make the process of semantic change a dynamic phenomenon (**robertinvanhove2008**). As individuals learn new words and meanings throughout their life, so does a language. As language users actively engage in processing and interpreting the language, the semantic history of words are woven into the texts that then survive time and are presented to us now. In the long run, a word is likely to convey a meaning completely different or unfathomable. For instance, “the quick and the dead”, quoted from the Bible, means “the living and the dead”, but the collective adjective “the quick” no longer makes sense in Present-Day English (**semanticincrowley2010**).

The nature of language is reflected in its use. In **sinclair1982reflections**, **sinclair1982reflections** envisions the possibility of “vast, slowing changing stores of text” and “detailed evidence of language evolution” (**renouf2002time**). In the recent years, a huge amount of historical text data have been digitized and made available to the public, and the use of digitized libraries as rich linguistic resources to observe how certain linguistic features are “assimilated” into the language becomes more and more feasible (**renouf2002time**). While recent studies have used time-sliced collections of texts to observe swift meaning changes, the digitalization of texts from earlier time periods opens up research opportunities that incorporates a corpus-driven approach to trace the diachronic development of words and their meanings (**kutuzov2018survey**; **tahmasebi2018survey**; **camacho2018survey**).



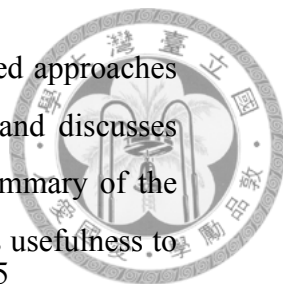
With the recent advances in Natural Language Processing (NLP) techniques, the changes in meaning over time can be to a great extent captured by representing discrete linguistic data as numeric vectors such as word embeddings, especially after the release of Word2vec (mikolov2013efficient), GloVe (pennington2014glove) and FastText (bojanowski2016enriching). For instance, for the study of semantic change of individual words across time, initial efforts have been put into generating word embeddings from different time spans and explore whether semantic change occurs based on the neighboring words of the target word from each time period.

As the pioneering computational-historical investigation in Mandarin, the monosyllabic word 家 *jiā* ‘home’ is selected as a case study in this thesis. The concept of home is an ancient, seemingly familiar and encompassing, but tangible one. Various humanities disciplines have sought to grasp the full picture. Defined by the Oxford English Dictionary (OED), the word *home* is “the place where a person or animal dwells” (“**homeinoed**”, **homeinoed**). As one of the earliest 1% entries to be included in the OED, this word has 35 main senses and 214 total senses—Home is a physical space, a place where we feel a “sense of belonging [and] comfort”, and even a person’s “country or native land.” In Mandarin Chinese, the MOE Revised Mandarin Chinese Dictionary defines its translated equivalent 家 *jiā* ‘a’s “a place where family members live together (眷屬共同生活的場所)”, “a private property (私有財產)”, and “people in certain professional fields (經營某種行業或具有某種身份的人)” (“**jiainmoe**”, **jiainmoe**). Yet, how is the concept of home encoded linguistically? Specifically, how its diachrony interacts with synchrony and variations is the main concern of this study.

From the perspective of corpus-based computational linguistics, research questions are invoked as to how the concept of home is properly computationally represented? What words are semantically related to this concept? and how are these words co-construct the meanings of home, and how this concept comes into shape through the lens of time.

The remainder of this thesis is organized as follows. An theoretical overview and reflections of lexical semantic change in general, the concept of home in literature, as well as the diachronic word embeddings techniques are given in Chapter ???. Chapter ??? introduces the preprocessing issues, and the proposed corpus-based clustering method and distributed semantic representation models for the study. The development of word-level and sense-level word representations brings to the fine-grained analyses and

generalizations of semantic change. Chapter ?? describe how the proposed approaches are evaluated, and showcase analyses made possible by our approach, and discusses their successes and limitations. Finally, Chapter ?? concludes with a summary of the contributions and with considerations on the future works as well as on its usefulness to linguistic investigations and other social-cultural applications.



0.5



0.5


## Chapter 2

### Related works

#### 2.1 Lexical semantic change

Language is dynamic; it changes in the passage of time. Previous studies have shown that lexical semantic change is both linguistically and socially motivated (**kutuzov2017tracing**; **kutuzov2018survey**; **hamilton2016cultural**).

Semantic change can be broadly understood as the “reanalysis” of a word (**fortson2017approach**), and recognizing different types of semantic change does not entails an absolute distinction of a certain type, but outlines the research foci of previous studies (**fortson2017approachtraugott2017semantic**). **bloomfield1933language** classification of semantic change highlights the denotative (broadening/narrowing), connotative (degeneration/elevation), intensity (hyperbole), figurative (metonymy/metaphor), and relational (synecdoche) aspects of a lexical item that undergoes semantic change. In **semanticincrowley2010**, types of semantic change are distinguished from the forces. The former includes broadening, narrowing, bifurcation (split), and shift, and the latter includes hyperbole, metaphor, euphemism, interference, folk etymology, and hypercorrection. Whether an instance of semantic change is bifurcation or shift is determined by the absence of the original sense. Semantic shift is reflected in the cognate words from target languages, which do not come to have the new meaning. In terms of hyperbole, words in constant use become more and more neutral. Interference describes the semantic relations of synonyms or homonyms; other word are in place to avoid confusion in communication.

The main types of semantic change —of which e.g. **traugott2017semantic** offers historical examples are as follows (quoted from (**giulianelli2019lexical**)): 

1. broadening (or generalisation): the extension of the range of concepts designated by a term, 0.5
2. narrowing (or specialisation): the contraction of the range of concepts designated by a term,
3. metaphorisation: the conceptualisation of one referent in terms of another, guided by analogical reasoning and implying an unspoken simile,
4. metonymisation: a meaning transfer from one word to another, guided by spatial, temporal or causal contiguity between the two referents,
5. amelioration: the acquisition of or shift towards a positive connotation,
6. pejoration: the acquisition of or shift towards a negative connotation.

**traugott2001regularity** also noted that meaning change often occurs in the direction from concrete to abstract. Originally, a lexical item bears contentful meaning. During grammaticalization, grammatical or procedural meaning is enriched although the contentful one might persist.

Depending on the initial step of investigation, semantic change can be approached from a semasiological and onomasiological perspectives (**geeraerts1997diachronictraugott2001regularity**). A semasiological perspective highlights the direction from linguistic expression to concept, so meaning change is studied under the consideration of a lexeme in a fixed, predetermined form. Conversely, an onomasiology perspective starts from concept to linguistic expression, and thus meaning change is framed within a given concept expressed by a set of alternative words. Nonetheless, both of the two complementary paths lead to such important topics in lexicology as polysemy and sense relations. Semasiologically, when a lexeme undergoes semantic change and additional meanings are gained, the different senses might gradually be perceived as unrelated to one another by the language users. That is, the lexeme first becomes polysemous, and then homonymous (**traugott2001regularity**). Onomasiologically, on the other hand, focuses on synonyms, nearsynonyms, and

namegiving to connect lexical items with sense relations that exist and develop under a concept over time (**geeraerts1997diachronic**).

Polysemy, for instance, goes hand in hand with the semasiological view. It is described as “families of related meanings” in **traugott2001regularity**, and serves as a foundation of generalizations of semantic change with recurring patterns. The coexistence of older and newer meanings in a lexical item, along with the influence of multiple meanings on one another, brings about the dynamics of “saliency” (**traugott2001regularity**). Being polysemous with more than one single semantic reading is not only necessary but also omnipresent. In particular, synchronic polysemy is pointed out as an integral component among the driving forces of lexical semantic change, a phenomenon that is often explored in a diachronic vein (**robertinvanhove2008**).

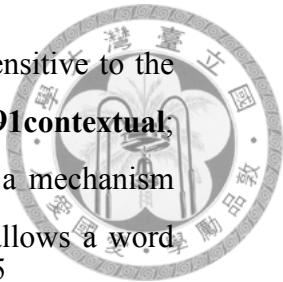
As a topic that has long interested scholars in semantics and historical linguistics, semantic change is a complicated phenomenon resulting from an interplay of polysemy, with subjectification (**traugott2001regularity**), prototypicality (**geeraerts1997diachronic**), and other contributing factors. Semantic change has been extensively studied because linguistic variations of language use are pervasive in the synchronic settings, and are amplified in a diachronic scope (**semanticincrowley2010; bowern2019semantic**). The term “brachychrony” is even coined by **mair1998corpora** to refer to a time span of 10 to 30 years, indicating how the change of a linguistic feature can be delineated within a short time frame.

The Invited Inferencing Theory of Semantic Change (IITSC) is proposed by **traugott2001regularity** to account for the actuation of meanings through recognition of different stages of a linguistic expression depending on whether intended meanings are coded or crystallized into commonly used implicatures. In other words, the degree of conventionality reflects the stages in which an expression is during a certain period of time. The more conventional or less context-specific an expression is, the more crystalized or coded the meaning is conveyed through this expression, which indicates that the expression has evolved in the later stages of the IITSC. Importantly, the meaning of an expression is not limited to only one, but a second reading often becomes more and more readily accessible as the coded meaning, and is acceptable by the language community. For example, through expressions of temporal sequence, invited inferences of causality can arise. Over time, semantic change follows a path from coded meanings to

utterance token meanings to utterance type, pragmatically polysemous meanings (GIINs) to new semantically coded meanings. That is, a new meaning emerges as a creative, innovative instance of language use by an individual and does not yet spread to a wider language community, but remains more idiosyncratic. Slowly, it is likely that the new meaning is acquired socially with strengthened pragmatic impact, the expression is then pragmatically polysemous. The final stage of the evolution cycle is for the expression to be semantically polysemous or coded, with the new meaning being the dominant or salient reading.

**geeraerts1997diachronic** puts forwards a conceptual framework that describes semasiological change motivated by the prototypicality theory. Extensionally, members of a semantic category do not have equal representativeness or typicality of the category, and their membership can even be uncertain if the member is highly peripheral. Intensionally, meanings of less typical members are received from the more salient meanings and can overlap, yet the salient meanings are not determined solely from one single cluster of attributes. Generally, the synchronic semantic structure of lexical categories echoes with the diachronic semasiological change. Diachronically, the more salient the meaning, the more stable it is. When semantic change takes place, the expansion of referential range denoted by a meaning is extended from the prototypical center to the peripheral area. Consequently, the peripheral area will have less and less in common with the prototypical center. It is also possible that a meaning of a lexical item is a combination of features that do not belong to the same cluster at all. Meanwhile, considering the uncertain boundaries to be drawn for a lexical item, its semantic history might involve discontinuous appearance of an identical meaning that is temporally unrelated to each other rather than resulting from textual evidences that do not survive time.

Under this conceptual framework, the flexibility of meaning construction relies on the adaptability and dynamics of human cognition that groups and regroups meanings to meet the need of cognitive efficiency. Building upon the distinction between speaker-oriented and hearer-oriented process to avoid possible communicative misunderstanding in phonology, this framework adopts a similar notion that homonymic clashes are resolved with opportunities of semantic change, including a tendency toward prototypicality and morphological transparency while striking a balance for as many morphological uses as possible.



For language speakers, the construction of meanings is flexible and sensitive to the context of use, in which ambiguity is resolved or cancelled (**miller1991contextual**; **harris1954distributional**). Additionally, the operation of metonymy is a mechanism that plays a practical role in meaning construction, for this mechanism allows a word to carry referential and conceptual meanings simultaneously (**hilpert2019historical**; **nerlich2001serial**). From the perspective of semantic change, an understanding of metonymic change, specifically, builds upon the familiarity of the culture in which the language is spoken, and therefore the attested examples in literature exhibit a rich diversity (**fortson2017approach**). Yet, the semantic history of a word might also unfold beyond the intuition of the language users. It is recognized that synchronically distinct meanings, which speakers of the given time period find conceptually related, might suggest otherwise, as in *bachelor*, for a relationship exists between “experiencing” and “evoking”, and *actually*, “unexpectedness” and “elaboration” (**traugott2001regularity**). On the other hand, synchronic convergence is also likely, as shown in instances of folk etymology, but not as common cross-linguistically.

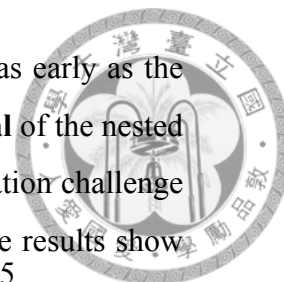
To measure semantic change quantitatively, frequency and collocational patterns allows for exploratory insights. If the word studied is one of the words with the highest frequencies, but stable, the establishment of a “collocational profile” for each character can be identified (**firth1957modes**).

Recently, the application of computation to larger sets of words across longer periods of time enables the generalization of regularities on semantic change (**hamilton2016law**). Semantic change driven by technological innovations are prominent examples, while shifts of meanings with linguistic cause tend to occur relatively more slowly (**hamilton2016law**). The changes encompass changes to “core meanings of words” or “subtle shifts of cultural associations” (**hamilton2016cultural**). The term “brachychrony” is even coined by **mair1998corpora renouf2002time** to refer to a time span of 10 to 30 years, indicating how the change of a linguistic feature can be delineated within a short time frame.

For Classical Chinese, **li2020evolution** used the dependency parser trained on Kyoto Corpus of the Four Books to explore change of syntactic categories of Classical Chinese, yet a character-based analysis is adopted due to the segmentation issue of pre-modern Chinese. However, contrary to the assertion that pre-modern Chinese is



mostly monosyllabic, the disyllabic development of Chinese has started as early as the Han dynasty (**zhou2009**; **chang2008**), but the proposal by **lee2012** of the nested multi-level segmentation is able to reflect the complicated word segmentation challenge for languages like (pre-modern) Chinese (**li2020**). However, the results show that tokenizers such as MeCab-Kanbun and Stanza segment words by characters, and verbs like 吃 ‘eat’ or 食 ‘eat’ might be tagged as noun. <sup>0.5</sup>



## 2.2 The concept of home in literature

The concept of home has been extensively studied in (environmental) psychology, sociology, anthropology, architecture, and other fields of study (**samanani2019house**; **mallett2004understanding**; **moore2000placing**; **sixsmith1986meaning**). Specialized topics on homelessness, journeying, migration, gender, and aging are also discussed. Previously, the meanings and concept of home are explored through questionnaires, interviews, and by examining quotes and literary works. When described using language, this concept becomes intertwined with such words as home, house, dwelling, and family, with these words used interchangeably (**mallett2004understanding**; **sixsmith1986meaning**). Nonetheless, home is “not only of belonging but also of potential alienation when attempts to make home fail or are subverted” (**samanani2019house**). The emphasized aspects of different word choices from literature can be summarized as follows:

1. House: physical space, reification of material circumstances and home concept organization through its layout, furnishings, renovation, and decoration (**samanani2019house**). For instance, Bourdieu compares how Kabyle people see the pair of light and dark to public and private, and asserts that a house “reflect[s] structured worldview” and “reproduce[s] it” (**samanani2019house**). Furthermore, materiality facilitates the development of a sense of belonging (**moore2000placing**).
2. Family: a structured social unit of living. A family is symbolic of marriage, kinship, togetherness, and homeliness (**samanani2019house**). A household is established through the process of homemaking, and the feeling of rootedness, safety, and value is thus deepened (**samanani2019house**; **moore2000placing**). On top of



that, marriage consolidates the concept of home through physical renovation and expansion of the house. From generation to generation, reproduction of class and gender differences is also strengthened or challenged (**samanani2019house**; **mallett2004understanding**).

0.5

The most detailed analysis is provided by **sixsmith1986meaning**. The co-existing relationships of home is plotted as three regions from questionnaire responses, as shown in Figure ?? (**sixsmith1986meaning**).

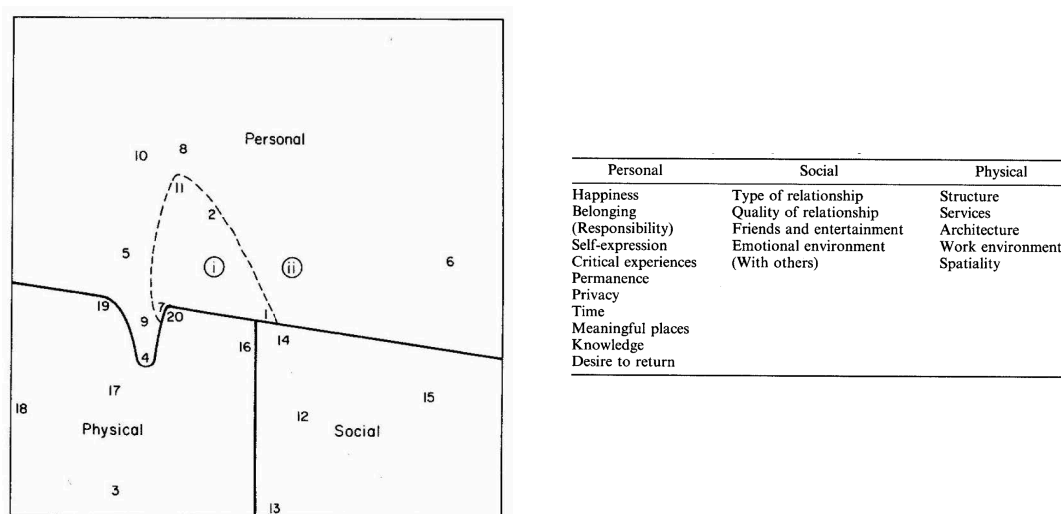


Figure 2.1. The concept of home split into 3 regions (“Personal”, “Physical”, and “Social”). The spatial distribution of the 20 categories are yielded from Kendall’s Tau correlation between the types and meanings of home defined by participants (Adopted from **sixsmith1986meaning**).

Culturally, the concept of home in Taiwan as a physical space has undergone changes caused by the sway of the world order (沈孟穎 **2015** 台灣現代住宅設計之轉化). Traditionally, *heyuan* houses are common architectural forms reflecting Chinese analogy of an abode to an extension of the human figure and Chinese cultures of calligraphy and sculpture. Later, influenced by Japanese power, Japanese-Western Eclectic style was introduced to Taiwan, and 街屋 *jie-wu* ‘street house’ transforms the architectural landscape by incorporating the commercial use into the residential function. This hybridization is embodied and preserved in places like Dihua Street and Dadaocheng Area.

Linguistically, **wang2005jia** have discussed the morphological development of *jiā* in pre-modern Chinese.

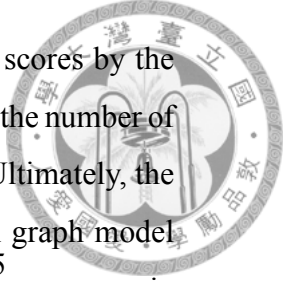


## 2.3 Diachronic word embeddings

Semantic change is a manifestation of language use in both conventional and creative ways by the language community, making textual data temporal-dependent in essence (kutuzov2018survey). As more attention is paid to the design of diachronic corpora and digitalization of historical text, a gap bridge and rapid advancements are seen in investigating semantic change in a data-driven way, especially from a distributional semantic perspective like diachronic word embeddings (kutuzov2018survey; tahmasebi2018survey; hamilton2016law; jawahar2019contextualized). Diachronic word embeddings make it possible to formulate or test hypotheses or laws of semantic change, establish temporal word analogy or relatedness, as well as discover semantic relations that are also changing over time. In hamilton2016cultural, linguistic drift and cultural shift can be also distinguished and measured based on diachronic word embeddings, with the latter restricted to a smaller set of neighboring words. With a growing interest in this research topic, insights have been made to highlight some key and challenging aspects of semantic change modeling (kutuzov2018survey; tahmasebi2018survey; camacho2018survey).

### 2.3.1 Topic-Over-Time

Besides vector space models, topic modelings are also widely applied to the study of semantic change, e.g., Topic-Over-Time TOT (wijaya2011understanding) and hengchen2017phd. In practice, probability distribution is computed for each word in the vocabulary of a specific time period, and word senses are derived from the topic distribution. For instance, the word *gay* used to act as an adjective meaning happiness or cheerfulness, yet it shifts to refer to homosexuality; the word *awful* comes to have less intensity and negativity in meaning, and the k-means clustering shows that words with the highest tf-idf scores do not belong to the same clusters, indicating that these words are diverse in meaning and the word *awful* is then used as an adverbial intensifier with general meaning (wijaya2011understanding); for the word *mouse*, by decreasing the *k* in k-means, two clusters can be merged, and the last cluster represents the additional meaning acquired with the word *mouse*. Therefore, Latent Dirichlet Allocation (LDA) or topic modeling accompanied by clustering method is insightful when we examine the topic



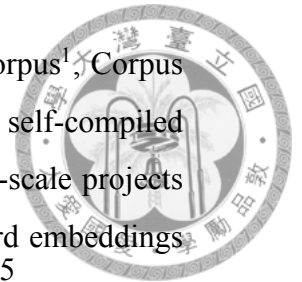
density of each cluster of a given time period, top words with the highest scores by the selected metric (i.e., tf-idf scores), merging of clusters by means of adjusting the number of clusters, as well as links between words from the co-occurrence network. Ultimately, the evolution of dynamic networks, specifically temporal exponential random graph model (ERGM)(**robins2007introduction**) is proposed to model the network of co-occurrence in a diachronic vein.

In contrast, topic models are also used to yield topics that are most common in a given time period in order to anchor words that should be evaluated for the results (**antoniak2018evaluating**). By so doing, the number of topics set for the identification of anchoring words are much larger than that for Topic-Over-Time (TOT) so that the computed mean probability is based on as diverse topics as possible.

### 2.3.2 Diachronic word embeddings

The topic of semantic change has directed attention to the design of corpus used as input for diachronic word embeddings. In Natural Language Processing, word embeddings are commonly added to the last layer of a deep learning model to translate discrete linguistic data to continuous numeric vectors. On the other, another line of research, referred to as “corpus-centered” approach, focuses on the use of word embeddings as evidence for certain linguistic features or cultural characteristics (**antoniak2018evaluating**). Unsupervised lexical semantic change detection refers to the task of tracing semantic change based on diachronic word embeddings trained on time-sliced textual data or (sub)corpora. The modeling rests on the assumption that change in meaning is captured if change in word co-occurrences is identified. One of the crucial steps is the collection of text and its temporal information in order to build word embeddings of different time epochs. Diachronic corpus is subject to the lack of certain documents that are difficult to survive time and thus missing, and hard to expand. The presence and absence of documents, along with a smaller or less balanced corpus, has called for techniques like bootstrapping to mitigate the issue of variability (**antoniak2018evaluating**). The division of time periods, or the granularity, is also decided in the meantime of corpora compilation. Typically, the more recent the text is created, the more refined or specific the time units are set (**kutuzov2018survey**). Among the diachronic textual data currently available,

the main source includes but not limited to the Google Books Ngrams Corpus<sup>1</sup>, Corpus of Historical American English (COHA)<sup>2</sup>, Project Gutenberg Corpus<sup>3</sup> and self-compiled corpora with text from newspapers and online social media. While large-scale projects have led to the release of various pre-trained word embeddings, new word embeddings continue to be trained to allow for more diversity and richness of the textual contents, and to adapt to specific research questions to be answered. This trend pertains to the definition of “diachronic”, which highlights the characteristics of the source data with long stretch of time, and even from a long time ago in history.



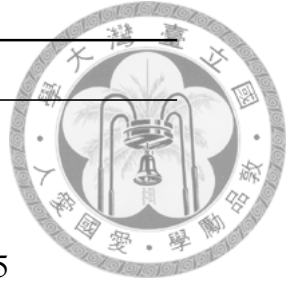
Regarding conversational diachronic corpus, (giulianelli2019lexical) uses the r/LiverpoolFC corpus, which contains 40 million words from posts on the English football team Liverpool from 2011 to 2017. Each utterance is annotated with a timestamp, and the dataset includes binary annotations of change on 100 selected words by 26 r/LiverpoolFC users themselves. The compilation of this corpus is based on sufficiently high temporal granularity, enabling detection of abrupt shifts, the language use of a specific community. However, it is non-uniformly distributed, and thus it is more difficult to study changes in some of the time periods when a few user posts are generated.

---

<sup>1</sup><http://books.google.com/ngrams>. A comprehensive review of diachronic corpora is provided by tahmasebi2018survey

<sup>2</sup><https://www.english-corpora.org/coha/>

<sup>3</sup><https://www.sketchengine.eu/project-gutenberg-corpus/>



Literature	Use cases	
<b>kulkarni2015statistically</b>	apple, tape	
<b>hamilton2016law</b>	gay, broadcast, awful*	
<b>hamilton2016cultural</b>	actually, must, promise, gay, virus, cell	0.5
<b>kutuzov2017tracing</b>	war, peace, stable	
<b>rodha2017panta</b>	πνεῦμα ‘breath’ → ‘spirit’ (Ancient Greek)	
<b>yao2018dynamic</b>	apple, amazon, obama, and trump	
<b>rudolph2018dynamic</b>	intelligence, iraq, jobs, prostitution	
<b>antoniak2018evaluating</b>	marijuana	
<b>hu2019diachronic</b>	please, alien	
<b>rodina2020elmo</b>	провальный ‘a place where the surface collapsed inward’ or ‘loss of consciousness’ → ‘failed’ (Russian)	

\*See also sense shift based on earlier literature with corpus data in **hamilton2016law**

Table 2.1. Example case studies from literature

### 2.3.3 Laws of semantic change

Diachronic word embeddings can be used to discover more possibilities of unknown change cases and underlying causes of general semantic change (**hamilton2016cultural**; **kutuzov2017tracing**; **heuser2017word**). In **hamilton2016cultural**, it is concluded that linguistically-driven semantic change occur more slowly than socially-motivated phenomenon. The invention of new technologies serves as prominent examples of cultural drift, as in *apple* and *cell*. **kutuzov2017tracing** exemplifies how social events such as armed conflicts are traced by monitoring word associations with “anchor words” like *war*, *peace*, and *stable*. Lists of words with the highest similarity scores or analogous pairs of words are analyzed to verify the results of diachronic word embeddings. In **hamilton2016cultural**, the results of linear regression shows that a local measure of this partial list is sufficient to account for the phenomenon of a cultural drift. Another example is how *president* becomes closer to *Obama* during his term, as well as *Israel’s Prime Minister* and *Christopher Nolan, The Dark Knight, 2008* (**rosin2017learning**) by finding

continuous peaks of lowest distance between vectors with dataset YAGO2<sup>4</sup> that contain temporal relations of named entities.

On top of that, based on the self-similarity scores of the English lexicon between 1850 and 2009, **dubossarsky2015bottom** find that lexical semantic change positively correlates with the centroid of a word's cluster, which is symbolic of the word's prototype, hence the "law of prototypicality." In **xu2015computational**, near-synonyms are shown to change in parallel, and thus the law of parallel change is more favorable than the law of differentiation. The law of conformity and innovation are put forward by **hamilton2016law**; the former posits that observed frequency positively correlates with the rate of semantic change, while the latter asserts that semantic change is positively influenced by a word's polysemy, the number of a word's senses, in controlled frequency. However, different conclusions exist given different experiment settings and source data, so no consensus has been reached regarding a wider generalization of semantic change in more languages building upon diachronic word embeddings.

Additionally, if time-specific embeddings are separately trained, the embeddings are randomly initialized, and it is necessary to align them in the same vector space (**hamilton2016law**). Thus, the alignment of embeddings leads to the comparability of cosine similarity scores of words from different time periods. To project separately trained word embeddings, linear transformation, distance-preserving projection, second-order embeddings that consist of vectors of word's similarities to all other words in the shared vocabulary of all models are used. The most widely adopted alignment algorithm is proposed by **hamilton2016law**, who utilizes second-order embeddings and orthogonal Procrustes transformations at the same time. Another line of research resorts to jointly learning word representations of all time periods by incrementally updating the model. Furthermore, the hierarchical softmax function is introduced to improve the efficiency of the updating.

In addition to alignment of separately trained embeddings, temporal referencing (TR) (**dubossarsky2019timeforchange**) is proposed to mitigate the noise issue induced by alignment. Because of alignmnet, the results, especially low-frequency words, are influenced by noises (**dubossarsky2019timeforchange**; **dubossarsky2019timeout**). However, the lack of widely-accepted evaluation procedures have made it difficult to learn

---

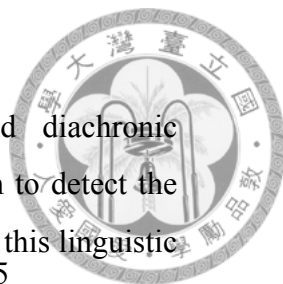
<sup>4</sup>The latest version is released in 2020.

more about the noises invited by vector space alignment (**dubossarsky2019timeout**).

Nonetheless, the scarcity of ground-truth test data has made it difficult to evaluate the employed approach. The rating-based and dictionary-based collection of evaluation data are met with low inter-rater agreement of recruited annotators and/or inaccessibility of sources from the time period of interest (**tang2018state**). **kutuzov2020uio** reveal that the results based on the test data can be distinctively varied across different languages. In contrast, evaluation datasets for Present-Day English are available, as well as translations and crowd-sourced human-annotated datasets in Mandarin Chinese. In downstream tasks, the importance of constructing temporal-aware embeddings as input data is acknowledged in the form of domain adaptation (**huang2019neural**). Temporal adaptation is introduced as a form of domain adaptation to diachronic word embeddings and proves effective in the task of document classification (**huang2019neural**).

Another challenge, namely the “meaning conflation deficiency”, is brought up by **camacho2018survey**. Previously, word embedding technique is first implemented by **mikolov2013efficient** in **mikolov2013efficient**. The embeddings models such as Continuous Bag-Of-Words (CBOW), Skip-gram with negative sampling (SGNS), Singular value decomposition on Positive Pointwise Mutual Information (SVD-based PPMI) are static, for only one vector is generated to represent each word type in the diachronic textual data. Word-level vector representations do not account for the context of the keyword. Therefore, two words are likely to move closer toward each other in vector space not necessarily because they become semantically closer, possibly because one of the words undergoes meaning change on the sense level. Due to the static nature of word embeddings, **hu2019diachronic** point out that the results do not show which sense has changed, and which remains stable, if not at a “coarse-grained” level. While static word embeddings rely on the analysis of neighboring words with the keyword to determine the presence or absence of meaning change, contextualized word embeddings mapped tokens to a possibly infinite sets of data points, allowing various methods to depict the subset of data. Pre-trained language models like ELMo and BERT are dynamic and contextualized. Multiple embeddings can be extracted to represent a word in various contexts, thus allowing different senses of a word to be distinguished. It is possible to produce mappings between contextualized word representations and sense descriptions from external linguistic resources (e.g. the Oxford English Dictionary)





(**hu2019diachronic**).

Notwithstanding, although context-independent and contextualized diachronic embeddings are proposed and explored in an increasing body of research to detect the presence of semantic change, which models are more capable of capturing this linguistic phenomenon remains an on-going topic that calls for evaluation methods for diachronic embeddings. It is debatable whether simpler models results in better performance (**schlechtweg2019wind**). Firstly, datasets like DUREl (Diachronic Usage Relatedness)<sup>5</sup> are established based on human ratings (**schlechtweg2018diachronic**) and word injection (**schlechtweg2019wind**), which is based on similar concepts like domain-specific word sense disambiguation or term ambiguity detection, inspired by term extraction and synchronic version of SUREl (Synchronic Usage Relatedness)<sup>6</sup> where variation lies in sense divergence across domains for research topics like online language analysis. However, evaluation data are scarce (**wevers2020digital**), hand-picked attested examples from literature or dictionaries with tags like “obsolete” (**hamilton2016cultural**) have proven that automatic semantic change detection is able to capture semantic change (See Table ??) (**schlechtweg2019wind**), but results still vary depending on test or evaluation data that are currently available. The result of **schlechtweg2019wind** shows that SGNS with orthogonal Procrustes alignment achieves the highest performance based on the DUREl dataset, whereas topic modeling has the least correlation with the examined dataset. Furthermore, the results in **schlechtweg2019wind**; **dubossarsky2017outta** shows that cosine distance (global neighborhood distance) outperforms local neighborhood distance under the condition of aligned embeddings, and the results of topic modeling is sensitive to corpus size and frequency of the target words, which make it a less desirable method in this study, as pre-modern Chinese texts might not reflect accurate counts of types and tokens.

Instead of sense inventories, various clustering algorithms are resorted to induce senses of target words, including K-Means, Gaussian mixture models (**giulianelli2019lexical**).

In comparison with other approaches of semantic change detection, diachronic word embeddings exhibit a stronger explanatory power than frequency-based methodologies such as raw and relative frequency counts, collocational analysis (**kutuzov2018survey**). Indeed, it is convenient to manipulate word vectors, but past literature also presents the

<sup>5</sup><https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/durel/>

<sup>6</sup><https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/surel/>

results and analysis in combination of the above two or more approaches to generalize the underlying principles of semantic change or echo with the proposed linguistic hypotheses (tahmasebi2018survey).

0.5



### 2.3.4 Historical/diachronic corpora

The compilation of corpora to include historical texts and annotations enables more detailed linguistic analysis. Examples include the Corpus of Historical American English (COHA, 1810-2000)<sup>7</sup>, A Representative Corpus of Historical English Registers (ARCHER, 1600-1999)<sup>8</sup> Royal Society Corpus (RSC, 1665-1996)<sup>9</sup>, Corpus of Late Modern English Texts (CLMET, 1710-1920)<sup>10</sup>, Hansard Corpus (1803-2005)<sup>11</sup>, among many others.

In Chinese, the number of diachronic corpora is relatively scarce, including Sheffield Corpus of Chinese<sup>12</sup> and Academia Sinica Ancient Chinese Corpus (中央研究院古漢語語料庫, hereafter ASAC Corpus)<sup>13</sup> (**wei1997corpus**). The ASAC Corpus is divided into 3 sub-corpora based on the development of Chinese syntax, namely Old Chinese subcorpus (上古 from pre-Qing to pre-Han), Middle Chinese subcorpus (中古 from Late Han to the Six Dynasties), and Early Mandarin Chinese subcorpus (近代 from Tang to Qing) to offer a synchronic sketch and a basis for diachronic comparisons. In the Academia Sinica Tagged Corpus of Early Mandarin Chinese (中央研究院近代漢語語料庫), raw texts are available from the Western Han dynasty to the Pre-Qing dynasty, with part of the texts imported from Scripta Sinica (漢籍全文資料庫計畫). It is believed that corpora creation is the foundation for a more thorough and accurate depiction for data collection during the establishment of lexical databases.

<sup>7</sup><https://www.english-corpora.org/coha/>

<sup>8</sup><https://www.projects.alc.manchester.ac.uk/archer/>

<sup>9</sup><https://fedora.clarin-d.uni-saarland.de/rsc/>

<sup>10</sup><https://perswww.kuleuven.be/~u0044428/>

<sup>11</sup><https://www.english-corpora.org/hansard/>

<sup>12</sup><https://www.dhi.ac.uk/scc/>

<sup>13</sup><http://lingcorpus.iis.sinica.edu.tw/early/>



## 2.4 Visualizing semantic change

In view of the scale of data, semantic change modeling is evaluated on two grounds—the combination of statistical testing and visualizations, as well as classification tasks (**tang2018state**). In addition to the exploration of linear relationships such as word analogies, high-dimensional visualization techniques are employed to assess the results of word representation learning (**liu2017visual**). Visualization of diachronic data allows researchers to explore any target word to see how the data changes along with time.

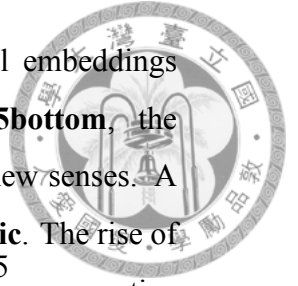
To visualize the results, vectors originally trained in high-dimensional space are transformed and projected in two or three dimensions. Principal Component Analysis (PCA) and t-distributed Stochastic Neighboring Embedding (t-SNE) (**vandermaaten2008tsne**) are two common methods of dimensionality reduction. Only the most influential dimensions are retained using the former approach, while the latter reflects more geometrical structure of the high-dimensional data. However, the exploration of the internal structure and properties of an embedding is generally non-interactive (**smilkov2016projector**). In **smilkov2016projector**, Google releases the Embedding Projector under the TensorBoard framework, which provides users with many interactive functionalities such as zooming, filtering, inspection of data points with metadata created in the table format by users (**smilkov2016projector**).

**coenen2019visualizing** recognizes the adaptability of BERT to various downstream tasks and the possibility of the language model to extract useful features from raw textual data. To understand the internal structure of BERT and how discrete linguistic units are translated into continuous numeric vectors, **coenen2019visualizing** use UMAP visualization of the token vectors and nearest-neighbor classifier. Semantically, fine-grained sense information is encoded in BERT, even in low-dimensional subspace. **coenen2019visualizing** conclude that both semantic and syntactic information are encoded in the contextualized embeddings in “complementary subspaces.” Yet, an attention-based model like BERT does not necessarily “respect semantic boundaries when attending to neighboring tokens, but rather indiscriminately absorb meaning from all neighbors.” (**coenen2019visualizing**)

It is summarized in **tang2018state** that the novelty of a sense can be understood as the change in sense distribution of different time intervals. The diachronic sense

distribution can be visualized based on both word-level and sense-level embeddings (**dubossarsky2015bottom**; **hu2019diachronic**). In **dubossarsky2015bottom**, the distance of a word's centroid is pinpointed to find out the emergence of new senses. A trajectory of sense evolution is graphically represented in **hu2019diachronic**. The rise of a new sense can be depicted in company with other senses in a competitive or cooperative relationship. Also (**gonen2020simple**).

However, the division of time periods, or the granularity, examined in previous studies, especially those on laws of semantic change, is restricted to the nineteenth century onward. Additionally, to trace semantic change of pre-modern Chinese, we need to account for the disyllabic development of words. Therefore, we aim to analyze both pre-modern and modern Chinese texts, which would be the first attempt to apply both computational and statistical models to explore the interplay between disyllabic development of words and semantic change in Chinese.





0.5

## Chapter 3

# Methods

### 3.1 Data collection and preprocessing

As early as the year of **sinclair1982reflections**, **sinclair1982reflections** already envisioned the possibility of having “vast, slowly changing stores of text” that provide “detailed evidence of language evolution” (**renouf2002time**). Since then, the importance of digitally storing both historical and modern textual data has been widely recognized in the study of corpus linguistics (**renouf2002time**). As **renouf2002time** mentioned, “we need the past in order to understand the present. An amalgamation would increase the scope, timespan and continuity of resources, whilst lessening the inconvenience of having to switch from one corpus and set of tools to another.” Among the existing corpora, written texts comprise a major portion of the corpus compilation efforts, and thus it is a turning point to explore the diachrony of the data along with more recently available texts from historical periods.

To construct a diachronic corpus in this study, texts of pre-modern and modern Chinese are collected from the Chinese Text Project (中國哲學書電子計畫, hereinafter CTEXT) (**sturgeon2019ctext**)<sup>1</sup> and Academia Sinica Balanced Corpus of Modern Chinese (中研院漢語平衡語料庫, hereinafter ASBC) (**chen1996sinica**)<sup>2</sup> respectively. The data from the aforementioned sources are sequential in time and large in size, which allows for a diachronic view of how the concept of home evolves.

---

<sup>1</sup><https://ctext.org/>

<sup>2</sup><http://asbc.iis.sinica.edu.tw/>

Firstly, the Chinese Text Project is an open-access digital library that collects pre-modern Chinese texts with time spanning from 1046 B.C. of the Western Zhou dynasty to 1949 A.C. of the Republican era (**sturgeon2019ctext**). Since the number of texts available from each era varies, the time periods with the highest number of texts, namely the Tang, Song, Yuan, Ming, and Qing dynasties, are included to construct the sub-corpora of pre-modern Chinese in this study. The texts and their metadata are retrieved from the Chinese Text Project (CTEXT) digital library using `ctext`<sup>3</sup>, a Python API (Application Programming Interface) wrapper of the same name developed by **ctextapi**.

Apart from the provision of the API access, the CTEXT project website is informative of how textual data and metadata are stored in the retrieved format<sup>4</sup>. Since the original prints are scanned and converted into the machine-readable format using the OCR (Optical Character Recognition) techniques, multiple versions of a text are likely to be produced through the employment of different OCR techniques, only one version representative of a set of texts is selected<sup>5</sup>, or, if needed, all versions are retained to help discern the differences in the converted texts. For example, to obtain frequencies of characters in different time periods, it is necessary to exclude duplicate counts, while the differences are kept intact during the training of word embeddings. On the document level, the corpus composition is summarized in Table ??.

---

<sup>3</sup><https://pypi.org/project/ctext/>

<sup>4</sup><https://ctext.org/instructions/wiki-formatting>

<sup>5</sup>Among a set of documents, the version labeled with the tags “TEXTDB” (the texts are selected in the main library), “WORKSET” (the texts are specified as representative of a group of documents), “OCR\_CORRECTED” (the texts are corrected through the community efforts), “OCR\_MATCHED” (the texts can be referenced to the part of the scanned page) in the metadata is treated as representative according to the instructions on the CTEXT project website. In the case where no tags are provided, the version with the largest file size is selected.

Time span (A.C.)	Number of texts	Number of unique texts
618 – 907 (Tang)	956	623
960 – 1279 (Song)	2,998	2,145
1271 – 1368 (Yuan)	991	742
1368 – 1644 (Ming)	4,248	3,497
1644 – 1911 (Qing)	9,669	7,719
Total	18,862	14,726

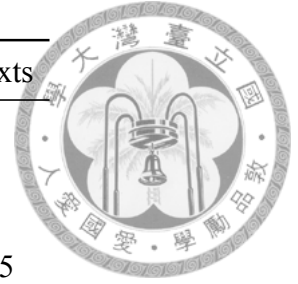


Table 3.1. Data composition of the CTEXT corpus

The source of textual data for modern Chinese is Academia Sinica Balanced Corpus of Modern Chinese (ASBC). The Academia Sinica Balanced Corpus of Modern Chinese (ASBC) contains articles from the year of 1981 to 2007. The corpus is well-balanced across genres and carefully segmented and POS tagged, which is considered representative of the language use of modern Chinese. Therefore, the choice of CTEXT and ASBC suits the language settings for this study.

The cleaning task is proceeded as described below:

- (1) The raw text is cleaned by (a) removing commentaries and marginal notes, (b) segmenting the text into two levels of chunks to indicate possible sentence and word/phrase boundaries according to the list of punctuations in the Instructions, and (c) extracting Chinese characters encoded in Unicode.
- (2) Chinese words are not delimited by space, nor is a conventional punctuation system adopted in pre-modern Chinese text. As a consequence, the punctuations should be viewed as symbols to mark 句讀 *jùdòu* ‘pauses or breaks’. Only the symbols specified in the website’s instructions are treated as indications of sentence boundaries, namely the newlines, full-width periods (。), and vertical bars (|). During the preprocessing, the set of punctuation marks used for phrase-level segmentation include the CJK Symbols and Punctuations, their half-width counterparts, and variants listed in the Unicode Standard <sup>67</sup>.

<sup>67</sup><https://unicode.org/charts/PDF/U3000.pdf>

<sup>7</sup>While the texts are in the units of characters in this study, dependency parsers for classical Chinese include UD-Kanbun (**yasuoka2019universal**) (<https://pypi.org/project/udkanbun/>) and Stanza in StandfordNLP (**qi2020stanza**) (<https://stanfordnlp.github.io/stanza/>).

- (3) To extract Chinese characters, Unicode range between U+4e00 and U+9fff are retained for basic Chinese characters, and variants or rare characters are captured from the Unicode range of CJK Extensions (**moran2018unicode**).

0.5

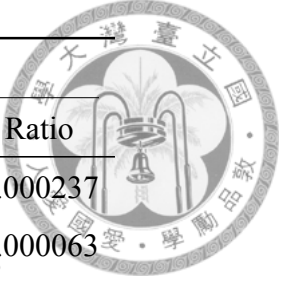


- (4) Text surrounded by quotation marks indicates conversations, sayings, or allusions, and is not removed during the preprocessing. On one hand, conversations are an integral part of the text; on the other, sayings and allusions reveal what is still in use or understandable in the time period of their appearance.
- (5) One of the difficulties in processing pre-modern Chinese lies in the word segmentation issue. This is particularly troublesome given the disyllabic development of Chinese. The overview of type and token counts of texts from the time-sliced corpora is as Table ?? and Table ??.

Corpus	Time span (A.C.)	All texts		
		Tokens	Types	Ratio
CTEXT	Tang	104,885,709	12,301	0.000117
	Song	449,371,130	17,219	0.000038
	Yuan	104,568,204	11,926	0.000114
	Ming	714,954,827	17,098	0.000024
	Qing	1,610,859,963	29,189	0.000018
ASBC	1981 – 2007	15,004,528	6,954	0.000463
ASBC (segmented)		8,934,360	66,021	0.007390

Table 3.2. Token and type counts of the diachronic corpora





Corpus	Time span (A.C.)	Selected texts		
		Tokens	Types	Ratio
CTEXT	Tang	48,701,732	11,549	0.000237
	Song	259,441,083	16,279	0.000063
	Yuan	59,572,917	11,336	0.000190
	Ming	517,074,764	16,657	0.000032
	Qing	1,137,949,237	21,878	0.000019
ASBC	1981 – 2007	NA	NA	NA
ASBC (segmented)		NA	NA	NA

Table 3.3. Token and type counts of the diachronic corpora

After the completion of preprocessing, this study proceeds to a preliminary exploratory data analysis with the bootstrapping method proposed by **lijffijt2016bootstrap** to reduce the influence of uneven distribution of linguistic features in texts and provide a more solid ground for the quantitative analysis.

The bootstrapping method is a process of multiple resampling in which a random sample of texts from a corpus is taken and placed back to the pool in a repetitive manner. In each resampling cycle, the value of the statistic of interest is noted and further generalized. The bootstrap test proposed by **lijffijt2016bootstrap** is as below.

$$p = \frac{\sum_{i=1}^N H\left(\text{freq}(q, T^i) - \text{freq}(q, S^i)\right)}{N}, \quad (3.1)$$

$$\text{where } H(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0.5 & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$p_{two} = 2 \times \min(p, 1 - p) \quad (3.2)$$

$$p^* = \frac{p_{two} \times N + 1}{N + 1} \quad (3.3)$$

In contrast to the bootstrapping method, tests like chi-square and log-likelihood ratio tests are “based on the assumption that all samples are statistically independent

of each other” (lijffijt2016bootstrap), yet words within a text are not independent in nature, and thus lijffijt2016bootstrap proposes to apply tests like Mann-Whitney U-test or bootstrapping methods to compare difference in word frequency. In terms of the assumption on independence, this relation exists at the level of texts rather than individual words using the bootstrapping method. Additionally, the bootstrap provides a more conservative  $p$ -value than those by bag-of-words-based methods, while the use of higher cut-off values in the chi-square or log-likelihood ratio tests do not correct the bias.

To understand the frequency distribution of characters in a diachronic view, the bootstrapping test is performed with 1K samples of 50 texts from the 500 texts of the Tang and Qing dynasties. The results are shown in Figure ??.

As Figure ?? shows, regarding frequency change for characters in this study, the trend is mostly a flat line. Characters with drastic change in observed frequency tend to belong to rare or historical characters. Additionally, among the 22 981 characters that have appeared in at least one dynasty, 12 233 characters are seen in both the Tang and Qing dynasty, and 404 of them receive a  $p$ -value at less than .05. In other words, 3.30 % of the chracters in use between the Tang and Qing dynasties change in their observed frequency.

Specifically, although the relative frequency of *jiā* slightly increases from 1260.92 to 1609.15 (The raw frequencies are 61 420 and 1 831 222), the difference in the use of the character is not statistically significant:  $p=0.5404595$ , 1k samples. Consequently, the use of *jiā* does not change in frequency, and is regarded as being stable in use.

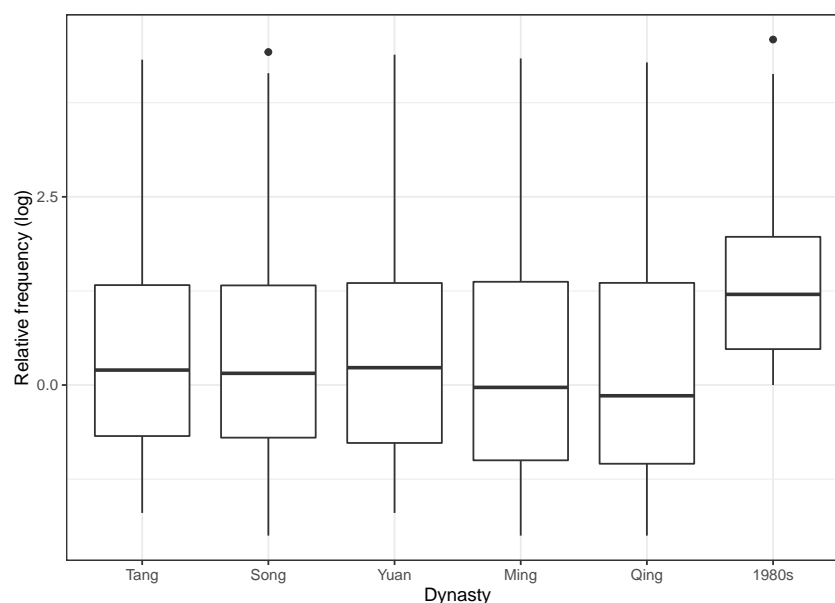


Figure 3.1. Frequency distributions of characters from the Tang dynasty to the 1980s

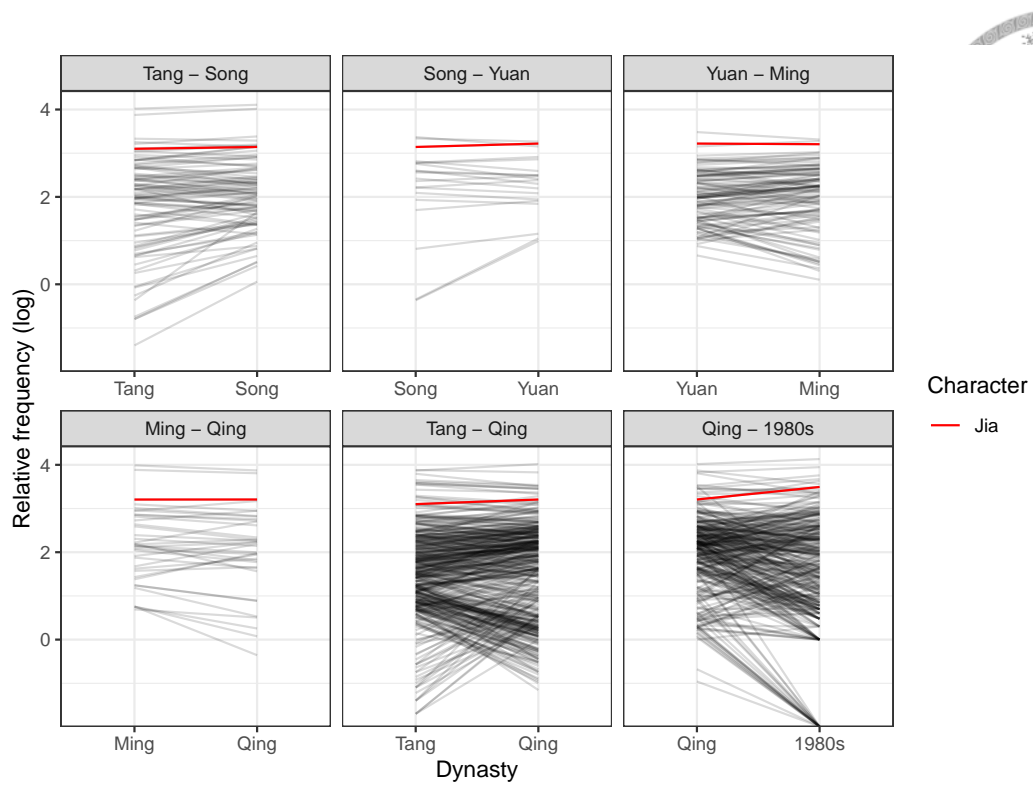


Figure 3.2. Frequency change derived from the bootstrapping test on characters between the Tang and Qing dynasty

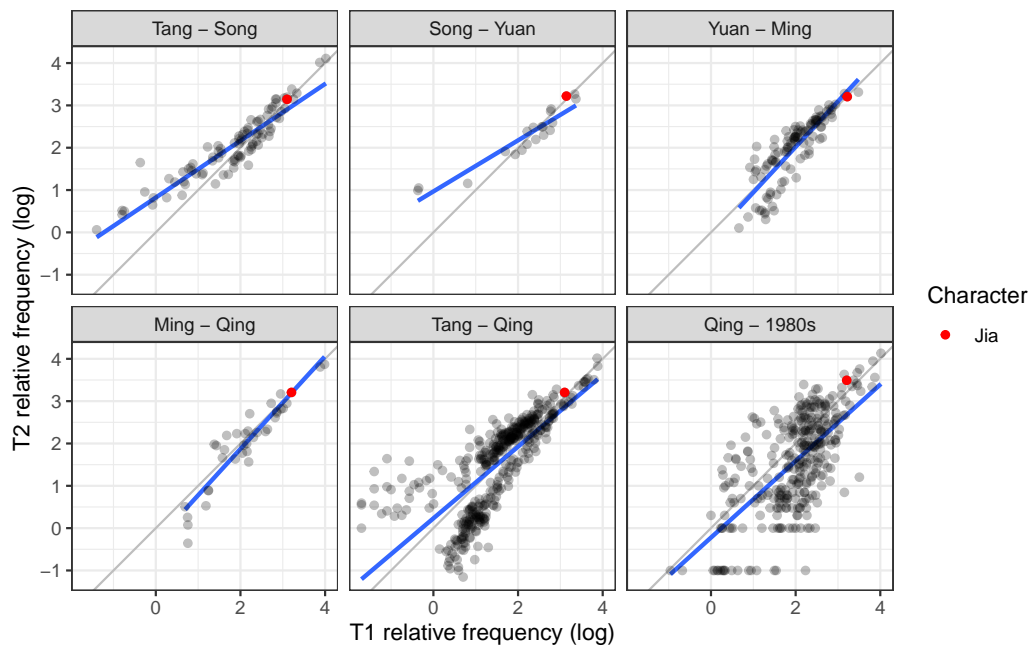
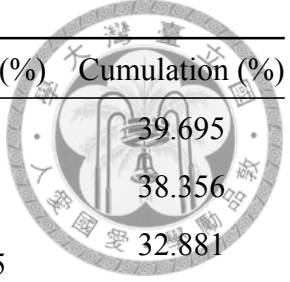


Figure 3.3. Frequency change derived from the bootstrapping test on characters between the Tang and Qing dynasty



Time period	Rank	Absolute frequency	Relative frequency	Percentage (%)	Cumulation (%)
Tang	139	61,420	1,260	0.129	39.695
Song	118	359,761	1,389	0.142	38.356
Yuan	91	98,883	1,659	0.170 <sub>0.5</sub>	32.881
Ming	87	830,135	1,605	0.163	29.568
Qing	92	1,831,222	1,609	0.163	29.395
1980s	41	46,661	3,110	0.311	25.551

Table 3.4. Frequency information of *jiā* from the Tang dynasty to the 1980s

For frequency information from other sources, see Appendix ??.

To investigate the semantic change of *jiā*, both word-level and sense-level analyses are employed.

## 3.2 Word-level embeddings

To learn what observations are supported by linguistic data in the three sub-corpora, embeddings are generated with Word2Vec in the Python gensim package, and the linguistic data from different time periods are separately trained. Additionally, as suggested by **li2019word**, character-based methods are likely to produce a more desirable results than word-based ones at some times, especially when the input data are “vulnerable to the presence of out-of-vocabulary (OOV) words,” and the words will thus be removed or left out from the subsequent computing process. To address the problem arising from word segmentation, character-based word embeddings are also generated for texts from pre-modern time, with the hyperparameter of window size set to 1 for both the precontext and postcontext. The choice of an immediate vicinity reflects the uni-syllabification of pre-modern Chinese. However, it is not to conclude that word segmentation is unnecessary, but that alternatives exist. It is also worth noting that not all word tokens are retained from the sources, as indicated by the percentage in parenthesis of the table. In this study, words of which frequency is lower than 5 are filtered out and not used for word embeddings. In addition, because unlike English, words are not separated with space in Chinese, the prediction capabilities of word embeddings can be hindered by the properties of each language. That is also likely to be the reason for which the number of word tokens

are far higher in the CTEXT sub-corpus than that of the other two sub-corpora.

In terms of separately trained word vectors, vector alignment is based on Procrustes analysis by **hamilton2016law**<sup>8</sup>. After the training of Word2Vec embeddings, embeddings are imported to TensorBoard to visualize the data points (**smilkov2016projector**), and further analyzed in the discussion section.

In addition to the word embeddings trained on the whole corpus, a bootstrapping without replacement approach is adopted (**antoniak2018evaluating**). While the fixed model indicates the baseline, algorithmic variability, i.g., random initiations, random negative sampling, random subsampling of tokens in documents (**antoniak2018evaluating**). Following **antoniak2018evaluating**, for each time period, 50 iterations are performed. For each iteration of resampling, a model is built on the  $N$  randomly selected documents ( $N = 150$  for pre-modern documents and  $N = 0.2$  of the documents in ASBC) in contiguous sequence. An ensemble of embeddings are generated with the results averaged over the bootstrap samples.

To evaluate the stability of the bootstrap samples, 20 query words are selected. Firstly, in each time-specific corpus, 100 most frequent words serve as candidate words. The selection of the 20 query words is determined by the results of the LDA modeling with 200 topics and words with the highest mean probabilities across all topics, so the query words can be regarded as words that are general in the given time period. In addition, the bootstrapping is carried out along with the calculation of cosine similarity scores between the query words and the other words to look for a tipping point of stabilization, which results in a bootstrapped model of word embeddings. We then average over the bootstrap samples to yield more reliable results in this study. 20 nearest neighbors are selected from the fixed settings.

Before the degree of semantic change is measured, a filtering of mid-frequency characters is conducted, for highly frequent characters are not “content-bearing” (**hamilton2016cultural**; **rodda2017panta**). Afterwards, the similarity of semantic vectors across time periods is compared using correlations; namely the similarity between T2 (the time period of interest) and T1 (the previous time period). Besides computing on the original vectors, alternatively called first-order embeddings, we resort to second-order embeddings composed of a full or partial list of neighboring words to the keyword.

<sup>8</sup><https://github.com/williamleif/histwords>

Specifically, the top 25<sup>9</sup> shared neighbors in the rank order of T2 are selected to form second-order local embeddings, which are said to capture swift word usage change as a consequence of cultural change in **hamilton2016cultural**.

0.5



### 3.3 Sense-level embeddings

In addition to character-level embeddings, contextualized embeddings are extracted to retrieve sense-level representations based on the diachronic corpus in this study. The sense-level representations are described as “sense representations” in **hu2019diachronic** and “usage representations” in **giulianelli2019lexical**, for the pre-trained language model allows for the extraction of a possibly infinite number of embeddings depending on the context of the input, and the embeddings reflect the authentic language use and distinguishes the usages in group to simulate the sense distribution. The chosen pre-trained language model is bert-base-chinese (**devlin2018bert**) with HuggingFace’s PyTorch Transformer framework, which is a Transformer architecture with 12 layers, 768 hidden units, 12 heads, and 110M parameters, and is trained on both Traditional and Simplified Chinese text from Wikipedia and BookCorpus with masked training and next sentence prediction task. Conventionally, the final or last 4 hidden layers are used as the token embeddings, which is followed by the averaging of multiple embeddings of a target word, yielding a 768-dimensional vector to represent the target word being studied. For senses with multiple example sentences, the corresponding sense representations are an aggregated vector.

Regarding degrees of semantic change, global and local measures are applied with different indices such as correlation and Jensen – Shannon divergence. The lower the score, the higher the degree of semantic change (**hamilton2016law**). Jensen – Shannon divergence is used in **giulianelli2019lexical**. Time is not identified when the token representations are extracted.

---

<sup>9</sup>In **hamilton2016cultural**, the range between 10 and 50 is recommended as their results reflect.

### 3.4 The variability-based neighbor clustering method (VNC)



To begin with, word-level analysis is performed using the Variability-based neighbor clustering (VNC) method (**gries2012variability**) and the Word2Vec algorithm (**mikolov2013efficient**). Proposed by **gries2012variability**, the VNC method is used to divide the development of a linguistic phenomenon into sequential periods based on the input data of each time span. Previous techniques like cluster analysis and principal component/factor analysis do not take the temporal ordering of data into consideration. As a hierarchical agglomerative clustering method, data points that are similar, homogeneous and temporally adjacent are grouped together. In other words, the variability between temporally continuous data points determines whether they are put in groups or not. The resulting groupings can be graphically represented with a dendrogram and further analyzed.

If the data is sparsely distributed, the VNC method can be applied prior to data analysis. The VNC method can also be conducted and repeated to remove noise by finding out anomaly clusters that are not merged with other subgroups, and therefore minimize the influence of the outliers. For example, if a year-by-year dataset is available to study the decline of a linguistic phenomenon, and the VNC periodization method reveals a number of one-year clusters, they are the anomalies and can be excluded from subsequent analyses.

The choice of amalgamation rules includes two common similarity measures, namely standard deviations and Euclidean distance. Typically, the former is applied to numerical data, and the latter is suited for vector data, which makes the VNC method especially useful even if a linguistic phenomenon does not change in frequency, but in other distributional ways. In addition, the merging of two neighboring time periods is based on the chosen amalgamation rule such as the average of values.

In this study, the distributional approach is based on the quantitative information of word co-occurrences drawn from the time-sliced sub-corpora. Association measures are applied to quantify the strength of word co-occurrences, or the “collocability” of words studied (**gablasova2017collocations**). Particularly, the LogDice score is standardized and scaled, and thus comparable across corpora (**rychly2008lexicographer**; **gablasova2017collocations**). To construct the vector data of the keyword *jiā* for each

time slice, the frequency of the keyword and its collograms, the unigrams before and after the keyword (**gablasova2017collocations**), are first calculated, and the LogDice score of each collogram is then computed. Collograms that do not appear consecutively across all time slices are filtering out, and the LogDice scores of the shared collograms form a vector per time slice. Eventually, the LogDice vectors of all time slices is structured as a matrix. Two matrices are prepared for cases where collograms occur before and after the keyword, as well as another one regardless of the position of the collograms. Building upon the matrices, the VNC method is performed and the dendrogram is plotted using the R script offered on the Lancaster Stats Tools Online (**brezina2018statistics**)<sup>10</sup>.



---

<sup>10</sup><http://corpora.lancs.ac.uk/stats/toolbox.php>





0.5

## Chapter 4

# Results and Discussion

Diachronic embeddings, which is trained for the purpose of tracing the change of word representations in vector space models, are met with challenges in how the training is evaluated. In this study, the trained embeddings are first examined in interactive interface in order to explore the structure of the diachronic embeddings. Furthermore, analogical reasoning and bootstrapping methods are employed as an attempt to pinpoint the properties of embeddings that might be influenced by the source data. From this perspective, the “bias” in an embedding is interpreted as a “feature”, not a “bug” (**wevers2020digital**).

## 4.1 Evaluation on analogical reasoning

Analogical thinking and context-dependent evidence lay a cognitive ground for the studies of semantic change (**traugott2017semantic**). The training of word embeddings are evaluated based on intrinsic and extrinsic evaluations. In terms of vector space models, analogical thinking is associated with the directionality of vectors that represent words in pairs or in groups. While tasks like similarity scoring and analogical reasoning belong to types of intrinsic evaluation methods, the analogical reasoning is more adaptable to historical data in this study, for it is criticized that evaluation datasets mainly consists of geographical entities that would be non-existent in historical time periods (**wevers2020digital**; **li2018analogical**). Despite its popularity, wide application, and the much effort into the expansion of datasets, the analogical reasoning task is not adaptable for diachronic or historical word embeddings (**wevers2020digital**).

The CA8 dataset<sup>1</sup>, created by **li2018analogical**, is adopted to extract semantic relations, specifically analogies, in the trained diachronic character-level embeddings. While a variety of datasets and translated versions are available for the purpose of analogical reasoning, the CA8 dataset is characteristic of its attempt to not rely heavily on geographical names and proper nouns in the target analogical pairs. On the contrary, 8 relational types are included. Additionally, among the 1,307 analogical pairs in the type “nature,” 282 of them are single-character word pairs (or 1-gram, as categorized by **li2018analogical**), and the semantic relations are rich and elemental, including “number, time, animal, plant, body, physics, weather, reverse, color” (**li2018analogical**). It is the two reasons that enable the possibility to extract the semantic relations in pre-modern Chinese texts.

$$b' = \arg \max_{d \in V} (\cos(b', b - a + a'), \quad (4.1)$$

$$\text{where } \cos(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|}$$

$$\arg \max_{b' \in V} \frac{\cos(b', b) \cos(b', a')}{\cos(b', a) + \varepsilon}, \varepsilon = 0.001 \quad (4.2)$$

By solving the pair-based 3CosAdd and 3CosMul objectives (**levy2014linguistic**), it is found that 26 and 35 pairs are consistently identified across all time periods within smaller (window size set to 1) and larger (window size set to 5) window sizes. For example, pairs like 東-西: 左-右 ‘east-west:left-right’, 真-假: 左-右 ‘real-fake: left-right’, and 冷-熱: 南-北 ‘cold-hot:south-north’ are solved in all time periods, and the pair 冰-水: 雪-雨 ‘ice-water:snow-rain’ is also stably analogous except in 1980s. However, it has not yet been feasible to extract semantic relations with set-based objectives like 3CosAvg, for the mean of a set of vectors from the source and target single-character words under the same category defined in the dataset do not yield more analogical pairs in this study.

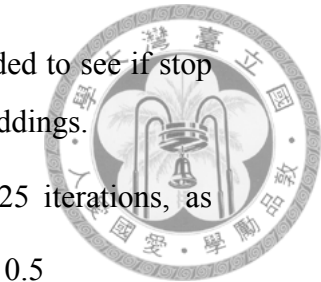
## 4.2 Stability of BOOTSTRAP diachronic embeddings

The first five common query words are ‘公’, ‘君’, ‘國’, ‘太’, ‘官’ for pre-modern Chinese, and ‘二’, ‘官方’, ‘發生’, ‘兼’, ‘且’. While some query words like ‘兼’ and ‘且’

<sup>1</sup> <https://github.com/Embedding/Chinese-Word-Vectors>

might be considered stop words and otherwise removed, they are included to see if stop words are asserting more impact on the stability of BOOTSTRAP embeddings.

The results show that the bootstrap samples become stable after 25 iterations, as suggested in **antoniak2018evaluating**.



0.5

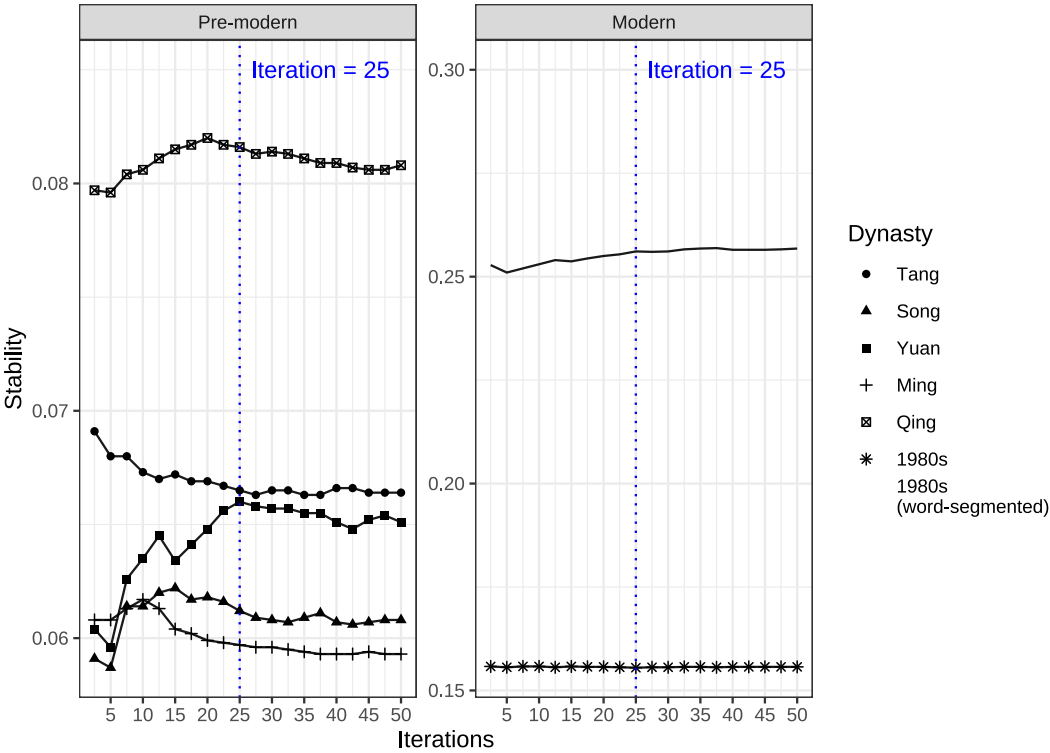


Figure 4.1. Mean stability over iterations based on query words extracted from LDA topic models and 20 nearest neighbors from fixed embeddings

For the stability plot of individual query terms, see Appendix B.

### 4.3 Collocational-based approach

The results of the VNC periodization are plotted as dendrograms (See Figure ??, Figure ??, and Figure ??)

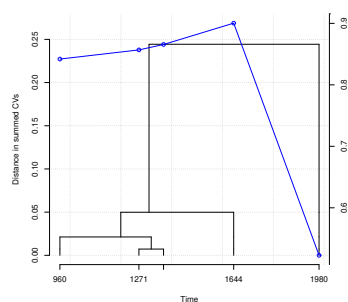


Figure 4.2. VNC periodization of collograms occurring before *jiā*

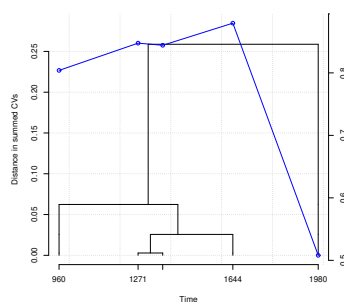


Figure 4.3. VNC periodization of collograms occurring after *jiā*

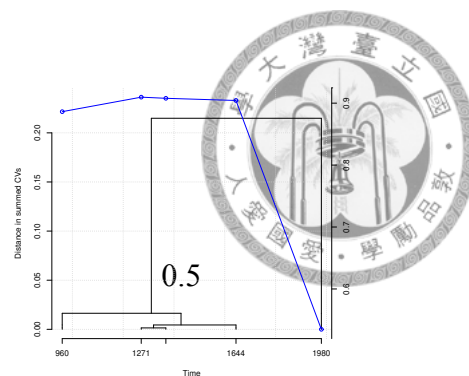


Figure 4.4. VNC periodization of collograms occurring with *jiā*

The correlation between the Qing dynasty and 1980s shows a drastically decreasing trend compared to that of its predecessor, the Ming dynasty and the Qing dynasty, marking a distinct new stage of development. Furthermore, the flattening of the line at 2 clusters in the scree plot suggests no subgroups are identified. It is generalized from the results of the VNC method that while modern Chinese is drastically different from pre-modern Chinese, the timeframe from the Tang dynasty to the Qing dynasty shows that each dynasty is dissimilar from one another and cannot be merged, even for the shortest dynasty Yuan. The granularity of diachronic data is not equally partitioned.



## 4.4 Diachronic word embeddings

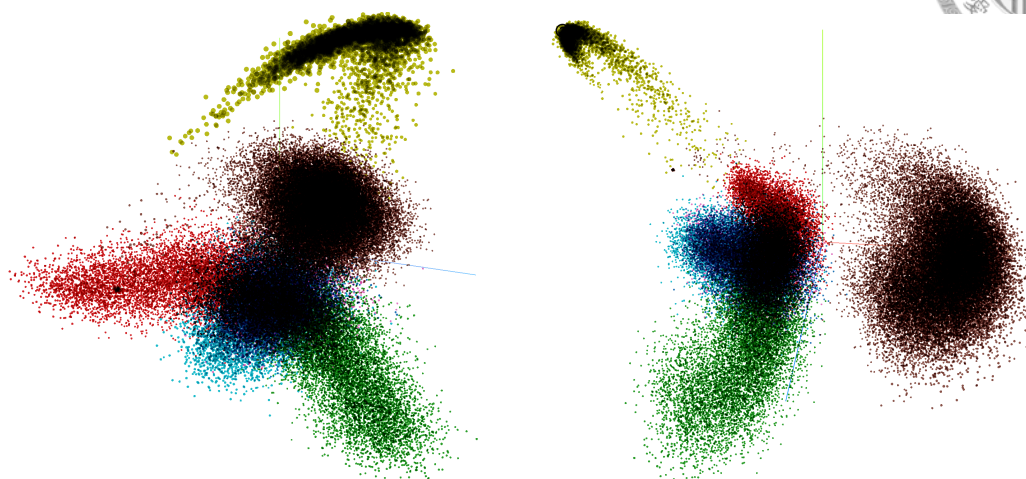


Figure 4.5. Snapshot of PCA Embedding Projector in TensorBoard

\* Total variance described: 34.6%.

Tang (dark blue); Song (red); Yuan (pink); Ming (sky blue); Qing (green); 1980s (brown); 2010s (mustard).

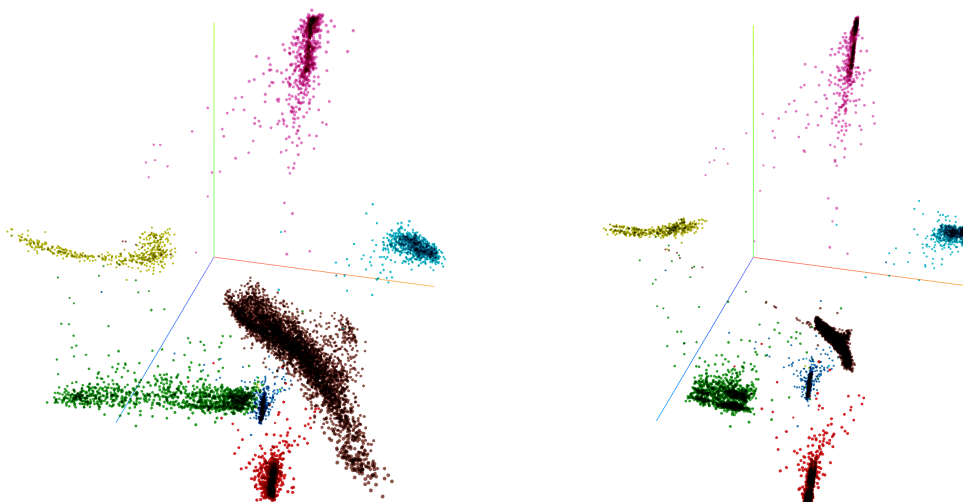


Figure 4.6. Snapshot of t-SNE Embedding Projector in TensorBoard

\* Perplexity: 74; learning rate: 10

Iteration: 67 (left panel); 102 (right panel)

Tang (dark blue); Song (red); Yuan (pink); Ming (sky blue); Qing (green); 1980s (brown); 2010s (mustard).

After word embeddings from Tang dynasty to Qing dynasty are generated, 10 words with the highest cosine similarity scores of *jia* are extracted from each dynasty. Character-based results are shown in Fig. 1, and word-segmented results are provided in

the Appendix. It is found that character-based word embeddings yield a set of words with meanings that are closer to the definitions listed in the OED and MOE dictionaries.



0.5

Nonetheless, it is probable that zhong ‘burial mound’ tops the list because it could be coded for its resemblance of strokes to jia, or because the word was also used to refer to the eldest male offspring in the family, as in jia-zhong and zhong-fu ‘wife of the eldest male offspring.’ From the perspective of nearest neighboring words (**hamilton2016cultural**), the core meanings of jia remains stable from pre-modern time, indicating a strong association with the family clan and the role of a wife, as in zu and qi. Secondly, the words li ‘village; neighborhood’ and cun ‘village; country’ are evident of the structured social unit of living from pre-modern time. However, the nearest neighboring words of li falls into the category of measurement units such as zhang ‘one-tenth of chi’ and chi, whereas zun is still closely linked to words like zhuang ‘village; town’ and xiang ‘lane; valley.’ Interestingly, the most semantically related words to jia in pre-modern Chinese time depicts the idea of home more as a social concept than a physical one. If such words as zhi ‘nephew’, zi ‘offspring’, and sao ‘sister-in-law’ are considered, it becomes clearer that word vectors are able to capture the cultural aspect of jia in pre-modern Chinese. Noticeably, on the list of most similar words are two words related to money—fu ‘to be wealthy’ and zi ‘to estimate (value).’ Although they do not appear as frequently as the aforementioned words, they are assigned higher similarity scores than shi ‘era; decades’ and guo ‘nation; feudal land’, which are thought of as one aspect of core meanings of jia, as in guo-jia ‘nation; state.’

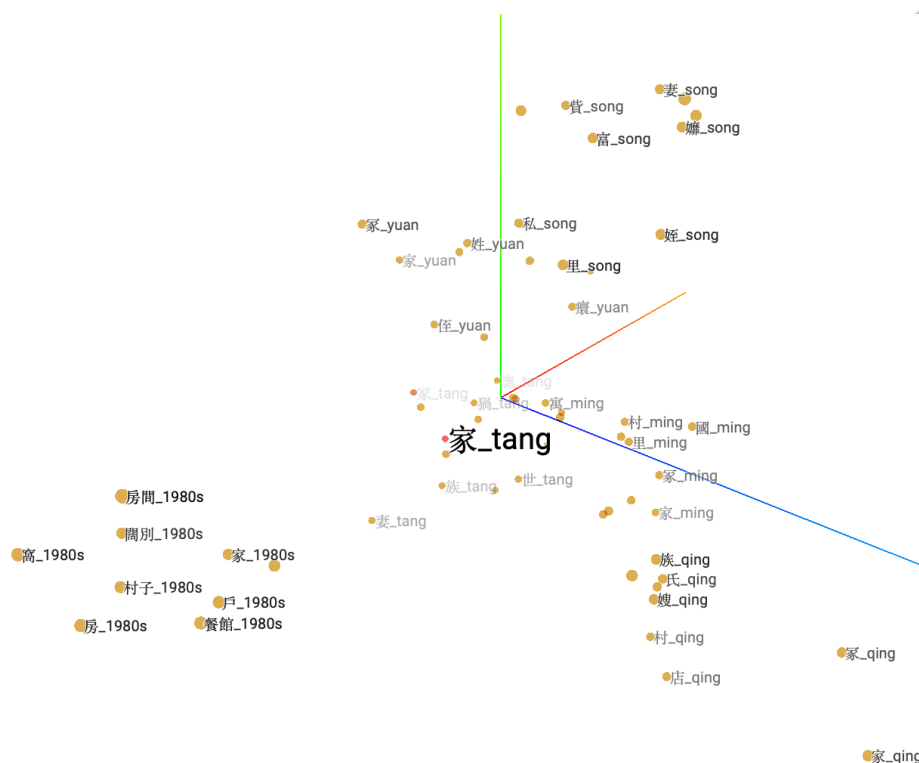


Figure 4.7. Neighboring words of *jiā* projected in a three-dimensional space

Time	Word	Nearest neighboring words									
Tang	家	冢, 族, 冢, 妻, 獨, 世, 富, 讎, 教, 國									
Song	家	冢, 族, 里, 富, 貲, 妻, 姪, 私, 貴, 嬖									
Yuan	家	冢, 族, 貲, 妻, 踰, 姓, 世, 侄, 瘵, 俵									
Ming	家	冢, 族, 妻, 豕, 里, 村, 寓, 富, 國, 產									
Qing	家	冢, 村, 族, 氏, 妻, 店, 子, 寓, 病, 嫂									
1980s	家	房間, 村子, 闊別, 戶, 家小, 酒店, 窩, 房, 旗下, 餐館									
	家庭	婚姻, 小家庭, 職業婦女, 家族, 單親, 兩性, 同儕, 貧苦, 上班族, 鄰里									
	家人	親友, 親人, 部屬, 親朋好友, 同事, 師長, 親戚, 父母親, 妻兒, 異性									
	家族	豪門, 母系, 氏族, 超人氣, 白種, 救星, 文化人, 族, 小家庭, 宗派									
2010s	家	離開, 感受, 要求, 安慰, 遇見, 聽到, 身上, 早上, 傷害, 陪伴									

Table 4.1. Neighboring words with the highest similarity scores to the words *jiā* , 家庭 *jiāting* ‘family/household’, 家人 *jiārén* ‘family members’, 家族 *jiāzú* ‘a family’s clan’.

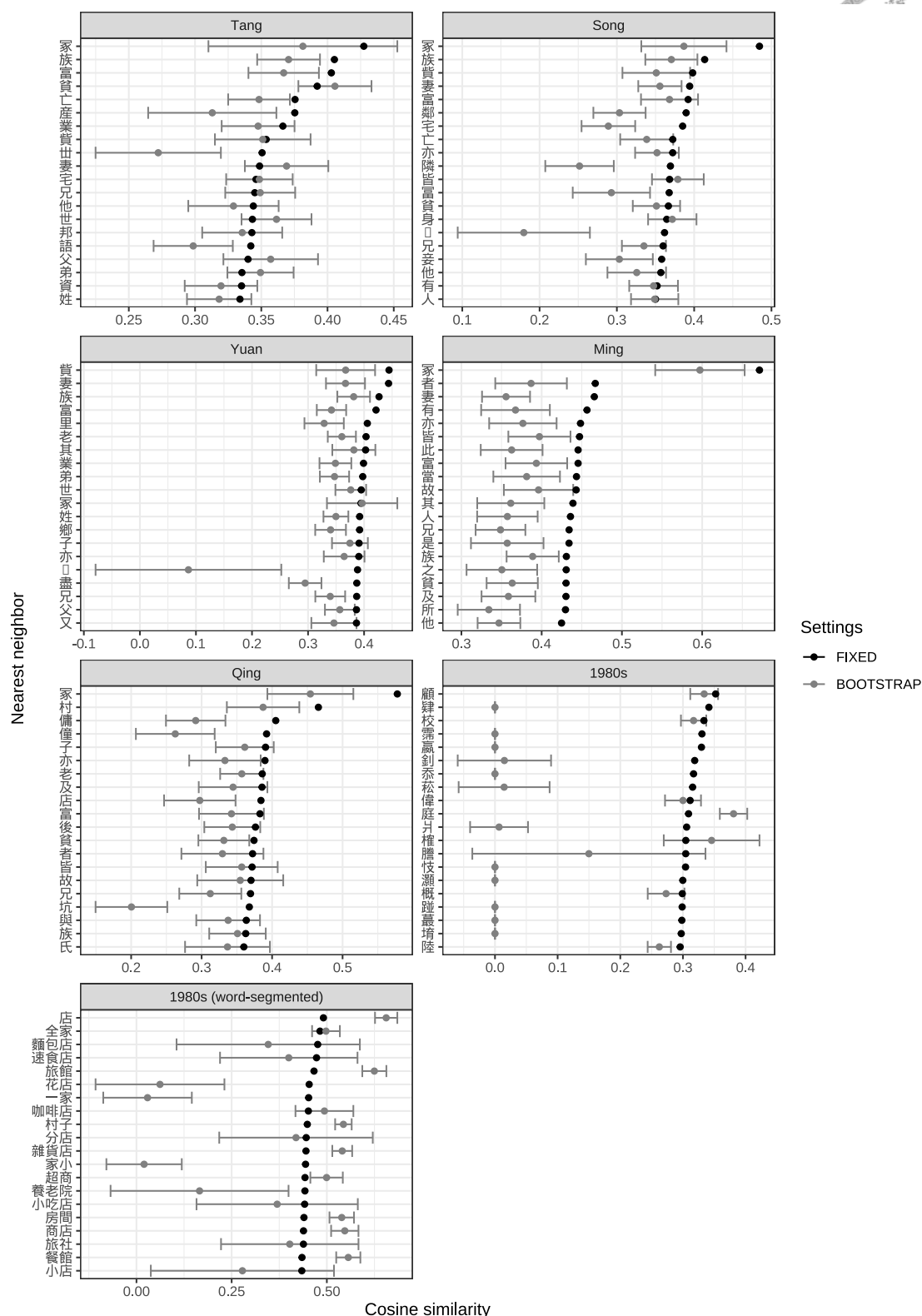


Figure 4.8. Nearest neighbors of *jiā* with means and standard deviations of cosine similarities derived from character-based embeddings in the FIXED and BOOTSTRAP settings. The 20 nearest neighbors are selected from the FIXED settings, and word-segmented embeddings are included for the time period of 1980s.



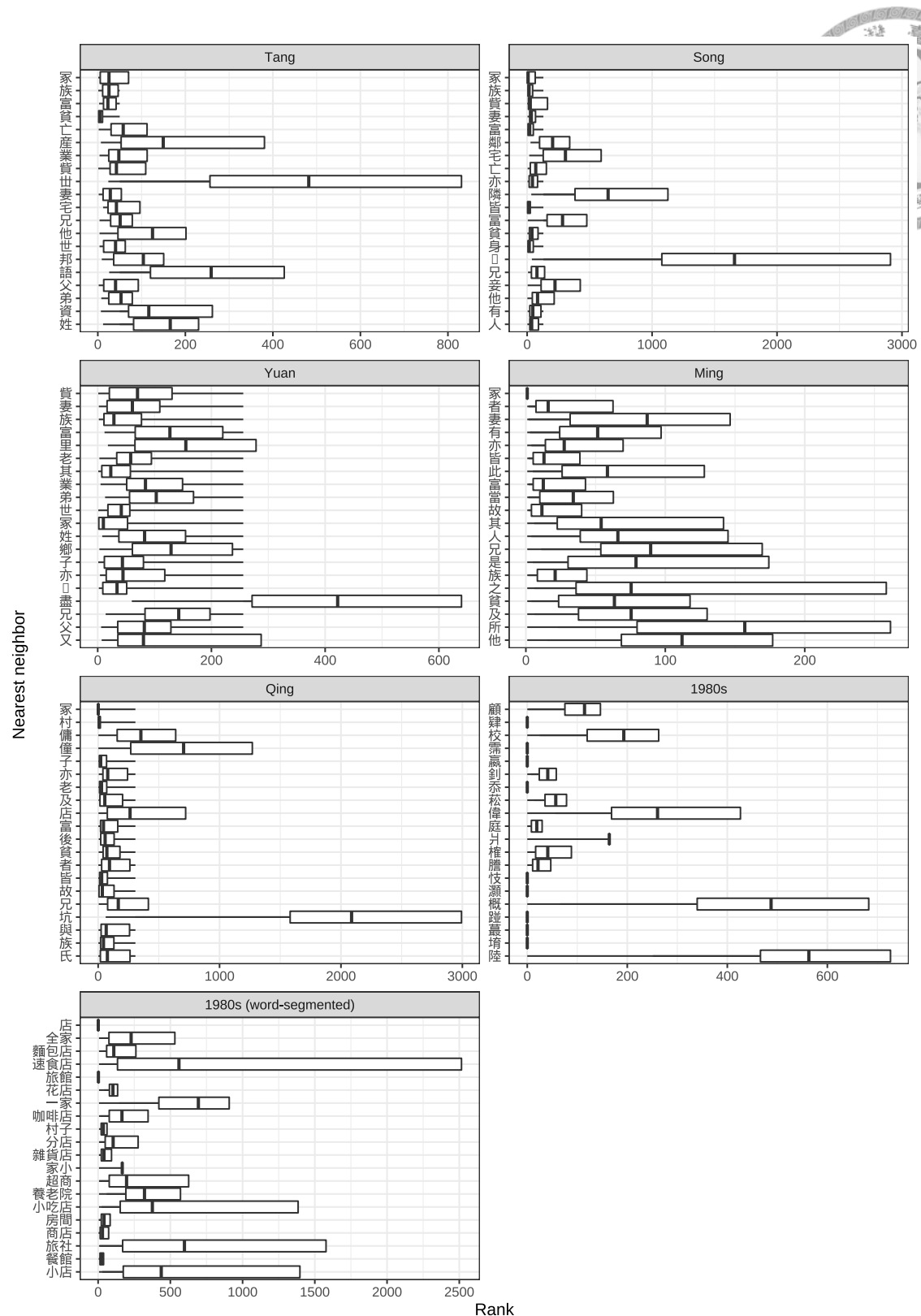


Figure 4.9. Nearest neighbors of *jiā* with changes in rank derived from character-based embeddings in the BOOTSTRAP settings. The 20 nearest neighbors are selected from the FIXED settings, and word-segmented embeddings are included for the time period of 1980s.

ASBC are representative of the concept of jia in the late 20<sup>th</sup> and 21<sup>st</sup> century. As Table ?? shows, cun-zi ‘village’ are still closely related to the concept of jia, appearing as one of its semantically most similar words in the vectors of both window size 1 and 5. Furthermore, more words carrying the meaning of family are seen on the list of ASBC, including jia-xiao ‘wife and children’, quan-jia ‘the whole family’, and yi-jia ‘(a) family’, yet zu and qi are no longer seen, which might reflect the shift of family clans as units of living to smaller household sizes and more equal status of each family member.

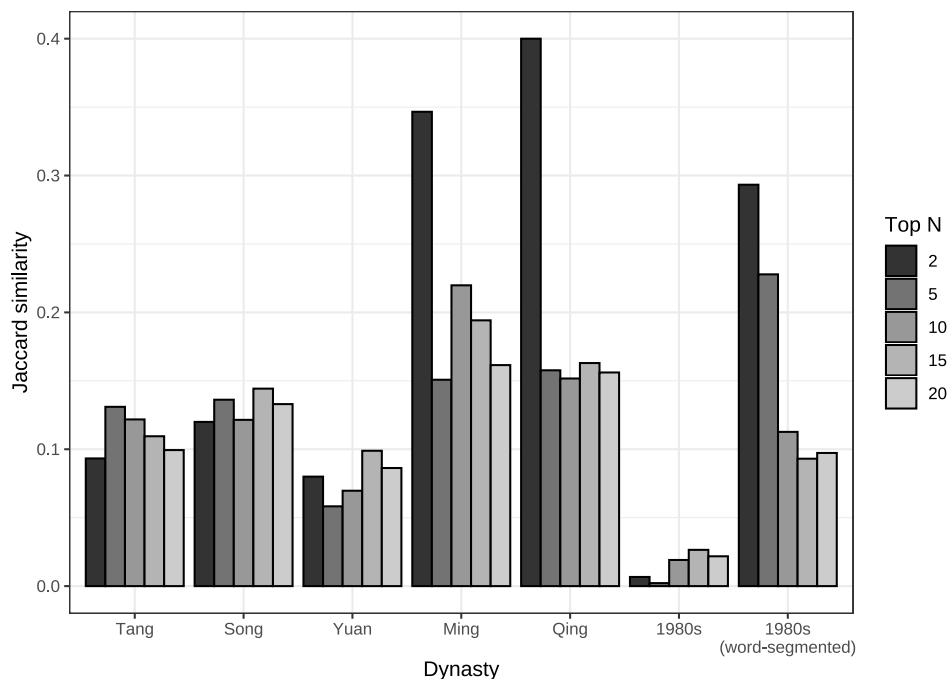


Figure 4.10. Mean of Jaccard similarities from top  $N$  nearest neighbors in the BOOTSTRAP settings. The higher the mean, the higher the degree of intersection for the nearest neighbors across the bootstrap iterations.

Secondly, not the word yu ‘apartment’, but hu ‘one-paneled door; household’, wo ‘nest; hiding place’, and fang ‘house; room’ are used to refer to jia as a physical space or unit of living. Because of the emergence of these alternative words, home evolves to be a private sphere (**mallett2004understanding**). These words highlight the physical aspect of meaning of jia and its characteristics under transformation. The word wo can be used either as a noun or a verb, and as a verb, it stresses that home is portrayed as a place where we feel cozy and at ease, and where we can “retreat and relax” (**mallett2004understanding**).

Interestingly, aside from wo as a verb, kuo-bie ‘to be separated for a long time’ is the only verb on the list of ASBC (**mallett2004understanding**; **samanani2019house**).

Besides, terms of commercial properties are spurring in the list of most similar words to jia, including jiu-dian ‘hotel’, can-quan ‘restaurant; bistro’, lu-quan ‘hotel’, xiao-chi dian ‘eatery.’ It is speculated that commercialization is accountable for this new trend, but it is also possible that jia starts to be used as a classifier, as in yi-jia-lu-quan ‘one hotel.’ Judging from the data in ASBC, it is seen that not only does the concept of jia changes across time, but the word use of jia changes as well, which is evident in more alternative word choices to refer to the concept of jia.



In the 21<sup>st</sup> century, the word jia is associated with a wider variety of words, mostly verbs. Unlike data from earlier time spans, the words are less semantically associated with the direct naming of a physical space or family unit, but because people engage themselves more and more often in describing their daily life and encounters, verbs like li-kai ‘to leave’, qan shou ‘to-feel’, shang-hai ‘to hurt’, and pei-ban ‘to accompany’ are assigned the highest probabilities to words of jia.

Although word embedding technique grows increasingly prevalent in the field of computational linguistics and natural language processing, it has been criticized for representing words with multiple meanings as one single vector, which is referred to as “meaning conflation deficiency” (**camacho2018survey**) To allow the algorithms to know different senses of the same word form, two main methods for sense embeddings are proposed. [21, 22] One is unsupervised as senses are “induced” from the training corpora; the other is knowledge-based, meaning external sense inventories, such as WordNet, are required to fine-tune the word vector models.

Since the keyword jia does not reveal how people are connected in this recent era, 2 other keywords are chosen to see if more insights can be gained. The words jia-ren and jia-ting can help us understand the social structure of home nowadays. As the above figure shows, the concept of jia is first depicted with a single word jia, and as time passes, jia is conceptualized with multiple other lexical items. In other words, in earlier time, different aspects of home are described by the character jia, yet these aspects are embodied with different words such as jia ren-ren and jia-ting in modern Chinese texts.



0.5

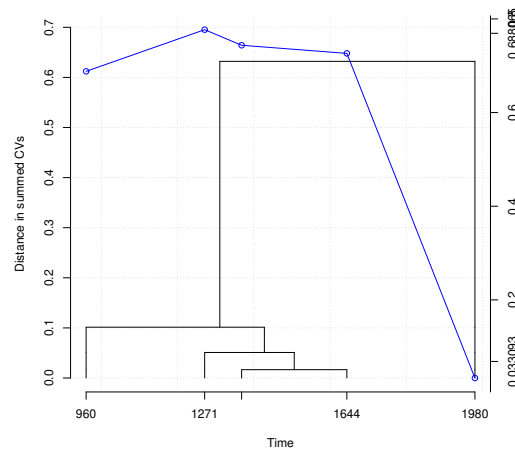


Figure 4.11. VNC results of word-level embeddings

## 4.5 Diachronic sense embeddings

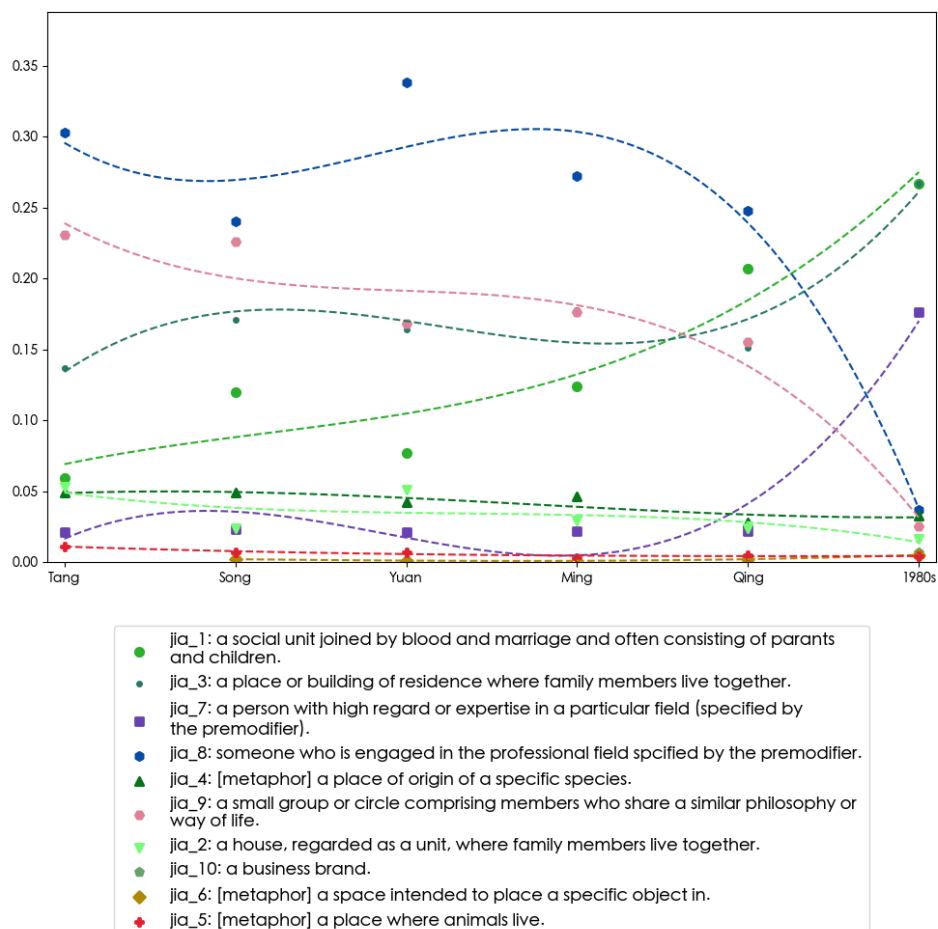
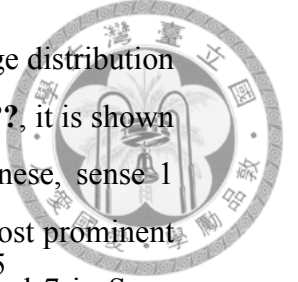


Figure 4.12. Diachronic interactions of senses



The extraction of contextualized embeddings allows for a sketch of usage distribution displayed by proportion and interactions of different senses. From Figure ??, it is shown that senses do compete and cooperate semantically. In present-day Chinese, sense 1 (family), sense 3 (house), and sense 7 (-ist) are shown to be three of the most prominent senses, yet sense 1 does not evolve in identical direction with sense 3 and 7 in Song and Ming. Instead, its rise of sense 1 has indicated that single-character words like *jiā* can be read as ‘family’, and combined with sense 3, they account for over 60 percent of the usage proportion, while sense 7 is only half of it. Interestingly, both sense 7 and 8 carry the meaning of describing someone’s profession, but the contextualized embeddings distinguish the two readings in terms of the percentage. Qualitatively, these are influenced by different schools of thought. Furthermore, it is comparatively rare for *jiā* to serve as adjective ‘domestic’, or sense 10 as categorial name.

The polysemy of a lexical item is addressed by constructing multiple contextualized token embeddings. Shades of meanings are reflected in the diversity of contextual use.

The results indicate that *jiā* enjoy far global distance but low local distance, and suddenly rises during 1980s.

## 4.6 Discussion

Following **hamilton2016law**, in which the evaluation is based on examples from previous works on semantic change and words with the “obsolete” tag in the Oxford English Dictionary (OED), dictionary entries are consulted to look for “舊時” and “古代” for attested examples to evaluate the trained diachronic word embeddings.

For example, 齒 *chǐ* ‘tooth’ used to carry the meaning ‘age (年齡)’ and ‘being of equal rank (並列)’ because age determination is made by numbering horses’ teeth, which emerges one each year, as in ‘子之齒長矣，不能事人 (You are long in the tooth)’ and ‘不敢與諸任齒 (I would not dare to take rank equivalent to yours)’; another example is 卑鄙 *bēi-bǐ* ‘despicable’, which is more neutral in connotation in the past (**wang1997gujinyi**). Dictionaries include **wang1997gujinyi**; **liu1992gujinyi**, which lists word entries with meanings that are distinctive between modern and pre-modern times. Detailed information relevant to semantic change is the number of disyllabic word entries, whether the word convey connotations with varying sentiment polarities, and whether certain senses fall

into disuse nowadays, which is valuable resources for the comparison with the results of computational methods.

The meanings are based on 漢語大字典, 漢語大詞典, 辭源, 辭海 as well as 現代漢語詞典 and 新華詞典 (both published by 商務印書館).

frequency data is derived from 在线古代汉语语料库字频数据<sup>2</sup> and 近代漢語語料庫詞頻統計<sup>3</sup>, which are the metadata from the 70-million-word Ancient Chinese Corpus (在线古代汉语语料库) by the Ministry of Education, China and Academia Sinica Tagged Corpus of Early Mandarin Chinese (近代漢語語料庫) by Academia Sinica, Taiwan.

The case study of *jiā* is based on the assumption that the time-sliced corpus might reflect the similar and different descriptions in language use. While words in Table ?? fall into the categories of technological innovations and ideologies, this study chooses *jiā* because of its linguistic and cultural characteristics. In pre-modern Chinese, *jiā* is associated with words that denote physical objects like house.

Because the corpus contains multiple versions of a document, some orthographically-similar characters rank top in terms of cosine similarity scores. However, if compared with the results from BOOTSTRAP samples, the scores are widest. In addition, the ranks vary widely in different iterations, and are a reliable indicator of neighbor analysis. For example, 貧 *pín* ‘poor;impoverished’ appear 43 times out of the 50 iterations as the top 20 closest neighbors, followed by 窶 *jù* ‘poor;impoverished’ also appear 26 times. Other closest neighbors include 族, 世, 妻, 冢, 富, 窶 *jù* ‘poor;impoverished’, 孀, 紉, 父 (all more than 15 times.)

As for the word 宅, the closest neighbors include 田 (48), 廨 (47), 居 (39), 園 (36), 墅 (36), 家 (35), and 廬 (14), filtering out 冢 (1). Compared with FIXED embeddings, the closest neighbors for the Tang dynasty include 廨, 田, 宇, 邸, 園, 營, 室, 塾, 寺, 住, 妝, 寓. Therefore, if neighbor analysis can be compared from two directions, it is likely to mitigate the issue arising from OCR errors?

The semantic history of linguistic units or expressions are far more unpredictable than data that contain seasonality. Regarding the closet neighbors for *jiā*, the results differ in a distinctive way, with a low percentage of overlaps between the FIXED embeddings and the BOOTSTRAP ones. In addition, before the diachronic character-based embeddings are constructed, a decision needs to be made on whether the different versions of a workset of

<sup>2</sup><http://corpus.zhonghuayuwen.org/resources.aspx>

<sup>3</sup><http://elearning.ling.sinica.edu.tw/jindai.html>

texts are to be included or excluded. Considering the fact that the documents are converted from scanned copies to the digital texts in UTF-8 encoding using the OCR technique, the FIXED embeddings reinforce the parts that are consistently recognizable and transformed into similar strings of characters. In other words, the inclusion of all versions in a workset of documents prevents misrecognized characters from taking up a significant portion of the word occurrence behavior. On the other hand, the word cooccurrence profile remains susceptible to orthographically highly similar characters, e.g., 家 and 冢, 人 and 入, and 怡 and 恰, and place the mistaken form as the close neighbors, oftentimes the closest neighbor.

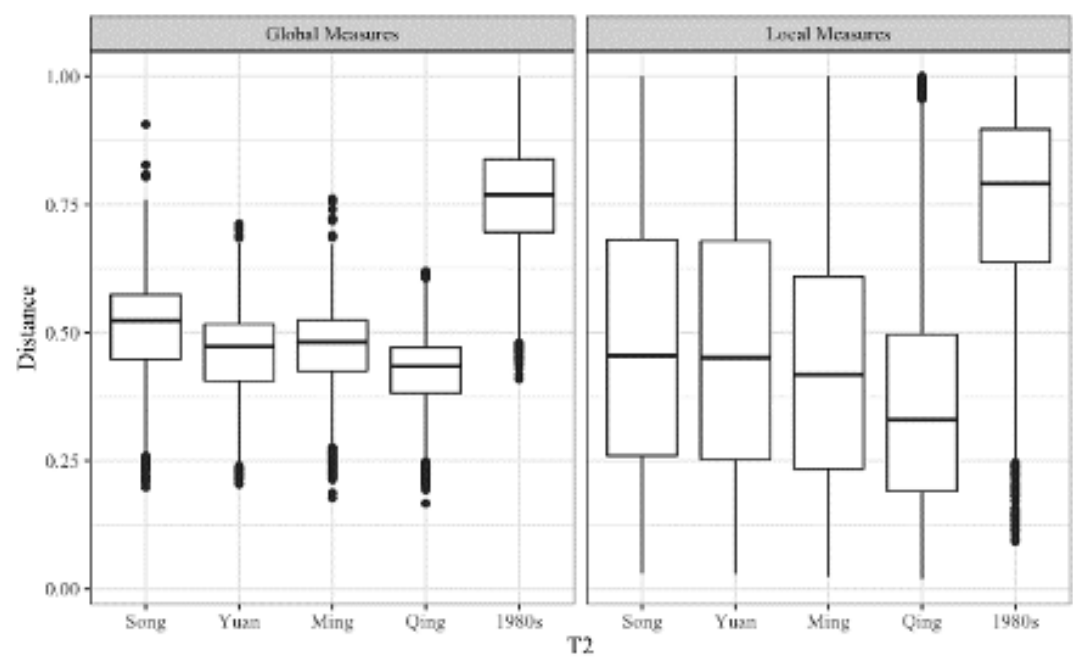
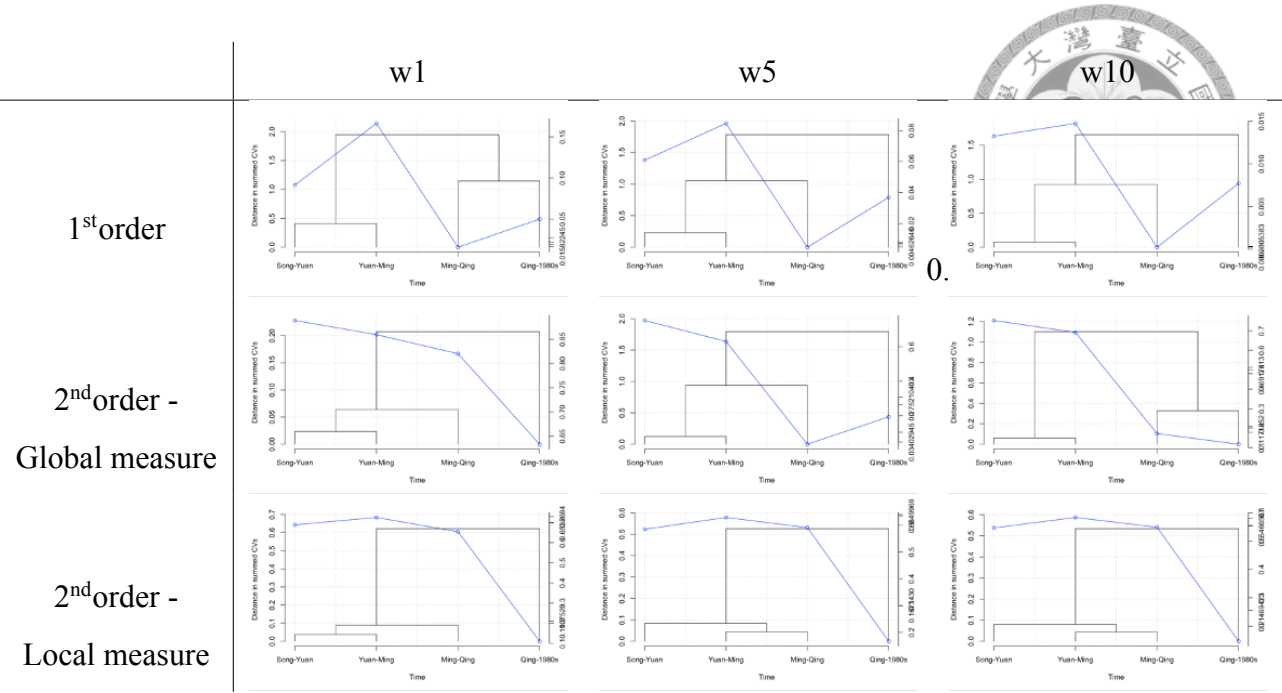


Figure 4.13. Distribution of degree of semantic change for global and local measures

Note: Below is the results for semantic change trends based on all of the character-based embeddings, change to example of *jiā* or group of words only







0.5

## Chapter 5

## Conclusions

In light of the growing interest in diachronic lexical semantic change, this thesis is a case-study investigation of *jiā* through a corpus-based approach. Language does not cease to change beyond the observable texts within the time frame of the chosen corpora, and to capture semantic change that might not be accompanied by change in frequency.

The evolution of *jia* is a compressed history of the Chinese society and the Chinese language. The analysis of word representations of *jia* serves as a starting point to pinpoint the core, stable meanings of the word, outlining the properties of a physical space and a structured social unit. While the emphasis has been put on the economic situation from pre-modern time, the word *jia* becomes less associated with individuated roles such as a wife, but more closely focused on the self, depicting personal memories of home leaving and returning.

With the advantage of distributional semantic models, the meaning conflation of home, house, and family can be explored as different components. Especially, premodern Chinese is distinguished from the current written form, uses different lexical items, and is mostly in the form of one syllable. The disparity results in the addition of new senses of the one-character *jia*, and aspects of meanings are encoded in different two-character words in modern time. In the field of corpus and computational linguistics, changes of word choice and the inclusion of more senses allow for a closer look at the texts in snapshots of specific time frames, while resonates with studies in other disciplines.

How polysemy of homophone is to be explored through external resources such as dictionary and negative examples **traugott2001regularity**. Cross-linguistic and

metalinguistic analyses are insightful. In addition, as change in meaning is ongoing, the detection of semantic change can be detected in progress.

As discussed in **giulianelli2019lexical**, the fine-tuning of large-scaled pre-trained language models like BERT does not yield satisfactory results of temporal-specific contextualized usage/token representations. As hinted by **giulianelli2019lexical**, the fine-tuning is based on classification task of recognizing the time period of a portion of documents, but the fine-tuned models might instead reflect the style of prominent authors of certain time periods, reering away from baseline representations. Faced with these problems, **kutuzov2020uio** also compares contextualized embeddings with context-independent ones, and find that for semantic change detection, context-independent embeddings are effective.

Semantic change modeling has profound impacts in linguistic analysis. As language is a dynamic phenomenon, a temporal-aware understanding is explored as a starting point. Following the examination of factors, sense evolution prediction, the interaction between semantic change and different linguistic, cultural factors can deepen our understanding, especially the aspects of polysemy and multi-word expressions. The task of meaning representation from the perspective of semantic change is especially rewarding toward how the modeling of meaning representation can be tweaked, with the complexity being justified, unlike in English, it is not always the case that preprocessing of compound words are taken into account from the beginning.

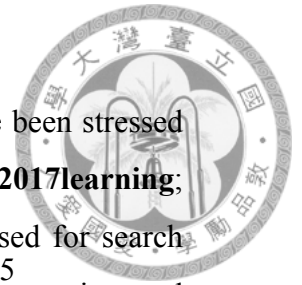
However, the character-based embeddings serve as a starting point to investigate the semantic development of Chinese, which is so distinctively different in pre-modern and modern time that calls for an integration of the disyllabic development of Chinese to account for the differences in different time periods. Recently, dependency parser of pre-modern Chinese has been released, yet the segmentation still split many disyllabic words into units of single characters. Nonetheless, through the analysis of different measures of semantic change, this study captures different aspects of semantic properties, and it is hoped that the results can lay an empirical basis of how single characters behave semantically by considering the time dimension of the textual data. In conclusion, this study aims to explore the word representations that are more dynamic than present application is populated for, and to show how word co-occurences can be revealing in terms of such a concept like home that is relatively stable but ever-evolving with the



passage of time.

The importance of temporal-aware, diachronic word embeddings have been stressed both for modern texts and historical ones (**huang2019neural**; **rosin2017learning**; **ruder2017word**). With the accumulation of texts in corpora that are used for search system, i.e., to answer “when” two terms are related to each other, query expansion, and weighted synonyms (**rosin2017learning**). It is by this aim that this study is motivated, and for the purpose of achieving more understanding of the properties of language use through the lens of time. Furthermore, the rate of change is another important issue so as to incorporate “time-sensitive” query expansion (QE) (**rosin2017learning**) to involve the time dimension of linguistic phenomenon more in this rising, flourishing field of study.

As researches combine textual data from various corpora or sources, the detection of semantic change and measurement of degrees of change helps compare not only time-specific needs, but also corpora of different types (**schlechtweg2019wind**), which becomes increasingly critical with an abundance of textual data presented to us nowadays. The analysis can be further explored by reaching out to other research disciplines and communities, and even the design and functionality of diachronic corpus itself.





## Appendix A

0.5

Time period	Word	Rank	Frequency	Percentage	Cumulation
Old Chinese	家 (NA3)	238	59	0.053	64.414
	Total	-	59	0.053	-
Pre-modern Chinese	家 (Nc)	31	10885	0.380	26.605
	家 (Nc)[+spo]	822	457	0.016	66.682
	家 (T4)	2777	113	0.004	81.827
	家 (Nes)	22890	4	0.000	97.318
	家 (Na)	41336	1	0.000	99.331
	家 (Nc)[+vrr]	41336	1	0.000	99.331
	家 (Nh)	41336	1	0.000	99.331
	Total	-	11462	0.400	-
Modern Chinese	家 (Nc)	193	2793	0.057	40.002
	家 (Nf)	299	1835	0.038	44.999
	家 (Na)	11546	36	0.001	86.357
	家 (Nc)[+spo]	25841	12	0.000	92.634
	家 (Na)[+spo]	70282	2	0.000	98.041
	家 (Nc)[+p2]	93826	1	0.000	99.208
	家 (Na)[+p2]	93826	1	0.000	99.208
	Total	-	4680	0.096	-



## Appendix B

0.5

Id	Analogy in Chinese	Analogy in English	SGNS w1		SGNS w10	
			Add	Mul	Add	Mul
1	冷-熱: 南-北	cold-hot: south-north	6	6	6	6
2	鬆-緊: 南-北	loose-tight: south-north	6	6	6	6
3	鬆-緊: 左-右	loose-tight: left-right	6	6	6	6
4	大-小: 南-北	big-small: south-north	6	6	6	6
5	大-小: 左-右	big-small: left-right	6	6	6	6
6	真-假: 左-右	real-fake: left-right	6	6	6	6
7	貧-富: 左-右	poor-wealthy: left-right	6	6	6	6
8	粗-細: 南-北	thick-thin: south-north	6	6	6	6
9	東-西: 左-右	east-west: left-right	6	6	6	6
10	上-下: 南-北	upper-lower: south-north	6	6	5	5
11	高-低: 南-北	high-low: south-north	6	6	5	5
12	寬-窄: 南-北	wide-narrow: south-north	6	6	-	-
13	深-淺: 南-北	deep-shallow: south-north	6	6	-	-
14	胖-瘦: 南-北	fat-slim: south-north	5	5	6	6
15	遠-近: 左-右	far-near: left-right	5	5	6	6
16	上-下: 左-右	upper-lower: left-right	5	5	6	6
17	東-西: 南-北	east-west: south-north	5	5	6	6
18	強-弱: 左-右	strong-weak: left-right	5	5	6	6
19	明-暗: 左-右	light-dark: left-right	5	5	6	6
20	冷-熱: 左-右	cold-hot: left-right	5	5	6	6
21	輕-重: 左-右	light-heavy: left-right	5	5	6	6

22	粗 - 細: 左 - 右	thick-thin: left-right	5	5	6	6
23	南 - 北: 左 - 右	south-north: left-right	5	5	6	6
24	冰 - 水: 雪 - 雨	ice-water: snow-rain	5	5	5	5
25	明 - 暗: 南 - 北	light-dark: south-north	5	5	0.5	-
26	攻 - 守: 買 - 賣	attack-defend: buy-sell	5	5	-	-
27	寬 - 窄: 左 - 右	wide-narrow: left-right	-	-	6	6
28	高 - 低: 左 - 右	high-low: left-right	-	-	6	6
29	強 - 弱: 南 - 北	strong-weak: south-north	-	-	6	6
30	動 - 靜: 左 - 右	moving-still: left-right	-	-	6	6
31	深 - 淺: 左 - 右	deep-shallow: left-right	-	-	6	6
32	前 - 後: 左 - 右	front-back: left-right	-	-	6	6
33	動 - 靜: 東 - 西	moving-still: east-west	-	-	5	5
34	輕 - 重: 南 - 北	light-heavy: south-north	-	-	5	5
35	胖 - 瘦: 左 - 右	fat-slim: left-right	-	-	5	5

