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

詞向量的語意變遷計算模型:
以「家」為例

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

陳蓓怡

Pei-Yi Chen

指導教授：謝舒凱博士

Advisor: Shu-Kai Hsieh, Ph.D.

May 2021

中華民國 110 年 5 月

doi:10.6342/NTU202004346

致謝辭

因為論文進度緩慢，致謝辭也像是感謝日記一樣，一點一滴記錄下來，感謝在撰寫論文的期間受到的各式幫助。

謝謝我的指導老師謝舒凱老師，在我提出想要以「家」的概念演變為題目時，就一路指引我從各個角度切入這個主題，在論文完成的過程中，對於我的任性總是無比包容，一步一步接近知識的殿堂，同時也不忘像個點燈人，照亮晦暗的角落。

謝謝我的口試委員謝舒凱老師、呂佳蓉老師以及張瑜芸老師，對論文提出許多珍貴的建議。

謝謝 LOPE 實驗室的大家，在掙扎時總有許多援手幫忙。Taco, Hanna, Sabrina, Roxanne, Richard, Freya, Ben, Don, Debby, Yolanda, Jessica, Yongfu, Andrea, 貿昌, Simon, Amber, 智堯, 鈺琳, 飛揚, Joy

慧宇, 殷縈, Sherry, Sam, Brian, Ree, CJ。每個學期結束的爬山行程給我很大的動力，遠離枕頭山 XD。

Alyssa, 海格房的室友們 Irene, Wave, Eva, 乃甄, 小玉姐姐, Rachel, Sylvia, Sandy，忙碌的日子裡看到你們不斷地前進。Tom, Leston，捎來關心，煮好吃的食物共食。

Miffy, Thomas, 姿妤學姊

謝謝冠鳴，總是了解我的個性，以及爸媽與弟弟還有家裡的三隻貓。

doi:10.6342/NTU202004346

摘要

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

語言，將所思所想傳遞、紀錄，並在說話者使用語言時，不斷被重塑與流傳 (Blank, 1999: 61)。從共時（synchronic）的角度來看，語意存在各種變異（variation），而在歷時（diachronic）的脈絡下，經過時間累積而則彰顯了各種的變遷。近年來的歷史詞彙語意研究，從詞意的改變、新舊字詞的興衰，探索其背後的運作機制與認知層面，已開始摸索出語意變遷（semantic change）的規律性（regularities）(Blank, 1999: 63)。語料庫作為語言使用的經驗素材，提供了我們從中觀察、歸納出可質化、量化的語言分析；而歷時語料庫更因應科技進步，結合了計算語言學界近年來的語言向量表徵、神經語言統計模型等新方式探求語意在時間洪流下的變動與趨勢。

然而在歷時語料中，有些詞彙並無明顯的詞頻變化，其多義行為亦造成研究者面對巨量資料時的困擾。本論文的目的，在於結合語料統計模型與計算語意學的表徵模型，探究漢語的語意變遷。從數位化的原始語料中，以共現（co-occurrence）分佈的趨勢發覺意義分布的異同，並從語境詞向量（contextualized word embeddings）將多義性（polysemy）的變動做形式表達。期待以量化的方式量測語意變遷的程度，並以質化分析輔證已知的例子，並發掘更多可能的例子與規律。我們以歷時語料庫（中國哲學書電子計畫 (Sturgeon, 2019)）與現代漢語語料庫（中研院漢語平衡語料庫 (Chen et al., 1996)）為語料來源，建立歷時詞向量並搭配詞彙資料庫，並參考 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來比較各時代向量的方法，求其相關係數和語意變遷程度之間的關聯。並從詞彙的意義分布與互動，描繪出不同詞意的消長與變動。此外，本研究也同

doi:10.6342/NTU202004346

時採用以變異程度為基礎的近鄰群聚分析法（Variability-based Neighbor Clustering, VNC）(Gries and Hilpert, 2012)，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整合研究可以使我們在經驗基礎上回溯驗證個別詞彙的意義變化，更進一步梳理整體的原理原則。詞彙反映人們對於新事物賦予新名的動機、社會概念的更迭也同時牽動詞彙之間的關聯。本研究的應用範圍更可擴及到詞彙與文化變遷的探索。

關鍵詞：語意變遷、歷時語意、向量表徵、階層式集群

doi:10.6342/NTU202004346

Abstract

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

Language is a means of communication through which ideas are conveyed, stored, and recorded, and in essence, constant change and evolution occurs as the speakers use the language with the passage of time (Blank, 1999: 61).

The dynamics of meaning construction is embodied in the emergence and loss of senses, as well as the split and shifts, which contributes to the different distributions and interactions of words, reflects the regularities and adaptability of the language, and the cognition and culture operating behind (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

doi:10.6342/NTU202004346

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

The diachronic corpus consists of texts from the following sources: the Chinese Text Project (Sturgeon, 2019) and Academia Sinica Balanced Corpus of Modern Chinese for modern Chinese (Chen et al., 1996). 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. Finally, this study will proceed with periodization analysis using the Variability-based Neighbor Clustering (VNC) method (Gries and Hilpert, 2012). As a hierarchical clustering method, it is bottom-up, as opposite to the decisive clustering, a comprehensive evaluation of the influence of the selected linguistic factors in this study is implemented to explore how the development of meaning construction can be understood under different stages. In sum, this study explores the phenomenon of semantic change in retrospect to derive the semantic development in diachrony. The computational/statistical modeling of historical lexical semantic change will shed new light on how the language community describes and makes sense of the society that is also constantly changing.

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

doi:10.6342/NTU202004346

Table of Contents

doi:10.6342/NTU202004346

doi:10.6342/NTU202004346

List of Figures

doi:10.6342/NTU202004346

doi:10.6342/NTU202004346

List of Tables

doi:10.6342/NTU202004346

Chapter 1

Introduction

1.1 Computational-historical Analysis of Semantic Change

Language is constantly changing and evolving. The emergence of new senses, the demise of old ones, and the polysemous nature of linguistic expressions make the process of semantic change a dynamic phenomenon (Robert, 2008). As individuals learn new words and meanings throughout their life, so does a language. As language users actively engage in processing and interpreting the language, the semantic history of words are woven into the texts that then survive time and are presented to us now. In the long run, a word is likely to convey a meaning completely different or unfathomable. For instance, "the quick and the dead", quoted from the Bible, means "the living and the dead", but the collective adjective "the quick" no longer makes sense in Present-Day English (Crowley and Bowern, 2010: 199).

The nature of language is reflected in its use. In 1982, Sinclair envisions the possibility of "vast, slowing changing stores of text" and "detailed evidence of language evolution"

doi:10.6342/NTU202004346

(as cited in Renouf, 2002). In the recent years, a huge amount of historical text data have been digitized and made available to the public, and the use of digitized libraries as rich linguistic resources to observe how certain linguistic features are "assimilated" into the language becomes more and more feasible (Renouf, 2002). 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).

With the recent advances in Natural Language Processing (NLP) techniques, the changes in meaning over time can be to a great extent captured by representing discrete linguistic data as numeric vectors such as word embeddings, especially after the release of Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2016). For instance, for the study of semantic change of individual words across time, initial efforts have been put into generating word embeddings from different time spans and explore whether semantic change occurs based on the neighboring words of the target word from each time period.

As the pioneering computational-historical investigation in Mandarin, the monosyllabic word 家 *jiā* 'home' is selected as a case study in this thesis. The concept of home is an ancient, seemingly familiar and encompassing, but tangible one. Various humanities disciplines have sought to grasp the full picture. Defined by the Oxford English Dictionary (OED), the word *home* is "the place where a person or animal dwells" ("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

doi:10.6342/NTU202004346

we feel a “sense of belonging [and] comfort”, and even a person’s “country or native land.” In Mandarin Chinese, the MOE Revised Mandarin Chinese Dictionary defines its translated equivalent 家 *jiā* ‘a’s “a place where family members live together (眷屬共同生活的場所)”, “a private property (私有財產)”, and “people in certain professional fields (經營某種行業或具有某種身份的人)” (“Jia”, 2015). Yet, how is the concept of home encoded linguistically? Specifically, how its diachrony interacts with synchrony and variations is the main concern of this study.

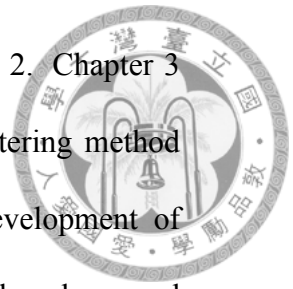
1.2 Research Questions

From the perspective of corpus-based computational linguistics, research questions are invoked as to how the concept of home is properly computationally represented? What words are semantically related to this concept? and how are these words co-construct the meanings of home, and how this concept comes into shape through the lens of time. In this study, the research questions are proposed as follows: (1) How can diachronic embeddings be applied to textual data of pre-modern Chinese? (2) Besides the linguistic factor of frequency change, how is semantic change of *jiā* be reflected in distributional ways? (3) What cultural implications can diachronic embeddings contribute to pre-modern Chinese? The research questions to be answered through a corpus-based case study approach along with diachronic word embeddings to investigate the evolution of meaning change in the target word *jiā*.

1.3 Organization of the Study

The remainder of this thesis is organized as follows. An theoretical overview and reflections of lexical semantic change in general, the concept of home in literature, as

doi:10.6342/NTU202004346



well as the diachronic word embeddings techniques are given in Chapter 2. Chapter 3 introduces the preprocessing issues, and the proposed corpus-based clustering method and distributed semantic representation models for the study. The development of word-level and sense-level word representations brings to the fine-grained analyses and generalizations of semantic change. Chapter 4 describe how the proposed approaches are evaluated, and showcase analyses made possible by our approach, and discusses their successes and limitations. Finally, Chapter 5 concludes with a summary of the contributions and with considerations on the future works as well as on its usefulness to linguistic investigations and other social-cultural applications.

doi:10.6342/NTU202004346

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).

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,

doi:10.6342/NTU202004346

and the latter includes hyperbole, metaphor, euphemism, interference, folk etymology, and hypercorrection. Whether an instance of semantic change is bifurcation or shift is determined by the absence of the original sense. Semantic shift is reflected in the cognate words from target languages, which do not come to have the new meaning. In terms of hyperbole, words in constant use become more and more neutral. Interference describes the semantic relations of synonyms or homonyms; other word are in place to avoid confusion in communication.

The main types of semantic change —of which e.g. Traugott (2017) offers historical examples are as follows (quoted from (Giulianelli, 2019: 6)):

(1) broadening (or generalization): the extension of the range of concepts designated by a term,

(2) narrowing (or specialization): the contraction of the range of concepts designated by a term,

(3) metaphorization: the conceptualization of one referent in terms of another, guided by analogical reasoning and implying an unspoken simile,

(4) metonymization: a meaning transfer from one word to another, guided by spatial, temporal or causal contiguity between the two referents,

(5) amelioration: the acquisition of or shift towards a positive connotation,

(6) pejoration: the acquisition of or shift towards a negative connotation.

Traugott and Dasher (2001: 81) also noted that meaning change often occurs in the direction from concrete to abstract. Originally, a lexical item bears contentful meaning. During grammaticalization, grammatical or procedural meaning is enriched although the contentful one might persist.

doi:10.6342/NTU202004346

Depending on the initial step of investigation, semantic change can be approached from a semasiological and onamasiological perspective (Geeraerts, 1997: 17; Traugott and Dasher, 2001: 25). A semasiological perspective highlights the direction from linguistic expression to concept, so meaning change is studied under the consideration of a lexeme in a fixed, predetermined form. Conversely, an onamasiological perspective starts from concept to linguistic expression, and thus meaning change is framed within a given concept expressed by a set of alternative words. Nonetheless, both of the two complementary paths lead to such important topics in lexicology as polysemy and sense relations. Semasiologically, when a lexeme undergoes semantic change and additional meanings are gained, the different senses might gradually be perceived as unrelated to one another by the language users. That is, the lexeme first becomes polysemous, and then homonymous (Traugott and Dasher, 2001: 25). Onamasiologically, on the other hand, focuses on synonyms, nearsynonyms, and name-giving to connect lexical items with sense relations that exist and develop under a concept over time (Geeraerts, 1997: 17).

Polysemy, for instance, goes hand in hand with the semasiological view. It is 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 coexistence of older and newer meanings in a lexical item, along with the influence of multiple meanings on one another, brings about the dynamics of "saliency" (Traugott and Dasher, 2001: 12). Being polysemous with more than one single semantic reading is not only necessary but also omnipresent. In particular, synchronic polysemy is pointed out as an integral component among the driving forces of lexical semantic change, a phenomenon that is often explored in a diachronic vein (Robert, 2008).

doi:10.6342/NTU202004346

As a topic that has long interested scholars in semantics and historical linguistics, semantic change is a complicated phenomenon resulting from an interplay of polysemy, with subjectification (Traugott and Dasher, 2001), prototypicality (Geeraerts, 1997), and other contributing factors. Semantic change has been extensively studied because linguistic variations of language use are pervasive in the synchronic settings, and are amplified in a diachronic scope (Bowern, 2019; Crowley and Bowern, 2010). The term "brachychrony" is even coined by 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.

The Invited Inferencing Theory of Semantic Change (IITSC) is proposed by Traugott and Dasher (2001: 34–40) to account for the actuation of meanings through recognition of different stages of a linguistic expression depending on whether intended meanings are coded or crystallized into commonly used implicatures. In other words, the degree of conventionality reflects the stages in which an expression is during a certain period of time. The more conventional or less context-specific an expression is, the more crystalized or coded the meaning is conveyed through this expression, which indicates that the expression has evolved in the later stages of the IITSC. Importantly, the meaning of an expression is not limited to only one, but a second reading often becomes more and more readily accessible as the coded meaning, and is acceptable by the language community. For example, through expressions of temporal sequence, invited inferences of causality can arise. Over time, semantic change follows a path from coded meanings to utterancetoken meanings to utterancetype, pragmatically polysemous meanings (GIINs) to new semantically coded meanings. That is, a new meaning emerges as a creative, innovative instance of language use by an individual and does not yet spread to a wider

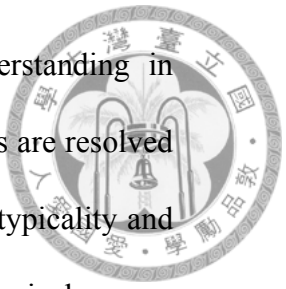
doi:10.6342/NTU202004346

language community, but remains more idiosyncratic. Slowly, it is likely that the new meaning is acquired socially with strengthened pragmatic impact, the expression is then pragmatically polysemous. The final stage of the evolution cycle is for the expression to be semantically polysemous or coded, with the new meaning being the dominant or salient reading.

Geeraerts (1997) puts forwards a conceptual framework that describes semasiological change motivated by the prototypicality theory. Extensionally, members of a semantic category do not have equal representativeness or typicality of the category, and their membership can even be uncertain if the member is highly peripheral. Intensionally, meanings of less typical members are received from the more salient meanings and can overlap, yet the salient meanings are not determined solely from one single cluster of attributes. Generally, the synchronic semantic structure of lexical categories echoes with the diachronic semasiological change. Diachronically, the more salient the meaning, the more stable it is. When semantic change takes place, the expansion of referential range denoted by a meaning is extended from the prototypical center to the peripheral area. Consequently, the peripheral area will have less and less in common with the prototypical center. It is also possible that a meaning of a lexical item is a combination of features that do not belong to the same cluster at all. Meanwhile, considering the uncertain boundaries to be drawn for a lexical item, its semantic history might involve discontinuous appearance of an identical meaning that is temporally unrelated to each other rather than resulting from textual evidences that do not survive time.

Under this conceptual framework, the flexibility of meaning construction relies on the adaptability and dynamics of human cognition that groups and regroups meanings to meet the need of cognitive efficiency. Building upon the distinction between speaker-oriented

doi:10.6342/NTU202004346



and hearer-oriented process to avoid possible communicative misunderstanding in phonology, this framework adopts a similar notion that homonymic clashes are resolved with opportunities of semantic change, including a tendency toward prototypicality and morphological transparency while striking a balance for as many morphological uses as possible.

For language speakers, the construction of meanings is flexible and sensitive to the context of use, in which ambiguity is resolved or cancelled (Miller and Charles, 1991/2007; Zellig, 1954/2015). Additionally, the operation of metonymy is a mechanism that plays a practical role in meaning construction, for this mechanism allows a word to carry referential and conceptual meanings simultaneously (Hilpert, 2019; Nerlich and Clarke, 2001). From the perspective of semantic change, an understanding of metonymic change, specifically, builds upon the familiarity of the culture in which the language is spoken, and therefore the attested examples in literature exhibit a rich diversity (Fortson IV, 2017: 649). Yet, the semantic history of a word might also unfold beyond the intuition of the language users. It is recognized that synchronically distinct meanings, which speakers of the given time period find conceptually related, might suggest otherwise, as in *bachelor*, for a relationship exists between "experiencing" and "evoking", and *actually*, "unexpectedness" and "elaboration" (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.

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).

doi:10.6342/NTU202004346

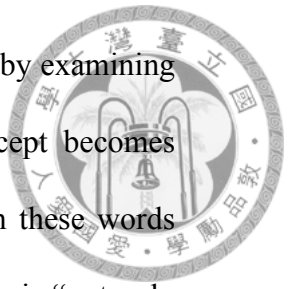
Recently, the application of computation to larger sets of words across longer periods of time enables the generalization of regularities on semantic change (Hamilton et al., 2016b). Semantic change driven by technological innovations are prominent examples, while shifts of meanings with linguistic cause 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 Mair (1998) Renouf (2002) to refer to a time span of 10 to 30 years, indicating how the change of a linguistic feature can be delineated within a short time frame.

For Classical Chinese, B. Li (2020) used the dependency parser trained on Kyoto Corpus of the Four Books to explore change of syntactic categories of Classical Chinese, yet a character-based analysis is adopted due to the segmentation issue of pre-modern Chinese. However, contrary to the assertion that pre-modern Chinese is mostly monosyllabic, the disyllabic development of Chinese has started as early as the Han dynasty (Zhang 張小平, 2008; Zhou 周俊勛, 2009), but the proposal by Lee (2012) of the nested multi-level segmentation is able to reflect the complicated word segmentation challenge for languages like (pre-modern) Chinese (as cited in B. Li, 2020). However, the results show that tokenizers such as MeCab-Kanbun and Stanza segment words by characters, and verbs like 吃 'eat' or 食 'eat' might be tagged as noun.

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

doi:10.6342/NTU202004346



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 are plotted as three regions from questionnaire responses, as shown

doi:10.6342/NTU202004346

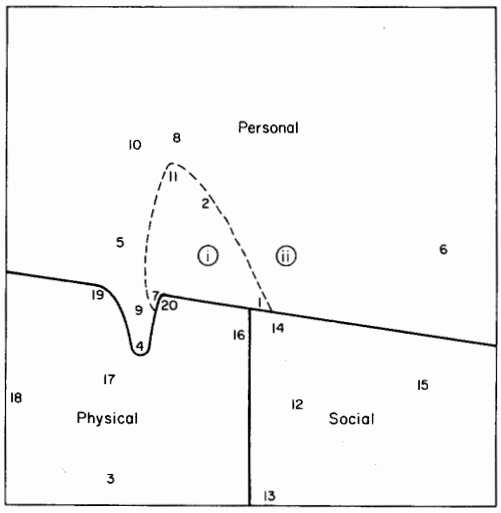
in Figure 2.1. The "diversity" of the meanings of home is the motivation behind the research by Sixsmith (1986). Home exists as a physical entity. Through styling and living, the house is transformed into a home. Home can even be used to describe any level of existential space, including neighborhood, town, city, and country, as well as having cultural expression attached to the meanings of home. The phenomenologically-based research collects empirical evidence from questionnaires under the framework of the "person-environment unity" that incorporates referents of places and actual experiences lived by people.

Through a method called "multiple sorting task", Sixsmith (1986) collects open-ended, participant-generated categories of home and sorting criteria. That is, the participants list categories of home and sort these categories according to a specific criterion they think of, and the procedure is repeated multiple times until all possible descriptions and orders have been attempted by the participants. This research is distinguished by the use of non-prescribed answers to depict the meanings of home from the perspective of the participants themselves. Through further transcription and categorization, the results have interwoven otherwise often disparate ideas of what home means statistically through multidimensional scaling technique, as shown in Figure 2.1.

Culturally, the concept of home in Taiwan as a physical space has undergone changes caused by the sway of the world order (Shen 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

doi:10.6342/NTU202004346

preserved in places like Dihua Street and Dadaocheng Area. Linguistically, Wang and Gou 王雲路, 郭穎 (2005) have discussed the morphological development of *jiā* in pre-modern Chinese.



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)).

2.3 Diachronic/historical Corpora

The compilation of corpora to include historical texts and annotations enables more detailed linguistic analysis. Examples include the Corpus of Historical American English (COHA, 1810-2000)[1], A Representative Corpus of Historical English Registers (ARCHER, 1600-1999)[2] Royal Society Corpus (RSC, 1665-1996)[3], Corpus of Late Modern English Texts (CLMET, 1710-1920)[4], Hansard Corpus (1803-2005)[5], among many others.

In Chinese, the number of diachronic corpora is relatively scarce, including Sheffield

[1]https://www.english-corpora.org/coha/
[2]https://www.projects.alc.manchester.ac.uk/archer/
[3]https://fedora.clarin-d.uni-saarland.de/rsc/
[4]https://perswww.kuleuven.be/ u0044428/
[5]https://www.english-corpora.org/hansard/

doi:10.6342/NTU202004346

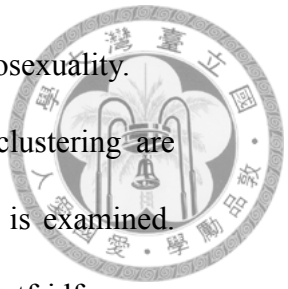
Corpus of Chinese[6] and Academia Sinica Ancient Chinese Corpus (中央研究院古漢語語料庫, hereafter ASAC Corpus)[7] (Wei et al. 魏培泉等, 1997). The Google books project for Chinese is not available until the year of 1950, and the latest date is 1999. The ASAC Corpus is divided into 3 sub-corpora based on the development 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 Topics-Over-Time (TOT)

Besides vector space models, topic models like Latent Dirichlet Allocation (LDA) are also widely applied to the study of semantic change, e.g., Wang and McCallum (2006), Wijaya and Yeniterzi (2011), and Hengchen (2017). As an extension to topic models, Topics-Over-Time (TOT) treats each year or each time slice as a document, and detects semantic change through top words used in documents and topics generated during the modeling. In practice, topic probability distribution is computed for each target word in the vocabulary of a specific time period, and word senses are derived from the topic distribution. Take the word *gay* as an example. When the number of topics is set to 2, a trend of topic distribution change can be seen for the word, which echoes with the meaning

[6]https://www.dhi.ac.uk/scc/
[7]http://lingcorpus.iis.sinica.edu.tw/early/

doi:10.6342/NTU202004346



change of the word from happiness or cheerfulness as an adjective, to homosexuality.

In companion with topic models, clustering methods like k-means clustering are insightful when the topic density of each cluster of a given time period is examined. The results of k-means clustering show that top words with the highest tf-idf scores do not belong to the same clusters, indicating that these words are diverse in meaning contribution. Specifically, the clusters for the word *awful* do not represent meaningful topics, which might be attributable to the fact that the word is used as an adverbial intensifier with general meaning (Wijaya and Yeniterzi, 2011). Over time, the word comes to be associated with negativity in meaning, and then less intensity is expressed through the use of the word. Another example is the word *mouse*. By decreasing the k in k-means, two clusters can be merged, and the last cluster represents the additional meaning acquired with the word.

Ultimately, the evolution of dynamic networks, specifically temporal exponential random graph model (ERGM) (Robins et al., 2007; Wijaya and Yeniterzi, 2011) is proposed to model the network of word co-occurrence in a diachronic vein. The word co-occurrence network illustrates the connections of words as nodes in the graph. A use case is to identify change of connotation in meaning for words such as *awful*, for the co-occurrence network would justify that no connections exist among the nodes and thus these words do not belong to the same topic. On top of that, the emergence or disappearance of sub-graphs is indicative of newly-acquired or lost meanings of a word. The setting of lower weighting for sub-graphs is also consistent with the possibility that the original meanings still prevail with the passage of time. In summary, the sketching of word profiles by selecting relevant metrics (i.e., tf-idf scores), the merging of clusters by adjusting the number of clusters, as well as the formation of the word co-occurrence

doi:10.6342/NTU202004346

network by building links and sub-graphs, have paved the way for early studies of semantic change.

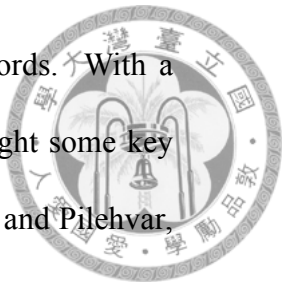
Recently, topic models continue to be used to explore the phenomenon of semantic change, yet with a different aim and approach. Topic models are used to yield topics that are most common in a given time period in order to anchor words that should be evaluated for the results (Antoniak and Mimno, 2018). By so doing, the number of topics set for the identification of anchoring words are much larger than that for the Topics-Over-Time (TOT) so that the computed mean probability is based on as diverse topics as possible.

2.5 Diachronic Word Embeddings

Diachronic word embeddings can be used to discover more possibilities of unknown change cases and underlying causes of general semantic change (Hamilton et al., 2016a; Heuser, 2017; Kutuzov, Velldal, et al., 2017).

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; Jawahar and Seddah, 2019; Kutuzov, Øvrelid, et al., 2018; Tahmasebi et al., 2018). Diachronic word embeddings make it possible to formulate or test hypotheses or laws of semantic change, establish temporal word analogy or relatedness, as well as discover semantic relations that are also changing over time. In Hamilton et al. (2016a), linguistic drift and cultural shift can be also distinguished and measured based on diachronic word

doi:10.6342/NTU202004346



embeddings, with the latter restricted to a smaller set of neighboring words. With a growing interest in this research topic, insights have been made to highlight some key and challenging aspects of semantic change modeling (Camacho-Collados and Pilehvar, 2018; Kutuzov, Øvrelid, et al., 2018; Tahmasebi et al., 2018).

Table 2.1. Example case studies of semantic change through computational analysis from literature

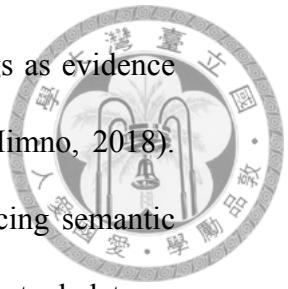
Literature	Use cases
Kulkarni et al. (2015)	apple, tape
Hamilton et al. (2016b)	gay, broadcast, awful, 病毒 'virus' *
Hamilton et al. (2016a)	actually, must, promise, gay, virus, cell
Kutuzov, Velldal, et al. (2017)	war, peace, stable
Rodda et al. (2017)	πνεῦμα 'breath' → 'spirit' (Ancient Greek)
Yao et al. (2018)	apple, amazon, obama, and trump
Rudolph and Blei (2018)	intelligence, iraq, jobs, prostitution
Antoniak and Mimno (2018)	marijuana
Hu et al. (2019)	please, alien
Rodina et al. (2020)	провальный 'a place where the surface collapsed inward' or 'loss of consciousness' → 'failed' (Russian)

* A list of attested historical shifts is provided in Hamilton et al. (2016b), and entries with the 'obsolete' tag in the Oxford English Dictionary (OED) are also considered informative of records of meaning shifts.

2.5.1 Word-level Embeddings

The topic of semantic change has directed attention to the design of corpus used as input for diachronic word embeddings. In Natural Language Processing, word embeddings are commonly added to the last layer of a deep learning model to translate discrete linguistic data to continuous numeric vectors. On the other, another line of research, referred

doi:10.6342/NTU202004346



to as "corpus-centered" approach, focuses on the use of word embeddings as evidence for certain linguistic features or cultural characteristics (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. 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 (Antoniak and Mimno, 2018). 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 Ngram Corpus[8], Corpus of Historical American English (COHA)[9], Project Gutenberg Corpus[10] 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.

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

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

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

doi:10.6342/NTU202004346

Regarding conversational diachronic corpus, (Giulianelli, 2019) 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.

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. Another example is how *president* becomes closer to *Obama* during his term, as well as *Israel's Prime Minister* and *Christopher Nolan, The Dark Knight, 2008* (Rosin et al., 2017) by finding continuous peaks of lowest distance between vectors with dataset YAGO2[11] that contain temporal relations of named entities.

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

[11] The latest version is released in 2020.

doi:10.6342/NTU202004346

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.

In addition to alignment of separately trained embeddings, temporal referencing (TR) (Dubossarsky, Hengchen, et al., 2019a) is proposed to mitigate the noise issue induced by alignment. Because of alignmnet, the results, especially low-frequency words, are influenced by noises (Dubossarsky, Hengchen, et al., 2019a,b). However, the lack of widely-accepted evaluation procedures have made it difficult to learn more about the noises invited by vector space aglinment (Dubossarsky, Hengchen, et al., 2019b).

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 in the form of domain adaptation (Huang and Paul, 2019). Temporal

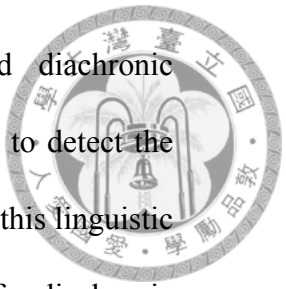
doi:10.6342/NTU202004346

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).

2.5.2 Sense-level Embeddings

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).

doi:10.6342/NTU202004346



Notwithstanding, although context-independent and contextualized diachronic embeddings are proposed and explored in an increasing body of research to detect the presence of semantic change, which models are more capable of capturing this linguistic phenomenon remains an on-going topic that calls for evaluation methods for diachronic embeddings. It is debatable whether simpler models results in better performance (Schlechtweg, Hätty, et al., 2019). Firstly, datasets like DURel (Diachronic Usage Relatedness)[12] are established based on human ratings (Schlechtweg, Walde, et al., 2018) and word injection (Schlechtweg, Hätty, et al., 2019), which is is based on similar concepts like domain-specific word sense disambiguation or term ambiguity detection, inspired by term extraction and synchronic version of SURel (Synchronic Usage Relatedness)[13] where variation lies in sense divergence across domains for research topics like online language analysis. However, evaluation data are scarce (Wevers and Koolen, 2020), hand-picked attested examples from literature or dictionaries with tags like "obsolete" (Hamilton et al., 2016a) have proven that automatic semantic change detection is able to capture semantic change (See Table 2.1) (Schlechtweg, Hätty, et al., 2019), but results still vary depending on test or evaluation data that are currently available. For example, the exploration of semantic change laws has proved influential in latest researches, the synchronic or "within-time-period" accuracy still has to rely on test data that are available for a certain period of time, which is the method used in Hamilton et al. (2016b) and the anchoring time period is 1990s for their diachronic corpora of Google Ngrams from 1800 to 2009. The result of Schlechtweg, Hätty, et al. (2019) shows that SGNS with orthogonal Procrustes alignment achieves the highest performance based on the DURel dataset, whereas topic modeling has the least correlation with the examined dataset.

[12]https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/durel/
[13]https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/surel/

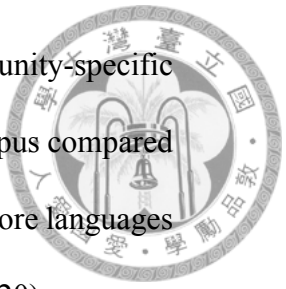
doi:10.6342/NTU202004346

Furthermore, the results in Dubossarsky, Weinshall, et al. (2017) and Schlechtweg, Hätty, et al. (2019) shows that cosine distance (global neighborhood distance) outperforms local neighborhood distance under the condition of aligned embeddings, and the results of topic modeling is sensitive to corpus size and frequency of the target words, which make it a less desirable method in this study, as pre-modern Chinese texts might not reflect accurate counts of types and tokens.

The BERT pre-trained language model can be used in companion 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, including 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

doi:10.6342/NTU202004346



data. Their method is shown to be effective on detection of short-term community-specific changes in word usages by including football data as the conversational corpus compared to diachronic corpus in their study. Their subsequent work is expanded to more languages and judgement data in the SemEval 2020 task (Kutuzov and Giulianelli, 2020).

Instead of sense inventories, various clustering algorithms are resorted to induce senses of target words, including K-Means, Gaussian mixture models (Giulianelli, 2019).

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).

2.6 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 Analysis (PCA) and t-distributed Stochastic Neighboring Embedding (t-SNE) (Van der Maaten

doi:10.6342/NTU202004346

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)[14].

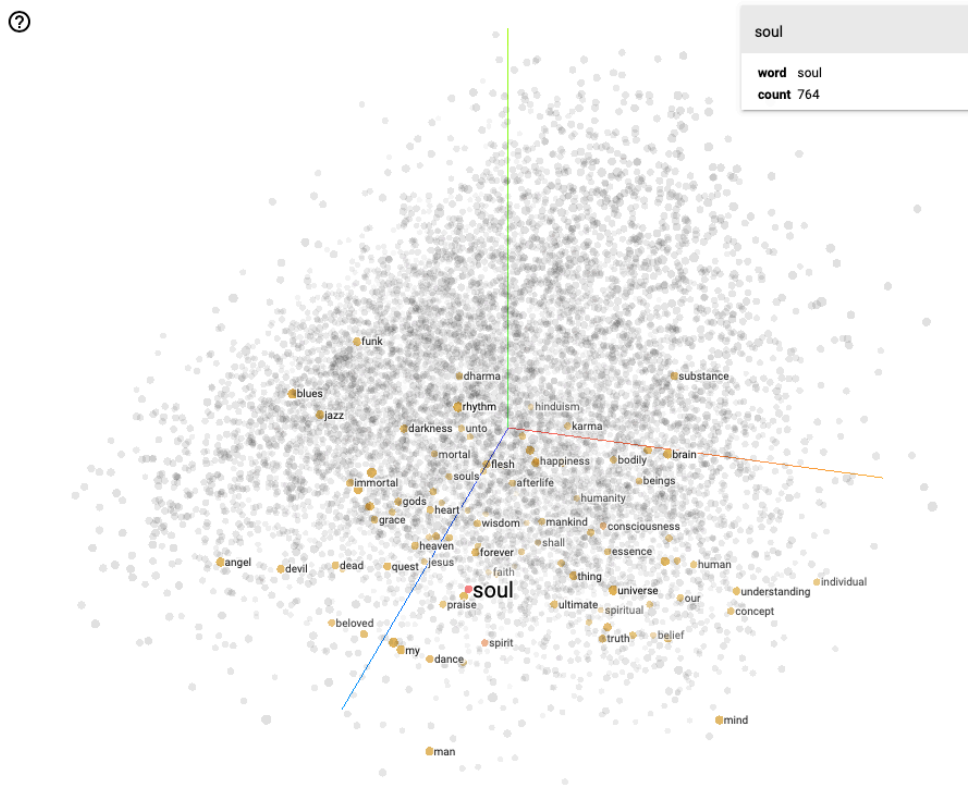


Figure 2.2. Visualizing Data using the Embedding Projector in TensorBoard

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

[14]Web-based demonstration of interactive graphical representation of high-dimensional embeddings via Google's TensorBoard Embedding Project can be visited at https://projector.tensorflow.org

doi:10.6342/NTU202004346

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, Tsvetkov, et al., 2015; Hu et al., 2019). In Dubossarsky, Tsvetkov, 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. Also (Gonen et al., 2020).

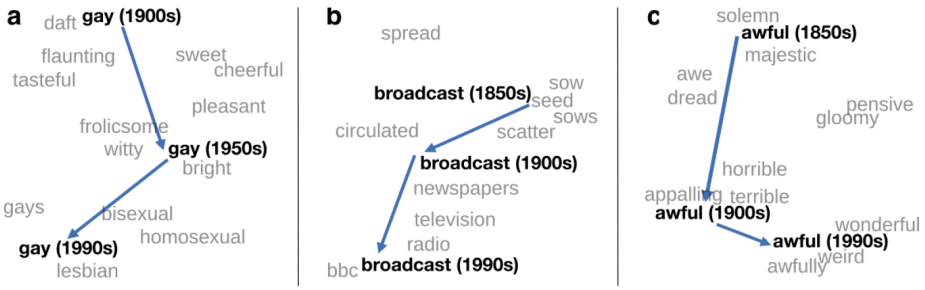


Figure 2.3. Two-dimensional visualization of semantic change for the word *gay*, *broadcast*, and *awful* in Hamilton et al. (2016b)

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

doi:10.6342/NTU202004346

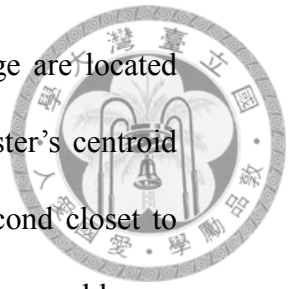
disyllabic development of words. Therefore, we aim to analyze both pre-modern and modern Chinese texts, which would be the first attempt to apply both computational and statistical models to explore the interplay between disyllabic development of words and semantic change in Chinese.

2.7 Laws of Semantic Change

The application of computational linguistics to historical semantics bears fruit of inspiring works on generalizations of semantic change. It is hoped that computational analyses would help explain the mechanisms of semantic change such that a level of maturity like laws of sound change could be reached. Instead of examining instances of semantic change, it becomes more and more computationally feasible and efficient to analyze this linguistic phenomenon on a large scale. The degrees of semantic change are quantitatively measured and how other linguistic factors play a role in semantic change is revealed, which motivates the research inquiries of testing the results against laws of semantic change that have been proposed in theory or observed on a smaller scale. Among them are the law of prototypicality (Dubossarsky, Tsvetkov, et al., 2015), the law of conformity (Hamilton et al., 2016b), and the law of innovation (Hamilton et al., 2016b). The competing laws of parallel change and differentiation are reviewed in Xu and Kemp (2015).

Based on the English lexicon between 1850 and 2009 in the Google Ngram corpus, Dubossarsky, Tsvetkov, 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." K-means clustering, with varying numbers of clusters, is applied to a list of most frequently used words in the Google Ngram corpus. Within

doi:10.6342/NTU202004346



the same cluster, words that undergo a higher degree of semantic change are located farther from the cluster's centroid. A further analysis reveals that a cluster's centroid has a stronger correlation than a prototypical exemplar, which is the second closet to the centroid. Observing this tendency, the authors argue that this finding enables an exploration of semantic change in relation to a hypothetical, abstract, non-lexicalized prototypical member. In addition, it is found that the correlation increases as the number of clusters increases, but drops once the number of clusters reaches a maximum, suggesting that the boundaries of semantic categories can be drawn. Therefore, this research offers a bottom-up analysis of the diachronic prototypical semantics with flexible boundaries of semantic categories to evaluate a large number of lexical items.

The laws of conformity and innovation are put forward by Hamilton et al. (2016b). The former posits that observed frequency negatively 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. Polysemy is measured through contextual diversity from the co-occurrence network of the trained diachronic embeddings, which is also the reason why the relationship between polysemy and the rate of semantic change is examined under controlled frequency due to the intrinsic correlation between the two variables. To evaluate the two laws, known instances of semantic change from literature and top 10 words from the experiment results are reviewed. The Spearman correlation is then calculated on a full scale between the rate of semantic change and the two linguistic factors, namely a word's observed frequency and number of senses, for 4 languages (English, German, French, Chinese), and 6 historical corpora (Google books in all genres for the 4 selected languages, an additional corpus from Google books in the fiction category, and COHA), which span 2 centuries (1800–2009) at the interval of

doi:10.6342/NTU202004346

a decade. Nonetheless, the judgement of synchronic accuracy has to rely on the time period of 1990s possibly because the test data for earlier time periods are rare, as pointed out to be an issue of diachronic semantics studies (Wevers and Koolen, 2020). Yet, the quantification of the rate of semantic change through computational analysis is inspiring. It is revealed that the two laws of semantic change are proportional in the form of a power law. Additionally, the consideration of controlled frequency is crucial because the results show complementary trends for the two laws of semantic change examined if untreated.

In Xu and Kemp (2015), near-synonyms are shown to change in parallel, and thus the law of parallel change is more favorable than the law of differentiation. That is, these two laws attempt to characterize whether words with similar meanings continue to be similar, or turn out to be divergent. Nouns, verbs, and adjectives are extracted from the Google Ngram corpus from 1890 to 1999. The selection of synonyms and control pairs is determined by external sources like thesauri and WordNet sense inventories and the computation of the Jensen-Shannon (JS) divergence scores, with a lower score representing a higher similarity in meaning. To further prevent bias from different lists of control pairs, the analysis by Xu and Kemp (2015) is conducted chronologically as well as in reserved time order. The degree of semantic change is then derived from the intersection of nearest neighbors between two time periods, and the results are consistently in favor of the law of parallel change against that of differentiation. Yet, although this research is often referenced in subsequent works on computational analyses of semantic change, the meaning vectors are not constructed from word embeddings, but frequency tables of the target word with words within 1 window size.

Not only are laws of semantic change examined by inspecting the relations between the degree of semantic change and various linguistic factors, the use of word embeddings

doi:10.6342/NTU202004346

itself as a method for diachronic linguistic studies are also evaluated. The results have shown to be helpful in the development of semantic change generalizations and principles. However, different conclusions might 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.

doi:10.6342/NTU202004346

Chapter 3

Methods

3.1 Data Collection and Preprocessing

As early as the year of 1982, Sinclair already envisioned the possibility of having "vast, slowly changing stores of text" that provide "detailed evidence of language evolution" (as cited in Renouf, 2002). Since then, the importance of digitally storing both historical and modern textual data has been widely recognized in the study of corpus linguistics (Renouf, 2002). As Renouf (2002) emphasizes, "we need the past in order to understand the present. An amalgamation would increase the scope, timespan and continuity of resources, whilst lessening the inconvenience of having to switch from one corpus and set of tools to another." Among the existing corpora, written texts comprise a major portion of the corpus compilation efforts, and thus it is a turning point to explore the diachrony of the data along with more recently available texts from historical periods.

To construct a diachronic corpus in this study, texts of pre-modern and modern Chinese are collected from the Chinese Text Project (中國哲學書電子計畫, hereinafter CTEXT)

doi:10.6342/NTU202004346

(Sturgeon, 2019)[1] and Academia Sinica Balanced Corpus of Modern Chinese (中研院現代漢語平衡語料庫, hereinafter ASBC) (Chen et al., 1996)[2] 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 to 1949 A.C. of the Republican era (Sturgeon, 2019). Since the number of texts available from each era varies, the time periods with the highest number of texts, namely the Tang (618 – 907 A.C.), Song (960 – 1279 A.C.), Yuan (1271 – 1368 A.C.), Ming (1368 – 1644 A.C.), and Qing (1644 – 1911 A.C.) dynasties, are included to construct the sub-corpora of pre-modern Chinese in this study. The texts and their metadata are retrieved from the CTEXT digital library using `ctext`[3], a Python API (Application Programming Interface) wrapper of the same name developed by Sturgeon (2017).

Apart from the provision of the API access, the CTEXT project website is informative of how textual data and metadata are structured in the retrieved format[4]. Since the original prints are scanned and converted into the machine-readable format using the OCR (Optical Character Recognition) techniques, multiple versions of a text are likely to be produced through the employment of different OCR techniques, only one version representative of a set of texts is selected following the instructions on the CTEXT project website[5], or, if needed, all versions are retained to help discern the differences in the converted texts. For

[1] https://ctext.org/

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

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

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

[5] Among a set of documents, the version labeled with the tags "TEXTDB" (the texts are selected in the main library/database), "WORKSET" (the texts are specified as representative of a group of documents), "OCR_CORRECTED" (the texts have been proofread and corrected through the community efforts), "OCR_MATCH" (the texts have been proofread and can be referenced to parts of the scanned document) in the metadata is treated as representative according to the instructions on the CTEXT project website. In the case where no tags are provided, the version with the largest file size is selected.

doi:10.6342/NTU202004346

example, to obtain frequencies of characters used in different time periods, it is necessary to exclude duplicate counts, while the differences are kept intact during the training of word embeddings. On the document level, the corpus composition is summarized in Table 3.1.

Table 3.1. Document composition of the CTEXT corpus

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

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

As instructed on the project website[6], the cleaning task for the CTEXT corpus is proceeded as described below:

(1) The raw text is cleaned by (a) removing commentaries and marginal notes, (b) segmenting the text into two levels of 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.

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

doi:10.6342/NTU202004346

(2) Chinese words are not delimited by space, nor is a conventional punctuation system adopted in pre-modern Chinese texts. As a consequence, the punctuations should be viewed as symbols to mark 句讀 *jùdòu* 'pauses or breaks'. Only the symbols specified in the project website's instructions are used to split the texts into sentences, 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, variants, and homoglyphs listed in the Unicode Standard[7,8].

(3) To extract Chinese characters, Unicode range between U+4E00 and U+9FFF are retained for basic Chinese characters, and variants or rare characters are captured from the Unicode blocks of CJK Extension A to F, CJK Compatibility Ideographs, and CJK Compatibility Ideographs Supplement[9]. The Unicode blocks serve as a way to find characters that tend to belong to a specific script (Moran and Cysouw, 2018). Due to the employment of OCR techniques, missing characters are indicated with filled black circles (●).

(4) Text surrounded by quotation marks indicates conversations, sayings, or allusions, and is not removed during the preprocessing. On one hand, conversations are an integral part of the text; on the other, sayings and allusions reveal what is still in use or understandable in the time period of their appearance.

(5) One of the difficulties in processing pre-modern Chinese lies in the word segmentation issue. This is particularly troublesome given the disyllabic development of Chinese. Therefore, the CTEXT corpus consisting of the cleaned

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

[8]While the texts are in the units of characters in this study, dependency parsers for classical Chinese include `UD-Kanbun` by Yasuoka (2019) (https://pypi.org/project/udkanbun/) and `Stanza` in StandfordNLP by Qi et al. (2020) (https://stanfordnlp.github.io/stanza/).

[9]The character-to-glyph issues of CJK (Chinese, Japanese, and Korean) characters are explained on the Unicode website (https://www.unicode.org/faq/han_cjk.html).

doi:10.6342/NTU202004346

texts has a character frequency profile that is distinctively different from the ASBC corpus. The overview of type and token counts of texts from the time-sliced corpora is summarized in Table 3.2 and Table 3.3.

Table 3.2. Token and type counts of the diachronic corpora in this study

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

Table 3.3. Token and type counts of the diachronic corpora in this study

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

doi:10.6342/NTU202004346

3.2 Exploratory Data Analysis (EDA)

After the completion of preprocessing, this study proceeds to a preliminary exploratory data analysis with the bootstrap test proposed by Lijffijt, Nevalainen, et al. (2016). The bootstrap test is a non-parametric test of statistical significance that is designed to minimize the influence of uneven distribution of linguistic features in texts and to provide a more solid ground for quantitative analyses on the comparison of (sub)corpora.

Prior to the introduction of the bootstrap method, bag-of-words methods like chi-squared and log-likelihood ratio tests rest on the assumption that all samples are statistically independent of each other and do not account for poorly dispersed words, hence the name (Lijffijt, Nevalainen, et al., 2016). Yet, words within a text are not independent in nature, and thus tests like Mann-Whitney U test or bootstrap test are more suitable to evaluate the differences in word frequencies of different corpora or time periods (Brezina, 2018; Lijffijt, Nevalainen, et al., 2016). In terms of the assumption on independence, this relation exists at the level of texts rather than individual words using the bootstrap method. Additionally, the bootstrap test produces a more conservative *p*-value than bag-of-words-based methods, which further prevents the use of higher cut-off values in the chi-squared or log-likelihood ratio tests given that the thresholds do not correct the bias resulting from the uneven distribution and high variance of word frequencies.

To perform the bootstrap test, which involves the process of multiple resampling with replacement, a random sample of texts from a corpus is taken and placed back to the original pool in a repetitive manner. In each resampling cycle, the value of the statistic of interest is noted and further generalized. The bootstrap test proposed by Lijffijt,

doi:10.6342/NTU202004346

Nevalainen, et al. (2016) to compute the *p*-value is conducted through the equations below.

$$p = \frac{\sum_{i=1}^{N} H\Big(freq(q, T^i) - freq(q, S^i)\Big)}{N}, \tag{3.1}$$

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

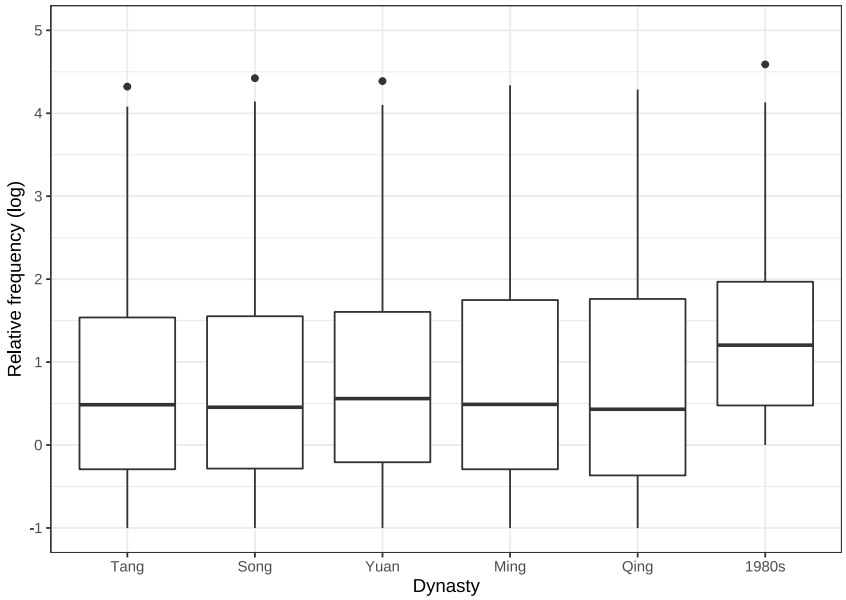
$$p_{two} = 2 \times min(p, 1 - p) \tag{3.2}$$

$$p* = \frac{p_{two} \times N + 1}{N + 1} \tag{3.3}$$

The frequencies of the word q in the two corpora T and S in a sample i are compared N times to derive the value $p*$ as the *p*-value for the bootstrap test. In Lijffijt, Säily, et al. (2012), the bootstrap test is employed to assess the diachronic stability of word frequency profile of the Corpus of Early English Correspondence in the seventeenth century.

In this study, to understand the frequency distribution of characters in a diachronic view, the bootstrap test is performed with 1*k* samples of 50 texts from the 500 texts of selected versions from the Tang dynasty to the Qing dynasty. The general distribution of character frequencies before the bootstrap test is illustrated in Figure 3.1 and Table 3.4, and the results are shown in Figure 3.2 and Figure 3.3.

doi:10.6342/NTU202004346



* The character with the highest relative frequencies in the Tang, Song, and Yuan dynasties, indicated as outliers in the boxplot, is the function word 之, which is replaced with its modern form 的.

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

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

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

* For frequency information from other sources, see Appendix A.

doi:10.6342/NTU202004346

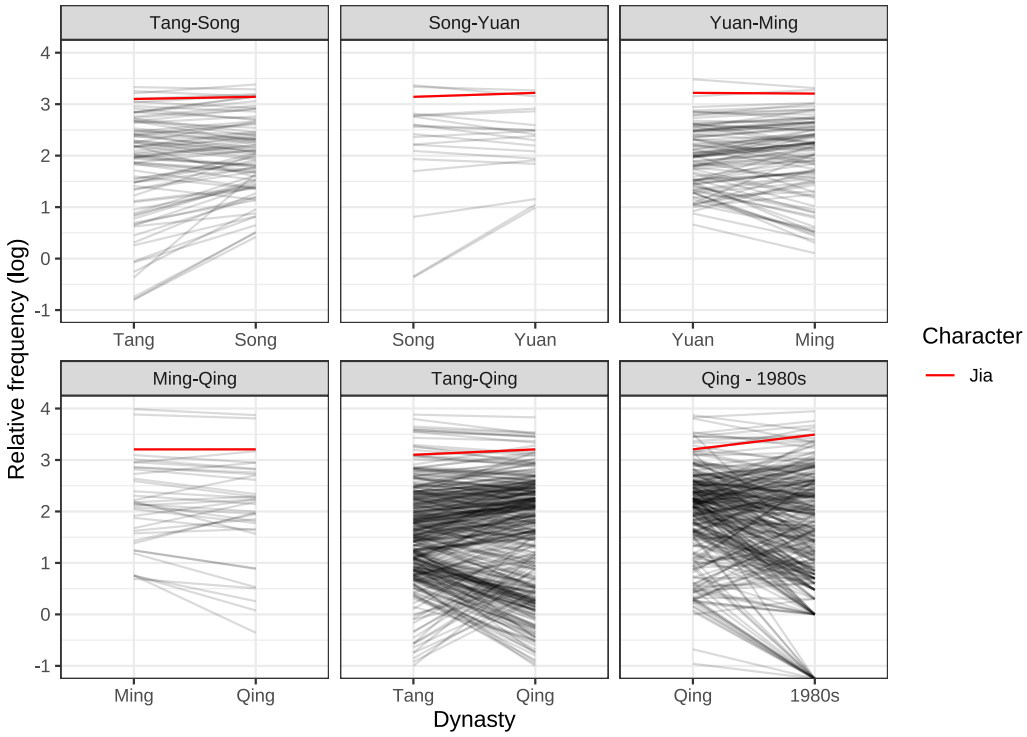


Figure 3.2. Frequency change with statistical significance derived from the bootstrap test on characters in comparison with *jiā* from the Tang dynasty to the 1980s

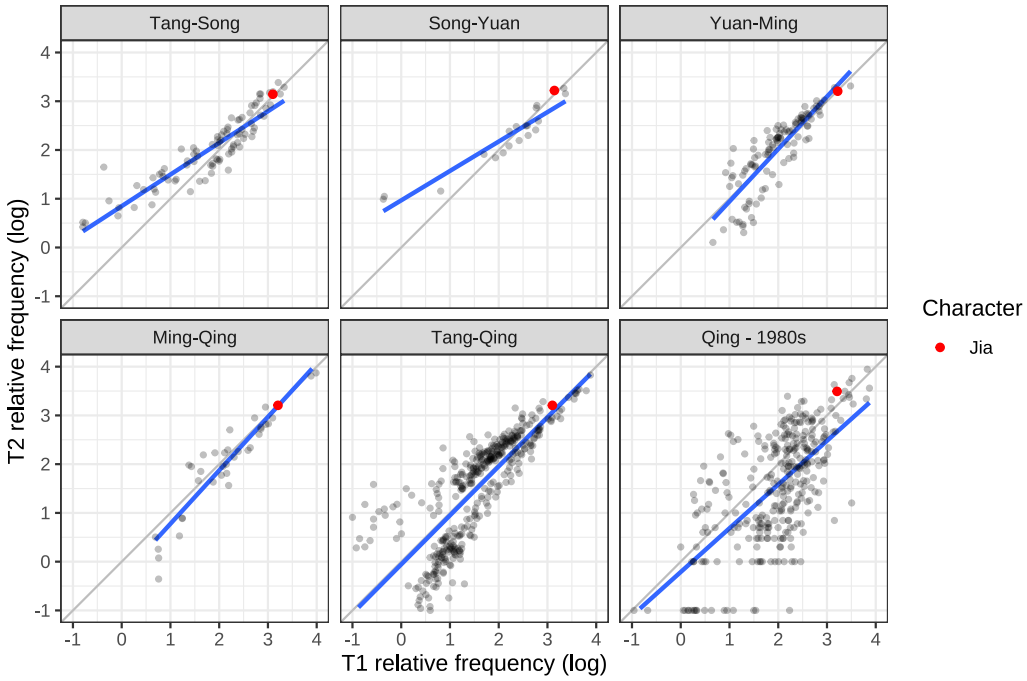
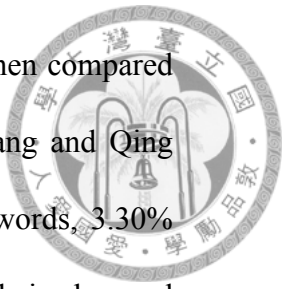


Figure 3.3. Frequency change with statistical significance derived from the bootstrap test on characters in comparison with *jiā* from the Tang dynasty to the 1980s

A total of 22,981 characters have appeared in at least one dynasty. The time period with

doi:10.6342/NTU202004346



the most instances of significant frequency change is the Qing dynasty when compared to the Tang dynasty. That is, 12,233 characters are seen in both the Tang and Qing dynasties, and 404 of them receive a *p*-value at less than .05. In other words, 3.30% of the characters in use between the Tang and Qing dynasties change in their observed frequency following the rejection of the null hypothesis. Regarding the direction of change in character frequency, both upward and downward trends of significant change can be witnessed for individual cases of characters in Figure 3.2, yet the trend lines in Figure 3.3 do not reflect an obvious tendency toward either direction. However, it is worth noting that between the Qing dynasty and the 1980s, a portion of data points fall in the bottom on the y-axis of the scatter plot, suggesting that these characters fall out of use in modern Chinese, but no such observation can be made between the Tang and Qing dynasties.

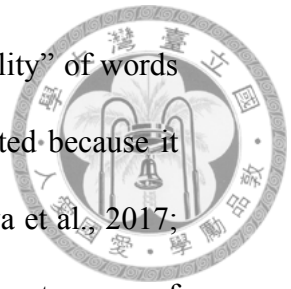
Specifically, the frequency profile of the character *jiā* reveals a stable use from the Tang dynasty to the Qing dynasty. Although the relative frequency of *jiā* slightly increases from 1,260 to 1,609 (The raw frequencies are 61,420 and 1,831,222 respectively), as shown in Table 3.4, the difference in the use of the character is not statistically significant: p=.5404, 1k samples. As a result, the bootstrap test fails to reject the null hypothesis that assumes no difference in the use of *jiā* between the two time periods, and similar results are found with the other combinations of time periods.

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

3.3 Collocation-based Approach

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

doi:10.6342/NTU202004346



applied to quantify the strength of word co-occurrences, or the "collocability" of words studied (Gablasova et al., 2017). Particularly, the logDice score is selected because it is standardized and scaled, and thus comparable across corpora (Gablasova et al., 2017; Rychlỳ, 2008). To interpret the logDice scores, a maximum of 14 represents cases of complete co-occurrences, but usually the scores are less than 10. A difference by 1 equals twice in co-occurrences frequency, and that by 7 means 100 times more frequent, according to Equation 3.4.

$$logDice = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y} \tag{3.4}$$

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 and those irregardless of the position, are first calculated, and the logDice score of each collogram is then computed. Collograms that do not appear consecutively across all time slices are filtered out, and the logDice scores of the shared collograms form a vector per time slice. Eventually, the logDice vectors of all time slices are structured as a matrix. Three matrices are prepared for pre-collograms, post-collograms, and all collograms of the keyword *jiā*.

3.4 Word-level Embeddings

To learn what observations are supported by linguistic data in the three sub-corpora, embeddings are generated with Word2Vec in the Python gensim package, and the linguistic data from different time periods are separately trained. Additionally, as suggested by 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

doi:10.6342/NTU202004346

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.

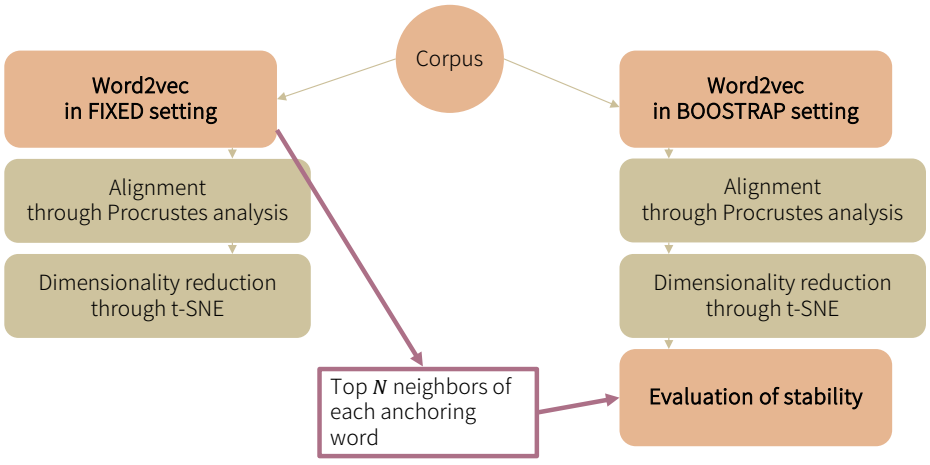


Figure 3.4. Workflow of word-level embeddings

In terms of separately trained word vectors, vector alignment is based on Procrustes analysis by Hamilton et al. (2016b)[10]. After the training of Word2Vec embeddings,

[10]https://github.com/williamleif/histwords

doi:10.6342/NTU202004346

embeddings are imported to TensorBoard to visualize the data points (Smilkov et al., 2016), and further analyzed in the discussion section.

In addition to the word embeddings trained on the whole corpus, a bootstrapping without replacement approach is adopted (Antoniak and Mimno, 2018). While the FIXED model indicates the baseline, algorithmic variability, i.e., random initiations, random negative sampling, random subsampling of tokens in documents (Antoniak and Mimno, 2018). Following Antoniak and Mimno (2018), for each time period, 50 iterations are performed. For each iteration of resampling, a model is built on the N randomly selected documents ($N = 150$ for pre-modern documents and $N = 0.2$ of the documents in ASBC) in contiguous sequence. An ensemble of embeddings are generated with the results averaged over the bootstrap samples.

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

Before the degree of semantic change is measured, a filtering of mid-frequency characters is conducted, for highly frequent characters are not "content-bearing" (Hamilton et al., 2016a; Rodda et al., 2017). Afterwards, the similarity of semantic vectors

doi:10.6342/NTU202004346

across time periods is compared using correlations; namely the similarity between T2 (the time period of interest) and T1 (the previous time period). Besides computing on the original vectors, alternatively called first-order embeddings, we resort to second-order embeddings composed of a full or partial list of neighboring words to the keyword. Specifically, the top 25[11] 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 Hamilton et al. (2016a).

3.5 Sense-level Embeddings

In addition to word-level embeddings, contextualized embeddings are extracted to retrieve sense-level representations based on the diachronic corpus in this study. The sense-level representations are described as "sense representations" in Hu et al. (2019) and "usage representations" in Giulianelli (2019), for the pre-trained language model allows for the extraction of a possibly infinite number of embeddings depending on the context of the input, and the embeddings reflect the authentic language use and distinguishes the usages in group to simulate the sense distribution. The chosen pre-trained language model is bert-base-chinese (Devlin et al., 2018) with HuggingFace's PyTorch Transformer framework, which is a Transformer architecture with 12 layers, 768 hidden units, 12 heads, and 110M parameters, and is trained on both Traditional and Simplified Chinese text from Wikipedia and BookCorpus with masked training and next sentence prediction task. Conventionally, the final or last 4 hidden layers are used as the token embeddings, which is followed by the averaging of multiple embeddings of a target word, yielding a 768-dimensional vector to represent the target word being studied. For senses with

[11] In Hamilton et al. (2016a), the range between 10 and 50 is recommended as their results reflect.

doi:10.6342/NTU202004346

multiple example sentences, the corresponding sense representations are an aggregated vector.

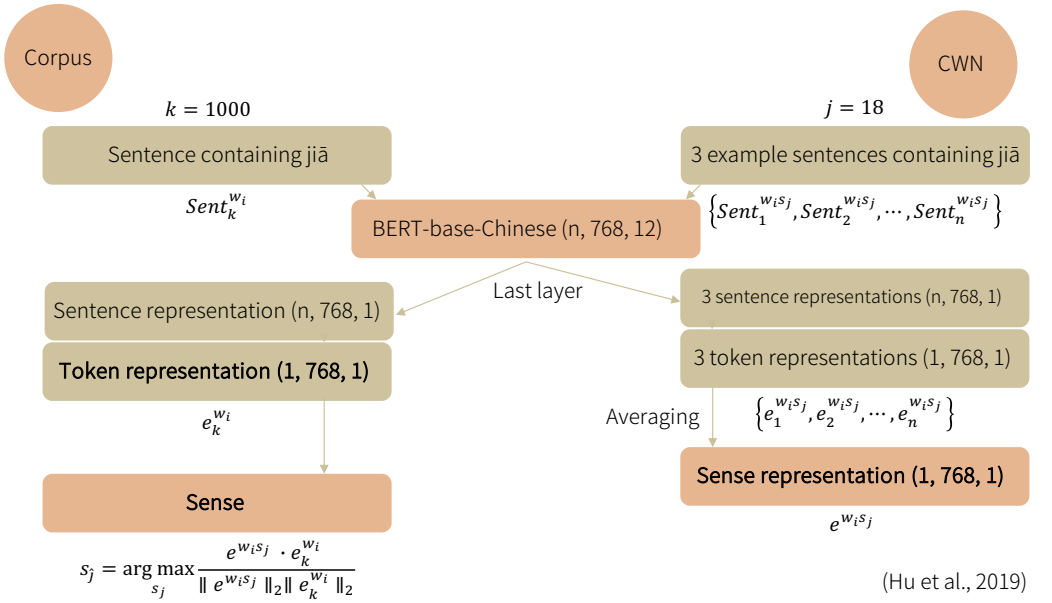


Figure 3.5. Workflow of sense-level embeddings

This study aims at inspecting the sense interaction of the keyword *jiā* from the Tang dynasty to the 1980s using the diachronic sense modeling technique proposed by Hu et al. (2019). Instead of using the senses defined in dictionaries, this study chooses the Chinese WordNet (CWN)[12] (Huang and Hsieh, 2010) as the references for the senses of the keyword *jiā*. Generally, linguistic resources like the Chinese WordNet (CWN) contain fine distinctions of senses for an entry of a word, and complete example sentences are consistently available under each entry. Regarding the entry of *jiā*, a total of 18 senses are listed under 2 lemmas, with 17 senses under one lemma and 1 under the other (See Appendix F). As the CWN is designed under the context of Modern Chinese, it is assumed that the meanings of *jiā* are pre-determined into 18 senses, which act as a foundation of diachronic sense modeling in this study. Thus, the senses of the keyword *jiā* are traced retrospectively, and further discussion is provided in Chapter 4.

[12]http://lope.linguistics.ntu.edu.tw/cwn2/

doi:10.6342/NTU202004346

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. To begin with, word-level analysis is performed using the VNC method (Gries and Hilpert, 2012) and the Word2Vec algorithm (Mikolov et al., 2013).

3.6 The Variability-based Neighbor Clustering Method (VNC)

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, and the order-preserving characteristic of the VNC method is crucially important for chronological variation research (Moisl, 2015). 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 serves as the basis of whether they are put in groups or not from a bottom-up fashion, as shown in the pseudo-code in Figure 3.6. The resulting groupings or periodization can be graphically represented with a dendrogram and further analyzed.

The amalgamation rules are based on two stages of similarity measures and linkage functions. Firstly, the choice of similarity measures includes standard deviation, Euclidean distance, correlation distance, among many others depending on the types of data for analysis. Typically, the former is applied to numerical data, whereas the latter is suited

doi:10.6342/NTU202004346

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 which the data are multidimensional. CV (coefficient of variation), also called RSD (relative standard deviation), can also be used to represent the standard deviation in the units of the mean. Secondly, the chosen linkage function determines the merging of two neighboring time periods. Particularly, the average linkage function, according to Equation 3.5, measures the distance between two clusters as the average distance between data points in the first cluster and those in the second cluster, and clusters with the smallest computed values are combined step by step in a bottom-up approach.

```
Given a table of n temporally distinct corpus parts where each corpus part
(i) is named by a different (average) year and
(ii) contains numerical information about a linguistic phenomenon …

1 repeat
2   for any two directly adjacent corpus parts
3     compute and store some measure of variability for their combined data
4   identify the two corpus parts with the smallest measure of variability
5   merge the data from these two corpus parts
6   assign a new name to the newly merged corpus part
7 until all recordings have the same name
```

Figure 3.6. Rationale of VNC in pseudo-code (Gries and Hilpert, 2012)

$$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j) \tag{3.5}$$

X_i: an observation from cluster 1

Y_j: an observation from cluster 2

$d(X_i, Y_j)$: distance between X_i and Y_j

Apart from the periodization of a linguistic phenomenon, the VNC method can be employed as a way for outlier detection and removal if the data is sparsely distributed. Prior to data analysis, the VNC method can be conducted and repeated to remove noise

doi:10.6342/NTU202004346

by finding out anomaly clusters that are not merged with other subgroups, and therefore minimize the influence of outliers. For example, if a year-by-year dataset is available to study the decline of a linguistic phenomenon and the VNC method reveals a number of one-year clusters, they are the anomalies and can be excluded from subsequent analyses.

Building upon various matrices, the VNC method is performed and the dendrogram is plotted using the R script offered on the Lancaster Stats Tools Online (Brezina, 2018)[13].

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

doi:10.6342/NTU202004346

Chapter 4

Results and Discussion

Diachronic embeddings, which is trained for the purpose of tracing the change of word representations in vector space models, are met with challenges in how the training is evaluated. In this study, the trained embeddings are first examined in interactive interface in order to explore the structure of the diachronic embeddings. Furthermore, analogical reasoning and bootstrapping methods are employed as an attempt to pinpoint the properties of embeddings that might be influenced by the source data. From this perspective, the "bias" in an embedding is interpreted as a "feature", not a "bug" (Wevers and Koolen, 2020).

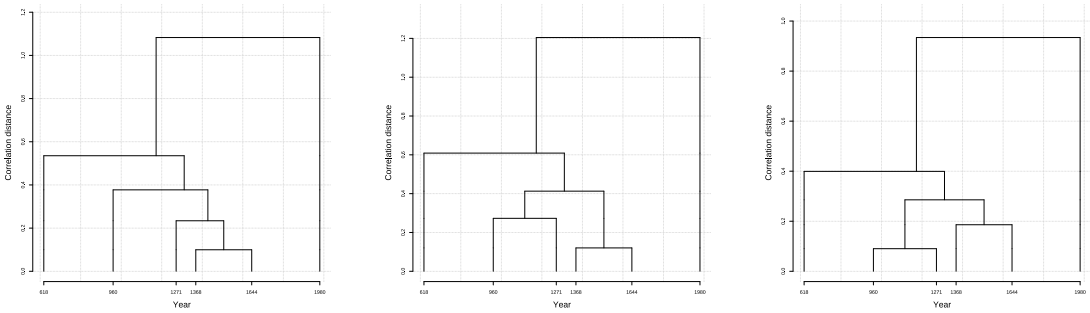
4.1 Collocation-based Approach

The results of the VNC periodization are plotted as dendrograms (See Figure **??**, Figure **??**, and Figure **??**, and the vector tables of collograms are provided in Appendix **??**, **??**, and **??**.

The correlation between the Qing dynasty and 1980s shows a drastically decreasing

doi:10.6342/NTU202004346

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.



(a) Pre-collograms (b) Post-collograms (c) All collograms

Figure 4.1. VNC periodization of collograms

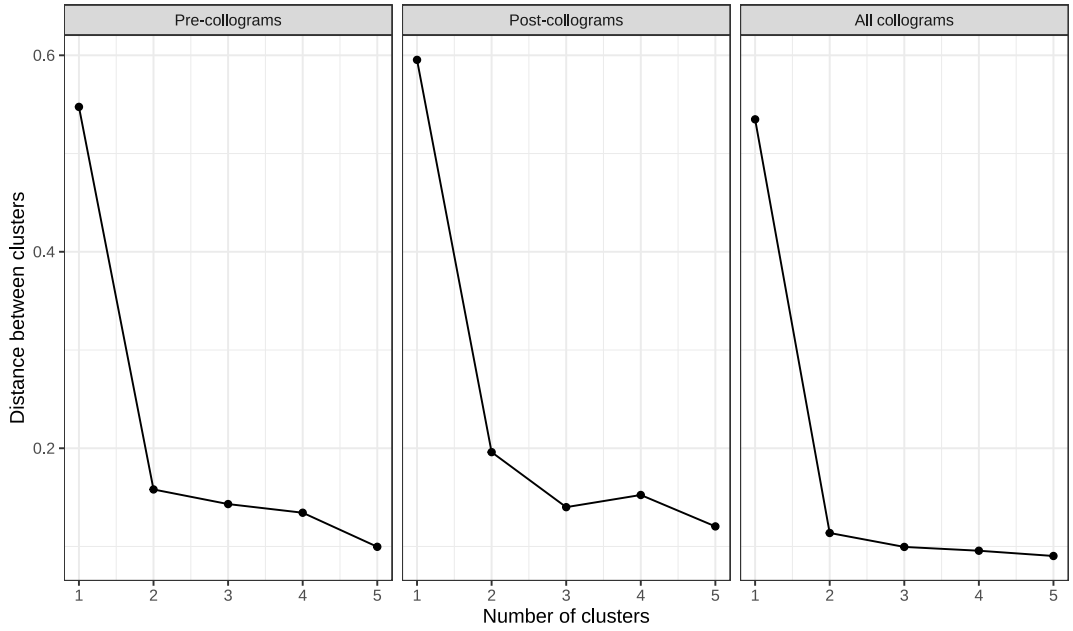


Figure 4.2. Screeplot for VNC periodization

doi:10.6342/NTU202004346

4.2 Word-level Embeddings

4.2.1 Evaluation on Analogical Reasoning

Analogical thinking and context-dependent evidence lay a cognitive ground for the studies of semantic change (Traugott, 2017). The training of word embeddings are evaluated based on intrinsic and extrinsic evaluations. In terms of vector space models, analogical thinking is associated with the directionality of vectors that represent words in pairs or in groups. While tasks like similarity scoring and analogical reasoning belong to types of intrinsic evaluation methods, the analogical reasoning is more adaptable to historical data in this study, for it is criticized that evaluation datasets mainly consists of geographical entities that would be non-existent in historical time periods (S. Li et al., 2018; Wevers and Koolen, 2020). Despite its popularity, wide application, and the much effort into the expansion of datasets, the analogical reasoning task is not adaptable for diachronic or historical word embeddings (Wevers and Koolen, 2020).

The CA8 dataset[1], created by S. Li et al. (2018), is adopted to extract semantic relations, specifically analogies, in the trained diachronic character-level embeddings. While a variety of datasets and translated versions are available for the purpose of analogical reasoning, the CA8 dataset is characteristic of its attempt to not rely heavily on geographical names and proper nouns in the target analogical pairs. On the contrary, 8 relational types are included. Additionally, among the 1,307 analogical pairs in the type "nature," 282 of them are single-character word pairs (or 1-gram, as categorized by S. Li et al. (2018)), and the semantic relations are rich and elemental, including "number, time, animal, plant, body, physics, weather, reverse, color" (S. Li et al., 2018). It is the two

[1] https://github.com/Embedding/Chinese-Word-Vectors

doi:10.6342/NTU202004346

reasons that enable the possibility to extract the semantic relations in pre-modern Chinese texts.

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

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

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

By solving the pair-based 3CosAdd and 3CosMul objectives (Levy and Goldberg, 2014), it is found that 26 and 35 pairs are consistently identified across all time periods within smaller (window size set to 1) and larger (window size set to 5) window sizes. For example, pairs like 東-西: 左-右 'east-west:left-right', 真-假: 左-右 'real-fake: left-right', and 冷-熱: 南-北 'cold-hot:south-north' are solved in all time periods, and the pair 冰-水: 雪-雨 'ice-water:snow-rain' is also stably analogous except in 1980s. However, it has not yet been feasible to extract semantic relations with set-based objectives like 3CosAvg, for the mean of a set of vectors from the source and target single-character words under the same category defined in the dataset do not yield more analogical pairs in this study.

4.2.2 Stability of BOOTSTRAP Diachronic Embeddings

The first five common query words are '公', '君', '國', '太', '官' for pre-modern Chinese, and '二', '官方', '發生', '兼', '且'. While some query words like '兼' and '且' might be considered stop words and otherwise removed, they are included to see if stop words are asserting more impact on the stability of BOOTSTRAP embeddings.

doi:10.6342/NTU202004346

The results show that the bootstrap samples become stable after 25 iterations, as suggested in Antoniak and Mimno (2018).

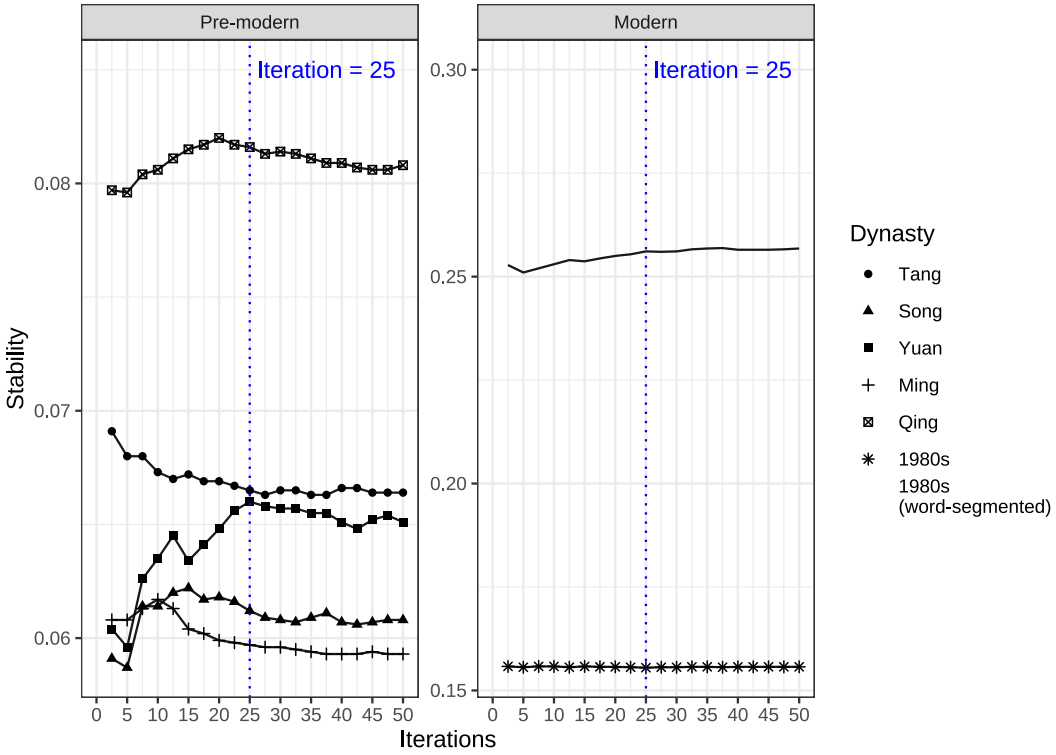


Figure 4.3. Mean stability over iterations based on query words extracted from LDA topic models and 20 nearest neighbors from FIXED embeddings

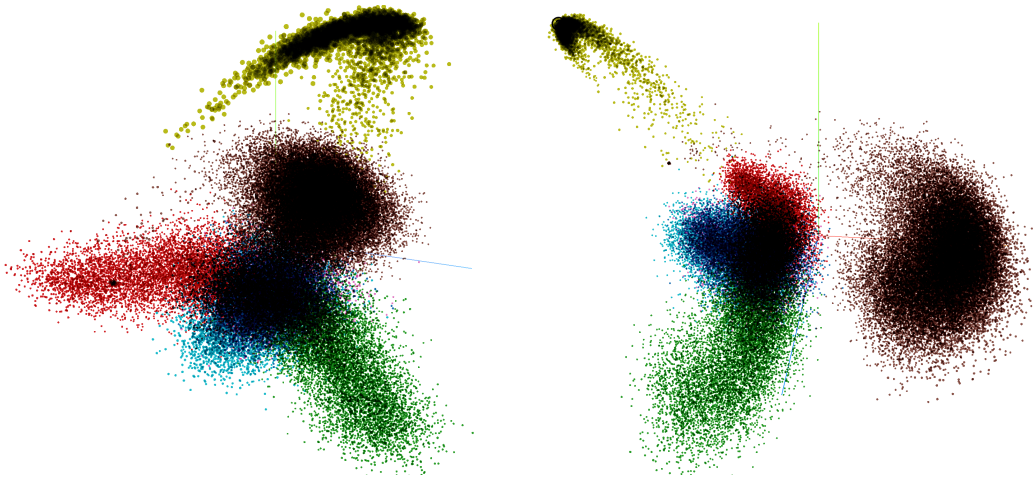
4.2.3 Diachronic Word Embeddings

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

doi:10.6342/NTU202004346

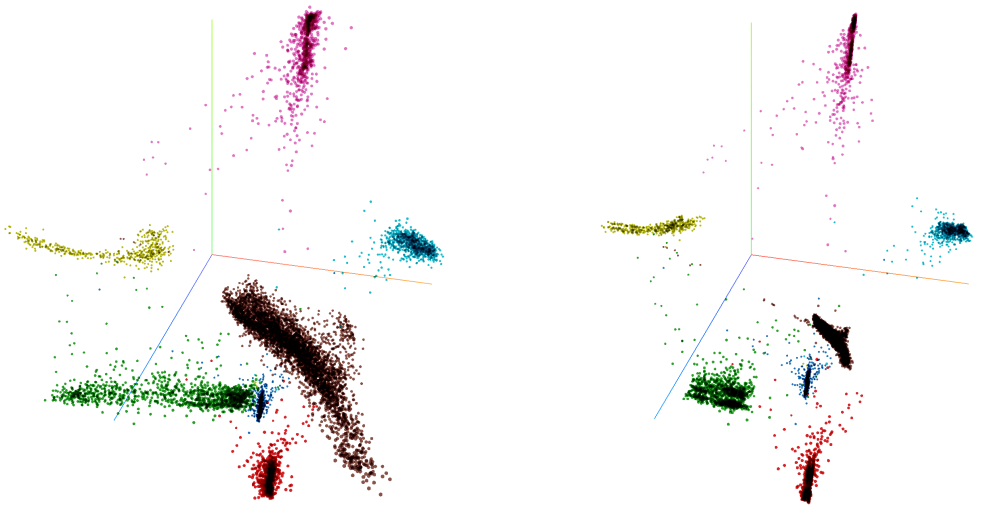
male offspring.' From the perspective of nearest neighboring words (Hamilton et al., 2016a), the core meanings of jia remains stable from pre-modern time, indicating a strong association with the family clan and the role of a wife, as in zu and qi. Secondly, the words li 'village; neighborhood' and cun '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.



* Total variance described: 34.6%
* Tang (dark blue); Song (red); Yuan (pink); Ming (sky blue); Qing (green); 1980s (brown); 2010s (mustard).

Figure 4.4. Snapshot of PCA Embedding Projector in TensorBoard

doi:10.6342/NTU202004346



* 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).

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

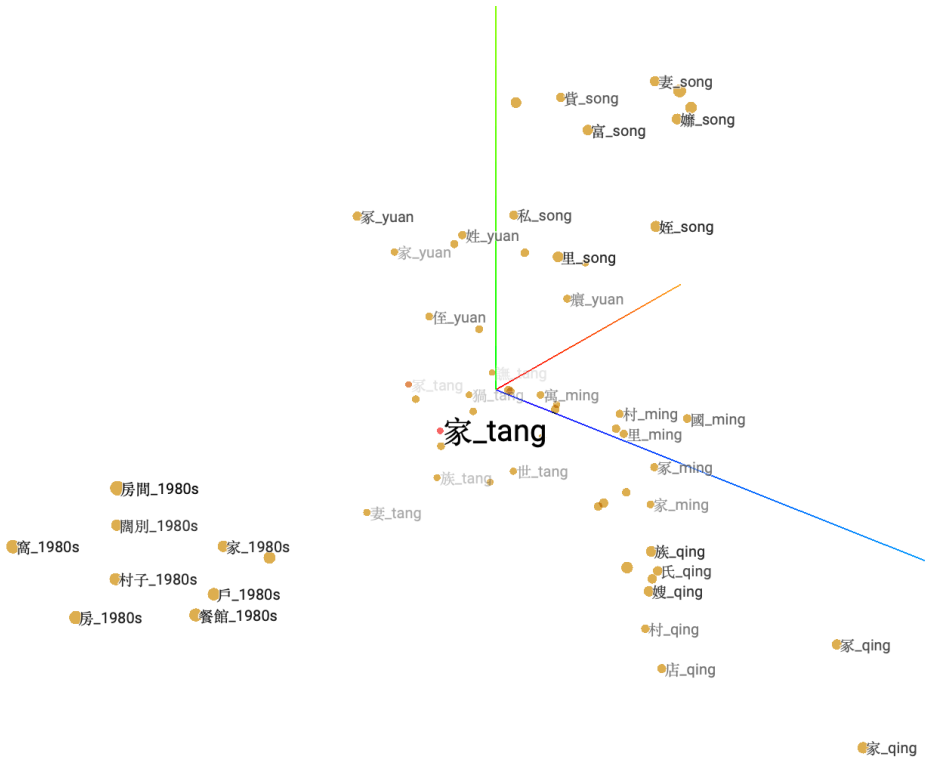


Figure 4.6. Neighboring words of *jiā* projected in a three-dimensional space

doi:10.6342/NTU202004346

Table 4.1. Nearest neighbors for modern

Dynasty	Top 20 nearest neighbors
Tang	冢, 族, 富, 貧, 亡, 産, 業, 貲, 世, 妻, 宅, 兄, 他, 世, 邦, 語, 父, 弟, 資, 姓
Song	冢, 族, 貲, 妻, 富, 鄰, 宅, 亡, 亦, 隣, 皆, 冨, 貧, 身, 舷, 兄, 妾, 他, 有, 人
Yuan	貲, 妻, 族, 富, 里, 老, 其, 業, 弟, 世, 冢, 姓, 鄉, 子, 亦, , 盡, 兄, 父, 又
Ming	冢, 者, 妻, 有, 亦, 皆, 此, 富, 當, 故, 其, 人, 兄, 是, 族, 之, 貧, 及, 所, 他
Qing	冢, 村, 傭, 僮, 子, 亦, 老, 及, 店, 富, 後, 貧, 者, 皆, 故, 兄, 坑, 與, 族, 氏
1980s	顧, 肄, 校, 霈, 嬴, 釗, 忝, 菸, 偉, 庭, 爿, 榷, 謄, 忮, 灝, 概, 踫, 蕞, 堉, 陸

Table 4.2. Nearest neighbors for modern

Keyword	Top 20 nearest neighbors
家	店, 全家, 麵包店, 速食店, 旅館, 花店, 一家, 咖啡店, 村子, 分店, 雜貨店, 家小, 超商, 養老院, 小吃店, 房間, 商店, 旅社, 餐館, 小店
家庭	主婦, 單親, 雙薪, 小家庭, 全職, 職業婦女, 寄養, 受虐, 教養, 婚姻, 養育, 主夫, 子女, 孤兒, 生計, 父母, 家計, 雙親, 小康, 貧苦
家人	親友, 妻兒, 親人, 爸媽, 親朋好友, 親戚, 鄰居, 左鄰右舍, 父母, 小佳, 公婆, 雙親, 阿眉, 娘家, 夫家, 村人, 訪客, 大弟, 團聚, 父母親
家族	母系, 後代, 豪門, 氏族, 公孫, 代代, 阿達, 甘迺迪, 鄭氏, 孤兒, 劉氏, 政商, 可米, 印地安, 文化人, 父系, 俄羅斯人, 回教徒, 共同體, 庇蔭

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.’

doi:10.6342/NTU202004346

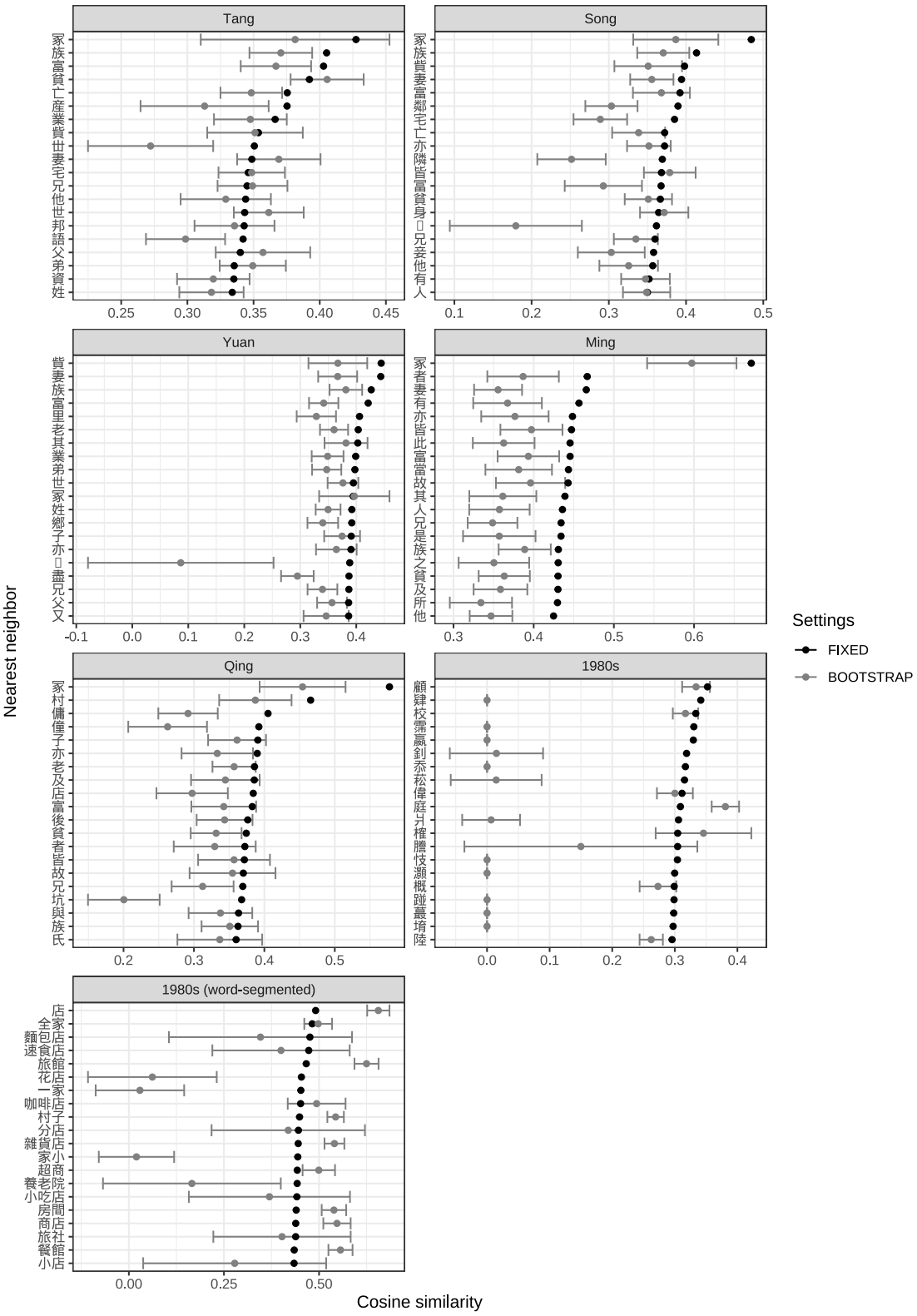


Figure 4.7. Nearest neighbors of *jiā* with means and standard deviations of cosine similarities derived from character-based embeddings in the FIXED and BOOTSTRAP settings. The 20 nearest neighbors are selected from the FIXED settings, and word-segmented embeddings are included for the time period of 1980s.

doi:10.6342/NTU202004346

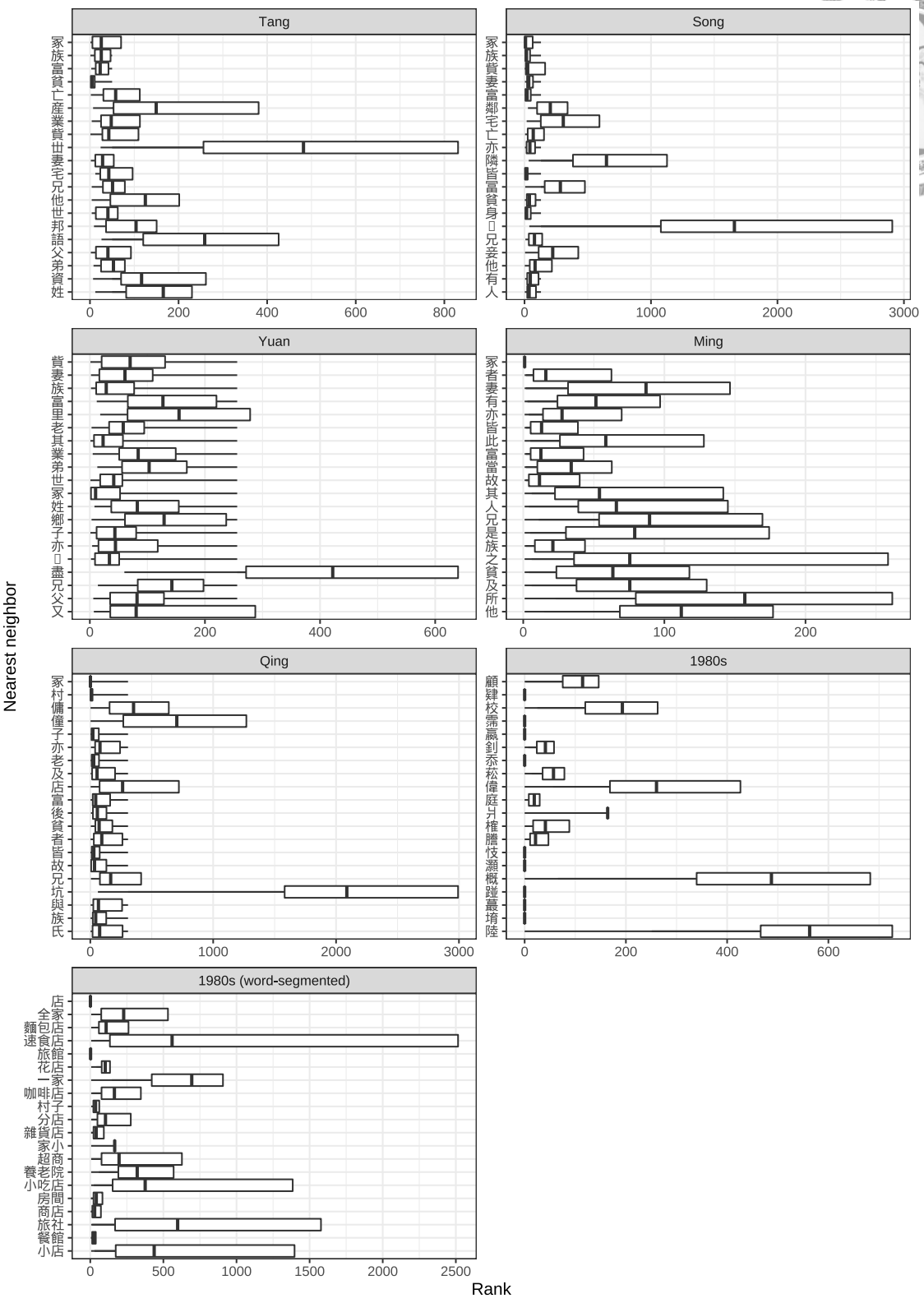


Figure 4.8. Nearest neighbors of *jiā* with changes in rank derived from character-based embeddings in the BOOTSTRAP settings. The 20 nearest neighbors are selected from the FIXED settings, and word-segmented embeddings are included for the time period of 1980s.

doi:10.6342/NTU202004346

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

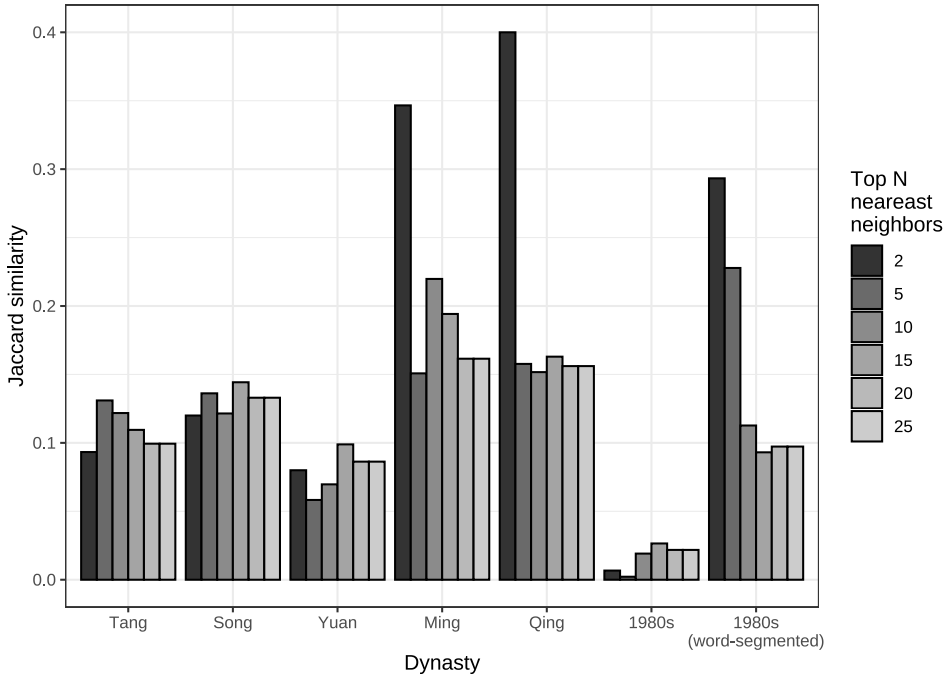


Figure 4.9. Mean of Jaccard similarities from top N nearest neighbors in the BOOTSTRAP settings. The higher the mean, the higher the degree of intersection for the nearest neighbors across the bootstrap iterations.

Secondly, not the word yu 'apartment', but hu 'one-paneled door; household', wo 'nest; hiding place', and fang 'house; room' are used to refer to jia as a physical space or unit of living. Because of the emergence of these alternative words, home evolves to be a private sphere (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

doi:10.6342/NTU202004346

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 (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.

doi:10.6342/NTU202004346

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.

4.3 Sense-level Embeddings

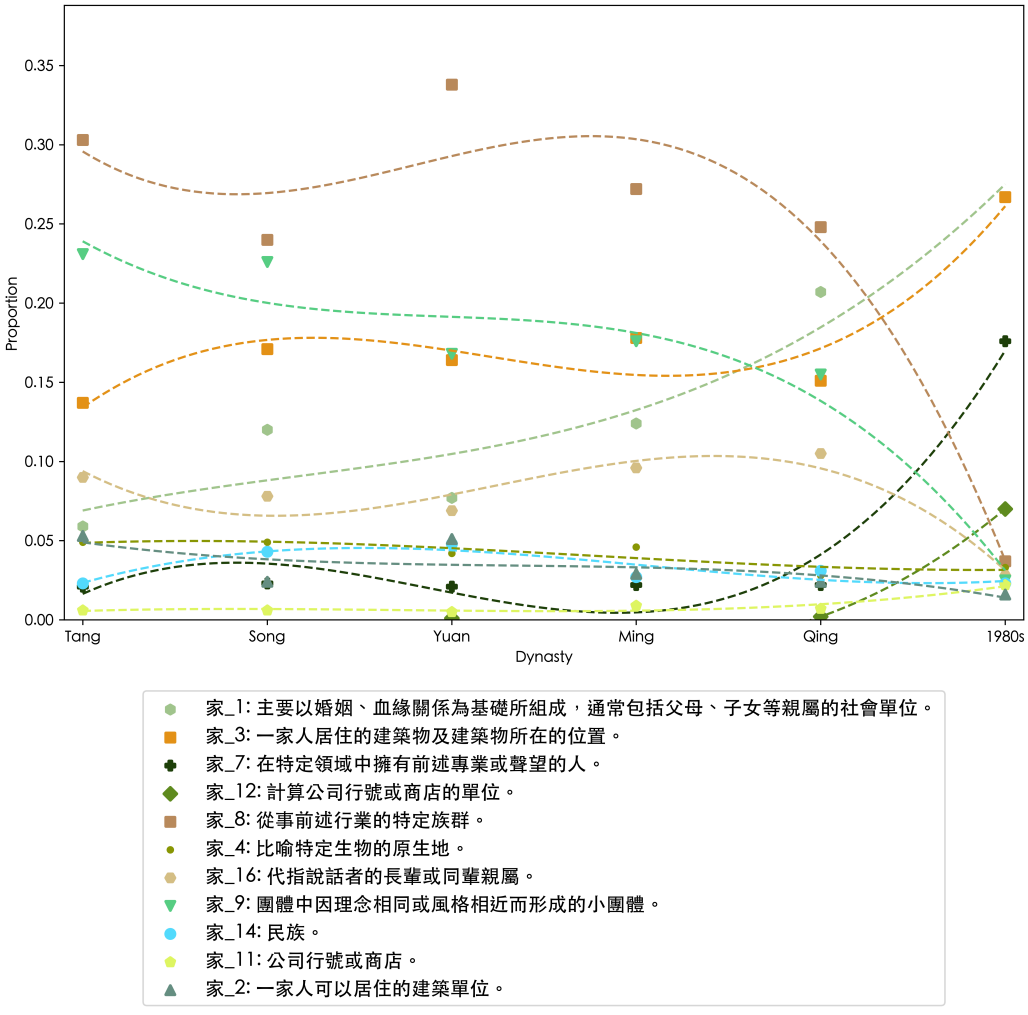


Figure 4.10. Diachronic interactions of senses

doi:10.6342/NTU202004346

The extraction of contextualized embeddings allows for a sketch of usage distribution displayed by proportion and interactions of different senses, as shown in Figure 4.10.

From Figure 4.10, it is shown that senses do compete and cooperate semantically. In present-day Chinese, sense 1 (family), sense 3 (house), and sense 7 (-ist) are shown to be three of the most prominent senses, yet sense 1 does not evolve in identical direction with sense 3 and 7 in Song and Ming. Instead, its rise of sense 1 has indicated that single-character words like *jiā* can be read as 'family', and combined with sense 3, they account for over 60 percent of the usage proportion, while sense 7 is only half of it. Interestingly, both sense 7 and 8 carry the meaning of describing someone's profession, but the contextualized embeddings distinguish the two readings in terms of the percentage. Qualitatively, these are influenced by different schools of thought. Furthermore, it is comparatively rare for *jiā* to serve as adjective 'domestic', or sense 10 as categorical name.

Firstly, Sense 1 and Sense 3 have the highest proportions in the 1980s. This prevalence follows a rapid growth in the use of the two senses. Nonetheless, the evolvement of Sense 3 is more non-monotonous than that of Sense 1 although the two senses start to share a similar upward trend from the Qing dynasty. It is also interesting that Sense 2 takes up around 2% to 5% (5.3%, 2.4%, 5.1%, 2.9%, 2.4% from the Tang to Qing dynasty respectively) of the overall proportion in pre-modern Chinese. Semantically, both Sense 2 and 3 can be used to refer to *jiā* as a physical entity, whereas Sense 1 describes *jiā* as a social unit, as in (1a) to (2c).

The fluctuation of Sense 3 might be a result of the distinguishment of this sense from a similar one, Sense 2, as exemplified in (2a). In contrast, the upward curve of Sense 1 is sharpe, which embodies the expression of *jiā* as a social unit. Yet, it is also found that the results of these three senses include sentences with bi-grams of *jiā* that do not

doi:10.6342/NTU202004346

have a corresponding pre-determined sense, e.g., 國 [家] *guójiā* ‘country; state’, 奴 [家] *nújiā* ‘your servant (humble self-reference by young female)’, as in (2d) and (1b). When a belonging sense is unavailable, it is challenging to disambiguiate the meanings given the token representations. It is especially attributable if an example sentence contains wider context information that obscures the results.

(1) a. 吾 [家 _1] 先人
‘My ancestors’
——— 至正直記 (Yuan)

b. 奴 [家 _1*] 把布接長
‘I, your servant, hold the clothes together to make it a long one’
——— 醒世恆言 (Ming)

(2) a. 始 [家 _2?] 咸陽焉
‘(He) originally settled down in Xianyang.’
——— 廣卓異記 (Song)

b. 余曾至其 [家 _3] 食
‘I once went to his house and had a meal there.’
——— 古清涼傳 (Tang)

c. 在住 [家 _3] 附近經常可以聽到嘰嘰嘰的蟲鳴聲
‘You will often hear the chirping insects near your house.’
——— ASBC (1980s)

d. 豈可於國 [家 _3*] 艱危之時而自圖安閒
‘How could I seek a carefree life for myself while the country is at stake.’
——— 建炎進退志 (Song)

In modern Chinese, Sense 7, 8, 9 are profession-related senses, as in 美聲 [家] *měishēngjiā* ‘bel canto singer’, 樵 [家] 和 獵 [家] *qiáojiā hàn lièjiā* ‘woodman and hunter’, and 儒墨兩 [家] *rúmòliǎngjiā* ‘the Confucian and Mohist schools’. Despite being semantically similar, only Sense 8 and 9 evolve cooperatively, whereas Sense 7 is seen to compete against these two senses. A steep downward curve is seen for both Sense 8 and Sense 9 in the Qing dynasty (from 24.8% and 15.5% in the Qing dynasty to 3.7% and 2.5% in the 1980s), while Sense 7 surges in use during the same time period, and continues to be more and more prominent in the 1980s (from 2.2% in the Qing dynasty to 17.6% in the 1980s). Professions like 醫 [家] *yījiā* ‘doctor’, 史 [家] *shǐjiā* ‘historian’,

doi:10.6342/NTU202004346

and 詩 [家] *shījiā* ‘poet’, as in (3a) through (3c), are mapped to Sense 7 in pre-modern Chinese, and a wide variety of occupations are included for the same sense in modern Chinese, which contributes to the rise and prominence of Sense 7 in the 1980s.

(3) a. 醫 [家 _7] 治痘斑之法
‘Doctors’ acne spot treatments.’
——— 痘疹心法要訣 (Qing)

b. 然史 [家 _7] 多是文詠之士
‘Oftentimes, historians are rather thought of as writers of poetry and prose.’
——— 孔氏雜說 (Song)

c. 此所謂詩 [家 _7] 之中道也
‘This is the so-called teaching of poets.’
——— 文鏡秘府論 (Tang)

d. 我是組織生態學 [家 _7]
‘I’m an organizational ecologist.’
——— ASBC (1980s)

e. 英文的同時有道教信徒和道士和道 [家 _7*] 的意思
‘In English, it can refer to Taoist followers, Taoist priests, and Taoism.’
——— ASBC (1980s)

(4) a. 莫孤負田 [家 _8] 瓦盆
‘Do not disobey the family precepts of a farmer’
——— 類聚名賢樂府群玉 (Yuan)

b. 窮人 [家 _8*] 的男子
‘A man from a poor family.’
——— ASBC (1980s)

(5) a. 自漢至明修輯者七十餘 [家 _9]
‘More than 70 names edited and compiled the works from the Han to Ming dynasty.’
——— 乾元秘旨 (Qing)

b. 漁翁不謂其出 [家 _9*] 人不宜食魚
‘The fisherman does not think he, as a monk, should abstain from having fish.’
——— 第十一尊杯渡羅漢 (Ming)

On top of that, regarding the proportion of usage, Sense 10, 11, and 12 consistently rank the lowest in pre-modern Chinese, but a sudden increase is witnessed for Sense 11 and 12 in the 1980s. Among all the 18 senses of *jiā*, Sense 12 is the only one acting as a classifier, or Nf in the ASBC tagset.

(6) 除了選擇一 [家 _10] 好的…
‘Apart from choosing one good brand…’
——— ASBC (1980s)

doi:10.6342/NTU202004346

(7) a. 但臺灣建築業者號稱上萬 [家 _11]
‘Yet it is claimed that there are up to ten thousands of construction companies in Taiwan.’
——— ASBC (1980s)

b. 頭一 [家 _11] 做生意就勿高興出來
‘The first time he opened up business, he was unhappy.’
——— 海上花列傳 (Qing)

(8) a. …列有四百五十 [家 _12] 值得信賴的商店
‘…lists out 450 brands that are trustworthy.’
——— ASBC (1980s)

b. 是省城第一 [家 _12?] 好主戶
‘They are the best settled household around capital.’
——— 歧路燈 (Qing)

Sense-level embeddings are capable of capturing fine-grained senses and their evolution, yet the contextual information provided from pre-modern Chinese sentences might not be sufficient enough to accurately map the token representations to their belonging senses from the pre-trained language model.

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

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

4.4 Discussion

Following Hamilton et al. (2016b), 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

doi:10.6342/NTU202004346

‘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 (Wang, Chen, and Zhao 王春庭, 陈顺芝, 赵明, 1997: 前言). Dictionaries include Liu and Hou 刘庆俄, 侯刚 (1992) and Wang, Chen, and Zhao 王春庭, 陈顺芝, 赵明 (1997), which lists word entries with meanings that are distinctive between modern and pre-modern times. Detailed information relevant to semantic change is the number of disyllabic word entries, whether the word convey connotations with varying sentiment polarities, and whether certain senses fall into disuse nowadays, which is valuable resources for the comparison with the results of computational methods.

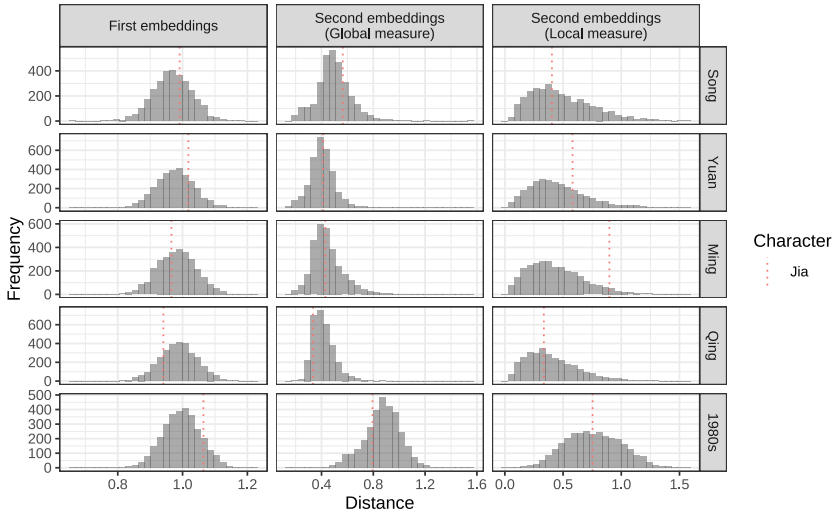


Figure 4.11. Distribution of degree of semantic change for global and local measures

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[2] and 近代漢語語料庫詞頻統計[3], which are the metadata from the 70-million-word Ancient Chinese Corpus (在线古代汉语语料库) by the Ministry of Education, China and Academia Sinica Tagged Corpus of Early Mandarin Chinese (近代漢語語料庫) by Academia Sinica, Taiwan.

[2]http://corpus.zhonghuayuwen.org/resources.aspx
[3]https://elearning.ling.sinica.edu.tw/jindai.html

doi:10.6342/NTU202004346

The case study of *jiā* is based on the assumption that the time-sliced corpus might reflect the similar and different descriptions in language use. While words in Table 2.1 fall into the categories of technological innovations and ideologies, this study chooses *jiā* because of its linguistic and cultural characteristics. In pre-modern Chinese, *jiā* is associated with words that denote physical objects like house.

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

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

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

doi:10.6342/NTU202004346

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

doi:10.6342/NTU202004346

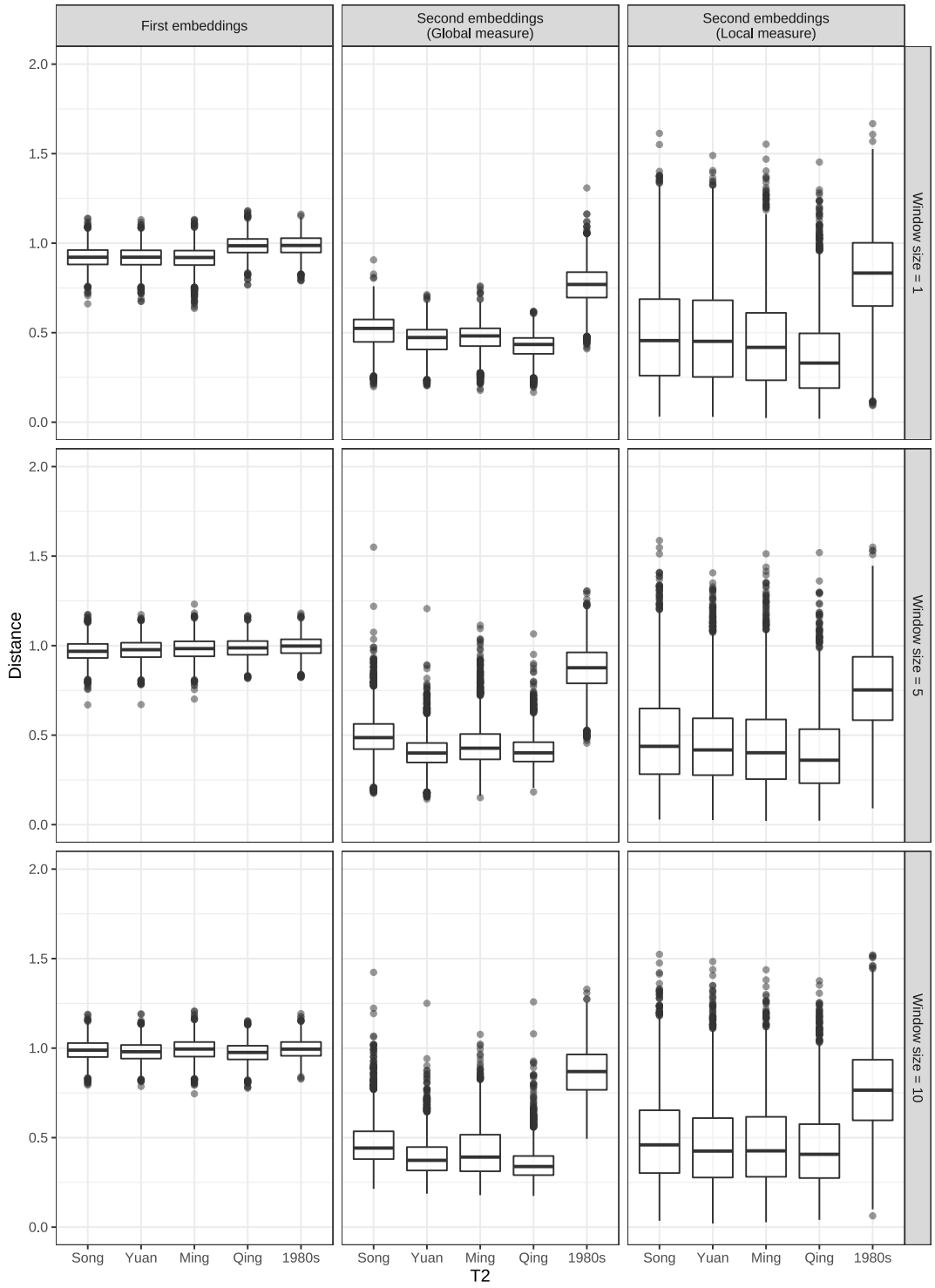
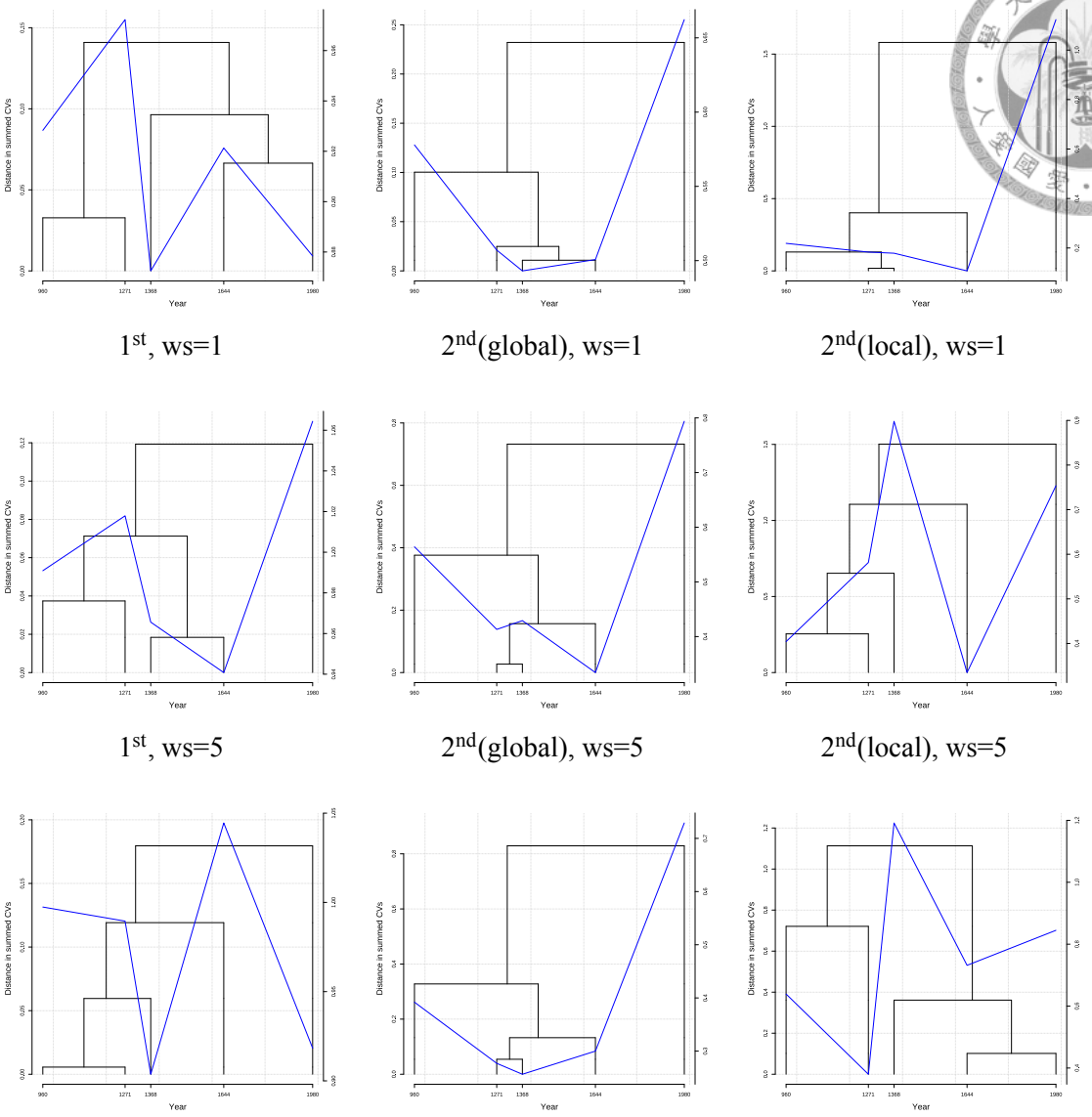


Figure 4.12. Distribution of degree of semantic change for global and local measures

doi:10.6342/NTU202004346



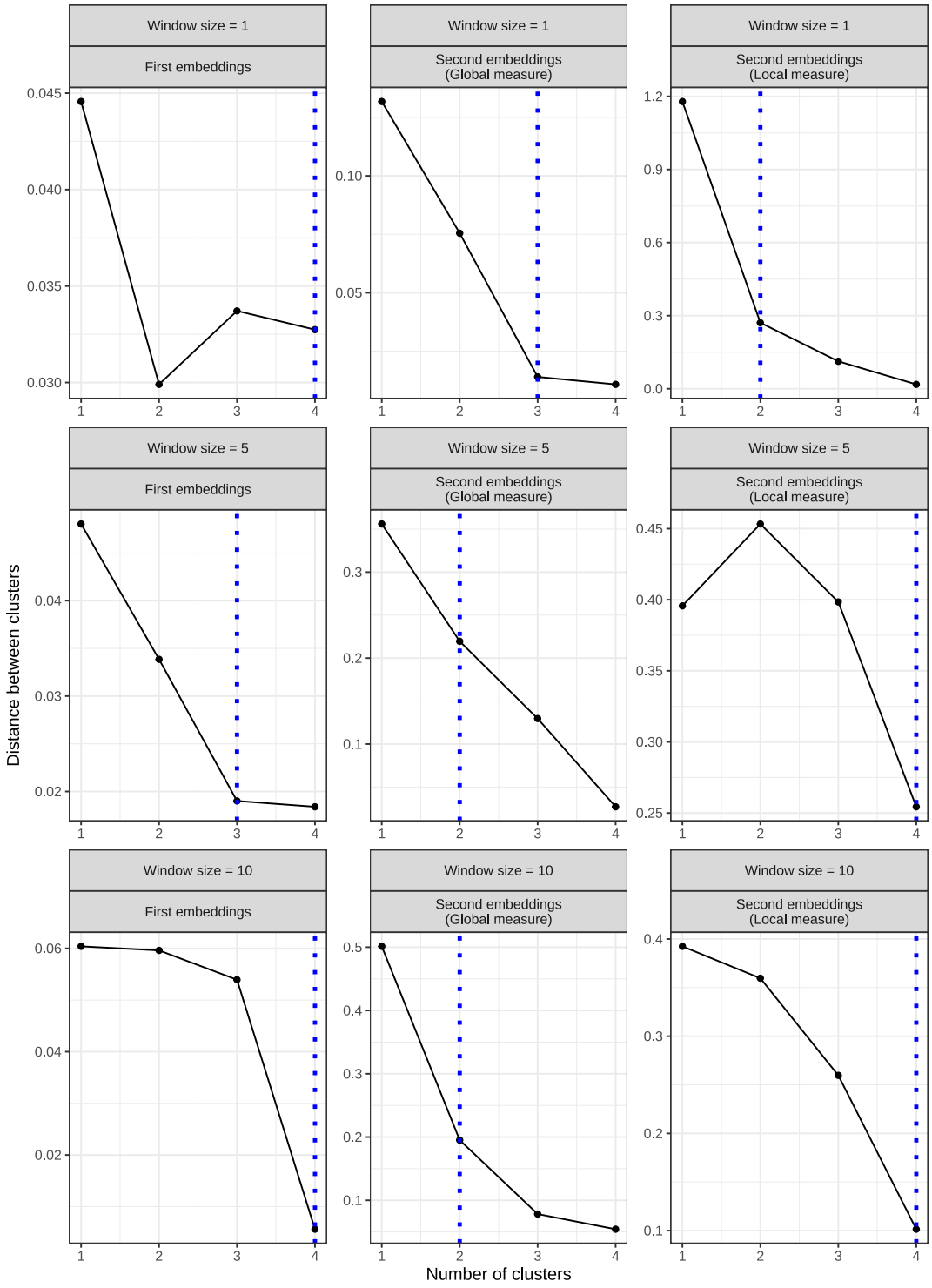
1st, ws=1 | 2nd(global), ws=1 | 2nd(local), ws=1

1st, ws=5 | 2nd(global), ws=5 | 2nd(local), ws=5

1st, ws=10 | 2nd(global), ws=10 | 2nd(local), ws=10

Figure 4.13. VNC periodization of global and local measures

doi:10.6342/NTU202004346



doi:10.6342/NTU202004346

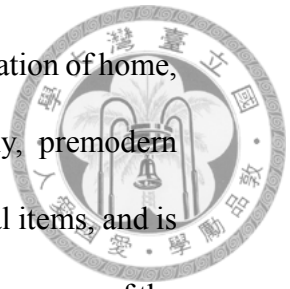
Chapter 5

Conclusions

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

The evolution of jia is a compressed history of the Chinese society and the Chinese language. As linguistic change may not necessarily be reflected in abrupt frequency change when the time stretch is long, a computational analysis can provide further insights into the phenomenon of semantic change by taking into consideration various linguistic factors. 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.

doi:10.6342/NTU202004346



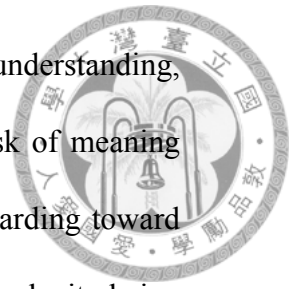
With the advantage of distributional semantic models, the meaning conflation of home, house, and family can be explored as different components. Especially, premodern Chinese is distinguished from the current written form, uses different lexical items, and is mostly in the form of one syllable. The disparity 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.

As discussed in Giulianelli (2019), the fine-tuning of large-scaled pre-trained language models like BERT does not yield satisfactory results of temporal-specific contextualized usage/token representations. As hinted by Giulianelli (2019), the fine-tuning is based on classification task of reconizing the time period of a portion of documents, but the fine-tuned models might instead reflect the style of prominent authors of certain time periods, reering away from baseline representations. Faced with these problems, Kutuzov and Giulianelli (2020) also compares contextualized embeddings with context-independent ones, and find that for semantic change detection, context-independent embeddings are effective.

Semantic change modeling has profound impacts in linguistic analysis. As language is a dynamic phenomenon, a temporal-aware understanding is explored as a starting point. Following the examination of factors, sense evolution prediction, the interaction between

doi:10.6342/NTU202004346

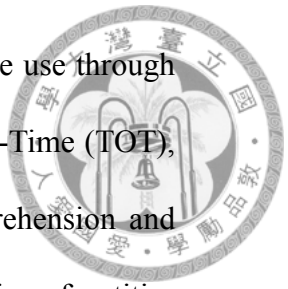


semantic change and different linguistic, cultural factors can deepen our understanding, especially the aspects of polysemy and multi-word expressions. The task of meaning representation from the perspective of semantic change is especially rewarding toward how the modeling of meaning representation can be tweaked, with the complexity being justified, unlike in English, it is not always the case that preprocessing of compound words are taken into account from the beginning.

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

The importance of temporal-aware, diachronic word embeddings have been stressed both for modern texts and historical ones (Huang and Paul, 2019; Rosin et al., 2017; Ruder, 2017). With the accumulation of texts in corpora that are used for search system, i.e., to answer "when" two terms are related to each other, query expansion, and weighted synonyms (Rosin et al., 2017). It is by this aim that this study is motivated, and for

doi:10.6342/NTU202004346



the purpose of achieving more understanding of the properties of language use through the lens of time. As already pointed out in the application of Topics-Over-Time (TOT), temporal-aware meaning representations are beneficial to reading comprehension and background settings, as well as event extraction that exhibit the dynamics of entities involved (Wijaya and Yeniterzi, 2011). Furthermore, the rate of change is another important issue so as to incorporate "time-sensitive" query expansion (QE) (Rosin et al., 2017) to involve the time dimension of linguistic phenomenon more in this rising, flourishing field of study.

As researches combine textual data from various corpora or sources, the detection of semantic change and measurement of degrees of change helps compare not only time-specific needs, but also corpora of different types (Schlechtweg, Hätty, et al., 2019), which becomes increasingly critical with an abundance of textual data presented to us nowadays. The analysis can be further explored by reaching out to other research disciplines and communities, and even the design and functionality of diachronic corpus itself.

doi:10.6342/NTU202004346

References

Antoniak, Maria and David Mimno. (2018). *Evaluating the stability of embedding-based word similarities*. *Transactions of the Association for Computational Linguistics*, *6*, 107–19.

Blank, Andreas. (1999). *Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change*. *Historical Semantics and Cognition*, *61*.

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*. https://arxiv.org/abs/1607.04606.

Bowern, Claire. (2019). *Semantic change and semantic stability: Variation is key*. *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*. *Journal of Artificial Intelligence Research*, *63*, 743–88.

Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. (1996). *Sinica Corpus: Design methodology for balanced corpora*. *Language*, 167–76.

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.

doi:10.6342/NTU202004346

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

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. (2019a). *Time for change: Evaluating models of semantic change without evaluation tasks*. In: *Cambridge Language Sciences Annual Symposium 2019: Perspectives on Language Change*.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. (2019b). *Time-Out: Temporal referencing for robust modeling of lexical semantic change*. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 457–470.

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.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. (2017). *Outta control: Laws of semantic change and inherent biases in word representation models*. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1136–1145.

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*. *The Handbook of Historical Linguistics*, 648–66.

Gablasova, Dana, Vaclav Brezina, and Tony McEnery. (2017). *Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence*. *Language learning*, *67*(S1), 155–79.

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.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. (2020). *Simple, interpretable and stable method for detecting words with usage change across cor-*

 doi:10.6342/NTU202004346

pora. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 538–555.

Gries, Stefan Th and Martin Hilpert. (2012). *Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics*. *The Oxford Handbook of the History of English*, 134–44.

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.

Hengchen, Simon. (2017). *When does it mean? Detecting semantic change in historical texts*. PhD thesis. Université Libre de Bruxelles.

Heuser, Ryan James. (2017). *Word vectors in the Eighteenth century*. *Digital Scholarship in the Humanities*.

Hilpert, Martin. (2019). *Historical linguistics*. *Cognitive Linguistics-A Survey of Linguistic Subfields*, 108–31.

Home. (2020). *The Oxford English Dictionary*. Last accessed: 2021-04-10. 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. https://doi.org/10.18653/v1/P19-1379.

Huang, Chu-Ren and Shu-Kai Hsieh. (2010). *Infrastructure for cross-lingual knowledge representation-towards multilingualism in linguistic studies*. *Taiwan NSC-granted Research Project (NSC 96-2411-H-003-061-MY3)*.

 doi:10.6342/NTU202004346

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.

Jawahar, Ganesh and Djamé Seddah. (2019). *Contextualized diachronic word representations*. In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 35–47.

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

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. (2015). *Statistically significant detection of linguistic change*. In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 625–635.

Kutuzov, Andrey and Mario Giulianelli. (2020). *UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection*. 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.

Lee, John. (2012). *A classical Chinese corpus with nested part-of-speech tags*. In: *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 75–84.

Levy, Omer and Yoav Goldberg. (2014). *Linguistic regularities in sparse and explicit word representations*. In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 171–180.

Li, Bai. (2020). *Evolution of part-of-speech in Classical Chinese*. *arXiv preprint arXiv:2009.11144*.

 doi:10.6342/NTU202004346

Li, Shen, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. (2018). *Analogical reasoning on Chinese morphological and semantic eelations*. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 138–143.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. (2016). *Significance testing of word frequencies in corpora*. *Literary and Linguistic Computing*, *31*(2), 374–97.

Lijffijt, Jefrey, Tanja Säily, and Terttu Nevalainen. (2012). *CEECing the baseline: Lexical stability and significant change in a historical corpus*. In: *Studies in Variation, Contacts and Change in English*. Vol. 10. Research Unit for Variation, Contacts and Change in English (VARIENG).

Liu, Qing-E and Gang Hou 刘庆俄, 侯刚. (1992). *hàn-yǔ cháng-yòng-zì gǔ-jīn-yì duì-bǐ zì-diǎn A comparative dictionary of* 汉语常用字古今义对比字典. hǎi-nán chū-bǎn-shè Hainan Publishing House 海南出版社.

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*. *IEEE transactions on visualization and computer graphics*, *24*(1), 553–62.

Mair, Christian. (1998). *Corpora and the study of the major varieties of English: Issues and results*. *The major varieties of English: Papers from MAVEN 97*, 139–58.

Mallett, Shelley. (2004). *Understanding home: A critical review of the literature*. *The sociological review*, *52*(1), 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*. https://arxiv.org/abs/1301.3781.

Miller, George A. and Walter G. Charles. (1991/2007). *Contextual correlates of semantic similarity*. *Language and Cognitive Processes*, *6*(1), 1–28.

Moisl, Hermann. (2015). *Cluster analysis for corpus linguistics*. Vol. 66. Walter de Gruyter.

doi:10.6342/NTU202004346

Moore, Jeanne. (2000). *Placing home in context*. *Journal of Environmental Psychology*, *20*(3), 207–17.

Moran, Steven and Michael Cysouw. (2018). *The Unicode cookbook for linguists*. Language Science Press.

Nerlich, Brigitte and David D. Clarke. (2001). *Serial metonymy: A study of reference-based polysemisation*. *Journal of Historical Pragmatics*, *2*(2), 245–72.

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.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. (2020). *Stanza: A python natural language processing toolkit for many human languages*. *arXiv preprint arXiv:2003.07082*.

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

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. (2007). *An introduction to exponential random graph (p*) models for social networks*. *Social networks*, *29*(2), 173–91.

Rodda, Martina A., Marco S. G. Senaldi, and Alessandro Lenci. (2017). *Panta Rei: Tracking semantic change with distributional semantics in Ancient Greek*. *Italian Journal of Computational Linguistics*, *3*(1), 11–24.

Rodina, Julia, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. (2020). *ELMo and BERT in semantic change detectionfor Russian*. *arXiv preprint arXiv:2010.03481*.

Rosin, Guy D., Eytan Adar, and Kira Radinsky. (2017). *Learning word relatedness over time*. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1168–1178.

Ruder, Sebastian. (2017). *Word embeddings in 2017: Trends and future directions*. https://ruder.io/word-embeddings-2017/.

 doi:10.6342/NTU202004346

Rudolph, Maja and David Blei. (2018). *Dynamic embeddings for language evolution*. In: *Proceedings of the 2018 World Wide Web Conference*, pp. 1003–1011.

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. http://doi.org/10.29164/19home.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. (2019). *A wind of change: Detecting and evaluating lexical semantic change across times and domains*. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 732–746.

Schlechtweg, Dominik, Sabine Schulte im Walde, and Stefanie Eckmann. (2018). *Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change*. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 169–174.

Shen, Meng-Ying and Zhao-Qing Fu 沈孟穎, 傅朝卿. (2015). *tái-wān xiàn-dài zhù-zhái shè-jì zhī zhuǎn-huà: yǐ 1920 nián-dài zhì 1960 nián-dài gōng-gòng (guó-mín) zhù-zhái wéi lì 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. *shè-jì xué-bào Journal of Design* 設計學報, *20*(4), 43–62.

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

Sixsmith, Judith. (1986). *The meaning of home: An exploratory study of environmental experience*. *Journal of Environmental Psychology*, *6*(4), 281–98.

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

Sturgeon, Donald. (2017). *ctext 0.265: Chinese Text Project API wrapper*.

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

 doi:10.6342/NTU202004346

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

Tang, Xuri. (2018). *A state-of-the-art of semantic change computation*. *Natural Language Engineering*, *24*(5), 649–76.

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

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

Van der Maaten, Laurens and Geoffrey Hinton. (2008). *Visualizing data using t-SNE*. *Journal of Machine Learning Research*, *9*, 2579–605.

Wang, Chun-Ting, Shun-Zhi Chen, and Ming Zhao 王春庭, 陈顺芝, 赵明. (1997). *gǔ-jīn-yì-yì bǐ-jiào cí-diǎn A comparative dictionary of* 古今异义比较词典. jiāng-xī jiào-yù chū-bǎn-shè Jiangxi Education Publishing House 江西教育出版社.

Wang, Xuerui and Andrew McCallum. (2006). *Topics over time: a non-Markov continuous-time model of topical trends*. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 424–433.

Wang, Yun-Lu and Ying Gou 王雲路, 郭穎. (2005). *shì-shuō gǔ-hàn-yǔ zhòng de cí-zhuì "jiā" On the suffix "jia" in ancient Chinese* 試說古漢語中的詞綴"家". *gǔ-hàn-yǔ yán-jiù Studies on Ancient Chinese* 古汉语研究, (1), 29–33.

Wei, Pei-Chuan, P. M. Thompson, Cheng-Hui Liu, Chu-Ren Huang, and Chaofen Sun 魏培泉, 譚樸森, 劉承慧, 黃居仁, 孫朝奮. (1997). *Historical corpora for synchronic and diachronic linguistics studies* 建構一個以共時與歷時語言研究為導向的歷史語料庫. *Computational Linguistics and Chinese Language Processing*, *2*(1), 131–45.

Wevers, Melvin and Marijn Koolen. (2020). *Digital begriffsgeschichte: Tracing semantic change using word embeddings*. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 1–18.

Wijaya, Derry Tanti and Reyyan Yeniterzi. (2011). *Understanding semantic change of words over centuries*. In: *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, pp. 35–40.

Xu, Yang and Charles Kemp. (2015). *A computational evaluation of two laws of semantic change*. In: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*, pp. 2703–2708.

 doi:10.6342/NTU202004346

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. (2018). *Dynamic word embeddings for evolving semantic discovery*. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 673–681.

Yasuoka, Koichi. (2019). *Universal dependencies treebank of the Four Books in Classical Chinese*. In: *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pp. 20–28.

Zellig, Harris. (1954/2015). *Distributional structure*. *Word*, *10*(2-3), 146–62.

Zhang, Xiao-Ping 張小平. (2008). *dāng-dài hàn-yǔ cí-huì fā-zhǎn biàn-huà yán-jiù Studies on Chinese lexicon development and change in contemporary times* 当代汉语词汇发展变化研究. qí-lǔ shū-shè Qilu Press 齐鲁书社.

Zhou, Jun-Xun 周俊勋. (2009). *zhōng-gǔ hàn-yǔ cí-huì yán-jiù gāng-yào Outlines of pre-modern Chinese lexicon* 中古汉语词汇研究纲要. bā-shǔ shū-shè Ba-shu Press 巴蜀书社.

 doi:10.6342/NTU202004346

Appendix A Frequency information of *jiā* from historical corpora constructed by Academia Sinica[1]

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

[1] Accessed from the frequency statistics of corpora compiled by Academia Sinica (https://elearning.ling.sinica.edu.tw/cwordfreq_index.html), including Academia Sinica Tagged Corpus of Old Chinese (中研院上古漢語標記語料庫), Academia Sinica Tagged Corpus of Early Mandarin Chinese (中研院近代漢語標記語料庫), and Academia Sinica Balanced Corpus of Modern Chinese (中研院現代漢語語料庫).

doi:10.6342/NTU202004346

Appendix B List of Matched Word Pairs from the Analogical Reasoning Task

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

doi:10.6342/NTU202004346

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

doi:10.6342/NTU202004346

Appendix C LogDice Scores of Collograms before *jiā*

Rank	Tang		Song		Yuan		Ming		Qing		1980s	
	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice
1	國	9.43	國	10.35	國	10.03	世	10.74	國	10.01	國	10.83
2	出	8.39	百	7.76	世	7.90	國	10.56	世	8.22	大	10.35
3	誰	8.25	氏	7.71	其	7.70	其	7.70	諸	7.68	專	9.80
4	百	7.80	邦	7.66	還	7.64	誰	7.64	其	7.60	回	9.42
5	還	7.76	室	7.64	百	7.63	還	7.56	張	7.24	學	8.91
6	世	7.76	其	7.62	誰	7.62	吾	7.26	誰	7.23	畫	8.87
7	起	7.71	還	7.59	諸	7.55	起	7.24	人	7.16	客	8.82
8	其	7.45	起	7.56	起	7.51	一	7.22	百	7.09	作	8.50
9	室	7.43	世	7.52	吾	7.43	百	7.15	我	7.07	術	8.29
10	邦	7.39	誰	7.39	室	7.19	漢	7.01	吾	6.96	人	8.03
11	田	7.35	諸	7.32	故	7.16	諸	7.01	一	6.83	儒	7.70
12	漢	7.19	吾	7.31	一	7.12	室	6.92	回	6.83	一	7.59
13	家	6.96	漢	7.28	仙	7.08	于	6.91	室	6.82	全	7.44
14	吾	6.95	一	7.15	君	7.04	邦	6.90	還	6.77	在	7.43
15	在	6.92	周	6.95	邦	7.03	人	6.90	在	6.75	玩	7.13

doi:10.6342/NTU202004346

Appendix D LogDice Scores of Collograms after *jiā*

Rank	Tang		Song		Yuan		Ming		Qing		1980s	
	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice
1	貧	8.62	傳	8.43	貧	7.57	忠	9.13	貧	8.12	庭	10.95
2	語	7.48	貧	7.91	世	7.32	恭	8.35	口	7.92	長	9.23
3	人	7.10	法	7.29	庭	7.24	毅	7.69	人	7.52	裡	9.16
4	家	6.96	世	7.05	藏	7.15	高	7.67	居	7.34	都	9.00
5	僮	6.91	屬	6.98	法	7.12	睿	7.66	莊	7.21	族	8.53
6	室	6.89	人	6.94	傳	7.09	貧	7.48	屬	7.18	屬	8.30
7	口	6.86	語	6.91	人	6.95	居	7.48	語	6.98	公	8.06
8	莊	6.63	藏	6.79	居	6.91	仁	7.25	藏	6.96	人	8.05
9	財	6.56	庭	6.75	之	6.74	顯	7.17	奴	6.87	鄉	7.87
10	有	6.51	居	6.68	學	6.73	人	7.14	堰	6.84	具	7.74
11	註	6.50	之	6.65	有	6.61	文	7.13	法	6.73	中	7.70
12	國	6.48	財	6.64	聲	6.56	世	6.86	禮	6.71	電	7.47
13	世	6.47	聲	6.64	者	6.48	屬	6.81	橋	6.69	的	7.27
14	聲	6.46	事	6.56	焉	6.47	之	6.79	的	6.67	安	6.87
15	事	6.44	有	6.55	家	6.42	法	6.69	之	6.52	認	6.81

doi:10.6342/NTU202004346

Appendix E LogDice Scores of Collograms with *jiā*

Rank	Tang		Song		Yuan		Ming		Qing		1980s	
	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice	1-gram	logDice
1	國	8.61	國	9.40	國	9.10	世	9.84	國	9.07	庭	9.95
2	貧	7.87	傳	7.80	世	7.64	國	9.61	世	7.59	國	9.85
3	出	7.48	世	7.30	人	6.94	忠	8.14	貧	7.35	大	9.42
4	誰	7.30	貧	7.18	其	6.87	恭	7.36	人	7.35	專	8.84
5	世	7.25	邦	7.04	貧	6.86	居	7.06	口	6.93	回	8.46
6	室	7.19	室	7.02	傳	6.81	人	7.02	居	6.82	長	8.25
7	家	6.96	百	6.88	百	6.80	其	6.87	諸	6.79	裡	8.17
8	百	6.93	其	6.75	誰	6.74	高	6.80	其	6.74	學	8.08
9	還	6.88	氏	6.74	還	6.73	誰	6.74	室	6.49	人	8.04
10	人	6.81	人	6.69	諸	6.69	貧	6.72	張	6.34	畫	7.93
11	起	6.81	還	6.68	居	6.67	毅	6.70	誰	6.3	客	7.85
12	邦	6.74	起	6.61	起	6.64	睿	6.66	莊	6.29	作	7.66
13	其	6.60	法	6.59	室	6.54	還	6.65	家	6.28	族	7.55
14	在	6.56	居	6.54	吾	6.53	一	6.53	大	6.28	屬	7.30
15	語	6.54	誰	6.50	法	6.53	室	6.46	我	6.26	公	7.27

doi:10.6342/NTU202004346

Appendix F Chinese WordNet (CWN) Senses of *jiā* and their Sense Evolution

Sense	Pos	Definition	Example sentences	Percentage (%) $N = 1000$					
				Tang	Song	Yuan	Ming	Qing	1980s
家_1	Na	主要以婚姻、血緣關係為基礎所組成，通常包括父母、子女等親屬的社會單位。	僕人問：「我們[家]的鵝，一隻會叫，一隻不會叫，殺哪隻比較好呢？」	5.9	12.0	7.7	12.4	20.7	26.7
家_2	Na	一家人可以居住的建築單位。	布拉格像[家]的旅館，走小而美溫馨路線。	5.3	2.4	5.1	2.9	2.4	1.6
家_3	Nc	一家人居住的建築物及建築物所在的位置。	公公是當時頗有權勢的人，[家]裡布置自然豪華，點心也特別精緻好吃。	13.7	17.1	16.4	17.8	15.1	26.7
家_4	Na	比喻特定生物的原生地。	國王企鵝的[家]在北極喔。	4.9	4.9	4.2	4.6	2.8	3.3
家_5	Na	比喻動物棲息的地方。	台灣首富郭台銘要替流浪狗蓋一個[家]了！	1.1	0.7	0.7	0.3	0.5	0.4
家_6	Na	比喻存放特定物品的專用空間。	最近想幫我的電腦換個[家]，原本的CASE是5年前還不太懂電腦買的雜牌，超重，散熱也不佳。	0.0	0.2	0.1	0.0	0.2	0.5
家_7	Na	在特定領域中擁有前述專業或聲望的人。	他認為宮崎駿更應該說是一位動漫[家]，他在動漫上的成就更甚於漫畫。	2.1	2.3	2.1	2.2	2.2	17.6
家_8	Na	從事前述行業的特定族群。	木和石頭搭建的亭裏，今天又放了皮袍和木柴，想又是燃[家]和獵家送來的東西了。	30.3	24.0	33.8	27.2	24.8	3.7
家_9	Na	團體中因理念相同或風格相近而形成的小團體。	每一[家]的拳法皆不相同，走步也不相同，好像基本的馬步是一定有的。	23.1	22.6	16.8	17.6	15.5	2.5
家_10	Na	企業品牌的專用名稱。	看了這麼多介紹，品牌繁雜，到底哪一[家]的好用？	0.0	0.0	0.0	0.0	0.0	0.7
家_11	Na	公司行號或商店。	這[家]真的很好吃耶！可是，新竹好像好店都不長留，爛店遺千年？	0.6	0.6	0.5	0.9	0.7	2.2
家_12	Nf	計算公司行號或商店的單位。	他看準新總統上任後，經濟會變好，近期準備再開第五[家]店。	0.0	0.0	0.1	0.0	0.2	7.0
家_13	Na	特定事件中地位相當的參與者。	在遊戲中除了盡量讓自己的出牌順利外，也同時要想辦法阻礙其他三[家]的出牌。	1.2	0.6	0.3	0.5	0.2	0.6
家_14	Na	民族。	此次造訪台中科學博物館，發現中庭的地方，擺了這麼一座真實的「雲南苗[家]吊腳樓」。	2.3	4.3	4.8	2.7	3.1	2.3
家_15	A	形容被人飼養的。	如果您願意支持我，您也可以幫助這些貓咪，讓牠們成為有人疼的[家]貓。	0.4	0.3	0.3	0.6	0.2	0.8
家_16	Nep	代指說話者的長輩或同輩親屬。	到那裡已經是中午了，未及開口，[家]叔就邀我們一起出去吃飯。	9.0	7.8	6.9	9.6	10.5	2.8
家_17	Na	屬於前述年齡層或性別的人。	妳是個女孩[家]，要有羞恥心，不要一直常常主動去找他。	0.1	0.2	0.2	0.7	0.9	0.6
Total				100.0	100.0	100.0	100.0	100.0	100.0

doi:10.6342/NTU202004346