東亞觀念史集刊　第七期
2014 年 12 月　頁 427-454

從文言到白話：

《新青年》雜誌語言變化統計研究

何立行、余清祥、鄭文惠*

摘　　要

現代漢語與古代漢語最重要的區別之一，乃是書面語以語體文爲主，又稱爲白話文，與古代的文言文相對。目前學者考察現代白話的發展，多推前至晚清傳教士所辦報刊，但使白話有效取代文言成爲主流書面語的是五四運動，而五四運動時期倡導白話文最力的莫過於《新青年》雜誌。過去學者主要以文本分析爲研究方法，就理論建樹、創作實踐和議論宣傳等方面，探討《新青年》雜誌對白話文通行和白話文學發展的貢獻。然而，漢語書面語由以文言爲主變爲白話當家，從文人學者在《新青年》一類的刊物上提出主張，到眞正在社會上普及，歷經了多長的時間？轉變的過程爲何？白話什麼時候取代了文言？如何證明？這些問題恐怕難以用傳統的研究方式來回答，因爲，再勤奮的研究者也無法以人力按時序遍讀五四前後現存的巨量文獻，一一區分文白計算消長。但我們能否藉助數位研究方法，另闢蹊徑，尋找答案？或許，從建立客觀（而非直覺）判讀文、白篇章的有效工具開始，是一個值得探索的方向。

* 作者何立行現爲清華大學中國文學系博士後研究員；余清祥現任政治大學統計學系教授；鄭文惠現任政治大學中國文學系教授。

本文以《新青年》全文共十一卷爲素材，透過統計方法比較各卷的異同，觀察語言轉換歷程，尋找可以建立客觀判讀文、白篇章的指標。使用的方法大致分爲兩類：監督學習（supervised learning）、非監督學習（un-supervised learning）。第一類先設定比較用的指標（或是變數、關鍵詞），再分析各卷的指標特性；第二類不預設比較標的，以不同角度探討文章風格，藉以找出區隔文、白篇章的關鍵因素。本研究的監督學習選用文、白的特定虛字，選擇虛字而非實詞作爲統計對象，乃是爲了將文章內容對語言形式的影響降至最低，驗證從慣用虛字區別文、白篇章的可行性。非監督學習的分析角度以用字、句子架構爲主要方向，因爲字彙多寡、使用頻率等統計數據，在比較文學中歷來都用以判斷寫作風格。

無論監督學習式的虛字分析或非監督學習式的用字習慣分析，都能反映出《新青年》初期與晚期文體的變化。就發展客觀判讀工具而言，以虛字爲指標也許較具潛力。值得注意的是，在總字數、不同字數、每句字數等的比較中，我們發現文言與白話有著明顯差異：文言篇章總字數少而用字多，白話篇章則是總字數多而用字少。明顯可看出白話文主要俾利於世俗啟蒙，因而總字數多而用字少；此外，我們或可借用生物多樣性的概念，追問文言、白話兩者內部生態系的差異；並進一步思考，在這樣的差異下，除了虛字、字彙總數及其使用比例之外，還有哪些具有成爲客觀區辨指標潛力的語言表徵，值得我們繼續開發。

關鍵詞：文體分析、五四運動、《新青年》、虛字分析、生物多樣性

From Classical Chinese to Modern Chinese:

A Study of Function Words from *Xin Qing Nian*

Ho Li-hsing, Yue Ching-syang, Cheng Wen-huei

Abstract

Is it possible for computers to tell whether a text was written in classical Chinese or vernacular modern Chinese? Can the new developments of digital humanities help find out the transformation of written Chinese language during the late Qing and early Republic? As previous scholars have pointed out, in the early stage of the history of modern Chinese, missionaries and reformists only used vernacular language as a tool to enlighten the public. Classical Chinese remained the standard written language until May Forth Movement in 1919, when *Xin Qing Nian* became the most influential publication. Throughout the last century, scholars have scrutinized the theoretical arguments and creative writing practices in *Xin Qing Nian* and several other progressive magazines to delineate the changing history of the language. But questions such as how long did it take for literati as well as the general public to adopt the vernacular language as the written standard, or how did the new standard spread from radical revolutionary magazines to other publications

like entertainment magazines or newspapers, remain unanswered. If we can teach computers to distinguish between classical and modern Chinese, it would be possible to bring in much more digitized texts in that period to study and to answer those questions. To achieve this goal, we adopt the concept of "genome mapping" to differentiate between classical and modern Chinese in this study.

We propose two approaches, supervised learning and un-supervised learning, to compare the differences in writing style between classical Chinese and modern Chinese. In addition to concepts and methods used in a lexical analysis, we also adapt the ideas in ecology. Supervised learning has long been used in linguistics to differentiate authorship via keywords. We choose ten function words for classical and modern Chinese each as the keywords, and we use Gini's index of volumes 1 and 11 from *Xin Qing Nian* to demonstrate the comparison.

There are no standard operating procedures for applying the un-supervised learning, and it is the main reason why this type of approaches is difficult to implement. In this study, we choose the diversity indices for un-supervising learning, for example, Gini's index, entropy, and Simpson's index, for measuring the statistical dispersion and evenness (or equality) of the words used. Based on our analyses, it seems that the later volumes (such as Volume 11) have lower species diversity, indicating that people can read articles without recognizing many words, which matches to the purpose of the May 4th Movement.

Keywords: Stylistic Analysis, May 4th Movement, La Jeunesse, Function Words Analysis, Species Diversity

從文言到白話：

《新青年》雜誌語言變化統計研究*

何立行、余清祥、鄭文惠

一、引言與研究目的

現代漢語與古代漢語最重要的區別之一，乃是書面語以語體文爲主，又稱爲白話文，與古代的文言文相對。

按照王力《漢語史稿》的分期，漢語的演變可以劃分爲四個時期：公元三世紀（五胡亂華）以前爲上古期，公元四世紀到十二世紀（南宋前半）爲中古期，公元十三世紀到十九世紀（鴉片戰爭）爲近代，二十世紀（五四運動以後）爲現代。兩分期之間常爲語言轉換的過渡階段，1840 年鴉片戰爭至 1919 年五四運動，即是漢語從近代轉入現代的過渡階段。[1]

---

* 本文於 2014 年 11 月 21-22 日於日本關西大學發表，係由日本關西大學、韓國翰林大學翰林科學院、政治大學「中國近現代思想及文學史專業數據庫（1830-1930）」計畫暨「觀念・事件・行動：中國近現代觀念形成與演變的數位人文研究」科技部整合型計畫共同主辦的「研究型數據庫與數位人文研究：東亞近現代觀念的形成與演變」國際工作坊。會議上承蒙沈國威教授、內田慶市教授提供寶貴意見，暨《東亞觀念史集刊》兩位匿名審查委員的指正，謹致謝忱！

1 王力：《漢語史稿》（北京：中華書局，1980 年，2004 年重印），頁 43-44。需要特別注意的是，此一分期爲語言分期，並非文學史分期，所以「近代」所指與一般文學史相當不同。

在上述四個時期的前三期中，書面語大抵以文言爲主。文言是一套與日常口語分離的獨立語言系統。由於與日常口語分離，文言既不受空間（地域方言）的限制，也不受時間（口語變化）的影響；相對於分支繁多、隨時變異的口語而言，文言是一個相對封閉而穩定的語言系統。另一方面，亦由於與日常口語分離，文言並不是任何人的母語，而是一種必須特別通過學習才能運用的語言；也因此，閱讀與書寫文言成爲專屬於知識階層的技能。1840 年鴉片戰爭後，中國受到列強武力和西方文明的衝擊，傳統的社會結構和生活方式都面臨了巨大的挑戰。新知識、新詞彙、新事物、新概念的輸入、新式傳媒的興起、大眾文化的興起、因應啟蒙革命的講演宣傳、傳教士宣教語言、文化翻譯的跨語際實踐等等，均衝擊著原本穩定的文言系統，漢語書面語開始發生巨大的轉變，歷經八十年左右，現代白話文方取代文言在漢語書面語中的地位。

目前學者考察現代白話文的歷史，多推前至晚清傳教士所辦報刊，[2]其後則有清末維新革命派文人繼起提倡，但使白話文有效取代文言文成爲主流書面語的是五四運動。[3]五四運動時期倡導白話文最力的莫過於《新青年》雜誌。過去學者主要以文本分析爲研究方法，就理論建樹、創作實踐和議論宣傳等方面，探討《新青年》雜誌對白話文通行和白話文學發展的貢獻。[4]然而，漢語書面語由以文言爲主變爲白話當家，從文人學者提倡實踐，到眞正在社會上普及，歷經了多長的時間？轉變的過程爲何？白話什麼時候大幅度的取代了文言？如何證明？這些問題恐怕難以用傳統的研究方式來回答，因爲，再勤

---

2 傳教士與現代白話文的關係，參見相關研究如段懷清：〈倪維思夫婦釋經布道文之語言實驗〉，《江蘇大學學報（社會科學版）》第 14 卷第 4 期（2012 年 7 月），頁 35-40。

3 David Der-Wei Wang, “Chinese Literature from 1841 to 1937” in Kang-I Sun Chang ed., *The Cambridge History of Chinese Literature. Volume II*. (Cambridge: Cambridge University Press, 2010).

4 參見黎錦熙：《國語運動史綱》（北京：商務印書館，2011 年）。

奮的研究者也無法以人力按時序遍讀五四前後現存的巨量文獻，一一區分文白計算消長。但我們能否藉助數位研究方法，另闢蹊徑，尋找答案？或許，從建立客觀（而非直覺）判讀文、白篇章的有效工具開始，是一個值得探索的方向。

白話文顧名思義是以日常口語爲基礎的書面語；實則漢語方言眾多，日常口語因地而異，現代白話文主要建立在漢語口語中的北方官話之上。由於中國歷史上政權統一的時期大多定都北方，廣義的北方話作爲中國通用口語爲基礎，由來已久。口語進入書面，在上古期已經可以看到，《論語》就是最好的例子，《史記》、《世說新語》亦雜用當時口語。然而相對於文言，仍非主流。南北朝駢文盛行之後，口語入書面的情況更少，除變文、判決書之類，中古期的書面語基本上由文言壟斷。直到近代期，宋代話本採用以北方話爲基礎的語體，元明清小說如《水滸傳》、《西遊記》、《儒林外史》、《紅樓夢》等，使用的也都是以廣義的北方話寫成而流通區域不限於北方的語體文。[5]現代期的北方日常口語基本上承續近代期，語音詞彙和部分句式雖有變異，穩定度雖不如文言，但整體而言系統是相承的。因此語體文也是相承的。這意味著，無論是傳教士爲了使信仰普及大眾而選擇以口語書寫宗教小說或文學，晚清文人爲了宣揚愛國理念提倡白話小說印製白話報，或是五四文人倡議文學改良推行白話文運動，他們使用的白話大部分既是對當時口語（北方官話）的模仿，亦前有所承，延續著近代期語體文的傳統。[6]換言之，鴉片戰爭以後的變化，與其說是文言文「演變」爲白話文，不如說是原本雙軌並行的兩個書面語系統在鴉片戰爭後地位與適用範圍產生變化，經過五四運動，白話文終於取

---

5 王力特別強調：「這裡的『北方話』是廣義的，不一定就是華北的話。例如《儒林外史》可能是用皖北話寫的。」王力：《漢語史稿》，頁46。

6 對當時口語的模仿，在五四語言論戰時期引起很大爭議，保守派抨擊白話文運動讓優雅的書面語墮落爲市井俚俗的「引車賣漿者言」。革新派回應的方式就是建立白話文學史，從古典文學中尋找優秀的語體文作品，其中最佳典範即是宋元話本和明清章回小說這些以近代期北方話寫成的文學作品。

代了文言文，成爲主流書面語。正因我們能將文言文與白話文視爲兩套堪稱穩定的書面語系統，嘗試爲兩者的差異建立客觀判別工具才不至於成爲一個耗時費力而難有進展的浩大工程。

我們初步的嘗試是以《新青年》全文共十一卷爲素材，透過統計方法比較各卷的異同，觀察語言轉換歷程，尋找可以建立客觀判讀文、白篇章的指標。使用的方法分爲兩類：監督學習（supervised learning）與非監督學習（un-supervised learning）。第一類先設定比較用的指標（或是變數、關鍵詞），再分析各卷的指標特性；第二類不預設比較標的，以不同角度探討文章風格，藉以找出區隔文、白篇章的關鍵因素。本研究的監督學習選用文、白的特定虛字。虛字就句法功能而言亦可稱爲虛詞、功能詞，本身不具完整詞彙意義，但卻具有語法意義或功能。漢語虛詞主要包括副詞、介詞、連詞、助詞、嘆詞、象聲詞等。本文選擇具有虛詞功能的字而非實詞作爲統計對象，乃是爲了將文章內容對語言形式的影響降至最低，驗證從慣用虛字區別文、白篇章的可行性。非監督學習的分析角度以用字、句子架構爲主要方向，因爲字彙多寡、使用頻率等統計數據，在比較文學研究中歷來都用以判斷寫作風格。

必須加以說明的是，無論監督學習的虛詞統計或非監督學習的用字、分句長短統計，都只是嘗試尋找可能的機器判讀指標，未敢宣稱能捕捉文白差異間本質性的文體因子。[7]《新青年》文白篇章增減的數據也並不能直接作爲白話取代文言的證據。本研究僅是爲了嘗試回答白話文運動史大問題所作的一個基礎性的準備工作。換言之，這只是一個尚在大膽假設並初步求證階段所作的前驅實驗。之所以選擇《新青年》作爲初步研究的對象，主要除了刊物本身在語言文學史上的關鍵地位，也是因爲刊物裡的白話文具有相當高的文體自覺，語言應具相當的典型性。[8]然而，倚賴菁英色彩強烈、作者多半具有高度文體

[7] 感謝匿名審查者的提醒。

[8] 以胡適爲領袖的《新青年》作者群對文體的自覺不僅在於區辨文白，更

自覺的同仁刊物《新青年》所研究出來的指標，很可能較近似於一種「領先指標」，亦即預兆而非呈現文白勢力消長的整體實況。若要眞實反映白話運動史的面貌，勢必有待後續一邊修訂在這個實驗裡建立的工具，一邊進行更大範圍的研究。

二、研究方法

本文以數位方法判別文言文、白話文的差異，切入點是以用於處理大數據（或海量資料；Big Data）的統計學習（Statistical Learning）理論，這類型方法的特徵是根據資料特性與問題定義，從資料中學習（或尋找）與問題有關的關鍵資訊，可搭配與問題相關的專業領域知識。統計學習理論並不執著於時髦或是複雜的計量模型，而是從描述資料的基本統計量（Descriptive Statistics）出發，包括樣本平均數、樣本變異數等，透過直方圖（Histogram）、箱型圖（Boxplot）之類的圖表找出資料的主要特徵。這種分析想法最早在1970年代由美國統計學家 Tukey 提出，也稱爲探索性資料分析（Exploratory Data Analysis, EDA），Tukey 認爲統計分析不應只是套用模型，更需找出資料的重要特性，如果要套用複雜模型，對症下藥才能事半功倍。

上述的統計學習與 “Data Driven”（資料驅動）的想法接近，和電腦程式撰寫的典範（programming paradigm）相關，[9]主張程式流程需與資料特性結合，文件處理程式語言（text-processing language）屬於資料驅動的範例之一。事實上，一般認爲 Tukey 提出的 EDA 也促成了

在於精鍊白話，要讓白話能成爲抒情言志的書面語言，而非僅只是載道啟蒙的工具。這正是他們承續而又有別於晚清啟蒙論者之處。參見李孝悌：〈胡適與白話文學運動的再評估——從清末的白話文談起〉，收於周策縱等：《胡適與近代中國》（臺北：時報文化出版公司，1991 年），頁 25-26。感謝匿名審查者提醒。

9 Wirfs-Brock, R. and Wilkerson, B., “Object-oriented design: a responsibility-driven approach”, in *Conference Proceedings on Object-Oriented Programming Systems, Languages and Applications* (New York: ACM,1989): 71-75.

統計計算（statistical computing）的發展，讓資料分析更爲即時、視覺化，使用者能根據需求修正分析方向，找出資料的重要趨勢、模式，以及判斷哪些觀察值屬於異常值（或是離群值，outliers）。統計學習大致可分爲兩種方式：監督學習（supervised learning）、非監督學習（unsupervised learning），主要差異在於使用者是否提供參考資訊。[10]

監督學習搭配使用者提供的資訊，以本研究判別文言文、白話文的文體爲例，除了大略可判斷《新青年》前幾卷接近文言文、後面卷次偏向白話文外，研究者還需要提供相關變數（例如：關鍵字詞），再套用數量模型驗證這些變數是否能區隔前後各卷。數位人文研究大多數算是監督學習。以最早應用數位技術的紅學研究爲例，像是趙岡與陳鍾毅以「兒」、「在」、「了」、「的」、「著」五個虛字比較《紅樓夢》前八十回、後四十回，[11] 余清祥則使用 17 個變數（包括詩詞字數、是否有「下回分解」等），[12] 都是由研究者指定變數。近年數位研究更加蓬勃發展，使用方法也多屬於監督學習，像是金觀濤等人也嘗試使用共現詞頻分析去探討「華人」觀念的起源。[13]

非監督學習比較著重於資料驅動，盡量減少人爲的主觀影響，透過 EDA 的初步資料分析發掘參考變數。以迴歸分析爲例，在確定研究目標的應變數（或是被解釋變數，dependent variable）後，通常會透過與應變數的關聯性（association），找出相關的解釋變數（或是自變數，independent variable）。監督學習由研究者指定自變數，非監督學習則由 EDA 等方法找出可能的自變數；換言之，非監督學習會比

---

[10] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer Series in Statistics. 2009.

[11] 趙岡、陳鍾毅：《紅樓夢研究新編》（臺北：聯經出版公司，1980 年）。

[12] 余清祥：〈統計在紅樓夢的應用〉，《國立政治大學學報》第 76 期（1998 年 6 月），頁 303-327。

[13] 金觀濤、邱偉雲、劉昭麟：〈「共現」詞頻分析及其運用——以「華人」觀念起源爲例〉，收於項潔等主編：《數位人文要義：尋找類型與軌跡》（臺北：臺灣大學出版中心，2012 年），頁 141-170。

監督學習多出幾個步驟，以資料找出可能的變數，一旦確定相關變數後，接下來監督學習、非監督學習的分析流程相同，決定最佳的迴歸模型爲目標。不過，EDA 方法多半需花費心力及時間，如同近年知名的資料採礦（或是資料挖掘，Data Mining），顧名思義，採礦本身得鉅細靡遺地篩選，避免一時疏忽造成功虧一簣。

監督學習的優點在於目標明確，執行上相對快速容易，但研究者挑選的變數扮演關鍵角色，分析結果常因變數不同而有不小差異。非監督學習的優勢在於資料驅動，只要趨勢明顯、資料量充足，多半能找出重要變數，然而篩選變數非常花費時間，有時也需要研究者提供意見回饋，對於電腦科技、統計分析等領域的依賴更大，跨領域整合更形重要。本文將兼採兩種分析想法，除了配合專業知識挑選重要變數外，也將以 EDA 的方式尋找可能的變數，但因爲非監督學習較缺乏標準操作程序，本文的非監督學習以中文字彙多寡、句子長短作爲搜尋方向。

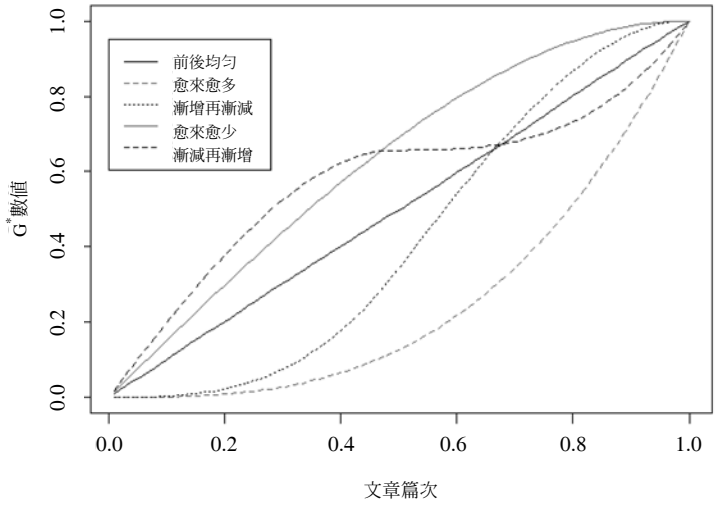
除了使用兩種分析想法外，本文仿效 Tukey 的 EDA 作法，以簡單易懂的敘述性統計量（平均數、變異數）爲主，找出文言文、白話文的差別。其中較大的變異數，以文字使用的角度詮釋，或可視爲某個時間、某些作者對特定字彙使用量較多，其他時間及作者並不常用，因此我們將以不均度的想法代替 EDA 中的變異數。常見的不均度指標包括吉尼係數（Gini’s index）、[14] 熵（Entropy; Shannon’s index）、Simpson’s Index 等，其中吉尼係數用於測量不均度，常見於經濟學中描述一個國家或地區的貧富不均度；熵原先用於化學、熱力學，在於測量某系統內的能量總數，後來也應用於描述系統的混亂程度，或是系統狀態的函數；Simpson’s Index 則用於生物學及生態研究，作爲描

[14] Forcina, Antonio, and Giorgi, Giovanni Maria. “Early Gini's contributions to inequality measurement and statistical inference,” *Journal Électronique d'Histoire des Probabilités et de la Statistique* [electronic only] 1.1 (2005): Article 3, 15 p., <http://eudml.org/doc/125313>.

述是否存在優勢物種（dominated species）。若以 p_i 代表某篇文章中第 i 個字彙的出現比例，則 Entropy 及 Simpson's Index 分別爲：

$$\text{Entropy} = -\sum_i p_i \log(p_i)$$

$$\text{Simpson's Index} = \sum_i p_i^2$$



圖一：G* 指數的幾種類型

除了這三種不均度測量值，我們也提出另一種不均度指標 G*，計算方式與吉尼係數類似，但吉尼係數需要先將上述 p_i 之類由小到大的排序，G* 則不需排序。因爲 G* 不需要排序，因此可反映一篇文章裡字彙的出現順序及頻繁與否，描述另一種不同的資訊，這個指數的定義將在第二節中詳述。圖一列出五種常見的 G* 數值曲線，接近對角線者表示某字彙出現維持穩定，始終在對角線下（上）者表示字彙使用愈來愈多（少），與對角線交錯的線條則是字彙一開始愈來愈多（少），但接著逐次減少（增加）。以這些圖形類型，可作爲判斷某些字彙在文章裡的出現特性。

三、監督學習的統計分析

如前一節所述，監督學習由研究者先決定關鍵字詞，再代入常見的分析方法，這類型方法可結合研究者的專業知識，根據問題需要、個人經驗學養選擇適當的關鍵字詞，然而也因爲對關鍵字詞的見解不同，即使相同資料庫及分析方法，可能產生截然不同的研究結果。爲了減少研究者主觀判斷的影響，本文除了現行多數研究會採用的監督學習外，也將考量可行的非監督學習之切入點，以下先說明本文監督學習的進行方式，以及選取關鍵字詞的緣由。

關鍵字詞的選取，主要考量文言文與白話文的文體差異。無論文言或白話，在表述任何內容時，必定需要使用功能詞，否則難以成句。換言之，任何篇章中必有一定數量之功能詞。有些功能詞爲文言與白話共用，有些則僅見於或較常用於文言或白話，這也就是爲何一般人提到文言首先想到的便是「之乎者也」。本文根據王力《漢語史稿》及楊寄洲、賈永芬編著的《漢語800虛詞用法詞典》，[15] 選出常用文言虛詞。爲了減少干擾因素，排除了雙音節虛詞。最終選定文言文、白話文常用虛字各十個爲主要研究對象，計算這些虛字在第1至第11卷使用比例的變化；觀察兩種文體的虛字在1919年五四運動前後，其出現特性是否有明顯不同。除了這些傳統的字詞分析外，本文也引進生態學的想法，以生物多樣性（species diversity）及演化（evolution）角度描述用字習慣的變遷，使用包括物種分布不均度等測量值，分析虛字與關鍵物種的關聯，下一節則以非監督學習角度切入，藉由字彙的豐富程度、分句長短等角度比較文言文、白話文的差異。

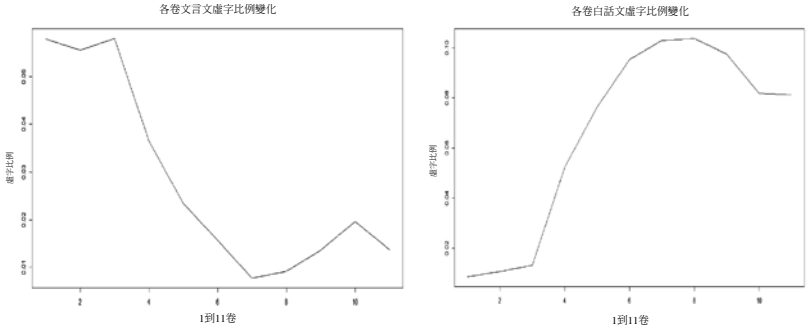
本文選擇的常見虛字如表一，其詞類等相關特性說明可參考附錄。這二十個虛字約佔所有文字的9.7%，文言文的虛字出現比例明顯較少，其中又以「之」、「無」的使用比例最高，分別佔了約1.7%

[15] 楊寄洲、賈永芬編著：《漢語800虛詞用法詞典》（北京：北京語言大學出版社，2013年）。

及0.4%；白話文虛字則以「是」、「的」最多，出現比例分別3.6%及1.4%。其中文言文虛字在前三卷的出現比例高於6%，隨後直線下降，在第6卷以後的比例不到2%；相反地，白話文虛字在前三卷的出現比例少於2%，之後迅速增加，第6卷之後的比例就不低於8%（圖二），從文言文、白話文虛字間的消長可看出《新青年》的文體變化。

表一：文言文、白話文的虛字及其出現比例

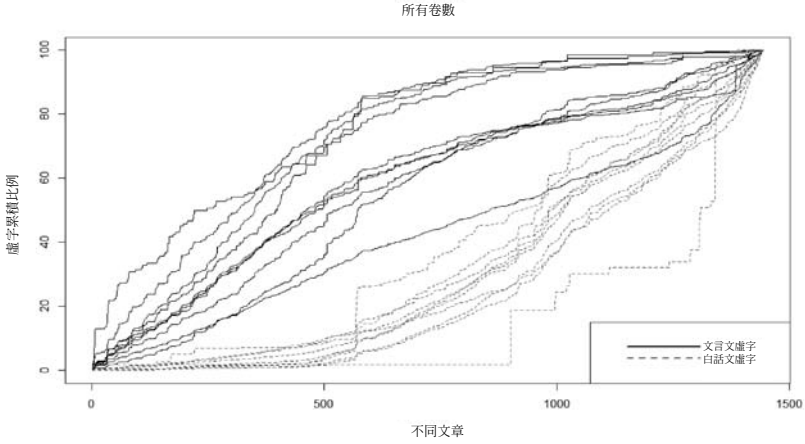
	文言文	白話文
虛字	矣乎焉歟哉耳豈之乃無	的是們個了和麼著嗎吧
出現比例	2.4%	7.3%



圖二：《新青年》各卷文言文（左圖）、白話文（右圖）虛字比例

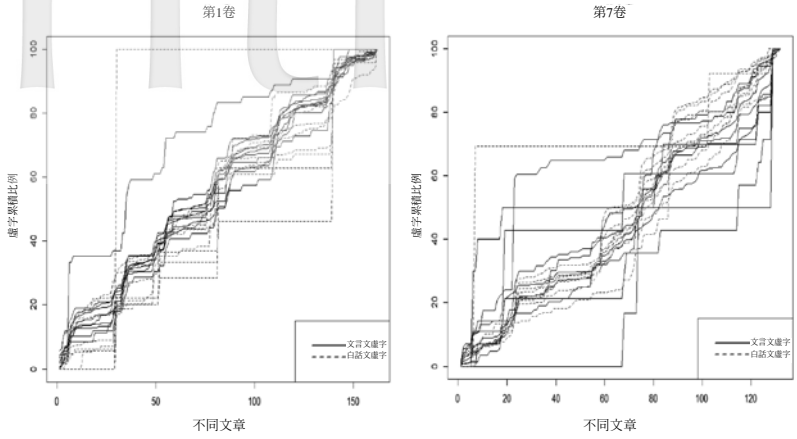
除了使用比例的變化，我們也使用 G* 指數衡量文言文、白話文虛字的使用趨勢。關於 G* 指數的定義，首先計算虛字在每卷使用總數及十一卷總數（記爲 $x_1, x_2, \ldots, x_{11}$ 及 $x=\sum_{i=1}^{11} x_i$），再計算累積各卷虛字所佔比例（記爲 $y_1=\frac{\sum_{i=1}^{1} x_i}{x}, y_2=\frac{\sum_{i=1}^{2} x_i}{x}, \ldots, y_{11}=\frac{\sum_{i=1}^{11} x_i}{x}$），由（$x_1, x_2, \ldots, x_{11}$）構成的曲線即爲 G* 指數曲線；G* 指數也可以文章發表先後順序計算，上述的 x_i 則是依時間編排的第 i 篇文章。圖三爲所有虛字在《新青年》依發表時間的 G* 指數，20 個虛字明顯可分爲兩群，實線爲文

言文虛字，除了一條曲線外（「之」），全部都在圖形上層，白話文虛字（虛線）則全部在下層，顯示文言文類型的虛字出現在較前面的卷次（時間較早），或是使用愈來愈少，白話文虛字在較後面的卷次出現愈來愈多。



圖三：所有虛字在《新青年》的 G* 指數

進一步比較二十個虛字在各卷次的特性，找出表一中的文言文、白話文的差異，因爲第7卷以後的結果大致與第7卷相似，在此僅以第1卷及第7卷爲代表，說明虛字隨時間的變化趨勢。圖四爲《新青年》第1卷、第7卷中，各十個文言文、白話文虛字的 G* 係數的數值，實線、虛線各代表文言文、白話文虛字，線條愈接近斜對角線者，表示該虛字使用頻繁且均勻。第1卷中實線較接近對角線，第7卷虛線較接近對角線，顯示第1卷的文言文虛字使用較爲頻繁，而第7卷的白話文虛字使用頻繁。更進一步分析，以第7卷爲例，其中文言文僅有「之」、「無」兩字出現較爲頻繁，白話文中僅有疑問詞「嗎」、「吧」較爲罕見，符合我們對《新青年》第7卷（以及第7卷之後）的文體較爲接近白話的認識。



圖四：第 1 卷、第 7 卷虛字的 G* 指數

表二：虛字在各卷的吉尼係數變化趨勢

第 1 到 11 卷的變化	文言文虛字個數	白話文虛字個數
明顯上升	6	0
明顯下降	0	9
上升、下降交錯	4	1

除了 G* 指數外，我們也計算虛字在各卷的吉尼係數，並比較係數的變化趨勢，以此判讀虛字的使用特性。因爲吉尼係數測量不均度，數值愈大代表虛字使用愈不平均，如果數值在隨著卷次上升，顯示使用愈加不均勻；反之，數值遞減則顯示虛字使用愈來愈均勻。表二爲虛字的吉尼係數變化趨勢統計，文言文虛字中有六個遞增、白話文則有九個遞減，代表文言文虛字的不均度上升，白話文虛字則漸趨均勻；另外，比較前三卷與後三卷的吉尼係數，十個文言文虛字全數上升。換言之，不均勻代表文言文虛字的使用變得較不常見，白話文虛字反而比較常見。這個結果與前述結果一致，亦即以本文選取的虛字爲判斷依據，大致可以印證《新青年》文體由文言列白話的轉變。以下將以非監督學習的角度，繼續探討文體的變化。

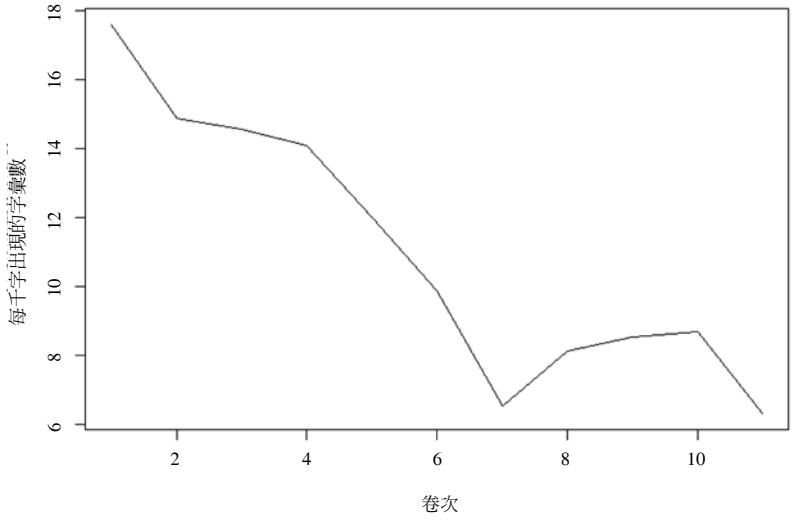
四、非監督學習的統計分析

有別於指定關鍵字詞的監督學習，非監督學習並無既定的方式，可爲數量分析增加不同可能性，卻也令研究少了可遵循的方向。本文針對中文寫作的特性，建議以字彙使用習性、句子結構兩個角度進行非監督學習的分析。若將中文數位分析類比於生態研究，則字彙可視爲生物的物種（species），套用描述生態豐富程度的物種多樣性（species diversity），亦即字彙多寡及其使用習慣則作爲判斷寫作風格的依據，這也是字彙使用比例常見於用字習慣的原因之一，齊夫法則（Zipf's law）即是知名範例。以下將先比較《新青年》雜誌各卷的字彙豐富程度。必須加以釐清的是，「字彙」在此係指單音獨體的中文字（character），而非音節、字數不拘的詞彙（word）。亦即，本文所統計的「字彙數」乃是有多少不同的中文字，而非「詞彙量」。

表三：《新青年》各卷字彙豐富程度

	總字數	字彙數	Simpson Index	Entropy
第 1 卷	248,833	4,379	0.004568	6.654036
第 2 卷	291,848	4,344	0.004500	6.649539
第 3 卷	290,038	4,227	0.004954	6.541824
第 4 卷	305,020	4,298	0.004172	6.539378
第 5 卷	343,519	4,125	0.004672	6.461579
第 6 卷	389,407	3,848	0.005749	6.348547
第 7 卷	586,942	3,850	0.006053	6.328604
第 8 卷	461,731	3,753	0.006035	6.320355
第 9 卷	437,748	3,745	0.005574	6.322103
第 10 卷	342,778	2,980	0.005700	6.177278
第 11 卷	489,223	3,093	0.005712	6.212699

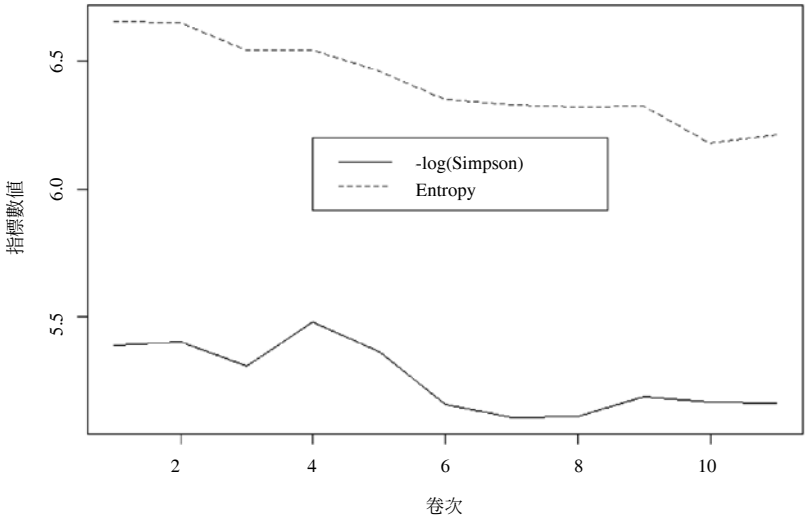
表三第2行、第3行列出《新青年》各卷總字數、字彙數。明顯可見各卷總字數大略隨時間遞增，但不同字彙數反而有下降的趨勢，第1卷總字數最少、字彙總數卻最多，第11卷總字數約爲第1卷的兩倍，但字彙總數僅有其三分之二。換算成每千字出現的新字彙數（圖四），更可看出字彙使用更加集中。這反映出《新青年》後期卷次的識字閱讀門檻低於前期，讀者需要識得的中文字較少。然而各卷字彙（character）使用的多樣性逐次降低並不意味著詞彙（word）量減少。事實上，總字數增多而字彙多樣性降低反映的應是白話文另一個重要特徵：雙音節的二字詞彙或多音節的多字詞彙大量增加。換言之，白話文有效利用有限的字彙組成多樣的詞彙，一方面降低識字閱讀門檻，另一方面卻能藉由排列組合大量擴充詞彙量。《新青年》後期卷次的讀者只要掌握了一定數量的中文字，不需學習新字，即可不斷學習新詞彙。



圖五：《新青年》各卷每千字出現的新字彙數

除了以字彙總數判斷用字的多樣性，每個字彙的使用比例也可用於比較用字習慣，我們採用和監督學習相同角度，以字彙不

均度作爲判斷寫作風格的指標。本文不均度考量 Simpson 指標及 Entropy（熵）指標，前者愈大代表愈集中（或是不平均），後者愈大代表愈平均；另外，若將 Simpson 指標取對數值，再取負號，也就是 -log(Simpson)，則與 Entropy 指標的方向一致，且數值約略相近。表二最後兩行是這兩個指標在第 1 卷至第 11 卷的數值，明顯可見第 1 卷的用字較爲平均，或是 Simpson 指標較小，Entropy 指標較大，而第 11 卷的指標顯示文字相對集中，或是如前一段所說的多樣性降低。圖六爲《新青年》各卷的兩個指標變化趨勢，皆可視爲多樣性的可能測量方法，兩者皆顯示多樣性隨時間降低，其中 Entropy 的變化較易辨識（可視爲比較敏感）。綜合字彙總數、字彙不均度，兩種分析均顯示《新青年》後期卷次的用字愈加集中（多樣性降低）。整體而言，字彙多寡與頻率的統計皆印證了《新青年》從文言到白話的轉變，同時反映出白話文善於利用既有文字組詞，有效降低閱讀門檻的語言特性。



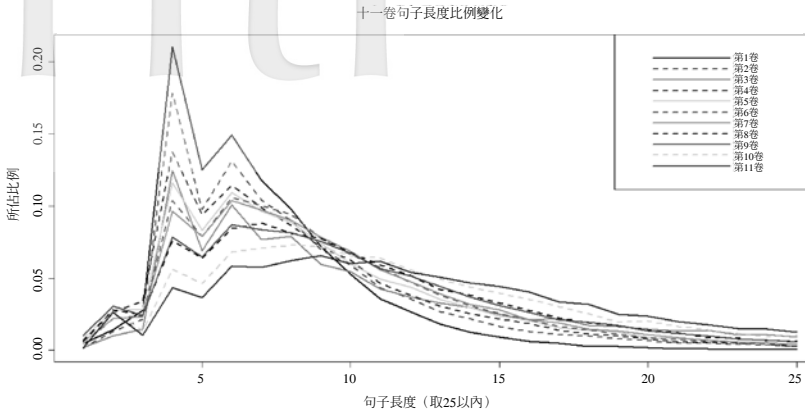
圖六：《新青年》各卷多樣性趨勢圖

除了字彙多寡與出現頻率，每個分句的長短也可用於觀察文體的

變化。在一般人的印象裡，文言「言簡意賅」且較白話文更富含「微言大義」。之所以形成這種印象，除了前述文言字彙多樣性和文白組詞方式的差異之外，可能還與一個文言完整句中的每個分句長度有關。亦即，一般而言，文言的每個分句字數應該不會太多，分句不會太長。以下我們將分析《新青年》各卷的分句長短，以計量方式驗證一般印象。在判定分句時，本文以幾個常見的標點作爲分斷依據，包括：「，」、「。」、「；」、「！」、「？」，每個分句的長度就是任兩個這些標點之間的字數。這樣一來，就文言文而言，相當於回歸傳統句讀，可以避免文本本身標點未規範化造成之混亂，以及在文本數位化過程中新式標點選用之爭議。事實上，《新青年》直至第7卷方統一標點符號使用規則。由於本文僅針對分句長度進行統計，暫不考慮不同標點符號的差異，標點符號的使用未規範化對統計結果並無影響。

在統計時，我們發現有些分句的字數長達40字以上，但爲了便於比較及節省篇幅，僅列出每個分句字數不多於25字的結果，長度大於25字的結果大致類似。在此以第1卷、第6卷、第11卷爲例，說明各卷的特性，其中第1卷以每分句四個字爲最多，而累積四、五、六個字的出現比例總共佔了約48%。第6卷則以每分句六個字最多，但要累積四、五、六、七、八個字，其比例才會達到48%；第11卷又以每分句九個字最多，累積六、七、八、九、十、十一、十二個字的比例約有46%。

圖七爲《新青年》第1卷至第11卷的每個分句長度不多於25個字的分布圖，從圖形明顯可看出每分句不多於十個字的比例，曲線從第1卷至第11卷逐次遞降，代表每分句字數較少者愈來愈少，這個結果與前一段的數值一致。圖六中也顯示每分句超過十個字的比例呈現逐卷上升的現象，從第1卷的13%迅速上升至第6卷的32%以及第11卷的58%。因爲每分句十個字以下的比例下降，每分句十個字以上比例上升，合併兩者後可判斷每分句平均字數從第1卷至第11卷的會逐卷上升，符合了文言文每分句字數較少的一般印象。



圖七：《新青年》各卷每句字數分布圖

五、討論與建議

以數位分析的角度探討文章風格的改變，是近年研究的熱門議題之一，本文嘗試以統計學習理論的監督學習、非監督學習，探討《新青年》雜誌前後各卷的文體，是否逐漸由文言文轉變爲白話文。監督學習中以各十個文言文、白話文的常用虛字，套用幾個不均度指標，發現除了文言中的「之」、「無」和白話中的「嗎」、「吧」之外，文言文虛字隨時間出現愈來愈少，白話文虛字則相反，在後面卷次出現愈來愈頻繁。出現的四個例外都與口語白話和書面白話之間的差異有關。「之」、「無」在白話口語罕用，但在書面語中顯然不排斥，形成與文言共用的現象。「吧」、「嗎」則多用於口語白話，或小說中的對白，少用於其他書面白話。

非監督學習不事先指定變數，由資料找出文體、風格等的寫作特性，這就是巨量資料的分析核心，也成爲資料驅動（Data Driven）。本文以字彙豐富程度（或是生物多樣性，species diversity）、不均度（evenness）等資訊，發現《新青年》雜誌前幾卷的字彙較爲豐富且出現比例較爲均勻，後幾卷雖然總字數較多，但不同字彙數較少，推

測與文白詞彙組成方式不同有關，也印證了白話有效降低字彙量閱讀門檻。除此之外，我們也發現前幾卷的每個分句所含字數較少，愈到後面幾卷，每個分句的平均字數增加也愈多。上述這些特性均顯示《新青年》雜誌前幾卷的文體較爲相近，從第6卷之後就轉變爲另一種特性。

白話取代文言成爲中文書面語主流的嬗變過程歷時久而過程複雜。本文爲了尋找可以量化的指標，將問題暫時簡化爲文言、白話二元對立，並以《新青年》此一文體自覺較高的雜誌爲分析對象。一般人以常識區辨一篇文章是文言文或白話文時，常用的判斷標準包括：兩種文體的常用的功能詞不同；文言文多艱深字，白話文少冷僻字；文言文單音節詞彙多於白話文，白話文雙音節或多音節詞彙多於文言文；文言文「言簡意賅」、白話文較爲「口語化」。本文以計量方式統計分析的結果大致吻合這些認知，《新青年》雜誌前幾卷的文體比較接近文言文，在五四運動之後，第6卷及其之後各卷更貼近白話文。這顯示本文所提出的數量分析想法（監督學習、非監督學習）確實可行，而且兩種切入點找出的特性不盡相同，實證分析時可交叉使用，輔助研究者驗證根據專家意見得出的推論。

限於篇幅，本文的分析仍有非常大的改進空間。現行的監督學習只以常見虛字作爲比較標準，未來應可嘗試其他變數，例如：文言文與白話文思想觀念關鍵詞的字數（音節數）。其中，文言文的思想觀念關鍵詞似乎以兩字（雙音節）組成較多，但白話文則未必，三個字及三個字以上的多音節思想觀念關鍵詞比比皆是。以後的研究或可嘗試其他作法。例如，發展決定思想觀念關鍵詞的分析方法，先以出現頻率找出潛在思想觀念關鍵詞，接著篩選出核心思想觀念關鍵詞，再以思想觀念關鍵詞比較兩種文體的異同。又或，統計多組文白同義實詞在各卷中數量的消長等等。[16] 當然，若能在逐次的比對傳教士文

[16] 感謝匿名審查者指出胡適在白話遊戲詩中曾列舉實詞古今用法差異，並建議思考此類實詞是否也可成爲統計對象。不過，由於實詞涉及語意，篇章

獻，及在《新青年》前後期的其他相關文獻，當能更突顯《新青年》從文言到白話轉變的歷史價值與代表性意義。

非監督學習進一步延伸研究的空間更大。本文僅考量字彙豐富程度（或是生物多樣性）、不均度、分句長度，接下來可嘗試較爲抽象的分析議題，像是如何數量化「言簡意賅」這個概念，以及文言文一字多義（或是同義字）的分析，甚至是將文法、詞性的想法套入數量分析，以具體的量化指數描述抽象的意念。或以本文對文白分句長度的統計爲例，最初的設想僅是希望驗證文言分句通常較白話分句爲短的一般印象。實則分句長短涉及文言、白話「小句合成體」（clause complex）「信息合成式樣」的不同。[17] 劉承慧針對漢語曲折形態不發達的特性，就句法層級以上的「規約化信息結構」（conventionalized information structure）提出了以「小句合成體」（clause complex）式樣作爲分析單位。「小句合成體」看似接近句法學上的「句子」（sentence），但並不像「句子」那樣根據印歐語系的傳統，以曲折形態（inflectional morphology）爲辨識基礎。「小句合成體」的操作型定義是：「依循語言中既定合成式樣所組成的單位，在現代書面作品由句號標註的長句子，若是依循既定式樣合成，即爲小句合成體」，「古代書面語沒有規範化標點符號，判別先秦小句合成體的不確定性更高，但是至少還有連接標記及其他意義合成的記號可以作爲判別的形式依據」。[18] 本文所統計的「分句」相當於此一操作型定義中的「小句」，是組成合成體的基本單位。而影響小句合成體「信息合成式樣」的除了本文用以隔斷「分句」的句讀或新式標點符號，還有表示事理

題材、內容的偏向勢必影響統計結果，以監督學習方式選定實詞詞彙時需審慎考量。一種可能的作法是結合監督學習與非監督學習，先以非監督學習方式找出單音節、雙音節高頻詞彙，再比對結果，尋找對應關係，再選定對應詞組，設爲統計項目。

[17] 參見劉承慧：〈先秦書面語的小句合成體——與現代書面語的比較研究〉，《清華中文學報》第4期（2010年12月），頁144。

[18] 劉承慧：〈先秦書面語的小句合成體——與現代書面語的比較研究〉，頁146。

關係的標記，例如：因果標記「因爲」、「所以」，轉折標記「然而」、「但是」等等。文白常用事理標記有同有異，使用頻率也大不相同，標記的差異、多樣性與使用頻率，或許亦會是我們進一步開發文白量化區辨方法時相當具有潛力的指標。

附錄：初選虛字詞類[19]

文言文：

矣〈助詞〉
乎〈助詞〉
焉〈疑問副詞〉〈助詞〉[20]
歟〈助詞〉
哉〈助詞〉
耳〈助詞〉
豈〈副詞〉
之〈助詞〉[21]
乃〈副詞〉
無〈副詞〉[22]

白話文：

的〈助詞〉
是〈副詞〉（〈動詞〉）[23]
們〈後綴〉
個〈量詞〉
了〈動態助詞〉、〈語氣助詞〉
和〈介詞〉、〈連詞〉
麼〈後綴〉

---

[19] 主要根據《漢語800虛詞用法詞典》（北京：北京語言大學出版社，2013年）。
[20] 亦常用作代詞。
[21] 亦常用作代詞，也有動詞等其他詞性用法。
[22] 亦可直接用作動詞，表否定。
[23] 「是」的詞性相當多樣。白話文中其實最常作動詞用，表肯定判斷，如「他是學生」。

著（着）〈助詞〉
嗎〈助詞〉
吧〈助詞〉

徵引書目

王　力：《漢語史稿》，北京：中華書局，1980 年，2004 年重印。

余清祥：〈統計在紅樓夢的應用〉，《國立政治大學學報》第 76 期，1998 年 6 月，頁 303-327。

周策縱等：《胡適與近代中國》，臺北：時報文化出版公司，1991 年。

段懷清：〈倪維思夫婦釋經布道文之語言實驗〉，《江蘇大學學報（社會科學版）》第 14 卷第 4 期，2012 年 7 月，頁 35-40。

項潔等主編：《數位人文要義：尋找類型與軌跡》，臺北：臺灣大學出版中心，2012 年。

楊寄洲、賈永芬編著：《漢語 800 虛詞用法詞典》，北京：北京語言大學出版社，2013 年。

趙岡、陳鍾毅：《紅樓夢研究新編》，臺北：聯經出版公司，1980 年。

劉承慧：〈先秦書面語的小句合成體——與現代書面語的比較研究〉，《清華中文學報》第 4 期，2010 年 12 月，頁 143-186。

黎錦熙：《國語運動史綱》，北京：商務印書館，2011 年。

David Der-Wei Wang, “Chinese Literature from 1841 to 1937” in Kang-I Sun Chang ed., *The Cambridge History of Chinese Literature*. Volume II. Cambridge: Cambridge University Press, 2010.

Forcina, Antonio, and Giorgi, Giovanni Maria. “Early Gini's contributions to inequality measurement and statistical inference,” *Journal Électronique d'Histoire des Probabilités et de la Statistique* [electronic only] 1.1 (2005): Article 3, 15 p., <http://eudml.org/doc/125313>.

Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer Series in Statistics. 2009.

Wirfs-Brock, R. and Wilkerson, B., “Object-oriented design: a

responsibility-driven approach". in *Conference Proceedings on Object-Oriented Programming Systems*, *Languages and Applications* (New York: ACM, 1989): 71-75.