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

論文題目
Thesis Title

陳蓓怡
Pei-Yi Chen

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

October 2020
中華民國 109 年 10 月

Abstract

This research proposes to investigate the topic of historical semantic change from the perspective of quantitative/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 (Blank, 1999:61). The dynamics of meaning construction is embodied in the emergence and losses 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 (Blank, 1999:63). 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 vector space 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-occurrence" of words in context, and the co-occurrence distribution represents the similarities and differences in meaning interactions. The diachronic corpus consists of texts from the following sources: the Chinese Text Project, Chinese Buddhist Electronic Text Association corpus for pre-modern Chinese, and Academia Sinica Balanced Corpus of Modern Chinese for modern Chinese. 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 Hamilton et al. (2016a) 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. Regarding the statistical modeling, Generalized Additive Models (GAMs) is used as the basis for quantitative analysis. The GAMs allows for the investigation of non-linear predictor effects without any predefined structure and the manner in which these effects develop over time. Finally, this study will proceed with periodization analysis using the Variability-based Neighbor Clustering (VNC) method. As a hierarchical clustering method, 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 project 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.

摘要

本研究欲從量化/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，歷史文本的電子化亦開展了更多與歷時語意相關的研究可能，進而從中分析、挖掘詞彙所蘊含的詞意。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，(Blank, 1999:61)。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，進而得出語意變遷（semantic change）的規律性（regularities）(Blank, 1999:63)。如果從共時（synchronic）的角度來看，語意存在各種變異（variation），而在歷時（diachronic）的脈絡下，經過時間累積而記錄著各式變遷。語料庫語言學以自然產生的語言使用資料為本，從中觀察、歸納出可質化、量化的語言分析，而歷時語料庫因應科技進步，計算語言學界亦已出現以詞向量（word embedding）、統計模型等方式探求語意在時間洪流下的變動與趨勢。多義性（polysemy）是語意變遷另一大成因，詞彙將各個概念、意義的分類以語言的形式表達，語言共時下的詞義關係，時常亦已存於歷時的發展。漢語的詞彙組成從單字詞走向雙字詞（disyllabic words），不僅可以讓我們探究同音異義（homophony）的影響、字詞間的詞意互動、雙字詞與複合詞（compound）的增長。對於語料較稀少的歷史主題，計算語意學與統計模型的方法可突破許多困境，因為原始語料為寶貴研究材料，除此之外，有些詞彙雖然並無明顯的詞頻變化，其指涉對象與意義內涵卻與以往大不相同，在詞向量等詞彙表徵方法蓬勃發展之時，歷史語意變遷亦是不可缺少的研究主題。在語料庫語言學的範疇，相關的研究主題被稱為「共現（co-occurrence）」方法，共現分佈的趨勢代表著意義分布的異同。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。本研究以歷時語料庫（如：中國哲學書電子計畫、中華電子佛典協會佛典集成）與現代漢語語料庫（如：中研院漢語平衡語料庫）為語料來源，以歷時的詞向量搭配詞彙資料庫，了解單音節至複音節詞彙的語意變遷程度，Hamilton et al. (2016a) 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，比較各時代向量的方法，其相關係數越低，語意變遷程度越高。此外，從詞彙的意義分布與互動，描繪出不同詞意的消長與變動。在統計模型的選擇上，以廣義相加模型（Generalized Additive Model, GAMs）作為語料分析的量化基礎，此方法針對變化劇烈且相互影響的資料，可得出時間維度下的語意變遷曲線，並進一步預測未來與回推過去此曲線的走勢。最後，本研究將採用以變異程度為基礎的近鄰群聚分析法（Variability-based Neighbor Clustering, VNC），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。計算語意學與歷史語意學的研究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、社會概念的更迭牽動詞彙之間的關聯。

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Language is constantly changing and evolving. The emergence of new senses, the demise of old ones, and the polysemous nature of lexical items make the process of semantic change a dynamic phenomenon (Robert, 2008). As individuals learn new words and meanings throughout their life, so does a language. 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 (Camacho-Collados and Pilehvar, 2018; Kutuzov, Øvrelid, et al., 2018; Tahmasebi et al., 2018).

"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 (Crowley and Bowern, 2010: 199).

As language users actively engage in interpreting and processing the language.

Renouf (2002) reflects on how textual data starts to be treated more than "a static entity." In 1982, Sinclair envisions the possibility of "vast, slowing changing stores of text" and "detailed evidence of language evolution" (as cited in Renouf, 2002). The use of digitalized libraries as rich linguistic resources to observe how certain linguistic features are "assimilated" into the language becomes more and more feasible (Renouf, 2002).

Additionally, the change in meaning is captured by translating discrete linguistic data into numeric vectors such as word embeddings, especially after the release of

Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2016). An initial attempt is to generate 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.

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" ("Home", 2020). 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* 'home' as "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)" ("Jia", 2015). Yet, how is the concept of home encoded linguistically? Specifically, how is diachrony interacts with synchrony and variations?

From the perspective of corpus-based computational linguistics, questions are invoked as to how the concept of home is understood by the computer? What words are semantically related to this concept? Because we live at home, far from home, or in a place we call home, and we are constantly searching for the meanings of home, this study aims to explore how semantically-related words construct the meanings of home, and how this concept comes into shape through the lens of time.

This paper is organized as follows. An overview and reflections of semantic change and diachronic word embeddings are given in section 2. The development of word-level and sense-level word representations brings to the fine-grained analyses and generalizations of semantic change. The topic of home is introduced in the second section. Section 3 describes how semantic change is captured and visualized.

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 (Hamilton et al., 2016a; Kutuzov, Øvrelid, et al., 2018; Kutuzov, Velldal, et al., 2017). In Hamilton et al. (2016a), linguistic drift and cultural shift are distinguished and measured based on diachronic word embeddings, with the latter restricted to a smaller set of neighboring words.

Depending on the starting-point of investigation, semantic change can be approached from a semasiological and an onamasiological perspective (Geeraerts, 1997: 17; Traugott and Dasher, 2001: 25). A semasiological perspective deals with meaning change of the fixed form of a lexeme, while an onamasiological one is framed within a given concept or domain expressed by a set of alternative words. Semasiologically, when a lexeme undergoes semantic change and additional meanings are gained, the different senses might gradually be perceived as unrelated to each other by the language users. That is, the lexeme first becomes polysemous, and then homonymous (Traugott and Dasher, 2001: 25). Onamasiology, on the other hand, focuses on synonyms, near-synonyms, and naming-gaving (Geeraerts, 1997: 17).

Semantic change can be broadly understood as the "reanalysis" of a word (Fortson IV, 2017: 650), 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 (Fortson IV, 2017: 650; Traugott, 2017). Bloomfield (1933) 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 Crowley and Bowern (2010: 199–205), 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.

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 (Traugott and Dasher, 2001: 81).

Polysemy, described as "families of related meanings" in Traugott and Dasher (2001: 11), and serves as a foundation of generalizations of semantic change with recurring patterns. The co-existence of older and newer meanings in a lexical item, and the influence of multiple meanings on one another, lead to the dynamics of "saliency" Traugott and Dasher (2001: 12). More than single semantic reading is not only necessary and omnipresent. Among the driving forces of lexical semantic change, synchronic polysemy is highlighted as the essential component (Robert, 2008). The construction of meanings is flexible and sensitive to the context of use (Miller and Charles, 1991/2007; Zellig, 1954/2015). Additionally, the mechanism of metonymy allows the co-existence of referential and conceptual meanings in the same word (Hilpert, 2019; Nerlich and Clarke, 2001). Specifically, an understanding of metonymic change builds upon the familiarity of the culture in which the language is spoken, which leads to the diversity of attested examples (Fortson IV, 2017: 649). Yet, it is recognized that synchronically distinct meanings, which spakers 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" Traugott and Dasher (2001: 13). On the other hand, synchronic convergence is also likely, as shown in instances of folk etymology, but not as common cross-linguistically. Nonetheless, semantic change is a complicated phenomenon resulting from not only polysemy, but also subjectification (Traugott and Dasher, 2001), prototypicality (Geeraerts, 1997), and other contributing factors. Linguistic variations of language use is omnipresent in the synchronic settings, but is amplified in a diachronic scope (Bowern, 2019; Crowley and Bowern, 2010).

Ambiguity is resolved or cancelled in context of use. Generalized invited inferences depending on whether intended meanings are coded or crystallized into commonly used implicatures. 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 polysemous (coded) meanings (Traugott and Dasher, 2001: 49).

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 (Firth, 1957).

The application of computation to larger sets of words across longer periods of time enables the generalization of regularities on semantic change (Hamilton et al., 2016b). Semantic change driven by technological innovations are prominent examples, while shifts of meanings with linguistic cause tend to occur relatively more slowly (Hamilton et al., 2016b). The changes encompass changes to "core meanings of words" or "subtle shifts of cultural associations" (Hamilton et al., 2016a). The term "brachychrony" is even coined by Renouf (2002)Mair (1998) 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.

2.2 The Concept of Home in Literature

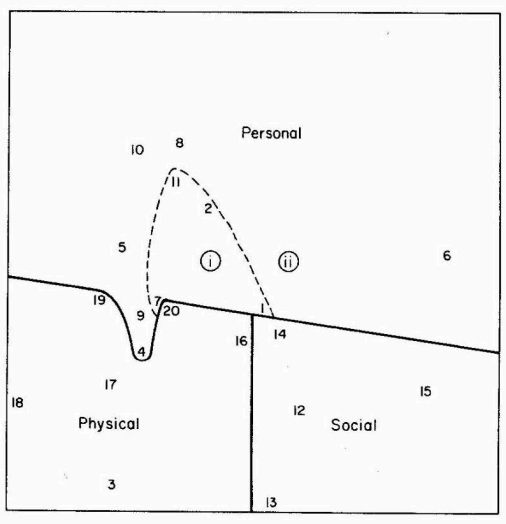
The concept of home has been extensively studied in (environmental) psychology, sociology, anthropology, architecture, and other fields of study (Mallett, 2004; Moore, 2000; Samanani and Lenhard, 2019; Sixsmith, 1986). Specialized topics on homelessness, journeying, migration, gender, and aging are also 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 (Mallett, 2004; Sixsmith, 1986). Nonetheless, home is "not only of belonging but also of potential alienation when attempts to make home fail or are subverted" (Samanani and Lenhard, 2019). 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 (Samanani and Lenhard, 2019). 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" (Samanani and Lenhard, 2019). Furthermore, materiality facilitates the development of a sense of belonging (Moore, 2000).

2. Family: a structured social unit of living. A family is symbolic of marriage, kinship, togetherness, and homeliness (Samanani and Lenhard, 2019). A household is established through the process of homemaking, and the feeling of rootedness, safety, and value is thus deepened (Moore, 2000; Samanani and Lenhard, 2019). 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 (Mallett, 2004; Samanani and Lenhard, 2019).

The most detailed analysis is provided by Sixsmith (1986). The co-existing relationships of home is plotted as three regions from questionnaire responses, as

shown in Figure 2.1 (Sixsmith, 1986).



Personal	Social	Physical
Happiness	Type of relationship	Structure
Belonging	Quality of relationship	Services
(Responsibility)	Friends and entertainment	Architecture
Self-expression	Emotional environment	Work environment
Critical experiences	(With others)	Spatiality
Permanence		
Privacy		
Time		
Meaningful places		
Knowledge		
Desire to return		

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 Sixsmith (1986)).

Culturally, the concept of home in Taiwan as a physical space has undergone changes caused by the sway of the world order (Chen and Fu 沈孟穎, 傅朝卿, 2015). 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.

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 (Kutuzov, Øvrelid, et al., 2018). 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 (Hamilton et al., 2016b; Kutuzov, Øvrelid, et al., 2018; Tahmasebi et al., 2018). With a growing interest in this research topic, insights have been made to highlight some key and challenging aspects of semantic change modeling (Camacho-Collados and Pilehvar, 2018; Kutuzov, Øvrelid, et al., 2018; Tahmasebi et al., 2018).

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 focuses on the use of word embeddings as evidence for certain linguistic feature. (Antoniak and Mimno, 2018) 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. 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 (Kutuzov, Øvrelid, et al., 2018). Among the diachronic textual data currently available, the main source includes but not limited to the Google Books Ngrams Corpus[1], Corpus of Historical American English (COHA)[2], Project Gutenberg Corpus[3] 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.

Diachronic word embeddings can be used to discover more possibilities of unknown change cases and underlying causes of general semantic change (Hamilton

[1]http://books.google.com/ngrams. A comprehensive review of diachronic corpora is provided by Tahmasebi et al. (2018: 38–41)

[2]https://www.english-corpora.org/coha/

[3]https://www.sketchengine.eu/project-gutenberg-corpus/

et al., 2016a). In Hamilton et al. (2016a), 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*. Kutuzov, Velldal, et al. (2017) 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 Hamilton et al. (2016a), 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.

On top of that, based on the self-similarity scores of the English lexicon between 1850 and 2009, Dubossarsky et al. (2015) 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." The law of conformity and innovation are put forward by Hamilton et al. (2016b); 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 (Hamilton et al., 2016b). 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 Hamilton et al. (2016b), 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.

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 (Tang, 2018). Kutuzov and Giulianelli (2020) 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 (Huang and Paul, 2019). Temporal adaptation is introduced as a form of domain adaptation to diachronic word embeddings and proves effective in the task of document classification (Huang and Paul, 2019).

Another challenge, namely the "meaning conflation deficiency", is brought up by Camacho-Collados and Pilehvar (2018). Previously, word embedding technique is first implemented by Mikolov et al. in 2013. 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, Hu et al. (2019) 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) (Hu et al., 2019).

The BERT pre-trained language model can be used in company with sense inventories or cluster analysis. Using the BERT pre-trained language model, Hu et al. (2019) track the evolution of 4881 English words from 1810 to 2009 in the Corpus of Historical American English (COHA), and visualizes the interactions of words' senses. The source texts from COHA are concordance lines which contain target words with a frequency of at least 10 times for over 50 consecutive years. Additionally, the sense identification task is performed by using example sentences in the Oxford English Dictionary (OED) as the knowledge base for similarity comparison with texts from COHA, and the total number of senses from the OED is 15836. Firstly, the last hidden layer of a target word's embedding is extracted from the pre-trained BERT language model. This token embedding is then compared with each sense representation retrieved from the OED word entry to determine which sense the target word belongs to. In Giulianelli (2019), the target words are collected from Gulordava and Baroni (2011) with annotated data on judgement task. Then, their cluster analysis reveals that types of semantic change can be identified, e.g., literal/metaphorical meanings, different senses of a polysemous word, words with different syntactic categories, and affixation. It is concluded that the change in sense distribution follows the "S shape" proposed in linguistics. Moreover, the actual uses of a certain sense can be inspected from the collected data. Their subsequent work is expanded to more languages and judgement data in the SemEval 2020 task (Kutuzov and Giulianelli, 2020).

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 (Kutuzov, Øvrelid, et al., 2018). Indeed, it is convenient to manipulate word vectors, but past literature also presents the 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 (Tahmasebi et al., 2018).

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)[4], A Representative Corpus of Historical English Registers (ARCHER, 1600-1999)[5] Royal Society Corpus (RSC, 1665-1996)[6], Corpus of Late Modern English Texts (CLMET, 1710-1920)[7], Hansard Corpus (1803-2005)[8], among many others.

In Chinese, the number of diachronic corpora is relatively scarce, including Sheffield Corpus of Chinese[9] and Academia Sinica Ancient Chinese Corpus (中央研究院古漢語語料庫, hereafter ASAC Corpus)[10]. 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.

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 (Tang, 2018). 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 (Liu et al., 2018). 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

[4]https://www.english-corpora.org/coha/
[5]https://www.projects.alc.manchester.ac.uk/archer/
[6]https://fedora.clarin-d.uni-saarland.de/rsc/
[7]https://perswww.kuleuven.be/ u0044428/
[8]https://www.english-corpora.org/hansard/
[9]https://www.dhi.ac.uk/scc/
[10]http://lingcorpus.iis.sinica.edu.tw/early/

Analysis (PCA) and t-distributed Stochastic Neighboring Embedding (t-SNE) (Van der Maaten and Hinton, 2008) 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 (Smilkov et al., 2016). In 2016, 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 (Smilkov et al., 2016).

Coenen et al. (2019) 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, Coenen et al. (2019) 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. Coenen et al. (2019) 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." (Coenen et al., 2019)

It is summarized in Tang (2018) that the novelty of a sense can be understood as the change in sense distribution of different time intervals. The diachornic sense distribution can be visualized based on both word-level and sense-level embeddings (Dubossarsky et al., 2015; Hu et al., 2019). In Dubossarsky et al. (2015), 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 Hu et al. (2019). The rise of a new sense can be depicted in company with other senses in a competitive or cooperative relationship.

Chapter 3

Methodology

3.1 Data Collection and Preprocessing

The textual data in this study is written text of pre-modern Chinese and present-day Chinese, and obtained from three sources, namely CTEXT (Sturgeon, 2019), Academia Sinica Balanced Corpus of Modern Chinese (ASBC) (Chen et al., 1996), and Dcard[1]. 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.

Firstly, the Chinese Text Project collects "pre-modern Chinese texts" with time spanning from 1046 B.C. of the Western Zhou dynasty onward (Sturgeon, 2019). Since the number of texts available from each era varies, and given that the higher the number of texts, the more diverse the content, the time frame with the highest number of texts is chosen to construct the sub-corpora of pre-modern Chinese in this study, and the database can be expanded if data of an earlier era is directly followed by data of the next era. As shown in Table 3.1, texts from Tang, Song, Yuan, Ming, and Qing dynasties are the largest in size, and are continuous in order, and thus are retrieved from the CTEXT database using ctext, a Python wrapper for the CTEXT Application Programming Interface (API) developed by Dr. Donald Sturgeon.

[1]https://www.dcard.tw/

Time span (A.C.)	Number of texts	Number of unique texts
618 – 907 (Tang)	956	623 (-333)
960 – 1279 (Song)	2998	2145 (-853)
1271 – 1368 (Yuan)	991	742 (-249)
1368 – 1644 (Ming)	4248	3497 (-751)
1644 – 1911 (Qing)	9669	7719 (-1950)
Average	3772	2945 (-827)

Table 3.1. Data composition of the CTEXT corpus

Corpus	Time span (A.C.)	All texts		Selected texts	
		Tokens	Types	Tokens	Types
CTEXT	Tang	1.0×10^8	1.2×10^4	3.9×10^7	
	Song	4.5×10^8	1.7×10^4	5.7×10^7	
	Yuan	1.0×10^8	1.2×10^4	4.0×10^7	
	Ming	1.8×10^8	1.5×10^4	7.1×10^7	
	Qing	6.4×10^6	1.5×10^4	8.4×10^7	
ASBC	1981 – 2007	9.2×10^6	2.1×10^5	N/A	N/A
Dcard	2011 – 2019	4.4×10^5	4.9×10^4	N/A	N/A

Table 3.2. Token and type counts of the diachronic corpora

Secondly, ASBC contains articles from the year of 1981 to 2007. The total number of segmented word tokens and type in the corpus is 8,940,871 and 66,562 respectively. In addition, not only are articles in ASBC temporal-labeled, but the texts are also carefully segmented, which makes it an ideal input data for word vector representations.

The third source of data is Dcard, an online social discussion platform established in December, 2011, in Taiwan. In order to retrieve posts with the most diverse topics Dcard API is utilized to retrieve posts from the section of most lately published articles, rather than posts from a number of selected channels.

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. Following the instructions[2], the preprocessing of raw texts is done as described below:

(1) The raw text is cleaned by (a) removing commentaries and marginal notes, (b) segmenting the text into two levels of chucks 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) To compile time-sliced subcorpora with equal size and relevant information, only one version is selected. The CTEXT digital library contains multiple versions of a text converted by different OCR (Optical Character Recognition) techniques, and the metadata includes tags that differentiate versions at varying degrees of OCR accuracy. In the case where no tags are provided, the version with the largest file size is selected, which is also the reason why text cleaning preceeds version selection.

Chinese words are not delimited by space, nor is punctuation systems 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 [3].

Unicode range between U+4e00 and U+9fff are retained and used to construct word embeddings.

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.

[2]https://ctext.org/instructions/wiki-formatting

[3]https://unicode.org/charts/PDF/U3000.pdf

After the completion of preprocessing, this study proceeds to a preliminary quantitative analysis using the R Quanteda library (Benoit et al., 2018). Since it is difficult to infer statistically significant frequency changes because linguistic resources of pre-modern Chinese are essentially insufficient and not of good quality, the bootstrapping method proposed by Lijffijt et al. (2016) is applied to reduce the influence of uneven distribution of linguistic features in texts and provide a more solid ground for the quantitative analysis. 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, as shown in Figure 3.1.

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.

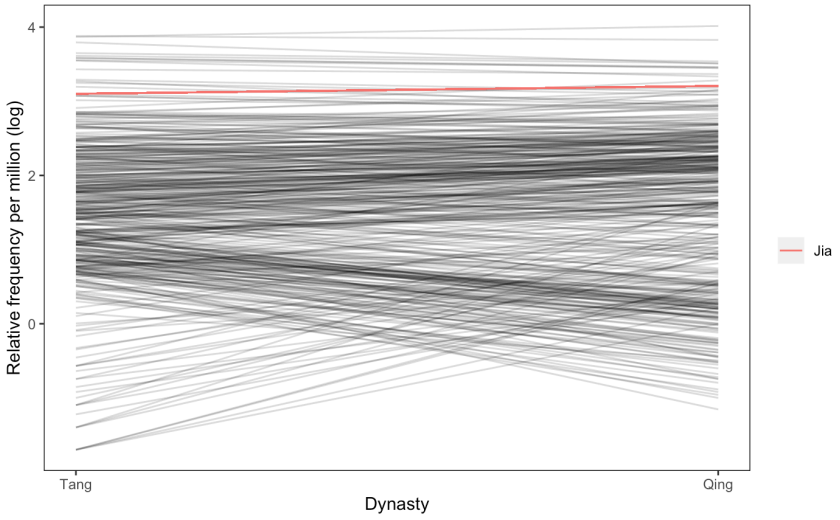


Figure 3.1. Frequency change derived from the bootstrapping test on characters between the Tang and Qing dynasty

* The line in red represents the frequency change of *jiā*.

3.2 The variability-based neighbor clustering method (VNC)

To investigate the semantic change of *jiā* , both word-level and sense-level analyses are employed. To begin with, word-level analysis is performed using the Variability-based neighbor clustering (VNC) method (Gries and Hilpert, 2012) and the Word2Vec algorithm (Mikolov et al., 2013). Proposed by Gries and Hilpert (2012), 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 outliners.

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 (Gablasova et al., 2017). Particularly, the LogDice score is standardized and scaled, and thus comparable across corpora (Gablasova et al., 2017; Rychlỳ, 2008). 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 (Gablasova et al., 2017), 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 (Brezina, 2018) [4].

3.3 Word-level Embeddings

In addition to the analysis by the VNC method, 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 Meng et al. (2019), 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

[4] http://corpora.lancs.ac.uk/stats/toolbox.php

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 Hamilton and Heuser on GitHub (Hamilton et al., 2016b). After the training of Word2Vec embeddings, embeddings are imported to TensorBoard to visualize the data points (Smilkov et al., 2016), and further analyzed in the discussion section.

3.4 Sense-level Embeddings

As for contextualized embeddings, the chosen BERT pre-trained language model is bert-base-chinese (Devlin et al., 2018), 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 (Hamilton et al., 2016b). Jensen–Shannon divergence is used in Giulianelli (2019). Time is not identified when the token representations are extracted.

Chapter 4

Results

4.1 Collocational-based approach

The results of the VNC periodization are plotted as dendrograms (See Figure 4.3, Figure 4.1, and Figure 4.2

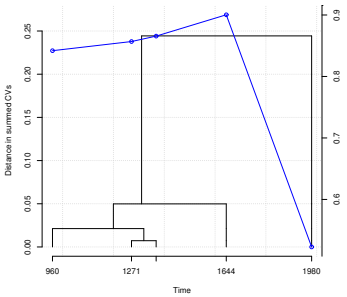


Figure 4.1. VNC periodization of collograms occuring before *jiā*

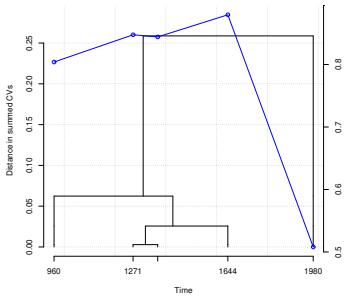


Figure 4.2. VNC periodization of collograms occuring after *jiā*

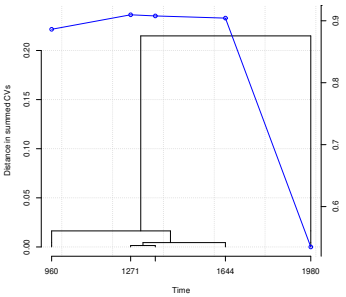


Figure 4.3. VNC periodization of collograms occuring with *jiā*

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.

4.2 Diachronic word embeddings

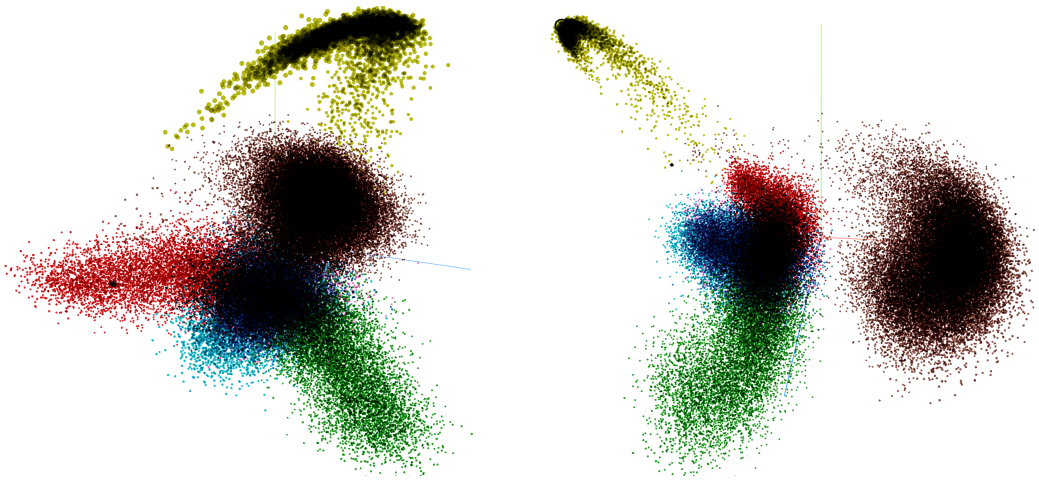


Figure 4.4. Snapshot of PCA Embedding Projector in TensorBoard

* Total variance described: 34.6%.

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

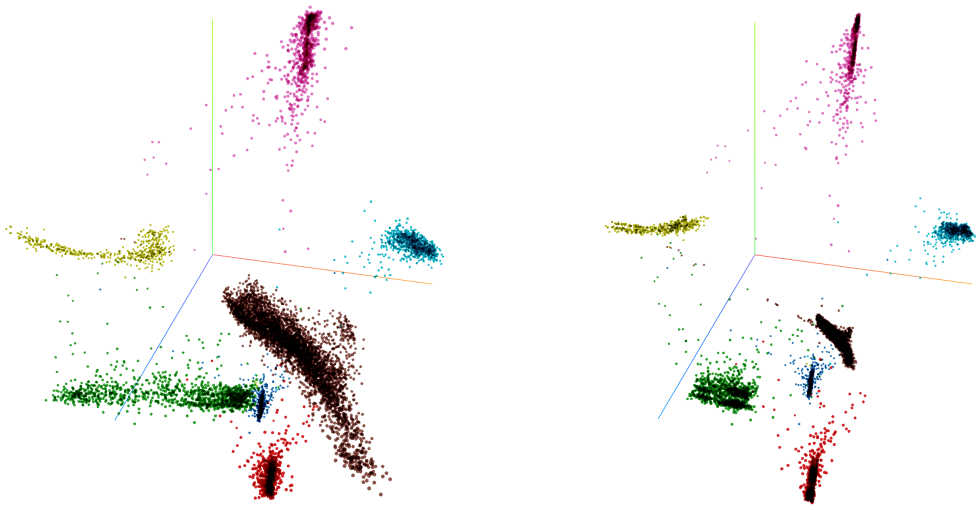


Figure 4.5. Snapshot of t-SNE Embedding Projector in TensorBoard

* 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.

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 (Hamilton et al., 2016a), 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.’

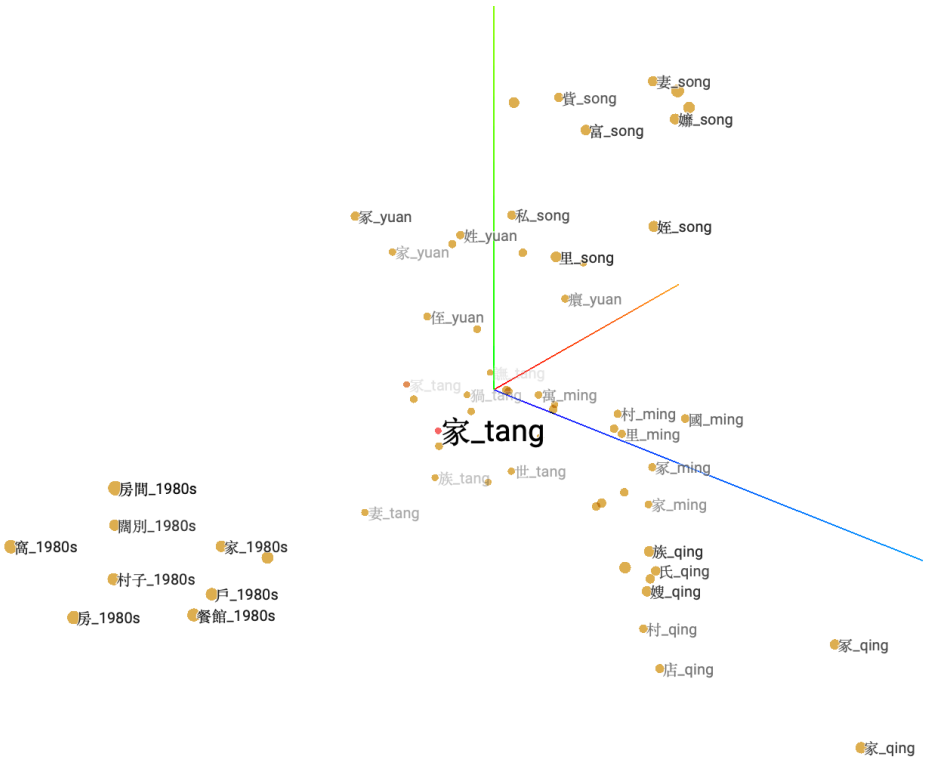


Figure 4.6. Neighboring words of *jiā* projected in a three-dimensional space

Time	Word	Nearest neighboring words									
Tang	家	冢,	族,	冢,	妻,	猧,	世,	富,	譕,	教,	國
Song	家	冢,	族,	里,	富,	貲,	妻,	姪,	私,	貴,	嬸
Yuan	家	冢,	族,	貲,	妻,	睮,	姓,	世,	侄,	癆,	侲
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 *jiātíng* 'family/household', 家人 *jiārén* 'family members', 家族 *jiāzú* 'a family's clan'.

ASBC and Dcard are representative of the concept of jia in the late 20th and 21st century. As Table 2 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.

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 (Mallett, 2004). 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" (Mallett, 2004).

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, and the concept of home as a “journeying” experience recurs in the Dcard corpus, as in li-kai ‘to leave’ (Mallett, 2004; Samanani and Lenhard, 2019). 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 21st 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 an d 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” (Camacho-Collados and Pilehvar, 2018) 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.

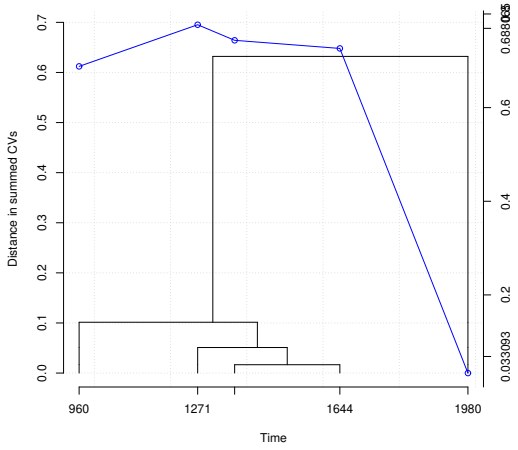


Figure 4.7. VNC results of word-level embeddings

4.3 Diachronic sense embeddings

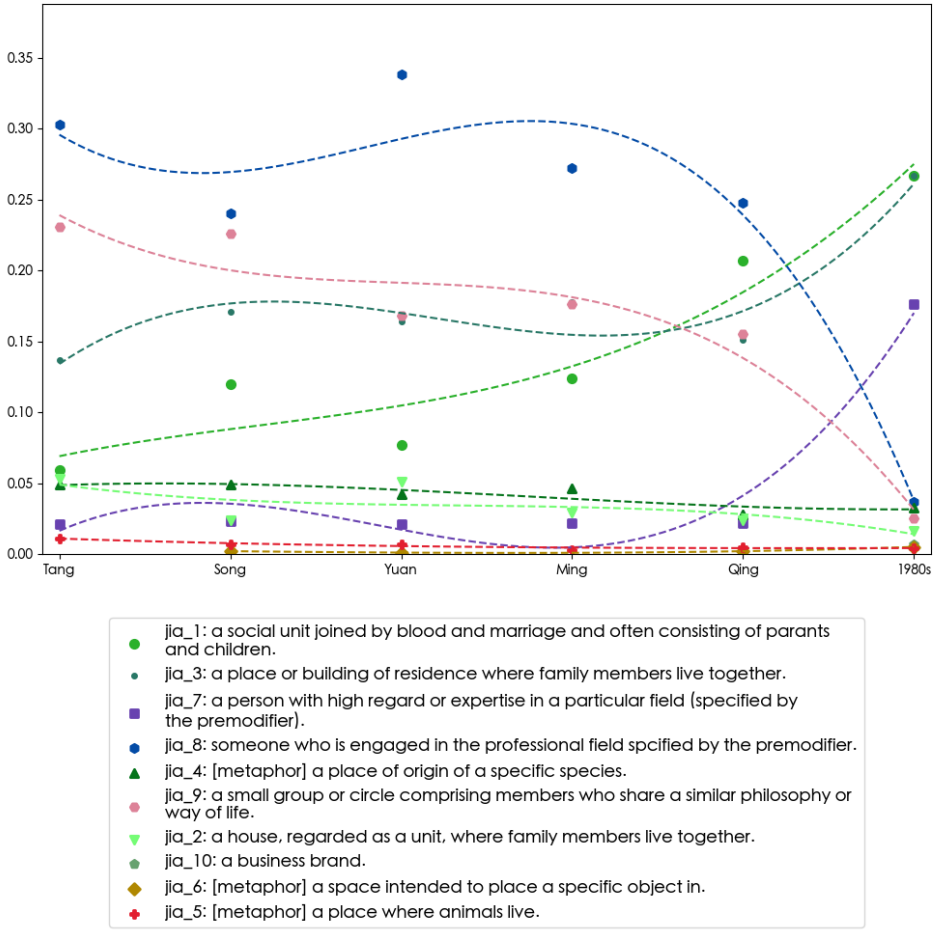


Figure 4.8. Diachronic interactions of senses

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.

Chapter 5

Discussion

Chapter 6

Conclusions

Renouf (2002) recognizes the importance of digitally storing both historical and modern textual data. "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" (Renouf, 2002). Although automatic data crawling and real-time processing makes it possible to build a dynamic, chronological web corpus of present-day language use, as written texts comprise a major part of existing corpora, it is a turning point to explore the diachrony of the data along with the lately available texts from historical periods.

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

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 Traugott and Dasher (2001: 15). 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.

Bibliography

Antoniak, Maria and David Mimno (2018). *Evaluating the stability of embedding-based word similarities*. In: *Transactions of the Association for Computational Linguistics* 6, pp. 107–119.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). *Quanteda: An R package for the quantitative analysis of textual data*. In: *Journal of Open Source Software* 3.30, p. 774.

Bloomfield, Leonard (1933). *Semantic change*. In: *Language*. Allen & Unwin. Chap. 24, pp. 425–443.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2016). *Enriching word vectors with subword information*. In: URL: https://arxiv.org/abs/1607.04606.

Bowern, Claire (2019). *Semantic change and semantic stability: Variation is key*. In: *arXiv preprint arXiv:1906.05760*.

Brezina, Vaclav (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Camacho-Collados, Jose and Mohammad Taher Pilehvar (2018). *From word to sense embeddings: A survey on vector representations of meaning*. In: *Journal of Artificial Intelligence Research* 63, pp. 743–788.

Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu (1996). *Sinica Corpus: Design methodology for balanced corpora*. In: *Language*, pp. 167–176.

Chen, Meng-Ying and Zhao-Qing Fu 沈孟穎, 傅朝卿 (2015). *Transformation of modern residential design in Taiwan: A case study on public housing projects from 1920s to 1960s*. 台灣現代住宅設計之轉化: 以 1920 年代至 1960 年代公共(國民) 住宅為例. In: *Journal of Design.* 設計學報 20.4, pp. 43–62.

Coenen, Andy, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg (2019). *Visualizing and measuring the geometry of BERT*. In: *Advances in Neural Information Processing Systems*, pp. 8594–8603.

Crowley, Terry and Claire Bowern (2010). *Semantic and lexical change*. In: *An introduction to historical linguistics*. 4th ed. Oxford University Press, pp. 199–216.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. In: URL: https://arxiv.org/abs/1810.04805.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman (2015). *A bottom up approach to category mapping and meaning change*. In: *Proceedings of the NetWordS Final Conference*, pp. 66–70.

Firth, John Rupert (1957). *Modes of meaning, papers in linguistics, 1934-1951*. Oxford University Press.

Fortson IV, Benjamin W (2017). *An approach to semantic change*. In: *The Handbook of Historical Linguistics*, pp. 648–666.

Gablasova, Dana, Vaclav Brezina, and Tony McEnery (2017). *Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence*. In: *Language learning* 67.S1, pp. 155–179.

Geeraerts, Dirk (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford University Press.

Giulianelli, Mario (2019). *Lexical semantic change analysis with contextualised word representations*. MA thesis. University of Amsterdam.

Gries, Stefan Th and Martin Hilpert (2012). *Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics*. In: *The Oxford Handbook of the History of English*, pp. 134–144.

Gulordava, Kristina and Marco Baroni (2011). *A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus*. In: *Proceedings of the 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp. 67–71.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016a). *Cultural shift or linguistic drift? Comparing two computational measures of semantic change*. In:

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). NIH Public Access, pp. 2116–2121.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016b). *Diachronic word embeddings reveal statistical laws of semantic change*. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1489–1501.

Hilpert, Martin (2019). *Historical linguistics*. In: *Cognitive Linguistics-A Survey of Linguistic Subfields*, pp. 108–131.

Home (2020). In: *The Oxford English Dictionary*. Last accessed: 2020-09-20. URL: https://www.oed.com/view/Entry/87869?rskey=OqFwzy&result=1#contentWrapper.

Hu, Renfen, Shen Li, and Shichen Liang (2019). *Diachronic sense modeling with deep contextualized word embeddings: An ecological view*. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3899–3908. URL: https://doi.org/10.18653/v1/P19-1379.

Huang, Xiaolei and J. Michael Paul (2019). *Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models*. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4113–4123.

Jia (2015). In: *The MOE Revised Mandarin Chinese Dictionary*. URL: http://dict.revised.moe.edu.tw/cgi-bin/cbdic/gsweb.cgi?o=dcbdic&searchid=W00000005502.

Kutuzov, Andrey and Mario Giulianelli (2020). *UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection*. In: URL: https://arxiv.org/abs/2005.00050.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal (2018). *Diachronic word embeddings and semantic shifts: A survey*. In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pp. 1384–1397.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid (2017). *Tracing armed conflicts with diachronic word embedding models*. In: *Proceedings of the Events and Stories in the News Workshop*, pp. 31–36.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2016). *Significance testing of word frequencies in corpora*. In: *Literary and Linguistic Computing* 31.2, pp. 374–397.

Liu, Shusen, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci (2018). *Visual exploration of semantic relationships in neural word embeddings*. In: *IEEE transactions on visualization and computer graphics* 24.1, pp. 553–562.

Mair, Christian (1998). *Corpora and the study of the major varieties of English: Issues and results*. In: *The major varieties of English: Papers from MAVEN 97*, pp. 139–158.

Mallett, Shelley (2004). *Understanding home: A critical review of the literature*. In: *The sociological review* 52.1, pp. 62–89.

Meng, Yuxian, Xiaoya Li, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li (2019). *Is word segmentation necessary for deep learning of Chinese representations?* In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 3242–3252.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient estimation of word representations in vector space*. In: URL: https://arxiv.org/abs/1301.3781.

Miller, George A. and Walter G. Charles (1991/2007). *Contextual correlates of semantic similarity*. In: *Language and Cognitive Processes* 6.1, pp. 1–28.

Moore, Jeanne (2000). *Placing home in context*. In: *Journal of environmental psychology* 20.3, pp. 207–217.

Nerlich, Brigitte and David D. Clarke (2001). *Serial metonymy: A study of reference-based polysemisation*. In: *Journal of Historical Pragmatics* 2.2, pp. 245–272.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). *Glove: Global vectors for word representation*. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Renouf, Antoinette (2002). *The time dimension in modern English corpus linguistics*. In: *Teaching and learning by doing corpus analysis*. Brill Rodopi, pp. 27–41.

Robert, Stéphane (2008). *Words and their meanings: Principles of variation and stabilization*. In: *From polysemy to semantic change: Towards a typology of lex-*

ical semantic associations. Ed. by Martine Vanhove. Vol. 106. John Benjamins, pp. 55–92.

Rychlỳ, Pavel (2008). *A lexicographer-friendly association score*. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN)*, pp. 6–9.

Samanani, Farhan and Johannes Lenhard (2019). *House and home*. In: *The Cambridge Encyclopedia of Anthropology*. Ed. by Felix Stein, Sian Lazar, Matei Candea, Hildegard Diemberger, Joel Robbins, Andrew Sanchez, and Rupert Stasch. URL: http://doi.org/10.29164/19home.

Sinclair, John (1982). *Reflections on computer corpora in English language research*. In: *Computer corpora in English language research*, pp. 1–6.

Sixsmith, Judith (1986). *The meaning of home: An exploratory study of environmental experience*. In: *Journal of environmental psychology* 6.4, pp. 281–298.

Smilkov, Daniel, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg (2016). *Embedding Projector: Interactive visualization and interpretation of embeddings*. In: URL: https://arxiv.org/pdf/1611.05469v1.pdf.

Sturgeon, Donald (2019). *Chinese Text Project: A dynamic digital library of premodern Chinese*. In: *Digital Scholarship in the Humanities*. URL: https://doi.org/10.1093/llc/fqz046.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt (2018). *Survey of computational approaches to diachronic conceptual change*. In: URL: https://arxiv.org/abs/1811.06278.

Tang, Xuri (2018). *A state-of-the-art of semantic change computation*. In: *Natural Language Engineering* 24.5, pp. 649–676.

Traugott, Elizabeth Closs (2017). *Semantic change*. In: *Oxford Research Encyclopedia of Linguistics*.

Traugott, Elizabeth Closs and Richard B Dasher (2001). *Regularity in semantic change*. Cambridge University Press.

Van der Maaten, Laurens and Geoffrey Hinton (2008). *Visualizing data using t-SNE*. In: *Journal of Machine Learning Research* 9, pp. 2579–2605.

Zellig, Harris (1954/2015). *Distributional structure*. In: *Word* 10.2-3, pp. 146–162.

Appendix A

Title of Appendix A

Appendix B

Title of Appendix B