利用文本採礦探討《紅樓夢》的後 40 回作者爭議

杜協昌 *

摘 要

《紅樓夢》全書共 120 回。一般公認前 80 回的作者是曹雪芹，但後 40 回的作者則存有爭議。從前學者們主要是透過可以考訂作者、時代、版本的材料，或者從內容前後的連貫性，來推斷後 40 回是否為他人所續。隨著電腦的出現，研究者開始利用量化的統計學方法分析前 80 回與後 40 回之間，是否在用字遣詞上存有顯著的差異。

這類統計方法，通常需先由研究者選定量化標的物（例如虛字頻率），然後再對這些標的物的分布進行統計檢定。有別於這樣的步驟，本論文運用文本採礦的技術，先讓電腦計算出可能有趣的候選字詞，然後再利用前後綴詞工具來觀察這些字詞的前後，經常相隨有哪些字。我們找到許多前人沒有注意到的字詞，它們在前 80 回與後 40 回的使用頻率上存在明顯的差異。例如前 80 回中有 34 回可看到「嬤嬤」一詞，但「嬤嬤」在後 40 回卻一次也沒有出現過。此外，分析「豈」的後綴字，我們發現該字有將近七成是被使用於「豈不」與「豈知」。有趣的是，「豈不」在前 80 回的出現頻率明顯偏高，而「豈知」卻僅在後 40 回出現。

我們認為，利用資訊工具可以有效幫助人文學者從文本中發掘新事證。我們的實驗結果，支持《紅樓夢》後 40 回作者並非曹雪芹的論點。

* 國立臺灣大學資訊工程系博士後研究員。

A Text-Mining Approach to the Authorship Attribution Problem of *Dream of the Red Chamber*

Hsieh Chang Tu

Abstract

Dream of the Red Chamber (*DRC*), written in the 18th century, is among the greatest Chinese classic novels. In 1921, Hu Shi (胡適) provided solid evidence to show that the first 80 chapters were written by Cao Xueqin (曹雪芹). He also attributed the remaining 40 chapters to Gao E (高鶚) (Hu 1921). While the first conclusion is commonly accepted, the second is still not settled.

Statistical methods that use certain pre-defined linguistic features, usually a set of function words, to check whether the feature frequencies in the first 80 chapters are significantly different from the last 40, have also been investigated. Interestingly, however, people came to different conclusions when choosing different features.

In this paper we propose a text-mining approach to the *DRC* author attribution problem. We define a mining function to find terms that clearly show discrepancies between the two corpuses. Some of the terms are semantic in nature, thus avoiding the pitfalls with the more syntactic function words approach. In addition to supporting the claim that the first 80 chapters and the last 40 were written by different authors, a somewhat surprising side result is the evidences that show Chapters 64 and 67, two chapters missing from the oldest existing edition, could also have been written by someone other than Cao Xueqin.

Our experiments show that making use of computer technology can help humanists find interesting facts and clues from text. These results support the statement that the last 40 chapters of *Dream of the Red Chamber* were not written by Cao Xueqin.

---

* Postdoctoral Researcher, Department of Computer Science and Information Engineering, National Taiwan University.

一、導論

《紅樓夢》是著名的中國古典小說，被許多學者認為具有高度的文學成就。現今流傳的《紅樓夢》全書共有 120 回。一般公認前 80 回的作者是曹雪芹，但後 40 回的作者則存有爭議。

《紅樓夢》的原名是《石頭記》。書中第一回交代《石頭記》的緣起時，就曾提到「曹雪芹於悼紅軒中，披閱十載，增刪五次，纂成目錄，分出章回」等情事。胡適（1921）在《紅樓夢考證》中，從袁枚的《隨園詩話》卷二、曹雪芹父親曹寅的相關史實、《八旗人詩鈔》裡幾首與曹雪芹有關的詩、以及《紅樓夢》內容的一些敘述，推斷曹雪芹確為此書的作者。胡適認為《紅樓夢》最初只有 80 回，一直到乾隆五十六年後才有 120 回的《紅樓夢》[1]；他並從張問陶的詩及註、以及俞樾的《小浮梅閒話》大膽推斷後 40 回的作者是高鶚。但由於續書並不是一件容易的事，而後 40 回不管在結構、伏線安排、以及詞句筆氣都與前 80 回非常近似，[2] 因此雖然後 40 回與前 80 回在情節發展上有許多不連貫的地方，仍有許多研究者認為《紅樓夢》全書都是曹雪芹所著。[3]

由於可供考證的史料稀少，加上近年資訊科技發展迅速，許多學者轉而利用文本的量化分析來判斷前 80 回（以下稱之為 A 文）與後 40 回（以下稱之為 B 文）是否出自相同作者之手。我們都知道不同作者的文筆有異。若這類文風差異可由文本中某些特徵（例如虛字或詩詞等）的出現頻率來區隔，那麼 A、B 之間如果發現相當明顯的特徵頻率差異，就可推測它們是由不同作者所撰。統計學的方法，可以讓我們在一般還算合理的假設（例如出現頻率滿足隨機且獨立抽樣的條件）下，推定 A、B 在特徵頻率之間是否存有顯著的差異。值得注意的是，雖然都是採用客觀的量化數據，但由於實驗設計、特徵選擇、機率分布假設、以及抽樣方式有所差異，學者們竟也常得出彼此相左的結論（參考第二節的相關文獻探討）。由於後 40 回不可能既是又不是曹雪芹所著，這提醒我們即使是運用客觀的量化數據，也依然必須檢視推論的前提是否確然成立，以避免在證據不夠充分的條件下做出錯誤的判斷。

這篇論文嘗試利用文本採礦（text-mining）的方法，來探討《紅樓夢》的前 80 回作者是否也撰寫了後 40 回。傳統文本分析的統計方法，必須先由人工篩選量化標的物（虛詞或詩詞數量等），然後再對這些標的物的分布進行統計檢定。我們依循前人的腳步進行類似的實驗，發現其結果僅能說明《紅樓夢》許多用字在不同情節

1　袁維冠（1978）指出：1959 年在山西發現的 120 回《紅樓夢稿》，說明早於程偉元與高鶚的刻本（乾隆五十六年、也就是 1971 年的程甲本）之前，就已經有 120 回本流傳於世。

2　參見太平閒人的《石頭記》讀法，參考文獻所列之《紅樓夢》三家（護花主人、大某山民、太平閒人）評本。

3　例如袁維冠（1978）就認為《紅樓夢》全 120 回都是曹雪芹所著。

發展上有顯著差異，卻仍不足以據此聲稱曹雪芹並非後40回作者。有別於傳統方法，我們先定義文本採礦函數，讓電腦計算出可能有趣的字詞，然後再利用資訊工具來觀察這些候選者是否真有價值。我們找到許多有趣的字詞，它們在前 80 回與後 40 回各章回的頻率分布上有著非常明顯的差異。這些發現支持《紅樓夢》後 40 回應為後人所續的看法。

二、相關文獻探討

作者歸屬（authorship attribution）是一種文件分類（text classification）的問題。給定一篇欲判定作者的文件 d，作者歸屬問題假設我們已知一個可能的作者集合，並且有這些作者所撰寫的一些文件集合 T。早期的作者歸屬方法，是先利用 T 對每位可能的作者建構該作者的寫作特徵圖（characteristic curve of composition）。如果文件 d 的特徵圖與某作者的特徵圖有顯著差異，就可據此推論 d 並非該作者所著。例如 Claude S. Brinegar（1963）就以字長（word length，英文單字所含的字元數量）的頻率分布，認定馬克吐溫（Mark Twain）並沒有寫 Quintus Curtius Snodgrass 的 10 篇信件。另一個有名的例子，是 Ronald Thisted 與 Bradley Efron（1986）利用生態學估計物種數目的方法，推定一篇 1985 年新發現的九節詩（nine-stanza poem）確實是由莎士比亞（Shakespeare）所著。近年來，學者們進一步利用統計學的主元素分析（principal components analysis）或機器學習（machine learning）方法，來分辨 d 最可能出自於哪位作者之手（Peng & Hengartner, 2001; Malyutov, 2006; Stamatatos, 2009）。當前許多在作者歸屬問題具影響力的論文，其焦點都是在通用性分類方法（general-purpose classifiers）的建構與效能評估上（Burrows, 2002; Hoover, 2004; Jockers & Witten, 2010）。基於幾項理由，我們並沒有採用這類通用性方法。首先，這些通用的分類方法並不假設文本的內容之間具有相依性，但《紅樓夢》在前 80 回與後 40 回的內容明顯具有強烈的相依性。其次，這些方法通常都利用出現頻率最高的字（most frequent words）或者虛字（function words）來進行作者風格的辨識，這與接下來將提到的一些前人做法類似。然而我們發現〔請參考第三節第（二）小節〕，《紅樓夢》不僅在前 80 回與後 40 回之間存在明顯的虛字頻率差異，其前 20 回（01-20）與中間 60 回（21-80）之間也存有相當大的用字差異。由於前 80 回都是由曹雪芹所著，僅說明《紅樓夢》在前80回與後40回之間的用字頻率有明顯差異，並不足以聲稱後 40 回為他人所作（否則，我們也可聲稱前 20 回與中間 60 回有不同的作者）。最後，這些方法通常僅對兩文本的差異給出一個或多個量化值，但由於這些數字通常並不具備直觀的意義，這使得人文學者很難利用這些數據對文本差異進行更進一步的檢視與討論。

趙岡和陳鍾毅（1975）在《紅樓夢研究新編》的第 6.2 節中，曾指出漢學者高本漢（Bernhard Karlgren）在 1952 年發表的論文，因出現頻率的分級過於粗糙而導致結論有誤。[4] 趙陳兩人從前 80 回（A 文）與後 40 回（B 文）各取樣 100 頁，統計「而、在、了、的、著」五個虛字在這些頁面所出現的頻率，然後利用 t-test 檢定出「而、在、的、著」四個虛字在 A、B 的平均頻率差異並非由偶然的機遇所造成。他們也深入分析 A、B 文中許多用字習慣的差異，例如「嗎、麼」、「我們、咱們」、「給、與」、語尾的「兒」字、以及讀音類似的「都、多」等。藉由虛字頻率分布的差異、以及對這些特殊用字的觀察，趙陳認為前 80 回與後 40 回絕非出於同一人之手。

陳炳藻與謝家浩（2003）在第三屆全球華文網路教育研討會中，簡要整理出陳炳藻在 1981 年所做的研究。該研究將欲進行比較的文本設計為《紅樓夢》01-40 回（X_1）、41-80 回（X_2）、81-120 回（X_3）、以及《兒女英雄傳》（X_4）等四組。接下來，以自行選析的五種詞類（名詞、副詞、形容詞、形動詞及虛字），計算兩組文本之間的統計相聯與相關係數（coefficients of association, θ, φ; coefficient of correlation, γ）。因為有五種詞類、三種相關聯係數，在每兩組文本之間可計算出 $5 \times 3 = 15$ 個係數。令 n(X, Y) 為文本組 X、Y 在這 15 個係數為負值的數量，它代表 X、Y 在這五組詞類中的「不相關程度」。陳炳藻的實驗發現 $n(X_1, X_2) = 2$、$n(X_1, X_3) = 4$、$n(X_2, X_3) = 2$、$n(X_1, X_4) = 10$、$n(X_2, X_4) = 8$，並據此聲稱 X_1、X_2、X_3 之間並沒有顯著差異——《紅樓夢》全書應為同一人所作。然而，《紅樓夢》與《兒女英雄傳》成書相隔近百年，敘事內容也明顯有差異，兩者之間存有多個負值相關聯係數應在預料之中。沒有考量《紅樓夢》續書人應會在寫作上極力模仿曹雪芹的文筆，而僅從《紅樓夢》01-40 回、41-80 回、81-120 回之間有相對少量的負值相關聯係數，就斷定它們都出自同一作者，在推論的過程上顯得草率而不夠嚴謹。

由於趙岡和陳鍾毅的研究僅從《紅樓夢》取樣 200 個頁面進行統計，余清祥（1998）依照趙陳的方法，以完整文本重新計算了一次。實驗結果發現「兒、在、了、的、著」五字中，「在、的、著」在前 80 回與後 40 回的出現頻率有顯著差異（趙陳的實驗「兒」字有顯著差異，應是取樣誤差所導致）。余清祥並利用章回中詩詞出現的數量與長短，以及每回結語用詞「下回分解」、「要知端的，且聽下回分解」等進行計量分析，認為前 80 回與後 40 回分屬不同作者。

何光國（2002）依據作品中「的、地、得」三字的頻率繪製分布線，聲稱這些分布線可以代表作者的寫作個性。他將「的」字的用法區分為定語助詞（例如：這來的便是閏土、圖書館的書）與句末語氣助詞（例如：這本書是我的、這樣做是可以的）兩類，發現這三字（四種用法）在《紅樓夢》前 80 回與後 40 回頻率的分布

4　高本漢聲稱《紅樓夢》前後都是由相同的作者所著。

比例極為類似，從而聲稱《紅樓夢》的著者只有曹雪芹一人。何光國的推論是有疑問的：即使這三字四用的頻率能反映著者的寫作風格，但該研究抽取的樣本太少且非隨機（僅抽取 49、60、70、80；81、90、100、110 共八回），僅從這些樣本來進行統計推論，其有效性是相當可疑的。

楊智傑等人利用文件中單字頻率的排名順序，設計了一個簡單的函數，它能夠有效地區隔任意兩篇文件的相似度（Yang, Peng, Yien& Goldberger, 2003）。他們將《紅樓夢》每 10 回視為一篇長文件，然後對這 12 篇長文件進行兩兩的相似度比對。他們發現紅樓夢 01-10、11-20 回兩者之間，21-30、31-40、41-50、51-60、61-70、71-80 回彼此之間，以及 81-90、91-100、101-110、111-120 回彼此之間的相似度都很高。另一方面，前 20 回（01-20）、中間 60 回（21-80）與後 40 回（81-120）之間的相似度差異就頗大。這表示《紅樓夢》的後 40 回在用字的頻率排序上，與前 80 回有頗大的差異。這些研究者還利用分群的方法將這 12 篇長文件組織成樹狀結構，並聲稱這些結果支持《紅樓夢》前 80 回與後 40 回分屬不同作者。

Joseph Rudman（1998）指出，許多非傳統方式的作者歸屬研究存有幾個嚴重的問題，導致它們的成果無法被廣泛地接受：(1) 因權宜考量採用非最合適的文本（通常會採用有電子檔全文的文本，但該文本可能並非最合適者）。(2) 經常沒能完整且中肯地引述先前該領域的研究成果。(3) 盲目地採取適用於其他領域的統計方法（例如 Efron-Thisted 的模型來自於蝴蝶的收集，Simpson's index 植基於生態系中共存物種的分布等等），這類方法通常假設文本特徵是均勻分布的，但是否成立卻未經過嚴格檢驗。(4) 主要的分析素材應盡可能接近原作的內容，因此需注意許多文本在流傳的過程中可能會被修改。(5) 不應以「我非該領域的專家」作為藉口。(6) 對採用方法的可能缺陷視若無睹（必須證明採用方法的前提假設是成立的，不能有意或無意去規避）。(7) 缺乏對誤差的討論，例如電子文本的錯字、統計結果的標準差、取樣可能並非隨機等。對於這些問題，該論文也提出一些可能的解決方式。

三、實驗方法

《紅樓夢》的後 40 回作者問題，有其本質上的困難性。一來足供考證的材料稀少，再者《紅樓夢》的版本眾多，各本之間的內容也經常有不小的差異。欲利用計量方法來探討後 40 回作者，必須先決定分析的版本。我們將在第（一）小節討論文本的選擇。

典型的字頻分析方法，是從欲分析的單字（例如事先挑選的虛字，或者文本中所有曾出現的單字）開始，計算這些單字在每一回出現的次數（稱之為頻率）。將

單字在各章回的出現頻率視為實驗數據，我們就可根據不同的假設[5]來檢驗前 80 回與後 40 回的字頻風格（在此風格指的是單字頻的分布狀況）是否存有顯著差異。顯著差異雖暗示前 80 回與後 40 回可能有不同的作者，但我們也不該忘記，小說後 40 回在情節上有頗大轉變（賈家由繁華轉為衰敗），這些字頻風格的差異有可能是因為敘事情節轉移所導致。我們在第（二）小節將會討論到這個問題。

第（三）小節描述我們如何利用文本採礦來尋找有趣的字詞。有別於前人所採用的步驟，必須先篩選量化的標的（例如文本曾出現的虛字或單字頻率），然後才能對這些標的進行統計分析。我們的方法，是先定義採礦函數（mining function），讓電腦利用這個函數計算出前 80 回與後 40 回出現篇數有明顯差異的字詞。我們用人力從這些「可能有趣的候選字詞」裡，挑選出不易受到情節敘事影響者，並對這些字詞進行較為深入的檢視與探討。這一小節的採礦結果，證實《紅樓夢》前 80 回與後 40 回之間，在許多用字遣詞上存有非常明顯的差異。

我們在第（四）小節做一些補充的實驗。我們說明前 20 回（01-20）與中間 60 回（21-80）之間的用字頻率雖也有不小差異，但這些差異可用敘事階段不同來解釋。此外，我們也利用前後綴詞工具，找出一些在前 80 回、前 20 回、中間 60 回的使用頻率均頗一致，但相較於後 40 回卻有明顯差異的字詞。

（一）文本的選擇

我們將《紅樓夢》簡單分為兩種系統：鈔本與刻本。鈔本因為內容保有脂硯齋的評語，因而也常被稱為脂評本。[6] 現存主要的鈔本有十多種，其中比較重要的有甲戌本（乾隆十九年，AD1754 年，現存 16 回）、己卯本（現存 42 回）、庚辰本（現存 78 回，前 80 回缺第 64、67 回）、王府本、楊藏本（即現今的《紅樓夢稿》，有 120 回）、甲辰本、列藏本、鄭藏本、卞藏本等等。刻本則主要有程甲本（乾隆五十六年，AD1791 年，由程偉元和高鶚整理後刻行，共 120 回）和程乙本（程高於隔年，也就是 AD1792 年重校的版本）。鈔本中只有楊藏本和王府本含有後 40 回的內容，[7] 但它們被發現的年代都晚於胡適質疑後 40 回作者的時間。因此，先於胡適就存在的後 40 回作者爭議，其「後 40 回」指的必然是程甲本或程乙本的刻本內容。至於前 80 回，目前公認甲戌本最接近曹雪芹原作，然該鈔本現僅存 16 回，通常也

5　例如假設同一位作者在每個章回所出現的單字頻率呈常態分布（趙岡、陳鍾毅，1975；余清祥，1998；何光國，2002），或者假設作者的用字偏好可用單字頻的排序來區隔（Yang et al., 2003）。這些方法都希望能以客觀的量化數據來代表作者的文風，但即使表面上看來其假設是合理的，其真實性仍應經過嚴謹的檢驗（Joseph Rudman, 1998）。

6　戚序本是個例外。它有脂硯齋的評語，卻是刻印而非手抄的。其內容僅有前 80 回。

7　王府本雖有 120 回，僅有前 80 回屬於脂評的鈔本系統，而後 40 回是配上的（參見百度百科 http://baike.baidu.com/view/530614.htm，上網日期：2012 年 9 月 27 日）。有人認為它是程甲本《紅樓夢》的階段性稿本（參見 http://www.openow.net/details/e13462.html，上網日期：2012 年 9 月 27 日）。

只能以庚辰本來遞補；至於庚辰本所缺的 64、67 兩回，一般也僅能據《紅樓夢稿》、程甲本或程乙本來補齊。

這裡衍生出一個問題。既然較近於曹雪芹原稿的庚辰本只有 78 回，那麼我們實驗所採用的「前 80 回」是不是也該從文本剔除 64、67 兩回、甚至把這兩回算在「後 40 回」的內容裡？現今流傳的文本，為了維持小說的整體性，通常會利用多個版本進行截長補短的工作。將 64、67 兩回從「前 80 回」剔除，雖可能讓內容更近於庚辰本，但我們的目的並非對庚辰本與程甲本進行用字比較，而是探討後 40 回是否曹雪芹所撰。為了方便討論，我們在實驗所使用的「前 80 回」還是包括有 64、67 兩回。當「加上這兩回」與「剔除這兩回」會對實驗結果的解釋造成較大影響時，我們會特別標註、或對此進行更深入的分析討論。

很幸運地，元智大學已在網路上提供《紅樓夢》的電子檔全文可資學術研究。[8] 根據該網站的說明，此電子版的內容來自於浙江文藝出版社印行、蔡義江所校注的《紅樓夢》。從該書的前言，[9] 我們可知其版本取捨，是以（脂評本為前 80 回底本的）俞平伯校本和紅研所新校注本為方向，再用現存的十餘種本子相互參校，擇善而從。根據此前言的說明，元智電子版的《紅樓夢》應已盡量保持曹雪芹原作的面貌，因此是相當適合本論文研究的文本。[10]

值得提醒的是，為了方便讀者閱讀，蔡義江也將一些不需保留原貌的字，換成現代漢語規範化的用法。例如《紅樓夢》原著中只有「他」字，蔡在書中將其分開為「他」、「她」或「它」；原著只有「那」字，蔡在書中將其分開為今日所用的「那」和「哪」等。此外，較陌生的異體字、另有別義的借用字等也都改用現代的字。例如小說中的「頑耍」、「頑笑」、「頑玩」，都改用現代的「玩耍」、「玩笑」、「遊玩」；而「舡」字則改為「船」字等等。另外，在實驗的過程中我們發現，截至 2013 年 1 月初，元智電子版的全文在第 82、84、86、88 回開頭都明顯漏失了幾個句子。雖然少了些許詞句應該不會對統計結果造成太大的影響，但為了減少失真，我們還是利用《紅樓夢校注》（曹雪芹、高鶚原著，馮其庸等校注，1984）來將這些明顯的缺漏補上。[11]

8 上網日期：2012 年 9 月 27 日。網址：http://cls.hs.yzu.edu.tw/hlm/read/text/text.asp

9 上網日期：2012 年 9 月 27 日。網址：http://cls.hs.yzu.edu.tw/hlm/read/intro/intro.htm

10 元智版的後 40 回似乎主要參考《紅樓夢稿》，其內容比較接近程乙本。

11 我們依循元智版對異體字的處理方式，在填補這幾個缺漏的句子時，將《紅樓夢校注》所用的「却」字改為「卻」。

（二）前 80 回與後 40 回的單字頻統計分析

統計學的計量方法具有堅實的學理基礎。若採用的統計方法其前提成立，則所得到的結論也將具有科學的客觀性與推理的有效性。其中，應用在文本分析的一種方法，是假設每位作者對用字各有其偏好，任何「不因敘事情節轉移而有大量用字頻率變化」的單字（尤其是虛字[12]），其在各章回的出現頻率會滿足某個特定平均數的常態分布。[13] 假設某單字在作者 α、β 的作品中，其出現頻率分別滿足平均數為 $\mu(\alpha)$、$\mu(\beta)$ 的常態分布。[14] 欲檢驗兩文本 A 與 B 是否出自不同作者 α 與 β，我們先計算單字在 A、B 各章回的出現頻率。若某字 w 在 A 各章回的頻率大致符合平均數為 $\mu(\alpha)$ 的常態分布，在 B 各章回的頻率接近平均數為 $\mu(\beta)$ 的常態分布，我們就可用 t-test 來檢定這兩個平均數會相等的機率。[15] 若兩個平均數會相等的機率很低（表示 w 在 A 各章回的頻率與 w 在 B 各章回的頻率有顯著差異），我們就可以聲稱 A 與 B 出自相同作者的可能性很低。必須注意的是，如果檢定後發現差異並不顯著，它僅表示我們不能排除 $\mu(\alpha)$ 與 $\mu(\beta)$ 相同的可能性，並不該據此就驟然推斷 A 與 B 出自於相同作者（不同作者仍可能有極為接近的平均數）。

表 1 列出實驗文本頻率最高的 100 字。[16] 除了少數單字因牽涉到常用的人名或指稱，[17] 導致其出現頻率易因敘事演進而產生相當大的差異，[18] 其他單字應不會因小說情節改變而在使用上產生太大的變化。我們將小說中最頻繁出現的「了」字於各章回出現次數列於表 2，並據該字在各頻率範圍的回數統計繪製出圖 1。由該圖可看出「了」字在前 80 回與後 40 回的字頻分布大致呈常態，因此我們的確可用雙尾 2-sample t-test 來檢定其平均數是否相等。我們取 p = 0.01，表示若檢定發現有顯著差異，則推定結果錯誤（也就是實際上兩者有相同平均數）的機率將小於 1%。對於「了」字而言，實際檢定後發現它的使用頻率在前 80 回與後 40 回並沒有顯著差異。我們在

12　虛字本身並沒有具體意義，也較不受敘事內容限制，例如「的」、「了」、「著」等。

13　在進行統計檢定前，應該先對分析對象背後所假設的母體有所了解。在此的母體是經由想像而非實際存在的，它包含了所有該作者以類似文風所生成的作品。也就是說，假想曹雪芹有許多著作，《紅樓夢》只是隨機抽樣的一份樣本，而單字在取樣章回的出現頻率滿足常態分布。剔除容易隨情節演進而產生使用量變化的單字，是因為這些單字可能因情節需要而密集出現在連續的數個章回，導致其頻率分布顯然偏離常態。

14　常態分布可由平均數與變異數所唯一決定（可參考相關的統計書籍，例如 Mendenhall, Beaver & Beaver, 1999）。為了簡化問題，我們假設相同單字在不同作者作品的常態分布僅在平均數有差異，而變異數則相同。

15　有關 t-test 的理論與實務應用，可參考相關的統計書籍（例如參考文獻 Mendenhall, Beaver & Beaver, 1999 或 Hogg & Tanis, 2006）。

16　另一種方式是計算字頻比例（字頻除以該回總字數的比例）最高的前 100 字。由於《紅樓夢》各回的字數相近，因此使用字頻或字頻比例所得的結果會十分接近。

17　例如人（夫人、襲人）、寶（寶玉、寶釵）、玉（寶玉、黛玉）、賈（賈母、賈政、賈璉等）、太（太太）、姐（鳳姐）、老（老太太、老爺）、母（賈母）等等。

18　人名出現的頻率很容易因敘事情節的演進而產生較大的變化。例如 98 回之後林黛玉已死，「黛玉」兩字的出現頻率當然會降低許多。

表 1 用星號「*」標出前 80 回與後 40 回頻率分布有顯著差異（p = 0.01）的字。表 1 被標上星號的共有 40 字。

表 1　《紅樓夢》字頻最高的前 100 字

	0+	10+	20+	30+	40+	50+	60+	70+	80+	90+
1	了	這	*子	*太	姐	都	看	今	呢	*此
2	不	*你	*又	好	*頭	*心	*如	*小	*忙	進
3	*的	去	賈	*在	聽	二	沒	問	*想	罷
4	*一	*著	*裡	*笑	就	事	*叫	*因	夫	倒
5	來	也	們	他	出	*老	*兩	奶	*爺	樣
6	道	玉	*見	家	*回	過	*到	等	才	*吃
7	人	有	只	上	知	還	母	鳳	*面	和
8	*是	兒	得	*她	*要	話	些	娘	*中	姑
9	說	寶	*那	*麼	*日	*起	時	可	王	正
10	我	*個	*太	大	*下	自	*之	什	打	*無

註：給定一個字，將該字最上方與最左方的數字相加，即可得到該字的字頻排名。例如「麼」的字頻排名是 30 + 9 = 39。字的前方若有「*」號，表示此字在前 80 回與後 40 回的頻率有顯著差異（雙尾 t 檢定，p = 0.01）。

表 2　《紅樓夢》總出現頻率最高的「了」字，其在各回的出現次數

	0+	10+	20+	30+	40+	50+	60+	70+	80+	90+	100+	110+
1	99	193	127	203	204	178	214	204	182	139	236	224
2	91	99	141	166	196	256	386	209	204	171	104	200
3	168	102	114	106	197	161	289	149	198	148	168	184
4	113	89	227	190	203	229	221	220	191	226	166	135
5	57	86	206	220	200	170	142	213	227	133	72	177
6	169	160	196	173	244	184	129	178	116	169	91	142
7	188	99	178	216	207	317	323	288	156	254	116	240
8	185	109	239	126	197	179	179	207	171	144	165	198
9	116	229	212	182	191	134	183	90	153	129	260	294
10	126	139	175	236	186	198	157	151	173	149	167	174

註：例如第 30 + 2 = 32 回的字頻是 166，第 110 + 8 = 118 回的字頻是 198。

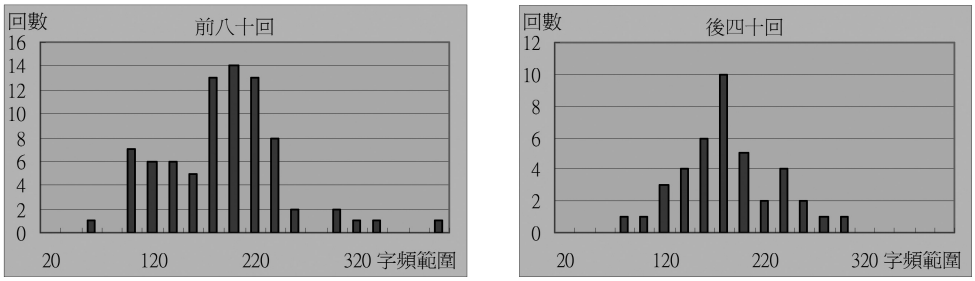

圖 1　將「了」字在《紅樓夢》各回的出現次數分組，計算有幾回其字頻落在 0-20、21-40、41-60、……、361-380、381-400 範圍內。例如前 80 回裡，字頻在範圍 101-120 的有 6 回（參考表 2：第 4、9、13、18、23、33 回）；而後 40 回裡，字頻落在範圍 101-120 的有 3 回（第 86、102、107 回）。這兩個圖形是為了說明，「了」的字頻分布在前 80 回與後 40 回都大致符合常態，因而可用 t-test 檢定其母體的平均數是否相等。

表 1 也驗證了趙岡、陳鍾毅（1975）與余清祥（1998）的實驗結果：在「兒、在、了、的、著」五個虛字中，「在、的、著」的字頻分布在前 80 回與後 40 回有顯著差異。趙岡、陳鍾毅與余清祥也據此推定《紅樓夢》的後 40 回作者並非曹雪芹。[19] 然而，《紅樓夢》在 80 回後敘事情節有頗大轉變（賈府由繁華轉衰敗），許多遣詞用字理當也與前 80 回有相當程度的差異。因此我們必須先了解：高頻字的使用頻率差異，是否僅發生在前 80 回與後 40 回之間？

我們仿照楊智傑等人（Yang et al., 2003）的處理方式，以 10 回為一個單位，將《紅樓夢》切分為 01-10 回、11-20 回、21-30 回、……、111-120 回等 12 個單位。在高頻字於各單位的章回字頻分布都接近常態的假設下，我們可利用 2-sample t-test 檢定單字在任兩個單位之間的頻率是否呈現顯著差異。給定任意兩個單位，我們計算表 1 所列的 100 個高頻字中，有多少字的頻率分布在這兩個單位間有顯著差異，並將其結果繪製於圖 2。

19　趙岡、陳鍾毅（1975）除了以字頻檢定用字差異外，另也指出許多前 80 回與後 40 回的用字歧異。該書第七章甚至以一個章節來討論「續書人究竟是誰」的問題。

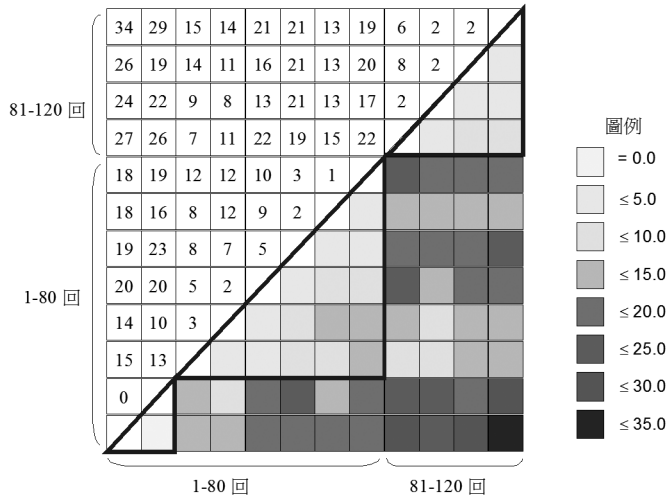

圖 2　以 10 回為一個單位，計算任兩個單位在前 100 高頻字中有幾個字的出現頻率有顯著差異。此圖的矩陣以右上到左下對角線為軸呈對稱。我們將顯著差異的單字數量列於左上方，而在右下方以顏色深淺來表示數量的多寡（較淺的方格表示兩單位間有較少出現顯著差異的單字）。例如，01-10 與 11-20 之間並沒有顯著差異的單字，71-80 與 81-90 之間有 22 個顯著差異單字，而 01-10 與 111-120 之間則存在 34 個顯著差異的單字。

圖 2 與楊智傑等人（2003）利用字頻排序方法所得的結果非常類似。從該圖可明顯看出，粗線框起的三個三角形其顏色都比較淡，表示頻率有明顯差異的字數並不多。具體地說，01-10 回與 11-20 回並沒有顯著差異的高頻字，而在 21-30、31-40、41-50、51-60、61-70、71-80 兩兩之間、以及 81-90、91-100、101-110、101-120 彼此之間，有顯著頻率差異的字數也不超過 12 個。此外，前 80 回的 8 個單位與後 40 回的 4 個單位之間，則幾乎都有超過 10 個（在 01-10 與 111-120 之間甚至有高達 34 個）呈現顯著差異的高頻字。值得注意的是，前 20 回（01-10、11-20）與中間 60 回（21-30、31-40、41-50、51-60、61-70、71-80）任兩個單位之間，在用字頻率上竟然都有 10 到 23 個高頻字在頻率分布上呈現顯著差異。這告訴我們，不僅前 80 回與後 40 回在用字上有顯著差異，前 20 回與中間 60 回之間也有許多常用字在頻率分布上明顯不一致。

為了知道前 20 回與中間 60 回的用字頻率差異，我們以表 1 的 100 個高頻字對這兩者進行平均數檢定[20]。我們很驚訝地發現，最常出現的「了」字雖然在前 80 回與後 40 回之間並沒有顯著的頻率分布差異，但該字的頻率分布卻在前 20 回與中間

20　這裡的前 100 高頻字本應取自前 80 回。由於前 80 回與完整 120 回的高頻字變化並不大，我們在此直接以 120 回的前 100 高頻字來分析。

60 回之間產生顯著差異！事實上，在前 20 回與中間 60 回之間，表 1 的 100 個高頻字中有高達 55 字在字頻的分布上呈現顯著差異。若我們僅因「在、的、著」三字在前 80 回與後 40 回的顯著差異，就推定它們來自不同作者，那同理是不是也該根據「了」（以及其他數十個高頻字）的頻率分布差異，推論前 20 回與中間 60 回是由不同人所作呢？

由於《紅樓夢》前 80 回都是曹雪芹所著，因此前 20 回與中間 60 回在常用字頻率上的不一致，告訴我們一件很重要的事：即使是同一位作者，在同一本書的不同階段，其常用字的使用頻率還是很可能會出現明顯差異。要合理解釋圖 2 的現象，我們可將《紅樓夢》分成 01-20、21-80、81-120 回等三個敘事階段。每個階段內只有少數常用字會在頻率分布上有顯著變化，而不同階段間則可能有許多常用字在頻率分布上有顯著差異。當然，這樣一來我們將無法僅由「前 80 回與後 40 回在許多用字的頻率上有顯著差異」就推斷「兩者的作者不同」。我們還需要其他的證據。

（三）文本採礦的探勘結果與討論

趙岡與陳鍾毅（1975）曾指出，《紅樓夢》後 40 回的續書人在寫作上極力模仿曹雪芹的語氣與用字習慣，而且在許多地方使用口語的程度是有過之而無不及。我們認為不同作者在用字遣詞上各有偏好，也同意模仿他人文筆並不可能達到百分之百一致。因此，若前 80 回與後 40 回確由不同人所著，應該可找到一些蛛絲馬跡，讓我們相信同一位作者不太可能在文本前後產生如此不一致的文筆運用。除了前人所批評的若干情節不連貫、前 80 回的文筆遠較優美等主觀判斷外，我們還希望能夠找到一些較為客觀的量化證據。

上節的實驗說明，若將《紅樓夢》的敘事分為前 20 回（01-20）、中間 60 回（21-80）、以及後 40 回（81-120）三個階段，那麼單字在每回出現的頻率可能會隨著敘事階段不同而產生明顯改變。在這種情況下，欲推論「前 80 回與後 40 回的作者不同」，我們必須同時說明 (1) 前 80 回與後 40 回在某些用字的頻率有非常大的差異，以及 (2) 前 20 回與中間 60 回在許多用字的頻率雖也有差異，但整體說來這些不一致並不算太強烈。接下來我們將運用採礦方法，有系統地從《紅樓夢》全文找出滿足條件 (1) 的字詞。至於 (2) 則留待第三節第（四）小節第 1 點再討論。

上一小節描述的統計方法，主要是利用到字頻（term frequency），也就是單字在各章回出現的頻率。現在我們將利用另一個量化資訊，也就是文頻（文件頻率，document frequency）來進行字詞的探勘。我們的目標，是希望能找到「在前 80 回出現於多個章回，後 40 回卻幾乎銷聲匿跡」或「前 80 回幾乎沒有出現，但後 40 回卻散見於多個章回」的字詞。這些字詞將暗示前 80 回與後 40 回的作者在用字習慣上

有相當大的差異。試想一下，若有一個不易因敘事情節轉移而有明顯使用量變化的字詞，它在前 80 回的多篇章回出現，卻在後 40 回裡銷聲匿跡，這難道不是一件頗奇怪、頗值得探究的事嗎？

我們先定義一些符號。令 D 為一個文件集合，[21] 我們用 $|D|$ 表示該集合的大小（也就是 D 中文件的篇數），而用 $d \in D$ 表示 D 裡的一篇文件。對任一字詞 t，我們用 $t \in d$ 表示 t 有出現在文件 d 之中。令 $D_t = \{d: t \in d, d \in D\}$ 表示集合 D 裡含有字詞 t 的文件子集合。於是，$|D_t|$ 表示集合 D_t 的文件數，也就是字詞 t 在集合 D 的文頻。給定一個字詞 t 與集合 D，我們定義 t 在 D 的平均文頻（average document frequency）為 $p_t(D) = |D_t|/|D|$。我們可以這樣說：平均說來 D 的任一文件有 $p_t(D)$ 的機會出現字詞 t；或者在 D 裡平均每 $1/p_t(D)$ 篇文件就可發現字詞 t 的蹤跡。

我們希望從《紅樓夢》全文裡找出「在前 80 回與後 40 回的平均文頻有明顯差異」的字詞。令 t 為任意一個字詞，X、Y 為任意兩個文件集合，我們定義採礦函數 $f_t(X, Y)$ 如下：

$$f_t(X,Y) = \frac{\max(p_t(X), p_t(Y)) + k}{\min(p_t(X), p_t(Y)) + k}$$

其中 k 為一個非負值常數。$f_t(X, Y)$ 的值表示，在比較文件集 X、Y 時，字詞 t 可能有趣的程度。舉例來說，令 A 表示《紅樓夢》前 80 回的文件，B 表示後 40 回的文件。若 $k = 0$ 且 $p_t(A) \geq p_t(B)$，則 $f_t(A, B) = p_t(A)/p_t(B)$，也就是字詞 t 在 A 與 B 平均文頻的比例。於是，$f_t(A, B)$ 越大表示 $p_t(A)$ 除以 $p_t(B)$ 的倍數越高，因而「在前 80 回與後 40 回平均文頻的差異」也越大。具有較大 $f_t(A, B)$ 的字詞 t，因為在 A 與 B 的平均出現篇數有明顯差異，因而較可能是有趣的。在採礦函數加上常數 k 的主要原因，是為了防止 $\min(p_t(X), p_t(Y)) = 0$ 時，產生 $f_t(X, Y) = \infty$ 的困擾。在以下的實驗中，我們都取 $k = 0.02$。

令 T_1 為《紅樓夢》全文出現的所有單字（unigram）集合，T_2 為全文出現的所有雙字詞（bigram）集合。[22] 對於 T_i, $i \in \{1, 2\}$ 的每個字詞 t，我們利用採礦函數 $f_t(A, B)$ 計算 t 在 T_i 的分數。表 3 列出 T_1 前 20 個高分的單字，而表 4 列出 T_2 的前 30 個高分雙字詞。以下我們分成兩點來討論其中一些有趣的發現。

21 在我們的實驗裡，把《紅樓夢》的每一回都視為一篇文件。

22 單字指的是全文中任一中文字元，而雙字詞指的是任意兩個不含標點符號的連續的中文字詞。例如句子「列位看官：你道此書從何而來？」所包含的單字為「列、位、看、官、你、道、此、書、從、何、而、來」，而雙字詞則為「列位、位看、看官、你道、道此、此書、書從、從何、何而、而來」。

1. 採礦所得的單字分析

首先要注意的是，若某字詞的出現次數太少，它的頻率分布將因隨機性較高而不具統計上的代表性。在這裡，我們只討論出現頻率高於 20 的字詞。從表 3 可知，單字詞「嬤、裡、展、胭、併」在前 80 回至少出現於 20 個章回（因此字頻也不小於 20），但在後 40 回裡卻一次也沒有出現。其中，「嬤」字僅以「嬤嬤」形式出現，我們將在下一點討論。「裡」字常與「裏」字通用，[23] 可能在傳抄或打字的過程中被替換，因此在未經過更仔細的檢視前，不宜貿然利用該字的文頻或字頻來判斷作者用字風格。值得一提的是，在實驗文本「裡」字共出現 109 次，其中有接近一半（54 次）出現在庚辰本所缺的第 67 回。這暗示我們第 67 回的多數「裡」字，或許出自於後人所增修，而非來自曹雪芹原稿。

表 3　採礦函數 $f_t(A, B)$ 對《紅樓夢》前 80 回與後 40 回單字詞的部分計算結果

	t	f_t	$\|A_t\|$	$\|B_t\|$
1	嬤	22.3	34	0
2	裡	22.3	34	0
3	嗎	22.2	1	28
4	展	17.3	26	0
5	疆	14.8	0	11
6	胭	14.8	22	0
7	併	13.5	20	0
8	困	12.9	19	0
9	苑	11.0	16	0
10	咧	11.0	2	19

	t	f_t	$\|A_t\|$	$\|B_t\|$
11	乳	10.4	15	0
12	宦	10.4	15	0
13	惦	9.9	2	17
14	蒸	9.8	14	0
15	苔	9.8	14	0
16	嘲	9.8	14	0
17	麵	9.1	13	0
18	匠	9.1	13	0
19	蕉	9.1	13	0
20	協	9.1	13	0

註：令 A、B 分別表示《紅樓夢》前 80 回與後 40 回的文件。我們令 k = 0.02，然後利用採礦函數 $f_t(A, B)$ 計算單字「可能有趣」的分數。此表列出得分最高的前 20 個單字。排名最前的「嬤」和「裡」雖然都出現於前 80 回的 34 個章回，但它們並不是都同時出現。[24]

23　《紅樓夢》中「裏」字出現的頻率遠高於「裡」字，而「房裏」、「房裡」等用法都曾出現。

24　「嬤」出現於第 3、5、8、16、18、19、20、23、24、26、29、35、36、37、39、40、42、43、44、45、48、51、52、53、54、57、63、71、74、75、76、77、79、80 回；而「裡」出現於第 5、6、7、9、12、19、21、22、24、25、26、27、28、29、30、34、35、36、37、39、40、42、43、50、51、52、55、59、60、64、67、71、73、77 回。

光從字面看一個單字，並不容易了解該字在小說中是以哪種形式出現。利用前後綴詞工具[25]，我們可以觀察某特定詞彙前後會有哪些字詞出現，其出現的頻率又是多少。換句話說，前後綴詞的分析結果可以幫助我們了解某指定字詞在小說中是如何被使用的。[26] 例如分析表 3「展」的後綴字（圖 3），我們可以知道「展」字主要被使用於「展眼」和「展開」，而《紅樓夢》中共有 13 個章回出現「展眼」一詞。運用類似的方法，我們知道「胭」字主要以「胭脂」的形式出現於 19 個章回，而「併」則多以「一併」的形式出現於另外 19 個章回。

term	df	tf
展眼	13	15
展開	5	6
展的	2	2
展拜	2	2

term	df	tf
展呈	1	1
展轉	1	1
展了	1	1
展其	1	1

term	df	tf
展些	1	1
展才	1	1
展奇	1	1
展幻	1	1

term	df	tf
展不	1	1

圖 3　《紅樓夢》裡「展」字僅出現於前 80 回。在此我們利用綴詞工具列出「展」的後綴字。第一列的 term 表示字詞，df 表示文頻，而 tf 則是詞頻的意思。從此圖可知「展」在《紅樓夢》主要的用法是「展眼」（df = 13 表示該詞出現於 13 個章回，tf = 15 表示它總共出現 15 次）與「展開」（df = 5，tf = 6）。

我們想強調的是，若《紅樓夢》全書都是曹雪芹所著，那麼一個字詞出現於前 80 回的多篇章回、卻在後 40 回銷聲匿跡，這種可能性是很低的。當然，如果多組字詞都出現這類狀況，其同時發生的機會必然更低。我們如果假設《紅樓夢》全書都是同一作者所著，且每回出現該字詞的機會均等，就可以從理論推算這種可能性的機率。[27] 例如，「展」字在前 80 回的平均文頻為 26/80 = 0.325，若假設它在全書的每一回都應有此機率出現，[28] 那麼（實際上）後 40 回並沒有出現的機率將僅有 $(1 - 0.325)^{40} = 0.00000015$。如果我們因敘事階段不同而放寬條件，假設「展」在後 40 回每一回出現的機率是 0.325 的一半，那麼後 40 回都看不到「展」字的機率就是 $(1 - 0.1625)^{40} = 0.0008$，可能性依然相當低。

25　該工具的使用方式，可參考 THDL（臺灣歷史數位圖書館）線上工具集，上網日期：2012 年 9 月 27 日。網址：http://thdl.ntu.edu.tw/SimpleTools/TermPat/TermPatSimpleUI.php

26　由於綴詞工具可輕易列出文本中某字詞前後所有可能的綴詞，它對於文本的派生詞研究應有相當大的助益。關於《紅樓夢》的派生詞研究可參考（陳俐后，2010）。

27　實際上，由於不同的敘事階段可能有不同的機率分布，這類機率試算的前提假設未必成立。在此進行這些試算，主要是為了強調（而不是證明）這些情況會發生的可能性是很低的。

28　「展」出現於第 1、5、6、8、15、18、19、23、25、27、28、34、37、39、43、47、48、53、56、62、68、69、70、71、76、77 回，其分布還算均勻，因此這裡的機率試算結果應可被參考或接受。

接下來我們討論兩個在前 80 回鮮少出現、在後 40 回卻經常可見的字：

(1) 在我們的實驗文本中，前 80 回僅在第 67 回出現一次「嗎」這個字：[29]
鳳姐笑道：「寶兄弟屋裡雖然人多……，我還捨得麻煩你嗎？我的姑娘！」比較《紅樓夢稿》、《紅樓夢校注》以及坊間據程甲本、程乙本刊印的《紅樓夢》，[30] 在這一段的描述文字都與實驗文本不同，且沒有出現「嗎」這個字：鳳姐笑道：「煩是沒的話。倒是寶兄弟屋裡雖然人多……，說你背地裡還惦著我，常常問我。這就是你盡心了。」由於前 80 回僅此段出現「嗎」字，我們懷疑實驗文本的這一段文字可能並非曹雪芹所撰。這結果也呼應趙岡和陳鍾毅（1975）的發現。他們說，《紅樓夢》前 80 回都是用「麼」而沒有「嗎」，後 40 回則兩字混用。趙陳還進一步比較《紅樓夢稿》的「正文」及塗改後的「改文」，發現只有「改文」才有出現「嗎」這個字。

(2) 表 3 另一個有趣的字是「咧」，它在前 80 回僅見於第 6、67 兩回，而在後 40 回卻有 19 個章回出現該字。若不考慮出現在第 67 回（庚辰本缺第 67 回）的「咧」，我們發現前 80 回的「咧」字都出現在第 6 回的某個段落：[31]
說著又推板兒道：「你那爹在家怎麼教你來？打發咱們作啥事來？只顧吃果子咧！」鳳姐早已明白了，聽她不會說話，因笑止道：「不必說了，我知道了。」因問周瑞家的道：「這姥姥不知可用了過早飯沒有呢？」劉姥姥忙道：「一早就往這裡趕咧，那裡還有吃飯的工夫咧！」

劉姥姥從《紅樓夢》第 6 回開始出場，而她在前 80 回裡只有這一段對話密集用到「咧」這個語尾助詞。然而這並不表示，第 6 回的這段文字必然經過後人修改潤飾。原因是，劉姥姥是位鄉下的老寡婦，曹雪芹可能特地用多個「咧」字強調其初見鳳姐時的心慌模樣。因此，用字的差異並不一定來自於不同的作者風格，而有可能是同一位作者對語境的特殊安排。這個例子也指出，我們的採礦方法有潛力協助研究者偵測出這類有趣的前後用語差異。

29　實驗文本的這段文字應是取自戚序本，請見參考文獻所列《戚蓼生序鈔本石頭記》，頁 2604-2605。

30　請參見參考文獻所列，廣文書局的《紅樓夢稿》、里仁書局的《紅樓夢校注》、以及正展出版社（程甲本）、建宏出版社（程乙本）印行的《紅樓夢》。以下章節在實驗時所對照參考的《紅樓夢稿》、《紅樓夢校注》、「程甲本」、「程乙本」均指這幾本書。

31　這段文字共有三處「咧」字。經比對程甲本和《紅樓夢稿》缺第三個「咧」，程乙本缺第一個「咧」，《紅樓夢校注》則三個「咧」都有。

2. 採礦所得的雙字詞分析

表 4 列出採礦所得的前 30 個雙字詞。

表 4　採礦函數 $f_t(A,B)$ 對《紅樓夢》前 80 回與後 40 回雙字詞的部分計算結果

	t	f_t	$\|A_t\|$	$\|B_t\|$		t	f_t	$\|A_t\|$	$\|B_t\|$		t	f_t	$\|A_t\|$	$\|B_t\|$
1	豈知	32.0	0	24	11	幸而	17.9	27	0	21	訴老	16.0	0	12
2	知端	27.9	43	0	12	聽如	17.9	27	0	22	可巧	15.4	54	1
3	未知	24.5	1	31	13	人皆	17.3	26	0	23	老嬤	15.4	23	0
4	一語	22.9	35	0	14	等語	17.3	26	0	24	角門	15.4	23	0
5	嬤嬤	22.3	34	0	15	是哪	16.6	25	0	25	這等	15.4	23	0
6	當下	22.3	34	0	16	彼時	16.0	24	0	26	名喚	15.4	23	0
7	皆是	21.0	32	0	17	起黛	16.0	0	12	27	日也	15.4	23	0
8	語未	20.4	32	0	18	照料	16.0	0	12	28	分明	15.4	23	0
9	取笑	19.8	30	0	19	外任	16.0	0	12	29	去吃	15.4	23	0
10	惦記	18.5	0	14	20	素來	16.0	0	12	30	海疆	14.8	0	11

註：令 A、B 分別代表前 80 回與後 40 回的文件集合。這張表列出取 k = 0.02 後，利用採礦函數 $f_t(A, B)$ 計算後分數最高的前 30 個雙字詞。表中 f_t 表示 $f_t(A, B)$，而 $|A_t|$、$|B_t|$ 分別表示字詞 t 在前 80 回與後 40 回所出現的文件篇數。

以下我們對其中一些有趣者進行較為深入的分析與討論：

(1) **豈知**：在後 40 回中「豈知」出現於 24 個章回，平均文頻高達 24/40 = 0.6。若假設這個詞彙以相同的機率隨機出現於每個章回，我們可以計算出後 40 回都沒有出現「豈知」的可能性小到令人難以置信。[32] 若從相似詞彙的觀點來比較，我們發現《紅樓夢》的前 80 回曾出現「誰知」、「那裡知道」、「哪裡知道」、「那知」、「哪知」、「焉知」這幾種與「豈知」有類似意義的用法 [33]。我們舉幾個例子如下：

32　機率試算的方式如下：首先，我們可假設該作者所寫的每一個章回，出現「豈知」的機率都是 0.6。那麼若前 80 回也是該作者所著，實際上這 80 回都沒有出現「豈知」的機率就僅有 $(1 - 0.6)^{80} = 1.46 \times 10^{-32}$。即使將條件放寬，假設每章回出現「豈知」的機率為 0.6 的一半，那麼前 80 回沒有「豈知」的機率依然僅有 $(1 - 0.3)^{80} = 4 \times 10^{-13}$，小於一兆分之一！

33　另一個與「豈知」有相近意義的詞彙是「安知」、「怎麼知道」或「怎知」。《紅樓夢》並沒有出現「安知」的用法；而前 80 回「怎麼知道」、「怎知」都不是出現在子句的開端，其前方通常接有名詞、代名詞或動詞（例如第 25 回：你怎麼知道他在那世裏受罪不安生？第 34 回：過後老太太不知怎麼知道了，說是珍大哥哥治的。第 74 回：太太怎知是我的？），其用法與後 40 回「豈知」都被置於子句開端並不相同。

（第 01 回）誰知此石自經煆煉之後，靈性已通，因見眾石俱得補天……

（第 62 回）誰知賈環聽如此說，便起了疑心。

（第 79 回）原不過是我一時的頑意，誰知又被你聽見了。

（第 61 回）只知雞蛋是平常物件，那裏知道外頭買賣的行市呢。

（第 53 回）你們山坳海沿子上的人，哪裏知道這道理。

（第 01 回）正嘆他人命不長，那知自己歸來喪。

（第 27 回）還認作是昨日中晌的事，哪知晚間的這段公案，還打恭作揖的。

（第 74 回）況且她們也常進園，晚間各人家去，焉知不是她們身上的？

對照幾個後 40 回「豈知」的例子：

（第 81 回）豈知那水裡的魚看見人影兒，都躲到別處去了。

（第 108 回）豈知寶玉只望裡走，天又晚，恐招了邪氣……

（第 120 回）今日又不肯叫人相伴。豈知到了五更，寒顫起來。

由於前 80 回類似「豈知」的詞彙以「誰知」的出現頻率最高（前 80 回裡共出現於 67 個章回，計 181 次，遠高於其他類似詞彙的頻率加總），我們可以推知曹雪芹通常是用「誰知」（而不是「豈知」）來表達「那裡知道、誰曉得」的意思。

(2) **知端**：前 80 回經常出現的「知端」這兩字，乍看之下不知道它究竟是怎麼被使用的。我們利用綴詞工具檢視其前後各一字，發現它其實是章回末「要知端的」、「要知端詳」、「不知端的」的一部分。在實驗文本後 40 回的章回末端，都沒有「知端」這樣的用法，而是以「未知如何」、「未知何事」、「要知後事如何」這類詞句來結尾。雖然「知端」一詞在實驗文本的 $f_t(A, B)$ 值很高，但經比對程甲本、程乙本以及《紅樓夢校注》，發現這些版本在章回末的結語文字經常頗有出入。這表示章回末「要知端的，下回分解」之類的詞句可能在小說流通的過程中經後人多次增補修訂。在未深入了解之前，並不適合利用這些章回末端詞語來分辨前 80 回作者是否也寫了後 40 回。

(3) **未知**：關於「未知」這個詞，多數都是出現在章回最末端「未知如何，下回分解」、「未知是誰，下回分解」、「未知何事，下回分解」等，少數出現在「還喜賈母、賈政未知」、「都哭過了。那衣衾未知裝裹妥當了沒有？」、「早把紅塵看破。只是自己的底裡未知」等句子中。在後 40 回中有 31 回可見到這個詞，比例非常高。有意思的是，前 80 回裡它僅在第 64 回以「未知

如何，下回分解」出現。因為庚辰本在第 64 回從缺，因此我們懷疑這個含有「未知」的詞句可能也非出自曹雪芹之手。

(4) **一語**：「一語」可見於前 80 回的 35 個章回，但它在後 40 回卻一次也沒有出現。利用綴詞工具檢視其後綴兩字（圖 4），可發現有 26 回出現「一語未了」，另各有 4 回出現「一語未完」、「一語提醒」、「一語不發」。這個雙字詞在前 80 回的平均文頻為 35/80 = 43.75%。若假設後 40 回與前 80 回的作者相同，且每回出現「一語」的機率也都是 43.75%，我們可計算出（實際上）後 40 回並沒有出現該詞的機率為 $(1 - 0.4375)^{40} = 1.01 \times 10^{-10}$。換句話說，若這些假設成立，後 40 回也是曹雪芹所著的機率將僅約百億分之一！

term	df	tf
一語未了	26	41
一語未完	4	4

term	df	tf
一語不發	4	4
一語提醒	4	4

term	df	tf
一語傳出	1	1
一語戳動	1	1

term	df	tf
一語倒把	1	1
一語未終	1	1

圖 4　「一語」僅出現於前 80 回。這裡利用綴詞工具分析「一語」的後綴兩字。最常出現的是「一語未了」，它出現在前 80 回的 26 個章回中，總共出現 41 次。

(5) **嬤嬤**：前 80 回共有 34 回出現「嬤嬤」一詞（總出現次數達 150 次），[34] 而後 40 回卻一回也沒出現過。在《紅樓夢》裡，「嬤嬤」的意思是有特殊地位的年老女僕或年紀大的奶媽兼女僕。「嬤嬤」因指涉到人物，其出現頻率容易因情節需要而出現於連續數個章回（或連續數個章回沒有登場）。但是《紅樓夢》中「嬤嬤」涵蓋了相當多人物（李嬤嬤、宋嬤嬤、賴嬤嬤、趙嬤嬤、王嬤嬤等），在 71-80 回「老嬤嬤」一詞更是密集出現於 71、74、75、76、77、79、80 等多個章回。[35] 另一方面，寧國府在第 105 回被查抄前，府內應該還有許多「嬤嬤」從事服務的工作。若後 40 回的作者也是曹雪芹，「嬤嬤」這個角色應不至於從第 80 回後就突然從偌大的賈府銷聲匿跡。換句話說，雖然無法僅憑「嬤嬤」的出現頻率推論後 40 回是否為後人所續，但這項採礦結果仍透露出這個字詞（或人物角色）在前後出現頻率的不協調性。

(6) **取笑**：前 80 回有 30 回出現「取笑」一詞，後 40 回則一次也沒出現。類似於「嬤嬤」是有意義的角色，「取笑」也是有意義的動作與行為，它是表示「招人譏笑」或「戲弄、開玩笑」的意思。「取笑」在前 80 回共出現 45 次，利用

34　這 34 回為：3、5、8、16、18、19、20、23、24、26、29、35、36、37、39、40、42、43、44、45、48、51、52、53、54、57、63、71、74、75、76、77、79、80。注意到，庚辰本所缺的第 64、67 兩回都沒有出現「嬤嬤」。

35　「老嬤嬤」出現於 3、8、18、23、24、29、40、42、43、44、48、51、52、53、54、63、71、74、75、76、77、79、80 共 23 回。

檢索系統查閱出現該詞彙的前後字句，可發現曹雪芹能夠相當生動地使用這個詞彙。「取笑」在後 40 回完全消失，顯示後 40 回的作者很可能並非曹雪芹。

(7) **可巧**：在前 80 回裡共有 54 回出現過「可巧」這個詞，其平均文頻高達 54/80 = 67.5%。另一方面，該雙字詞在後 40 回只出現於第 98 回，平均文頻僅 1/40 = 2.5%。事實上，若扣除人名「巧姐」，[36] 我們發現前 80 回非用於「巧姐」的「巧」字共出現 184 次，而後 40 回則僅有 13 次。這表示曹雪芹相當善用「巧」字，相對之下後 40 回的作者則很少用到這個字。

(8) 至於「當下」、「皆是」、「取笑」、「幸而」、「等語」、「彼時」等詞，它們都僅在前 80 回出現，且出現的篇數都至少有 24 篇（平均文頻至少 30%）。這表示曹雪芹對這些詞的使用頗為熟悉，而後 40 回的作者則不會使用這些詞彙。

最後，不要忘記《紅樓夢》本身曾經歷五次大增刪 [37]。由於這類增刪是全面性的（而不是僅對部分內容進行增刪），因此若 120 回都是曹雪芹一人所著，那麼多次出現於前 80 回的「一語、當下、皆是、取笑、幸而、等語、彼時」也應散見於後 40 回。同理，前 80 回並沒有出現，但後 40 回常見的「豈知、未知」等字詞，也從平均文頻的不一致性，支持兩者應有不同作者。

（四）字詞頻率差異的補充探討

上一小節已證實《紅樓夢》許多字詞在前 80 回與後 40 回之間的頻率分布存有明顯差異。我們在這一小節裡將進行一些補充探討，加強後 40 回確為後人所續的證據。我們在下文說明前 20 回（01-20）與中間 60 回（21-80）雖然也有幾個字詞在平均文頻上出現明顯差異，但可將此不一致視為敘事階段轉移所造成。接著，我們將繼續討論利用前後綴詞工具所獲得的一些有趣發現。

1. 前 20 回與中間 60 回的用字差異

令 C、D 分別表示前 20 回與中間 60 回的文件集合。我們利用採礦函數 $f_t(C, D)$ 比較 C 與 D 在平均文頻的差異，然後將前 10 高分的單字與前 20 高分的雙字詞列於表 5。表列單字「嫡、芬、覽、鯨」在《紅樓夢》小說的總出現次數都沒有超過 10 次，其頻率分布在統計上較不具代表性，我們在此略去而不加以討論。「儒、塾」兩字與敘事情節有關：「儒」牽涉到人名「賈代儒」，而小說前 20 回則對賈府家塾的相

36　「巧」字在前 80 回與後 40 回的文頻（字頻）分別為 70（198）、20（117），而「巧姐」在前 80 回與後 40 回的文頻（字頻）則分別為 2（5）、14（104）。

37　《紅樓夢》小說第一回：「後因曹雪芹於悼紅軒中，披閱十載，增刪五次，纂成目錄，分出章回」。

關事務著墨頗多。「詫」幾乎都是以「詫異」形式出現，「賬」用於「算賬」、「混賬」等，「仇」多用於「報仇」、「仇人」、「仇恨」等，而「兜」的使用方式則是「吃不了兜著走」、「兜肚」等。以「詫」字而言，若我們假設前 80 回每回出現該字的機率都是 18/60 = 0.3，那麼前 20 回找不到「詫」的機率是 $(1 - 0.3)^{20} = 0.0008$。這麼低的機率表示「詫」字的平均文頻在前 20 回與中間 60 回是有明顯差異的。

表 5　《紅樓夢》前 20 回與中間 60 回出現顯著差異的部分字詞

	t	f_t	$\|C_t\|$	$\|D_t\|$
1	嫡	18.5	7	0
2	詫	16.0	0	18
3	塾	16.0	6	0
4	儒	13.5	5	0
5	賬	13.5	0	15
6	仇	12.7	0	14
7	兜	11.8	0	13
8	芬	11.0	4	0
9	覽	11.0	4	0
10	鯨	11.0	4	0

	t	f_t	$\|C_t\|$	$\|D_t\|$
1	秦氏	23.5	9	0
2	園裡	21.8	0	25
3	鳳丫	18.5	0	21
4	了秦	18.5	7	0
5	春道	17.7	0	20
6	寶姑	17.7	0	20
7	不留	17.7	0	20
8	前頭	16.8	0	19
9	竟沒	16.8	0	19
10	想得	16.8	0	19

	t	f_t	$\|A_t\|$	$\|D_t\|$
11	大書	16.0	6	0
12	揚州	16.0	6	0
13	見秦	16.0	6	0
14	說秦	16.0	6	0
15	小蹄	16.0	0	18
16	詫異	16.0	0	18
17	們好	16.0	0	18
18	啐了	16.0	0	18
19	見林	16.0	0	18
20	太也	16.0	0	18

註：令 C、D 分別表示前 20 回與中間 60 回的文件集合。這裡列出 $f_t(C, D)$ 得分最高的 10 個單字和 20 個雙字詞。$|C_t|$、$|D_t|$ 分別表示 C、D 裡含有字詞 t 文件篇數。

表 5 許多排序在前的雙字詞都與人名或地名有關，並不適合拿來進行文本前後用詞頻率差異的比較。例如「秦氏」（賈蓉之妻）、「園裡」（小說從第 18 回之後才開始提及大觀園）、「鳳丫」（鳳丫頭）、「了秦」（隨了秦氏、忘了秦氏、見了秦鐘等）、「春道」（迎春道、探春道等）、「寶姑」、「揚州」、「見秦」（見秦鐘、見秦氏）、「說秦」（話說秦氏、話說秦鐘）等都牽涉到人地名。「大書」也與敘事階段相關，多被用於前 20 回介紹榮國府、太虛幻境、寧國府時，例如「匾上大書『敕造寧國府』五個大字」、「又有一副對聯，大書云：厚地高天，堪嘆古今情不盡」等。

「不留」的主要用法是「不留心、不留神、不留你」，它在中間 60 回的平均文頻為 20/60 = 0.333。若假設「不留」在前 20 回的每回也都有 0.333 的出現機率，我們可以計算出（實際上）它沒有出現在前 20 回的機率是 $(1 - 0.333)^{20} = 0.0003$。如果假設前 20 回因敘事階段與中間 60 回不同，僅有一半於 0.333 的出現機率，那麼前 20 回看不到「不留」的機率就提高到 0.026。至於「前頭」、「竟沒」和「想得」，它們在中間 60 回的平均字頻都是 19/60 = 0.317。在相同的假設下，這三個字詞（各別）沒有出現於前 20 回的機率會比 0.0003 還要高一些。

相較於表 3「展、胭、併」與「豈知、一語、當下」等字詞在前 80 回與後 40 回的強烈文頻差異，我們認為前 20 回與中間 60 回的文頻差異並不算太大，可用敘事階段不同來解釋。換句話說，這裡的實驗結果並不違背前 80 回都是曹雪芹所著的事實。

2. 其他有趣的發現

第三節第（三）小節的有趣字詞，主要是利用採礦函數自動找出，只有在最後階段才需經由人力過濾不合適（例如與人名或地名強烈相關）的字詞。這裡所討論的字詞，則並非由演算法所直接產生，多半是在利用綴詞工具檢視採礦所得字詞的過程所發現。

從第三節第（二）小節的討論，我們知道《紅樓夢》中出現最頻繁的字是「了」，也知道該字的頻率分布在前 80 回與後 40 回之間雖沒有顯著統計差異，但在前 20 回與中間 60 回之間卻存在顯著差異。有趣的是，利用綴詞工具分析「了」字的前一字（圖 5），我們發現不管是在前 80 回、前 20 回、或者中間 60 回，其詞頻（tf）最高者都依序是「聽了、去了、來了」；另一方面，後 40 回最頻繁出現的則依序為「來了、去了、聽了」。換句話說，前 80 回的作者似乎對「聽了」的偏好高於對「來了」的偏好，而後 40 回作者則相反。

(A) 01-80

term	df	tf
來了	80	687
聽了	79	761
去了	78	740
得了	77	203
見了	75	349
罷了	75	320
說了	72	301
完了	65	141
是了	64	173
好了	62	215

(B) 81-120

term	df	tf
來了	40	375
去了	40	352
聽了	40	294
說了	39	213
到了	39	130
見了	38	224
應了	37	92
好了	36	130
的了	35	102
是了	35	90

(C) 01-20

term	df	tf
聽了	20	168
來了	20	134
得了	19	42
去了	18	158
見了	18	76
罷了	16	57
看了	15	34
是了	14	31
有了	14	30
坐了	14	30

(D) 21-80

term	df	tf
去了	60	582
來了	60	553
聽了	59	593
罷了	59	263
說了	59	253
得了	58	161
見了	57	273
完了	52	115
好了	51	191
拿了	51	138

圖 5　利用綴詞工具分析「了」的前一字。這裡列出在 (A) 前 80 回、(B) 後 40 回、(C) 前 20 回、(D) 中間 60 回裡文頻與詞頻較高的前 10 個雙字詞，其中 df 表示文頻，而 tf 則代表詞頻。將這四者一併列出，可讓我們比較文頻或詞頻的排序，是否會因敘事階段的改變而產生變化。在這裡我們主要是觀察詞頻：在前 80 回、前 20 回、以及中間 60 回裡，詞頻最高的都依序是「聽了、去了、來了」，而後 40 回出現最頻繁的字詞則依序為「來了、去了、聽了」。

表 6 分析「道」的前綴字，從這裡可觀察到另一個有趣的現象。雖然「道」的字頻分布在前 80 回與後 40 回並沒有統計的顯著差異（表 1），但「笑道」、「嘆道」、「忙道」這三個雙字詞在前 80 回的詞頻明顯較後 40 回為高，而「問道」在前 80 回的詞頻卻比後 40 回的詞頻低。就敘事內容而言，僅以「後 40 回賈府開始衰敗」很難解釋這種不一致，因此這項結果也支持後 40 回並非曹雪芹所作。

表 6　以 10 回為一個單位，分析幾個雙字詞在每個單位的詞頻

	0+	10+	20+	30+	40+	50+	60+	70+	80+	90+	100+	110+
笑道	140	188	298	353	408	328	240	294	94	59	33	37
嘆道	9	10	9	8	5	8	1	17	2	3	2	7
忙道	7	5	5	15	23	20	9	17	1	0	1	1
問道	21	16	32	21	14	17	11	19	68	28	18	28

註：本表以 10 回為一個單位，列出「笑道」、「嘆道」、「忙道」、「問道」這四個雙字詞在每個單位的詞頻。表上第一列的 n+ 表示第 (n + 1) 到第 (n + 10) 回。例如 0+ 表示第 01-10 回，70+ 表示第 71-80 回。這張表顯示，「笑道、嘆道、忙道」在前 80 回的詞頻明顯較後 40 回為高，而「問道」則相反。

我們還發現，「忙上」這個雙字詞在前 80 回共出現於 34 個章回，且幾乎都以「忙上來」（文頻 24）、「忙上前」（文頻 7）、「忙上去」（文頻 5）的形式出現。「忙上」於後 40 回只在第 104 回出現過一次：

雨村疾忙上轎進內，只聽見人說……

這裡出現的「忙上」，應該是「疾忙」緊接「上轎」所導致，其用法不同於「忙上來、忙上前、忙上去」。換句話說，後 40 回的作者並不會使用「忙上來、忙上前、忙上去」這種語法。

另外，「豈」字在《紅樓夢》共出現 354 次，而這個字被使用在「豈知、豈不」這兩個詞就占了 246/354 = 69.4%，將近七成。先前已經討論過，「豈知」的平均文頻在前 80 回與後 40 回有明顯差異。有趣的是，從表 7 可知，「豈不」在前 80 回與後 40 回的使用頻率不但有明顯差異，而且還與「豈知」的頻率呈現反比的關係。此外，這兩個字詞在前 80 回、前 20 回、中間 60 回的平均文頻都很接近，但卻與後 40 回的平均文頻相差甚多。換句話說，在前 80 回與後 40 回有接近七成的「豈」字用法明顯不一致。

表 7　兩個含有「豈」的雙字詞：「豈知」和「豈不」在各階段的出現頻率

Term t	$\|A_t\|$	$p_t(A)$	$\|B_t\|$	$p_t(B)$	$\|C_t\|$	$p_t(C)$	$\|D_t\|$	$p_t(D)$
豈知	0	0%	24	60%	0	0%	0	0%
豈不	67	84%	13	33%	16	80%	51	85%

註：「豈知」、「豈不」在前 80 回 (A)、後 40 回 (B)、前 20 回 (C) 與中間 60 回 (D) 的文頻與字頻。$|S_t|$ 表示集合 S_t 的文件篇數，而 $p_t(S)$ 則表示字詞 t 在集合 S 的平均文頻。我們可以發現，這兩個字詞的 $p_t(A)$、$p_t(C)$、$p_t(D)$ 都很接近，但卻與 $p_t(B)$ 有頗大差異。

最後，我們檢視《紅樓夢》中所有出現「……不兩……」的詞句：

（第 03 回）已擇了出月初二日，小女入都，尊兄即同路而往，豈不兩便?

（第 03 回）況這林妹妹眉尖若蹙，用取這兩個字，豈不兩妙！

（第 17 回）待貴妃游幸時，再請定名，豈不兩全?

（第 33 回）若十分愛慕，老大爺竟密題一本請旨，豈不兩便?

（第 43 回）這已完了心願，趕著進城，大家放心，豈不兩盡其道。

（第 69 回）若天見憐，使我好了，豈不兩全?

（第 72 回）不如借了來，奶奶拿一二百銀子，豈不兩全其美。

（第 78 回）設若從那裡生出一件事來，豈不兩礙臉面。

（第 78 回）姐姐何不等一等他回來見一面，豈不兩完心願?

（第 98 回）咱們一心一意的調治寶玉，可不兩全?

很明顯地，在 80 回之前「不兩」的前綴字必然為「豈」。有趣的是，在 80 回後所出現的「可不兩全」，其意義與前 80 回曾出現的「豈不兩全」是一樣的。由此我們幾乎可以肯定地說，在第 98 回出現的「可不兩全」必然出自後人之手。

四、結論

《紅樓夢》後 40 回作者的爭議存在已久。由於可供考證的材料稀少，加上資訊科技的進步，近年許多學者轉而利用文本的統計資訊來解決這個問題。

跟隨前人的腳步，我們確認《紅樓夢》前 80 回與後 40 回之間，有許多用字在各章回的頻率分布存在顯著差異。我們還發現前 20 回（01-20）與中間 60 回（21-80）之間，有不少高頻字在兩者的頻率分布也存有顯著差異。依照每 10 回為一個單位的

字頻分布，我們發現《紅樓夢》可依敘事階段分為前 20 回、中間 60 回以及後 40 回。各個階段裡常用字的頻率較為一致，而相同單字在不同階段之間的頻率則可能存有相當大的差異。這項發現使得利用字詞頻率探討作者爭議變得複雜，因為除了舉證前 80 回與後 40 回在用字遣詞上有極大差異，我們還需說明前 20 回與中間 60 回雖也有用字差異，但相對之下較為一致。

我們選擇元智大學公開於網路的《紅樓夢》全文，以文本採礦方法找出前 80 回與後 40 回在使用頻率上有明顯差異的字詞。我們發現「嬤嬤」、「豈知」、「取笑」、「可巧」、「一語」、「當下」等用詞在前 80 回與後 40 回存在非常明顯的差異。值得注意的是，這當中有許多詞彙本身是具有意義的。我們認為，若無法以敘事階段移轉或隨機性來合理解釋這些差異，就證實前 80 回與後 40 回在用字遣詞上存有重要的差異。相較之下，前 20 回與中間 60 回之間的詞頻差異較小，而且有明顯差異的詞彙多半與小說劇情演進密切相關，因而可用隨機性或敘事階段移轉來解釋。

我們利用前後綴詞工具，發現在前 80 回、前 20 回與中間 60 回，「了」字最常依序被使用於「聽了、去了、來了」，這與後 40 回最常出現的「來了、去了、聽了」順序不同。分析「道」的前綴字，我們發現「笑道」、「嘆道」、「忙道」在前 80 回的頻率明顯較後 40 回為高；而「問道」則相反，它在前 80 回的詞頻比後 40 回低上許多。最後，我們發現「豈」字有接近七成是被使用於「豈知」與「豈不」。這兩個字詞在前 20 回與中間 60 回平均每回的出現次數都頗一致，但卻與後 40 回平均每回的出現次數有著相當大的差異。

在檢視採礦所得字詞的過程中，我們也發現一些證據，說明今存的《紅樓夢》第 64、67 兩回（庚辰本這兩回從缺），很可能也是後人所著。例如，「未知」在後 40 回中散見於 31 回，在前 80 回裡這個詞卻僅出現於第 64 回；而實驗文本「裡」字共出現 109 次，其中卻有接近一半（54 次）密集出現於第 67 回。

我們利用文本採礦所找出的有趣字詞，多半散見於《紅樓夢》各章回，不容易藉由人力閱讀來發現。這些結果說明妥善利用資訊工具，可以有效幫助人文學者從文本中發掘新事證。我們的實驗證實《紅樓夢》的用字遣詞在前 80 回與後 40 回存有許多明顯的差異，而這些不一致並無法用前後敘事階段不同來解釋。這些發現支持《紅樓夢》後 40 回並非曹雪芹所作的論點。

參考文獻

何光國，2002，〈從漢語白話文虛字“的、地、得”的運用論作者寫作個性——兼論《紅樓夢》著者問題〉，《傳統中國文學電子報》，第 131 期。

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

俞平伯，1923，《紅樓夢辨》，上海：亞東圖書館。另河洛出版社於 1970 年在臺影印初版。

胡適，1921，《紅樓夢考證》，《紅迷論壇》提供其全文、附記與跋，上網日期：2012 年 9 月 27 日。網址 http://www.hungmi.com/bbs/viewtopic.php?id=23

袁維冠，1978，《紅樓夢探討》，花蓮：自印。

曹雪芹、高鶚原著，馮其庸等校注，1984，《紅樓夢校注》，臺北：里仁書局。

曹雪芹、高鶚著，1994，《紅樓夢》（三家評本），臺北：建宏出版社。

曹雪芹、高鶚著，2003，《紅樓夢》（程甲本），臺北：正展出版社。

曹雪芹、高鶚編著，1994，《紅樓夢》（程乙本），臺北：建宏出版社。

曹雪芹著，1977，《紅樓夢稿》，臺北：廣文書局。

曹雪芹著，1977，《戚蓼生序鈔本石頭記》，臺北：廣文書局。

陳俐后，2010，《紅樓夢派生詞研究》，國立政治大學中國文學系碩士論文。

陳炳藻、謝家浩，2003，〈情感與理性：文學電腦統計與華文教材〉，第三屆全球華文網路教育研討會，臺北。

趙岡、陳鍾毅，1975，《紅樓夢研究新編》，臺北：聯經出版社。

Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association, 58*(301), 85-96.

Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship, *Literary and Linguistic Computing, 17*(3), 267-287.

Hogg, R. V. & Tanis, E. A. (2006). *Probability and Statistical Inference, 7th edition*, NJ, USA: Prentice Hall.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing, 19*(4), 453-475.

Jockers, M. L. & Witten, D. M. (2010). A Com Parative Study of Machine Learning Methods for Authorship Attribution. *Literary and Lingistic Computing, 25*(2), 215-223.

Malyutov, M. B. (2006). Authorship Attribution of Texts: A Review. *In General Theory of Information Transfer and Combinatorics*. Berlin Heidelberg: Springer-Verlag, 362-380.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (1999). *Introduction to Probability and Statistics,* 10th edition. USA: Duxbury Press.

Peng, R. & Hengartner, N. (2001). Quantitative Analysis of Literacy Styles. *Department of Statistics Papers*. Department of Statistics, UCLA.

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities, 31*, 351-365.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538-556.

Thisted, R. & Efron, B. (1986). *Did Shakespeare Write a Newly-Discovered Poem? Technical Report, 111*, CA, USA: Division of Biostatistics, Stanford University.

Yang, A. C.-C., Peng, C.-K., Yien, H.-W. & Goldberger, A. L. (2003). Information Categorization Approach to Literary Authorship Disputes, *Physica A*, *329*, 473-483.