Table of Contents

List of Figures . . . . . **ii**

List of Tables . . . . . **iii**

References **4**

List of Figures

List of Tables

歷時語料庫

早在 1982，語言學家 Sinclair 如此描繪了對於未來語料庫模樣的想像，文字的保存是大量的，其中是緩慢卻不斷變動的語料（“vast, slowly changing stores of text”），亦是對語言演化很詳細的紀錄（“detailed evidence of language evolution”）。

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)。

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。

(1) Corpus of Historical American English (COHA, 1810-2010)[1]

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

(2) A Representative Corpus of Historical English Registers (ARCHER, 1600-1999)[2]

(3) Royal Society Corpus (RSC, 1665-1869)[3]

(4) Corpus of Late Modern English Texts (CLMET, 1710-1920)[4]

(5) Hansard Corpus (1803-2005)[5]

(6) Sheffield Corpus of Chinese[6]

(7) Academia Sinica Tagged Corpus of Old Chinese (中央研究院上古漢語語料庫, from pre-Qing to pre-Han)[7], Academia Sinica Tagged Corpus of Middle Chinese (中央研究院中古漢語語料庫, from late-Han to the Six Dynasties)[8], and Academia Sinica Tagged Corpus of Early Mandarin Chinese (中央研究院近代漢語語料庫, from Tang to Qing)[9]. The division into 3 corpora is based on the development of Chinese syntax to offer a synchronic sketch of Chinese and a basis for diachronic comparisons. In the 3 Academia Sinica tagged corpora, raw texts are available, with part of the texts imported from Scripta Sinica (漢籍全文資料庫計畫). It is also worth noting that the Google Books project for Chinese is not available until the year of 1950, and the latest date is 2008. 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.

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，

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

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. (2016) 的全域鄰近詞法，以搭配詞的相似度數值組成二階向量（second-order embedding），提高語意表徵的精確度來比較各時代向量的方法，求其相關係數和語意變遷程度之間的關聯。並從詞彙的意義分布與互動，描繪出不同詞意的消長與變動。此外，本研究也同時採用以變異程度為基礎的近鄰群聚分析法（Variability-based Neighbor Clustering, VNC）(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。

一階向量由詞向量模型的原始數值組成，例如：以 Word2Vec 訓練而成的 300 維向量。Hamilton et al. (2016) 提出以二階向量計算語意變遷的程度，將某字詞與其鄰近詞（neighboring word）的相似度串連成數列，來代表這個字詞的語意表徵，更可依據是否取其所有鄰近詞，抑或是部分鄰近詞，細分成全域法（global measure）及部分法（local measure）[10]，因為以整個語言來看，語意是相對穩定的，而部分法可幫助我們抓取出語意變化較明顯的鄰近詞區段。

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，其應用範圍更可擴及到詞彙與文化變遷的探索。

[10]從Hamilton et al. (2016) 的研究結果中，發現 25 至 50 個鄰近詞即可。

References

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

Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. (1996). Sinica Corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*. Kyung Hee University, pp. 167–176. http://asbc.iis.sinica.edu.tw.

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*, 134–144. https://doi.org/10.1093/oxfordhb/9780199922765.013.0014.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. (2016). 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. http://dx.doi.org/10.18653/v1/D16-1229.

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

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.