國立臺灣大學文學院語言學研究所
碩士論文口試本 Graduate Institute of Linguistics

College of Liberal Arts

National Taiwan University
Master Thesis

論文題目
Semantic change in the diachronic perspective: 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.

November 2020
中華民國 109 年 11 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.

Keywords: Semantic change, diachronic semantics, meaning representations

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。

關鍵詞：語意變遷、歷時語意、向量表徵

Table of Contents

List of Figures

List of Tables

Chapter 1

Related works

1.1 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**). In **hamilton2016cultural**, 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 (**geeraerts1997diachronictraugott2001regularity**). 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 (**traugott2001regularity**). Onamasiology, on the other hand, focuses on synonyms, near-synonyms, and naming-gaving (**geeraerts1997diachronic**).

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.

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 (**traugott2001regularity**).

Polysemy, described as "families of related meanings" in **traugott2001regularity**, 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" **traugott2001regularity**. 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 (**robertinvanhove2008**). The construction of meanings is flexible and sensitive to the context of use (**miller1991contextual**; **harris1954distributional**). Additionally, the mechanism of metonymy allows the co-existence of referential and conceptual meanings in the same word (**hilpert2019historical**; **nerlich2001serial**). 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 (**fortson2017approach**). 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 *ac-*

tually, “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. Nonetheless, semantic change is a complicated phenomenon resulting from not only polysemy, but also subjectification (**traugott2001regularity**), prototypicality (**geeraerts1997diachronic**), and other contributing factors. Linguistic variations of language use is omnipresent in the synchronic settings, but is amplified in a diachronic scope (**semanticincrowley2010**; **bowern2019semantic**).

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 (**traugott2001regularity**).

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**).

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 **renouf2002timemair1998corpora** 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.

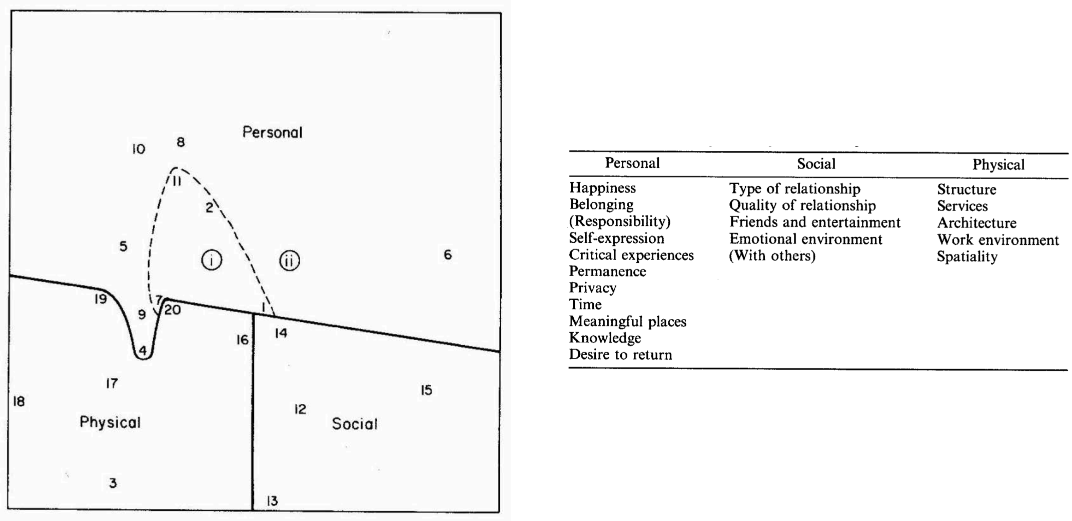
1.2 The Concept of Home in Literature

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**).

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**).



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 1.1. 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.

1.3 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**). 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**).

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[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.

[1] http://books.google.com/ngrams. A comprehensive review of diachronic corpora is provided by **tahmasebi2018survey**

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

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

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.

Literature	Use cases
kulkarni2015statistically	apple, tape
hamilton2016law	gay, broadcast, awful*
hamilton2016cultural	actually, must, promise, gay, virus, cell
kutuzov2017tracing	war, peace, stable
rodda2017panta	πνεῦμα 'breath' → 'spirit' (Ancient Greek)
yao2018dynamic	apple, amazon, obama, and trump
rudolph2018dynamic	intelligence, iraq, jobs, prostitution
antoniak2018evaluating	marijuana
hu2019diachronic	please, alien
rodina2020elmo	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 1.1. Example case studies from literature

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.

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.

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 (**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**).

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 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**).

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.

1.4 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

[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/

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 diachornic 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.

Chapter 2

Methodology

2.1 Data Collection and Preprocessing

renouf2002time 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” (**renouf2002time**). In **sinclair1982reflections**, **sinclair1982reflections** envisions the possibility of having “vast, slowly changing stores of text” that provide “detailed evidence of language evolution” (**renouf2002time**). As written texts comprise a major part of existing corpora, 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, 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)[1] (**sturgeon2019ctext**) and Academia Sinica Balanced Corpus of Modern Chinese (中研院漢語平衡語料庫, hereinafter ASBC)[2] (**chen1996sinica**) 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.

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

[1] https://ctext.org/

[2] http://asbc.iis.sinica.edu.tw/

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 chosen 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[3], a Python wrapper of the same name developed by **ctextapi**. Apart from the provision of the API (Application Programming Interface) access, the CTEXT project website is informative of how textual data and metadata are stored in the retrieved format[4]. Since multiple versions of a text are likely to be produced using different OCR (Optical Character Recognition) techniques, only one version labeled as representative of a set of texts is selected, or, if needed, all versions are retained to help discern the differences in the converted texts. In the case where no tags are provided, the version with the largest file size is selected. 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. The number of texts are summarized in Table **??**.

Time span (A.C.)	Number of texts	Number of unique texts
618 – 907 (Tang)	956	623 (-333)
960 – 1279 (Song)	2,998	2,145 (-853)
1271 – 1368 (Yuan)	991	742 (-249)
1368 – 1644 (Ming)	4,248	3,497 (-751)
1644 – 1911 (Qing)	9,669	7,719 (-1,950)
Total	18,862	14,726 (-4,136)

Table 2.1. Data composition of the CTEXT corpus

Secondly, 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. In addition, not only are articles in ASBC temporal-labeled, but the texts are also carefully segmented, which , which is considered representative of language use of modern Chinese. Therefore, the choice of CTEXT and

[3] https://pypi.org/project/ctext/

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

ASBC suits the language settings for this study.

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)

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 [5].

A dependency parser for classical Chinese is released based on historical corpus of Four Books (四書) (**yasuoka2019universal**), with dictionary-based tokenizer and POS taggers[6]. In this study, character-based embeddings are generated, and building upon the character-based embeddings, the results can serve as comparison basis with token-based ones.

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.

One of the difficulty in the language processing of pre-modern Chinese lies in the 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 **??**

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

[6] https://pypi.org/project/udkanbun/

Corpus	Time span (A.C.)	All texts		Selected texts	
		Tokens	Types	Tokens	Types
CTEXT	Tang	1.0×10^8	1.2×10^4	4.9×10^7	1.2×10^4
	Song	4.5×10^8	1.7×10^4	2.6×10^8	1.6×10^4
	Yuan	1.0×10^8	1.2×10^4	6.0×10^7	1.1×10^4
	Ming	7.1×10^8	1.7×10^4	5.2×10^8	1.7×10^4
	Qing	1.6×10^9	2.9×10^4	1.1×10^9	2.2×10^4
ASBC	1981 – 2007	8.9×10^6	6.6×10^4	N/A	N/A

Table 2.2. Token and type counts of the diachronic corpora

After the completion of preprocessing, this study proceeds to a preliminary quantitative analysis using the R Quanteda library (**quanteda**). 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 **lijffijt2016bootstrap** 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 **??**.

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.

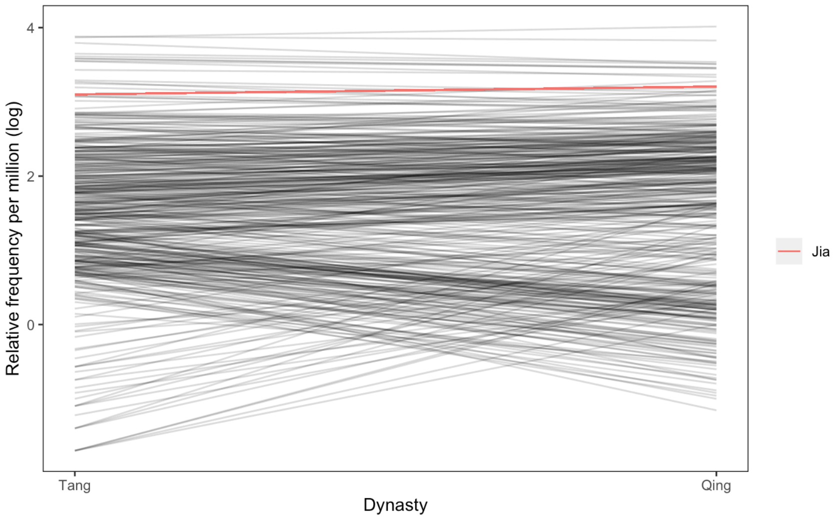


Figure 2.1. Frequency change derived from the bootstrapping test on characters between the Tang and Qing dynasty

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

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). Instead of 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. Specifically, the top 25[7] 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**.

2.2 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 (**gries2012variability**) and the Word2Vec algorithm (**mikolov2013efficient**). Proposed by **gries2012variability**, the VNC method is used

[7] In **hamilton2016cultural**, the range between 10 and 50 is recommended as their results reflect.

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 (**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**) [8].

2.3 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 **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 Hamilton and Heuser on GitHub (**hamilton2016law**). 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

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

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 stablization, 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.

2.4 Sense-level Embeddings

As for contextualized embeddings, the chosen BERT pre-trained language model is bert-base-chinese (**devlin2018bert**), 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.

Contextualized word representations, or usage representations, as termed in (**giulianelli2019lexical**), since the extracted representations reflect an usage-based data.

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.

Chapter 3

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 neural in connotation in the past (王 **1997** 古).

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[1] and 近代漢語語料庫詞頻統計[2]

[1] http://corpus.zhonghuayuwen.org/resources.aspx

[2] https://elearning.ling.sinica.edu.tw/jindai.html

Chapter 4

Conclusions

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 **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.

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.

Appendix A

Title of Appendix A

Appendix B

Title of Appendix B