利用文本採礦探討《紅樓夢》的後四十回作者爭議

杜協昌

2012-11-30 (DADH 2012)

論文報告大綱

- 前言：介紹什麼是「後四十回作者爭議」
- 回顧：前人對此問題的一些研究成果
- 實驗方法、結果與討論
  - 版本：說明實驗所用的文本
  - 字頻一致性檢測：字頻統計方法與討論
  - 文本採礦方法：利用電腦探勘可能有趣的字詞
  - 前後綴詞方法：利用工具找出其他的有趣字詞
- 結論

前言

- 《紅樓夢》是著名的中國古典小說
  - 共有 120 回
  - 前八十回的作者公認是曹雪芹
  - 後四十回的作者則一直存有爭議
- 對後四十回作者的幾種不同看法
  - 與前八十回相同，是曹雪芹所著。
  - 是高鶚的補遺續稿之作。
  - 續書者另有他人（既非曹雪芹、也不是高鶚）

探討「後四十回作者爭議」的途徑

- 利用文本之外的資訊
  - 困難：足供考證的材料稀少
- 利用文本內容
  - 語意表達、敘事內容的一致性（主觀論證）
    - 前八十回與後四十回在許多內容上並不連貫
    - 前八十回在文字的運用上遠較優美
  - 語法結構、用字遣詞的一致性（客觀數據）
    - 選擇合適的版本。
    - 選擇量化的標的物（虛字），對其進行統計分析
    - 問題：前人以不同的實驗方法，得出抵觸的結論！

一些前人的研究結論

- 胡適 (1921)
  - 利用考證與內容一致性，認為後四十回都是高鶚補的
- 高本漢 (Karlgren, 1952)
  - 利用統計方法，認為全書只有一位作者
- 趙岡、陳鍾毅 (1975)
  - 利用虛字統計方法，認為後四十回並非曹雪芹所作
- 陳炳藻 (1981)
  - 利用統計方法，計算相關聯係數，認為全書都是曹雪芹所著
- 余清祥 (1998)
  - 利用統計方法，探討虛字與詩詞數量等，認為作者至少有兩人
- 何光國 (2002)
  - 利用虛字統計方法，認為《紅樓夢》的作者只有曹雪芹一人
- 楊智傑 等人 (Albert C.-C. Yang et al.) (2003)
  - 利用文件的字頻排序計算相似度，認為後四十回非曹雪芹所作

實驗所選擇的文本

- 元智大學在網路上所提供的全文版
  - http://cls.hs.yzu.edu.tw/hlm/read/text/text.asp
  - 實體書為蔡義江所校注的《紅樓夢》
  - 該書序言提到其版本選擇：以俞平伯《紅樓夢八十回校本》和《紅研所新校注本》為主，參校各種本子，以盡量保持曹雪芹原作面貌。

實驗:《紅樓夢》用字的一致性測試

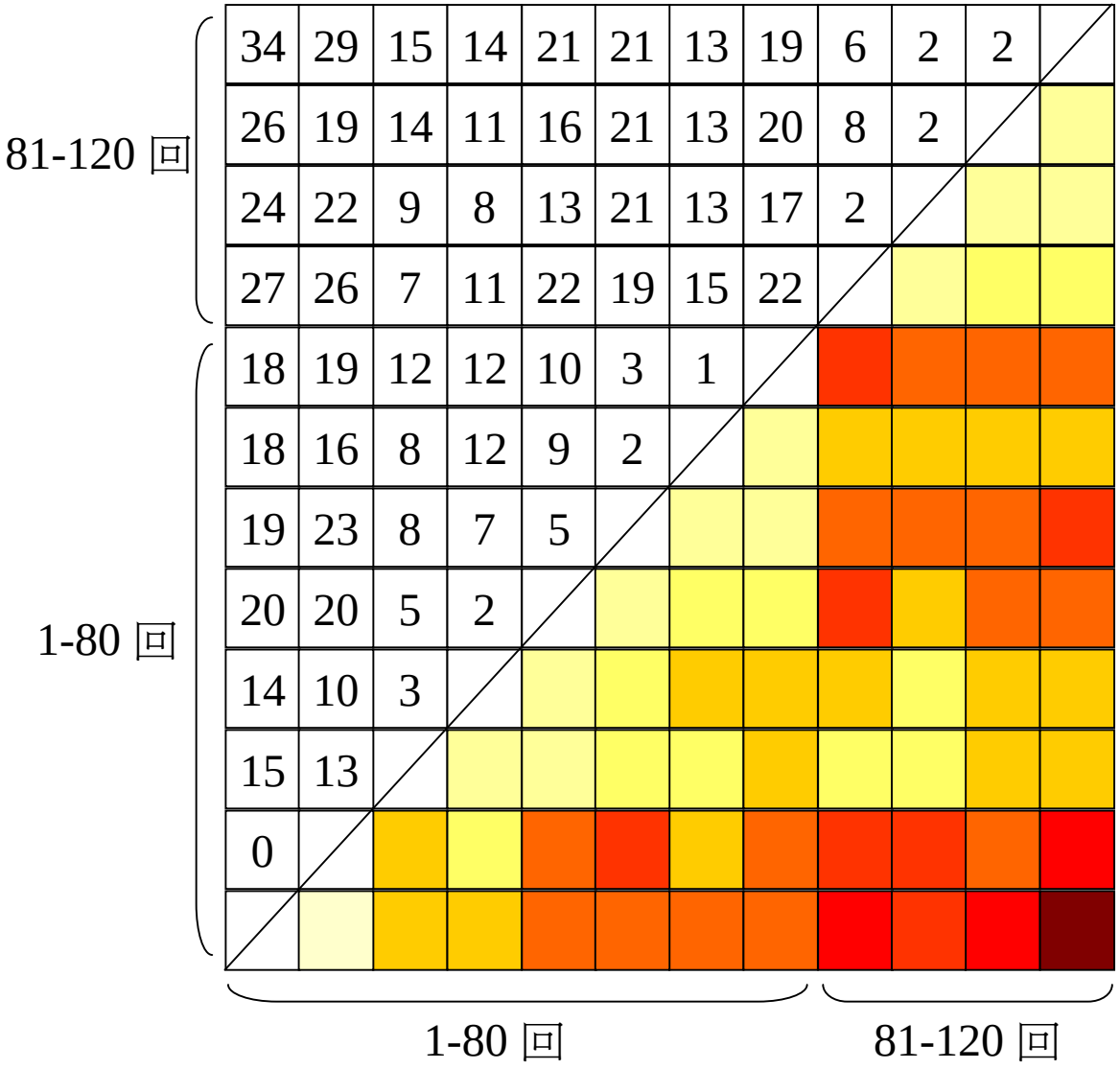
- 將 120 回分為 12 個單元，每個單元包含 10 回的內容
  - ❶ 計算《紅樓夢》中出現頻率最高的 100 個字
  - ❷ 對每個高頻字，計算其出現在各個單元（十個章回）的出現次數。
  - ❸ 對每個高頻字，利用假設檢定（取 p=0.01），測試其在兩單元的出現頻率是否有顯著不同。
- 統計任兩個單元之間，有多少高頻字在分佈上有顯著差異

出現頻率最高的 100 個字元

#	0+	10+	20+	30+	40+	50+	60+	70+	80+	90+
1	了	這	子	便	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	裏	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

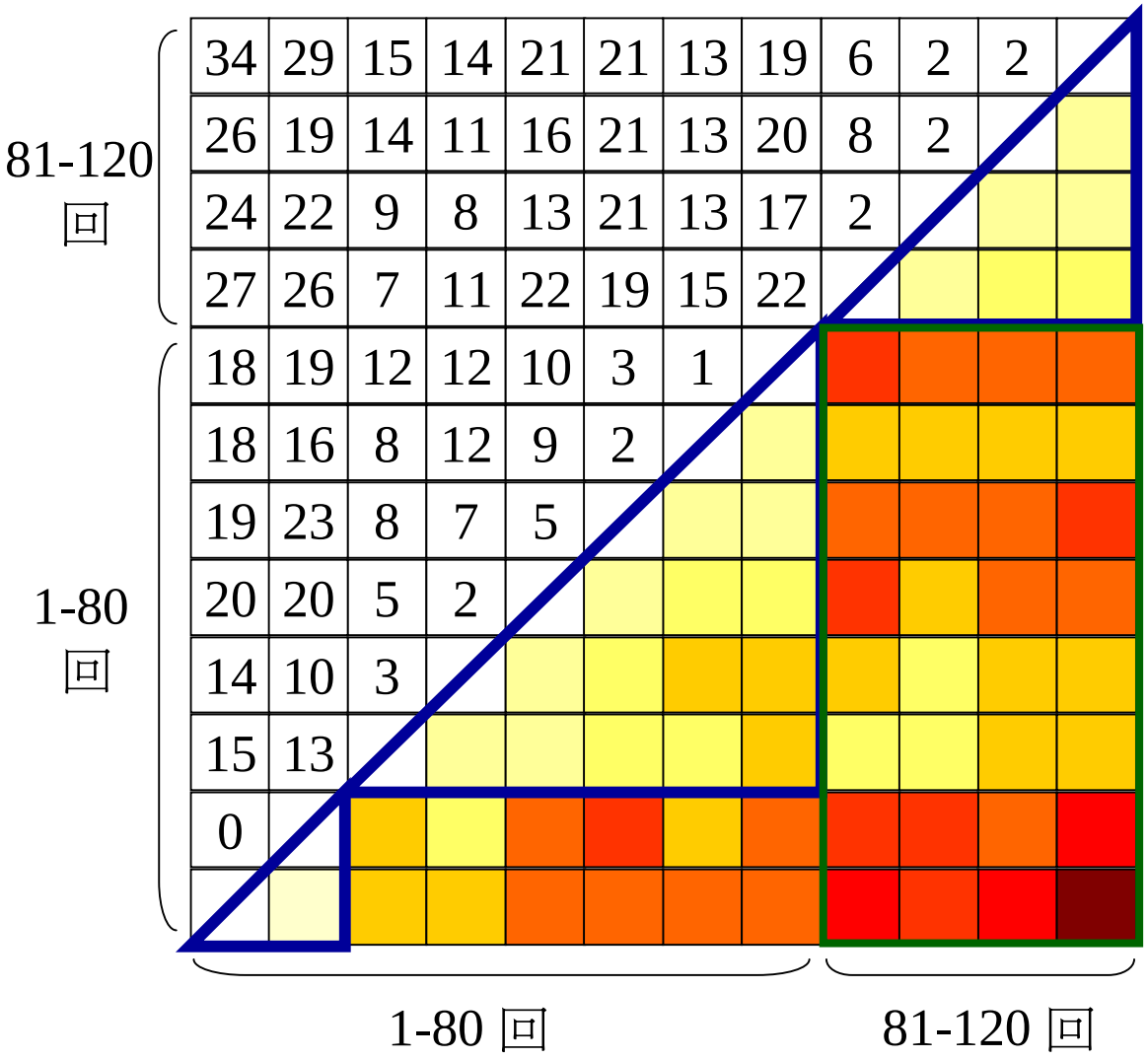
字頻有顯著差異的字數矩陣

統計兩兩單元之間，字頻有顯著差異的字元數



從《紅樓夢》的前 100 高頻字中，計算有幾個單字的使用頻率在 2-sample t-test 下具顯著差異 (two tails p=0.01)

解讀字頻測試的結果矩陣



- 01-20、21-80、以及 81-120 之間，用字較為一致（字頻分佈有顯著字頻的字數少）
- 前八十回與後四十回的用字差異較大（有顯著差異的字數多）
- 問題: 可否據此推論前八十回與後四十回出自不同作者？

我們的答案是「不行」

- 事實上，前 100 個高頻字中，有 55 個字（例如「了、也、著」等）在 01-20 回與 21-80 回的出現頻率有著顯著差異！
- 若僅憑前八十回與後四十回的字頻差異，就推論後四十回的作者並非曹雪芹...
  - 同理是否也該推論前二十回與中間六十回是由不同人所作？

字頻測試的矩陣告訴我們...

- 應該將《紅樓夢》分為三個敘事階段
  - 前二十回 (01-20)、中間六十回 (21-80)、以及後四十回 (81-120)
  - 相同敘事階段內的用字頻率較為一致，但不同敘事階段則可能有不小的差異
- 欲聲稱後四十回非曹雪芹所著
  - 除了前八十回與後四十回的字頻差異外，我們還需要更多的證據

想法：不同作者應有不同字詞偏好

- 若後四十回並非曹雪芹所著，應能從文本找出一些蛛絲馬跡，說明同一位作者不太可能在文本前後有著如此不一致的文筆運用
  - 同一作者在一本書中，即使創作前後期風格會有轉變，也不應有過多字詞在使用偏好上有差異。
  - 尤其《紅樓夢》經過五次大增刪，每次都應對全書進行檢視，在用字遣詞上更不應有過多的變化
- 問題是，該如何從文本中找出這些證據？

文本採礦

- 想法：利用電腦，找出在前八十回與後四十回「出現於任一章回的機率」有明顯差異的字詞
- 定義 $p_t(S)$ 表示「集合 S 中，任一篇文件可發現字詞 t 的機率」
  - 例如，若 S 代表前八十回，而「嬤」在前八十回出現於 34 個章回，其 $P_t(S)$ 就是 34/80 = 0.425
- 令 A, B 分別代表「前八十回」與「後四十回」，我們希望找出 $p_t(A)$ 與 $p_t(B)$ 有顯著差異的字詞 t。

文本採礦：實作

計算可能有趣的單字與雙字詞

- 令 A, B 分別代表「前八十回」與「後四十回」
- 令 t 為一個單字或雙字詞，我們用 $f_t(A,B)$ 代表 t 在 A,B 的差異程度

$$f_t(A,B) = \frac{\max(p_t(A), p_t(B)) + k}{\min(p_t(A), p_t(B)) + k},$$

  - 常數 k 是為了防止在 $p_t(A)=0$ 或 $p_t(B)=0$ 的情況下，產生 $f_t(A,B) = \infty$ 的困擾。在實驗中我們取 k=0.02 。
- 計算分數最高的單字 (unigrams) 與雙字詞 (bigrams)

採礦所得：前 10 高分的候選單字

No.	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

$|A_t|$ 表示 t 在前八十回的出現篇數
$|B_t|$ 表示 t 在後四十回的出現篇數

「嬤」可見於前八十回的 34 個章回，卻不見於後四十回的任一章回。因此

$$f_t(A,B)=\frac{\frac{34}{80}+0.02}{\frac{0}{40}+0.02}=22.25$$

此數值很高，表示其出現頻率差異大。

「嗎」僅見於前八十回的 1 個章回，而在後四十回出現於 28 章回。因此

$$f_t(A,B)=\frac{\frac{28}{40}+0.02}{\frac{1}{80}+0.02}=22.154$$

表示其出現頻率的差異頗大。

採礦所得單字討論：嗎

- 呼應了趙岡數十年前的發現：
  - 《紅樓夢》在前八十回都是用「麼」，而後四十回則是「嗎、麼」兩字混用
- 實驗文本前八十回「嗎」字僅在 67 回出現一次

> 鳳姐笑道：「寶兄弟屋裏雖然人多...，我還捨得麻煩你嗎？我的姑娘！」

- 《紅樓夢稿》、程甲本、程乙本此段都是

> 鳳姐笑道：「煩是沒的話。倒是寶兄弟屋裏雖然人多...，說你背地裏還惦著我，常常問我。這就是你盡心了。」

- 我們懷疑實驗文本這一段文字並非出自曹雪芹

採礦所得：前 10 高分的候選雙字詞

No.	t	f_t	$\|A_t\|$	$\|B_t\|$
1	豈知	31.0	0	24
2	知端	27.9	43	0
3	未知	24.5	1	31
4	一語	22.9	35	0
5	嬤嬤	22.3	34	0
6	當下	22.3	34	0
7	皆是	21.0	32	0
8	語未	20.4	31	0
9	取笑	19.8	30	0
10	惦記	18.5	0	14

「豈知」在前八十回並沒有出現，而在後四十回可見於 24 個章回。因此

$$f_t(A,B)=\frac{\frac{24}{40}+0.02}{0.02}=31$$

表示其出現頻率差異非常大。

「嬤嬤」可見於前八十回的 34 個章回，而在後四十回徹底消失。因此

$$f_t(A,B)=\frac{\frac{34}{80}+0.02}{0.02}=22.25$$

表示其出現頻率有頗大差異。

採礦所得雙字詞討論：嬤嬤

- 前八十回共有 34 回出現「嬤嬤」一詞
- 在《紅樓夢》裏，「嬤嬤」是指有特殊地位的年老女僕、或年紀大的奶媽兼女僕
- 《紅樓夢》中出現相當多的「嬤嬤」
  - 李嬤嬤、宋嬤嬤、賴嬤嬤、趙嬤嬤、王嬤嬤 等
- 在 71-80 回「老嬤嬤」更密集出現於 71, 74, 75, 77, 79, 80 等六個章回
- 由於寧國府在第 105 回被查抄前，府內應還有多位嬤嬤從事服務工作，因此在八十回後「嬤嬤」這個角色竟然徹底消失，顯得相當不協調

採礦所得雙字詞討論：取笑

- 「取笑」在前八十回出現於 30 個章回（45 次），後四十回則一次也沒有出現
- 前八十回共出現 3398 個「笑」字，而後四十回僅出現 585 個「笑」字
- 「取笑」是一種有意義的行為。這個詞的使用可能與敘事情節有關，但它（這種行為）在後四十回徹底消失，也透露出前八十回與後四十回在文筆上的不一致。

比較 C:01-20 與 D:21-80 的高分雙字詞

字詞	$f_t(C,D)$	$\|C_t\|$	$\|D_t\|$
秦氏	23.5	9	0
園裏	21.8	0	25
鳳丫	18.5	0	21
了秦	18.5	7	0
春道	17.7	0	20
寶姑	17.7	0	20
不留	17.7	0	20
竟沒	16.8	0	19
前頭	16.8	0	19
想得	16.8	0	19

- 秦氏在二十回前即已亡故
- 第 18 回元妃賜名後，才有「大觀園」的名稱
- 在二十回後，賈府長輩才開始頻繁用「鳳丫頭」提到「鳳姐」
- 「隨了秦氏」、「見了秦鐘」、「見了秦氏」等
- 迎春、探春、惜春在二十回後才開始活躍
- 丫鬟們稱寶釵為「寶姑娘」

中間六十回 (21-80) 裏，平均每三回就有一回出現「不留」。

- 雖有許多字詞在頻率上呈現明顯差異，但多半與人地名相關，且可用敘事情節演進來解釋。
- 整體而言，其字頻差異『遠小於』01-80, 81-120 回的差異

小結

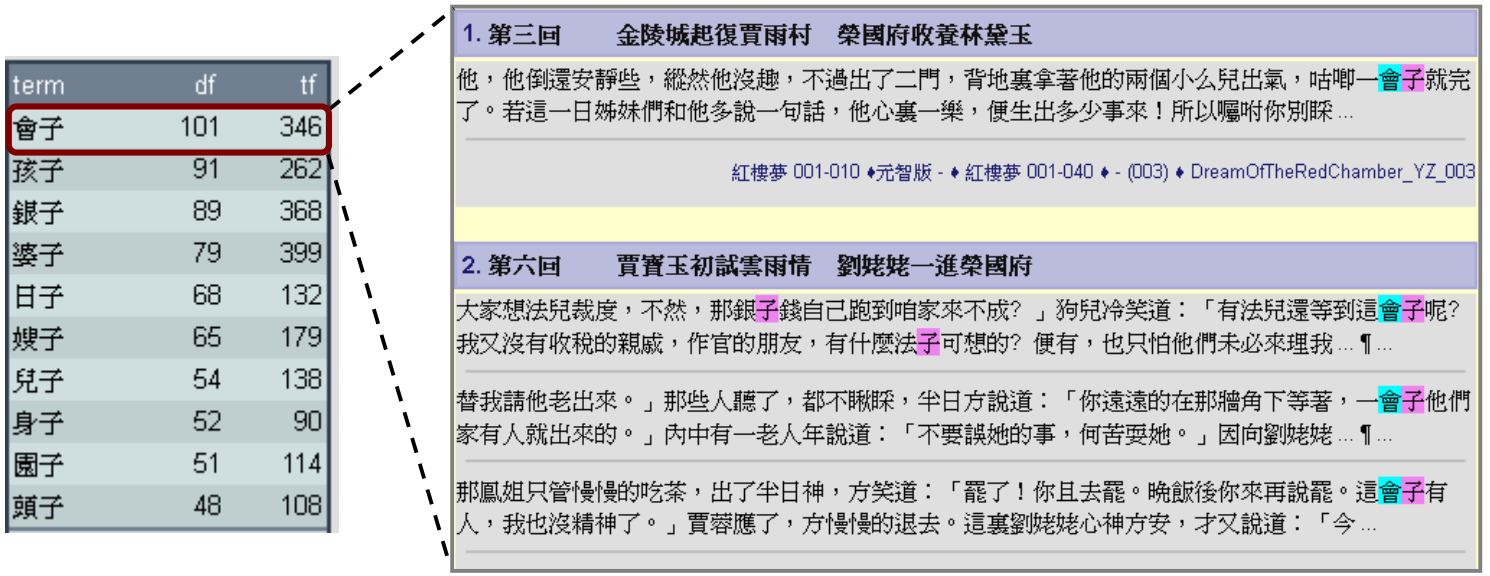
- 我們利用文本採礦函數
  - 系統化地找出許多「在前八十回與後四十回出現篇數有明顯差異」的單字與雙字詞
- 這些證據就足夠了嗎？
  - 我們還希望能多找出一些「在前二十回、中間六十回的出現頻率較為一致，但與後四十回的出現頻率則有明顯差異」的用詞

虛字頻率的方法，有時仍不充分

- 假設虛字「了」在 A, B 各出現 100 次
  - 在 A 通通都是以「看了」的形式出現
  - 在 B 通通都是以「聽了」的形式出現
- 則「了」在 A 與 B 的字頻當然沒有差異
- 但 A, B 在「看了」與「聽了」的頻率差異很大，表示它們在文筆運用上還是非常不同的
  - A 是利用視覺來傳遞訊息，B 則是用聽覺

前後綴詞工具

- 可觀察在一群文件中，某指定字詞的前後會接上哪些字
- 例如，可觀察在《紅樓夢》中，最常出現在「子」前面的是哪些字



term	df	tf
會子	101	346
孩子	91	262
銀子	89	368
婆子	79	399
日子	68	132
嫂子	65	179
兒子	54	138
身子	52	90
園子	51	114
頭子	48	108

1. 第三回　　金陵城起復賈雨村　榮國府收養林黛玉

他，他倒還安靜些，縱然他沒趣，不過出了二門，背地裏拿著他的兩個小么兒出氣，咕唧一會子就完了。若這一日姊妹們和他多說一句話，他心裏一樂，便生出多少事來！所以囑咐你別睬 ...

紅樓夢 001-010 ♦元智版 - ♦ 紅樓夢 001-040 ♦ - (003) ♦ DreamOfTheRedChamber_YZ_003

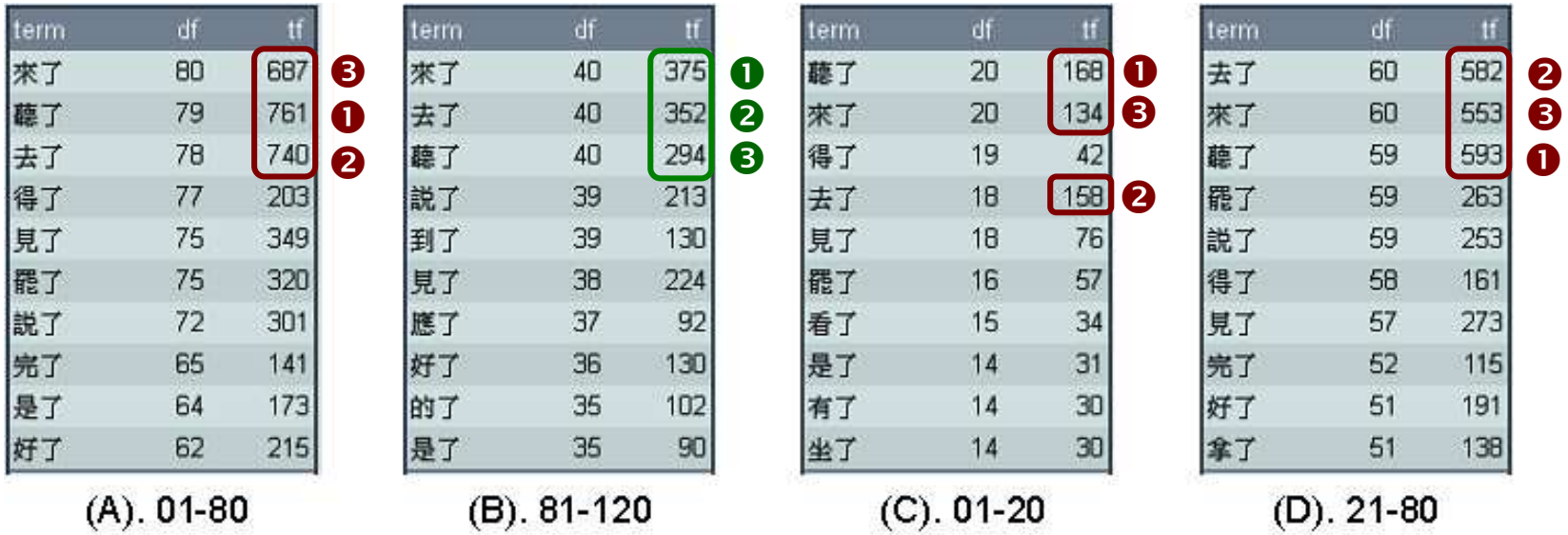
2. 第六回　　賈寶玉初試雲雨情　劉姥姥一進榮國府

大家想法兒裁度，不然，那銀子錢自己跑到咱家來不成？」狗兒冷笑道：「有法兒還等到這會子呢？我又沒有收稅的親戚，作官的朋友，有什麼法子可想的？便有，也只怕他們未必來理我 ... ¶ ...

替我請他老出來。」那些人聽了，都不瞅睬，半日方說道：「你遠遠的在那牆角下等著，一會子他們家有人就出來的。」內中有一老人年說道：「不要誤她的事，何苦耍她。」因向劉姥姥 ... ¶ ...

那鳳姐只管慢慢的吃茶，出了半日神，方笑道：「罷了！你且去罷。晚飯後你來再說罷。這會子有人，我也沒精神了。」賈蓉應了，方慢慢的退去。這裏劉姥姥心神方安，才又說道：「今 ...

「了」的前綴一字分析



term	df	tf
來了	80	687 ❸
聽了	79	761 ❶
去了	78	740 ❷
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

(A). 01-80

term	df	tf
來了	40	375 ❶
去了	40	352 ❷
聽了	40	294 ❸
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

(B). 81-120

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

(C). 01-20

term	df	tf
去了	60	582 ❷
來了	60	553 ❸
聽了	59	593 ❶
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

(D). 21-80

- 在 01-20、21-80、 01-80 回裏，詞頻最高的前三名都依序為「聽了、去了、來了」
- 在 81-120 回裏，詞頻最高三者的順序則相反，依序為「來了、去了、聽了」

笑道、嘆道、忙道、問道

	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

- 以十回為一個單位，列出「笑道」、「嘆道」、「忙道」、「問道」在每個單位的詞頻。
- 很明顯地，「笑道、嘆道、忙道」在前八十回的出現頻率遠較後四十回為高。
- 另一方面，「問道」在後四十回（尤其是 81-90 回）的詞頻則較高。

「豈」字用法：詞頻

term	df	tf
豈不	80	190
豈有	33	42
豈知	24	56
豈敢	7	8
豈可	7	7
豈是	6	6
豈肯	6	6
豈能	3	3
豈無	3	3

term	df	tf
豈非	3	3
豈許	2	2
豈容	2	2
豈止	2	2
豈只	2	2
豈得	2	2
豈獨	2	2
豈從	1	1
豈終	1	1

term	df	tf
豈煩	1	1
豈天	1	1
豈奈	1	1
豈全	1	1
豈招	1	1
豈神	1	1
豈意	1	1
豈遂	1	1
豈似	1	1

term	df	tf
豈在	1	1
豈人	1	1
豈令	1	1
豈係	1	1
豈畏	1	1
豈必	1	1

term t	01-20	21-80	01-80	81-120
豈	68	203	271	> 83
豈不	35	140	175	≫ 15
豈有	14	24	38	≫ 4
豈知	0	0	0	≪ 56

- 有 81.3% 的「豈」字被用於「豈不、豈有、豈知」
- 「豈不、豈有、豈知」在 01-20, 21-80, 01-80 的字頻都很接近，但卻與 81-120 的字頻相差甚多

⇨「豈」字是個文筆差異的好例證：據常理一位作者不可能在一本書的前後，對一個字有如此不一致的使用方式！

檢視所有「...不兩...」的詞句

3	已擇了出月初二日，小女入都，尊兄即同路而往，豈不兩便?
3	況這林妹妹眉尖若蹙，用取這兩個字，豈不兩妙！
17	待貴妃游幸時，再請定名，豈不兩全?
33	若十分愛慕，老大爺竟密題一本請旨，豈不兩便?
43	這已完了心願，趕著進城，大家放心，豈不兩盡其道。
69	若天見憐，使我好了，豈不兩全?
72	不如借了來，奶奶拿一二百銀子，豈不兩全其美。
78	設若從那裏生出一件事來，豈不兩礙臉面。
78	姐姐何不等一等他回來見一面，豈不兩完心願?
98	咱們一心一意的調治寶玉，可不兩全?

- 很明顯地，在八十回之前「不兩」的前綴字必然為「豈」
- 有趣的是，第 98 回所出現的「可不兩全」，其意義與八十回前曾出現的「豈不兩全」是一樣的 。
- 因此，幾可確定第 98 回的「可不兩全」出自後人之手

總結：與前人方法的比較

- 前人的統計方法
  - 先經人工先篩選量化的標的（虛詞），然後對這些標的的分佈進行統計檢定
- 本實驗所採的方法：
  - 文本採礦 (text mining)：利用電腦計算出「可能有趣的候選字詞
  - 前後綴詞系統：觀察高頻字的前後會接上哪些字，也可發現許多具頻率差異的有趣字詞

結論

- 我們的研究焦點：
  - 特別關心「沒有電腦則很難進行」的人文研究
  - 雖運用數位科技，但成果應以人文角度來檢驗
- 我們的實驗：
  - 說明採礦函數與綴詞系統，都是有效的工具
  - 發現許多字詞（有些是具意義的角色或行為）在前八十回與後四十回的使用上有顯著差異
  - 這些發現支持：《紅樓夢》的後四十回作者並非曹雪芹

補記

威斯康辛經濟系榮譽教授趙岡（左，-20211023）與夫人陳鍾毅女士