巨量資料與統計分析

政治大學統計系余清祥

2024年11月26日

第十一週：文字資料EDA

http://csyue.nccu.edu.tw

文字分析的範例

- 《哈利波特》作者 J.K.羅琳在2013年化名 Robert Galbraith 創作偵探小說《Cuckoo》，英國研究者透過**語意分析**技術，比對羅琳以前的寫作文本，發現寫作手法極為接近。

→金庸《天龍八部》（阿紫瞎眼）、曹雪芹與《紅樓夢》、博士論文的真偽。

- 文字分析的應用愈發多元

→**情感分析**、**關鍵片語擷取**、**語言偵測**、**實體辨識**。（註：Microsoft Azure文字分析 API）

文字分析文獻

- 趙岡與陳鍾毅(1980)

→使用統計分析判定《紅樓夢》作者用字習慣(前後各抽一百頁，每頁720字，小說全文有738,024字 ；2%抽樣)

→考慮五個虛字的使用習慣。

	前80回	後40回	t-檢定值
兒	2.60 (2.36)	3.99 (2.95)	3.677
在	2.98 (2.08)	3.95 (1.96)	3.392
了	7.62 (5.54)	17.71 (5.48)	0.166
的	12.25 (5.00)	14.81 (5.66)	3.391
著	4.83 (3.02)	6.57 (3.27)	3.910

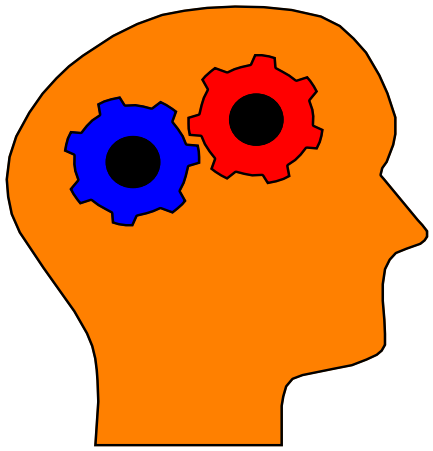
文字分析文獻（續）

(2) Mosteller and Wallace (1984)

→運用貝氏分析，探討擁護聯邦主義的論文(The Federalist Papers)作者。

→77篇中有12篇文章沒有定論

(可能是Hamilton或Madison所作)



文字分析文獻（續）

(3) Efron and Thisted (1976)

→估計莎士比亞(Shakespeare)的字彙總數及使用頻率

→用Poisson Process估計字彙

→在1987年推論1985年發現的一首詩，是莎士比亞所作

估計的範例—字彙總數

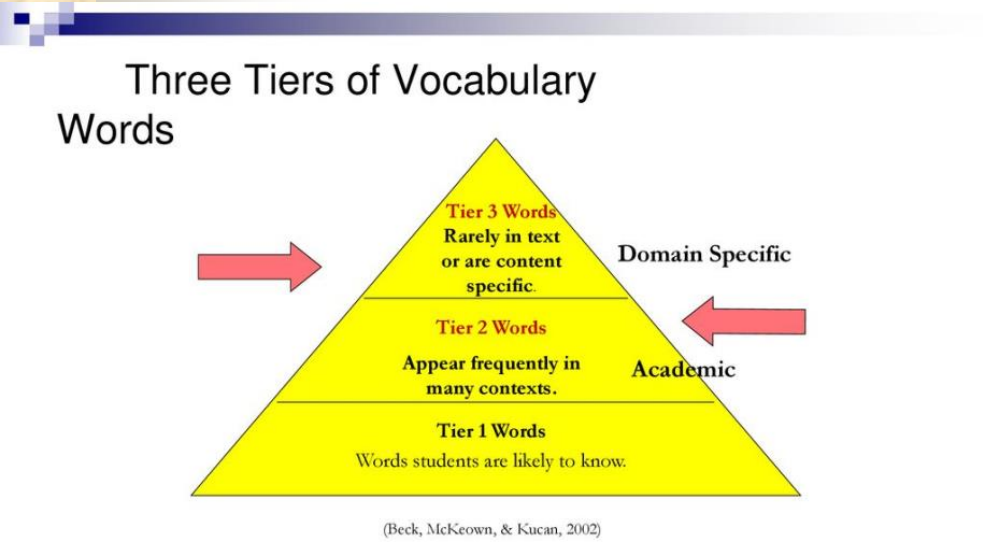
- 平常用的字彙數其實不多，認識四千多個中文（英文）單字大約足夠。

→問題：如何估計一個人的總字彙數？

- 知名範例：「莎士比亞知道多少單字」

→參考Efron and Thisted (1976 & 1987)

→Alan Turing

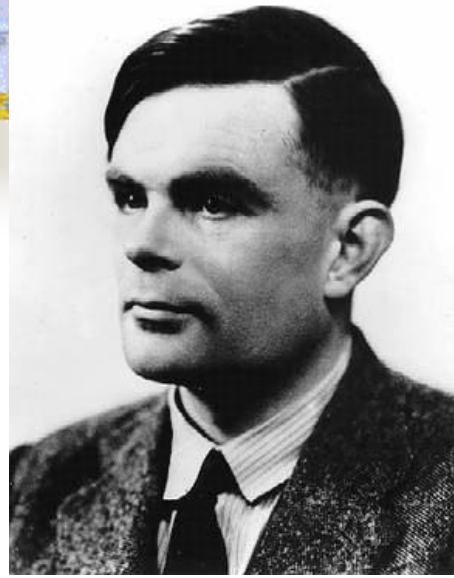


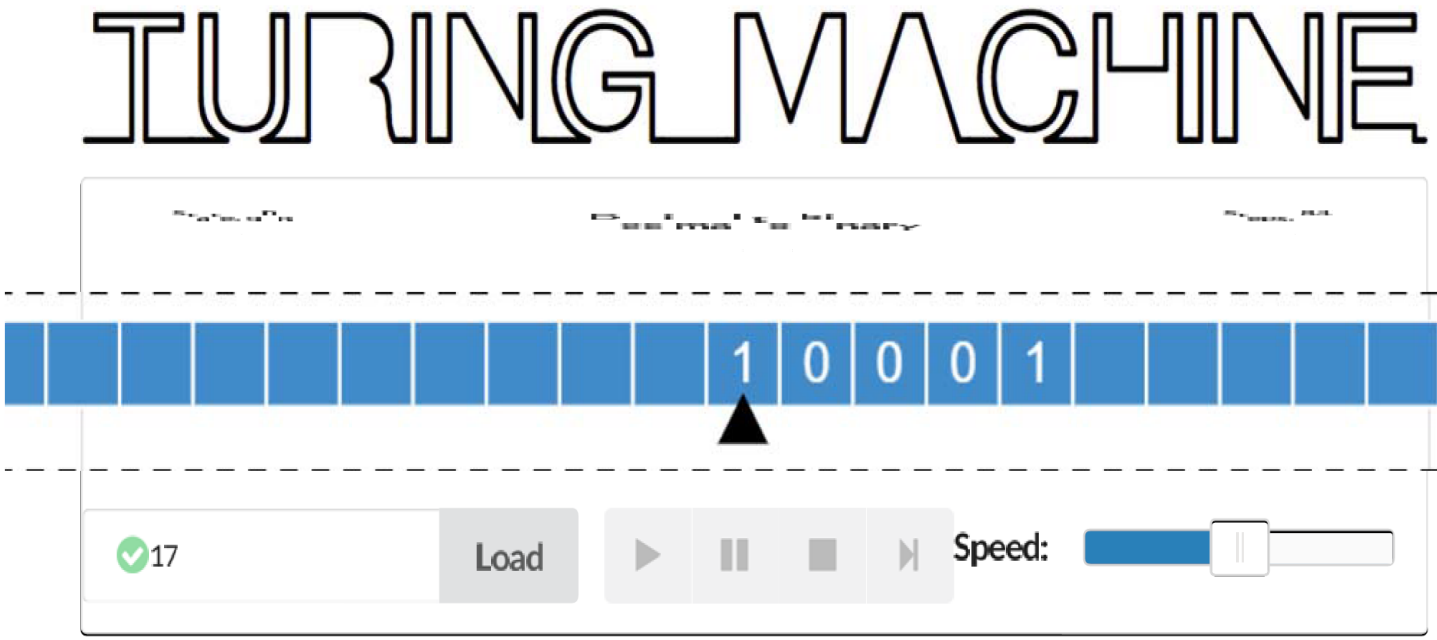
(Beck, McKeown, & Kucan, 2002)

https://ca-times.brightspotcdn.com/dims4/default/342d9a2/2147483647/strip/true/crop/2048x1363+0+0/resize/840x559!/quality/90/?url=https%3A%2F%2Fcalifornia-times-brightspot.s3.amazonaws.com%2Fc6%2Fd5%2F73e00db68cfa682f0de2df2d2545%2Fla-et-mn-the-imitation-game-movie-reviews-crit-001

https://slideplayer.com/slide/13963905/86/images/29/Three+Tiers+of+Vocabulary+Words.jpg

Turing Machine



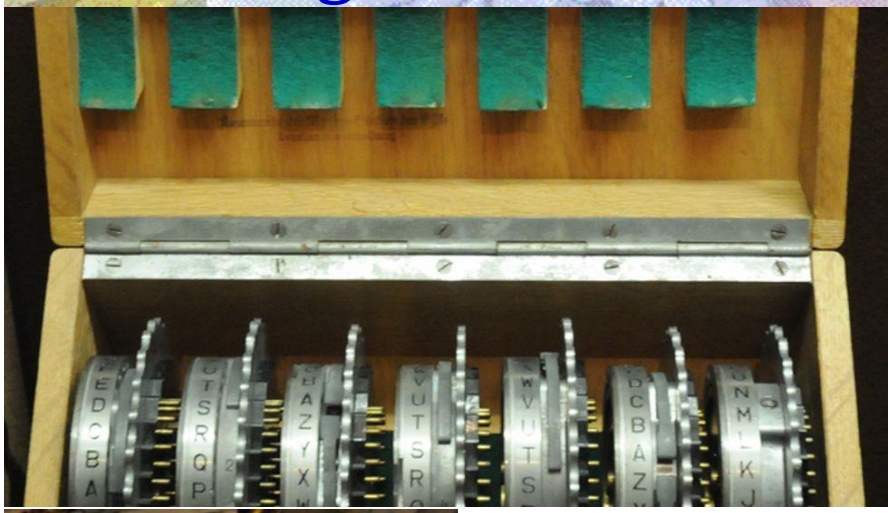


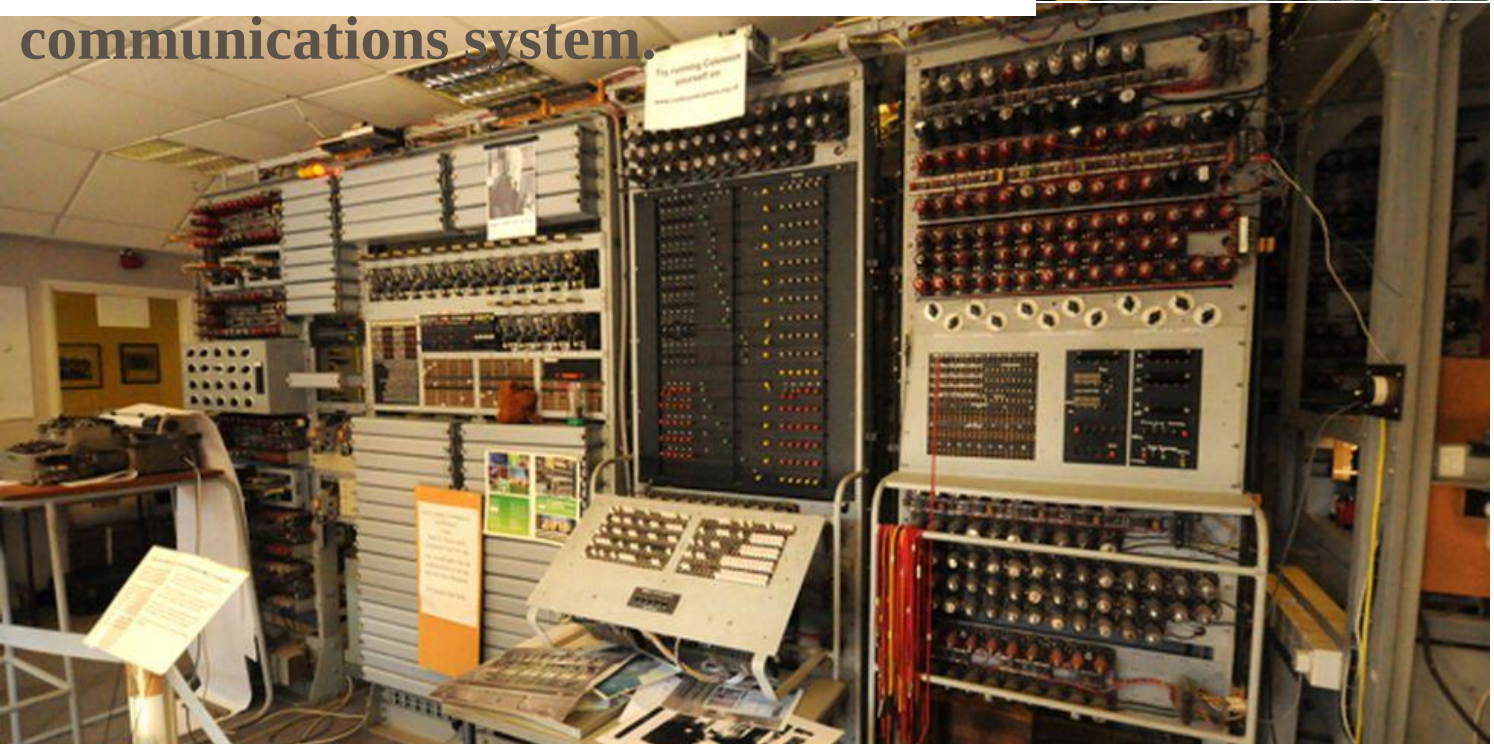
Source: https://turingmachinesimulator.com/

註：誰是Alan Turing、對電腦發展有什麼貢獻？

This is a rebuild of the famous Colossus Mark 2 machine that finally allowed the code breakers to quickly and efficiently break the high command's ciphers. For decades, since 1918, the Germans had been using Enigma cyphers as the core of their intelligence and military communications system.

Enigma rotors





https://cnet3.cbsistatic.com/img/f2L_cPTDlz5TH0OwdN5ZK_WX9DI=/980x551/2010/08/05/7e19b914-f0ed-11e2-8c7c-d4ae52e62bcc/Operational_rotors.jpg

https://cnet3.cbsistatic.com/img/YQ0CRrVg9Qz530e7SNDJux7MsN8=/980x551/2011/07/03/7e31eb6c-f0ed-11e2-8c7c-d4ae52e62bcc/Colossus_body_close-up.jpg

中文資料的分析策略

- 中文詞彙是由一個個的方塊字構成，更接近隨機樣本的概念，因此分析單位可定為單一字詞，比較單字的統計特性。
- 現代中文書寫大多為白話文，有別於20世紀之前使用的文言文。

→文言文、白話文特性差異頗大，分析策略也非常不同，舉凡標點符號、虛字都不一樣。以下範例以白話文為主。

文字資料的前置分析

- 處理文字資料時，統計分析包含以下步驟：
- Data Collection
- Text Parsing and Transformation

→摘錄、清理、由NLP定義變數等，包括斷句、篩選相關資料段落、定義關鍵字詞。

- Text Filtering

→挑選合適關鍵字詞。

- Text Mining

→Clustering, Classification, Association, and Link Analysis.

斷詞、關鍵詞與統計分析

- 關鍵詞(Keywords)猶如統計模型的變數，根據研究目的及問題定義，挑選適當變數以提分析的效率和準確性。

→傳統統計模型也注重解釋性。

- 中文分析先經過斷詞，接著再從中挑選出重要關鍵詞。

→白話文多以雙字詞（多字詞）表達觀念。

（註：字→詞→有意義的詞）

文言文、白話文的比較（維基百科）

比較	文言文	白話文/現代中國語文
長短	言簡意賅	較長篇
出處用法	書面語為主	「我手寫我口」為主，亦經修飾
語感	古雅精煉	通俗易明
文法詞組次序	彈性較大	次序明確
用詞1	單字已有獨立意思	二字詞為主
用詞2	一字多用	異字異用
用詞3	之	的
句末助語詞	已、矣、乎、也.....	了、吧、啊、嗎.....
標點	標點少而簡，句讀為主	標點繁多
經典例	《桃花源記》、《醉翁亭記》、《庖丁解牛》、《出師表》、《六國論》.....	魯迅《吶喊》自序、朱自清《綠》、冰心《紙船》、舒乙《香港：最貴的一棵樹》.....
流傳	限於曾學習文言的人，須有一定傳統文學修養，但可於東亞通行	一般中小學生也能看懂，廣傳於華文世界
習法	背誦為主，輔以字詞拆解	字詞拆解為主，文法分析輔助

中文的斷詞

- 白話文使用的字彙相對較少，闡述語意時多以雙字詞、或多字詞的方式出現，因此斷詞（或是找出有意義的關鍵詞）非常重要。
- 可透過事件及機率描述字與字間的關連(e.g., 雙字詞等關鍵詞)，例如：若「國家」是關鍵詞，兩者一起出現的機率較大（不獨立！）：

$$P(\text{「國家」}) >> P(\text{「國」}) \times P(\text{「家」})$$

註：兩單字若獨立，則

$$P(\text{「國家」}) \cong P(\text{「國」}) \times P(\text{「家」})$$

中文斷詞軟體

- 斷詞是自動化中文語意分析的關鍵步驟，類似統計分析的解釋變數、被解釋變數。

→斷詞後還需確認是否為關鍵詞（顯著變數！）

- 中文斷詞軟體不少，較受歡迎者包括中研院開發的斷詞系統(CKIP)、結巴(Jieba)，兩者的準確性都很高。

→大多數斷詞軟體仰賴訓練樣本（類似詞庫），如果目標文本與詞庫特性不同，斷詞結果通常不佳。（如：文言文、白話文）

幾個常見的中文斷詞系統

- 坊間有不少中文斷詞系統，僅舉出下列四種為代表：

- Jieba
- 中研院斷詞（舊版、CKIP）
- Stanford Word Segmentor
- Rwordseg

→系統的特色不同，考量可用於長篇的文本及執行效率，選擇結巴斷詞。

模型分析

- 透過數理模型描述觀察結果：

觀察現象 = 模型 + 誤差

或是

$y = f(x) + \text{error}$；觀察值 = 訊號 + 雜訊。

- 數量化模型的關鍵：

→量化目標值 y：定義問題！

→選取關鍵變數： $x_1, x_2, \ldots, x_p$

→建立量化模型：統計學習、機器學習。

關鍵詞篩選(Keyword Extraction)

- 關鍵詞如同迴歸分析中的顯著變數，便於詮釋、進一步探索。

→若套用機器學習模型，顯著變數未必是關鍵資訊。（註：不見得能分離出關鍵變數！）

- 使用關鍵詞可更有效率，加快運算速度。

→相對於中文斷詞，至今仍無有說服力的關鍵詞篩選軟體。

→關鍵詞與研究主體有關，學者也存有不少歧見，很難找到「標準答案」。

https://www.statistika.co/images/services/Exploratory%20Data%20Analysis%20-%20EDA%201000x468.jpg

單字（詞彙）的探索性分析

- 單字、詞彙的基本特性整理，包括單字及詞彙種類（即字彙數）及其出現機率（多項分配），以目視(EDA)觀察出基本特性與差異，做為導引後續分析的參考。
- EDA著重於描述文字的使用特性，如果考量的是二元分類等CDA估計，可以把所有單字/詞彙列為解釋變數（不考慮個別解釋！）。

→只考慮分類準確率，可能會因為變數太多，需考量資料縮減/維度縮減。

單字特性的其他測量方式

- 用於EDA的敘述性統計量也可用於描述單字、詞彙的特性，像是百分位數、累積機率分配(CDF)、PDF等。

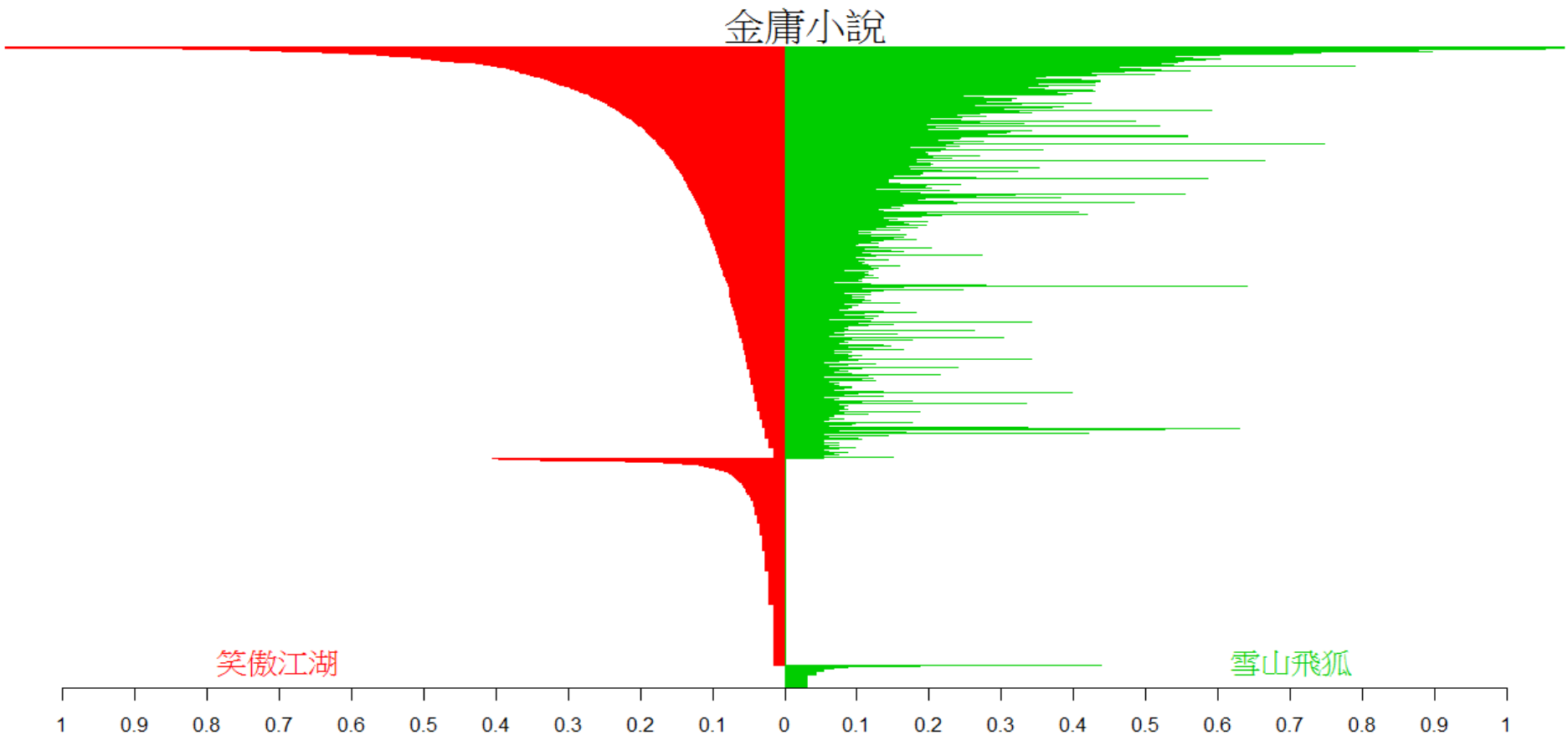
→兩兩樣本的比較可參考相關係數。

- 字彙豐富度(Richness)是另一種描述方式，常見的統計量為Type-Token Ratio (TTR)，也就是需要多少字才會出現一個新字彙。

→常用字彙大致固定（中文約四千多個字彙），TTR會隨著總字數增加而下降（標準化？）。

物種分佈比例（或族群結構）

		母體數（總字數）	物種種類數(不同字彙數)	共有物種數(共有字彙數)
金庸小說	笑傲江湖	420546	3690	2457
	雪山飛狐	106628	2591	



單字（詞彙）的探索性分析（續）

- 除了分配函數外，或可引進其他領域的測量值作為比較標準，例如：不均度（或物種豐富度、生物多樣性等），可用於描述單一母體的特性、或是比較兩個母體間的關連。

→吉尼係數、熵(Entropy) = $-\sum_i p_i \log(p_i)$ 等。

註：生物（醫學）、經濟相關指標。

- 單從資料驅動找不出頭緒，參考專家意見也是可行作法。（註：機器學習/人工智慧仍有不足之處！）

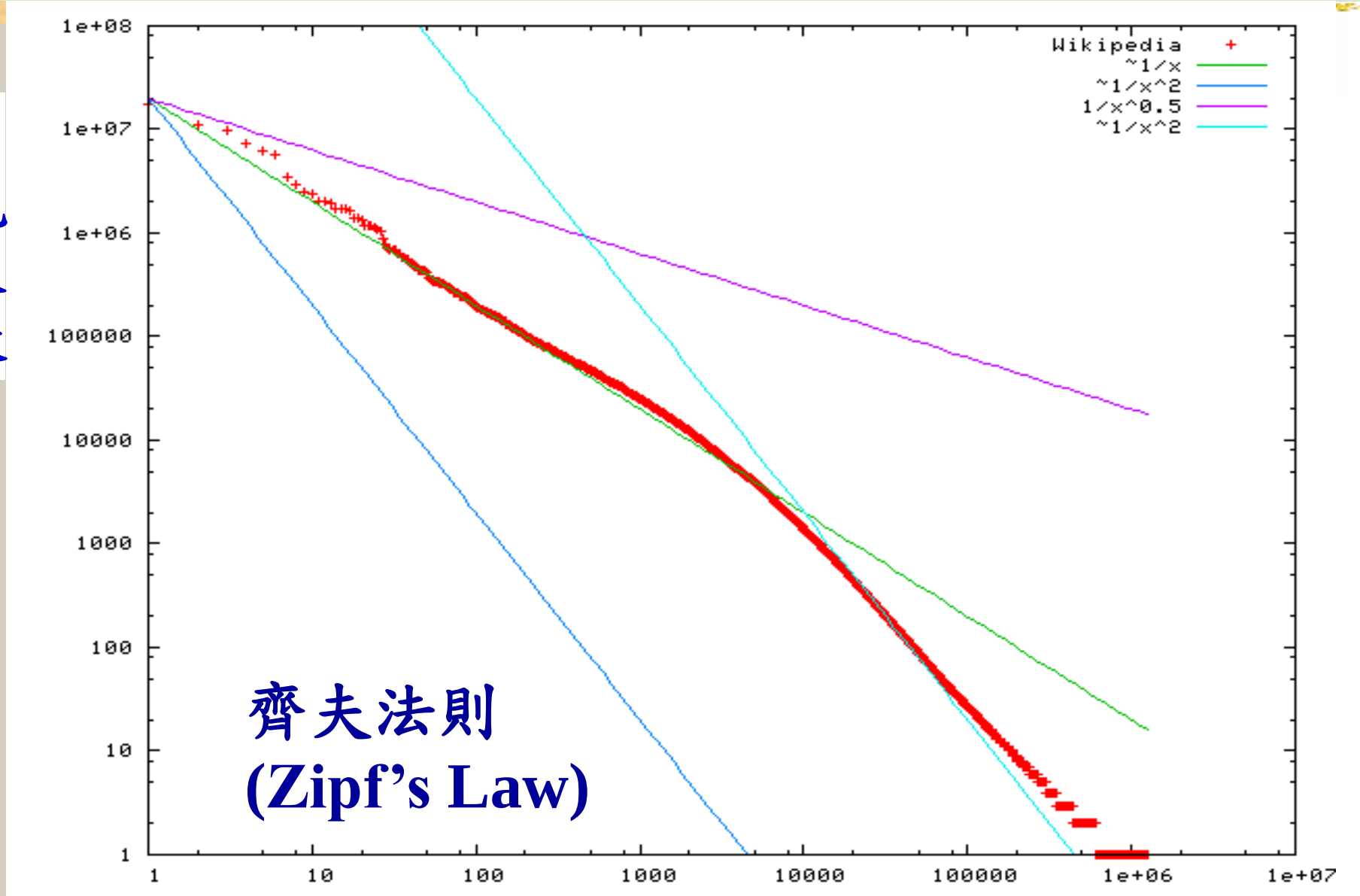
圖表與模型分析

- 可除了EDA、兩樣本間的比較外，也可透過模型描述字彙的出現特性，例如：齊夫法則(Zipf's Law)是知名的模型，也是量化語言學(Linguistics)的常見研究工具。

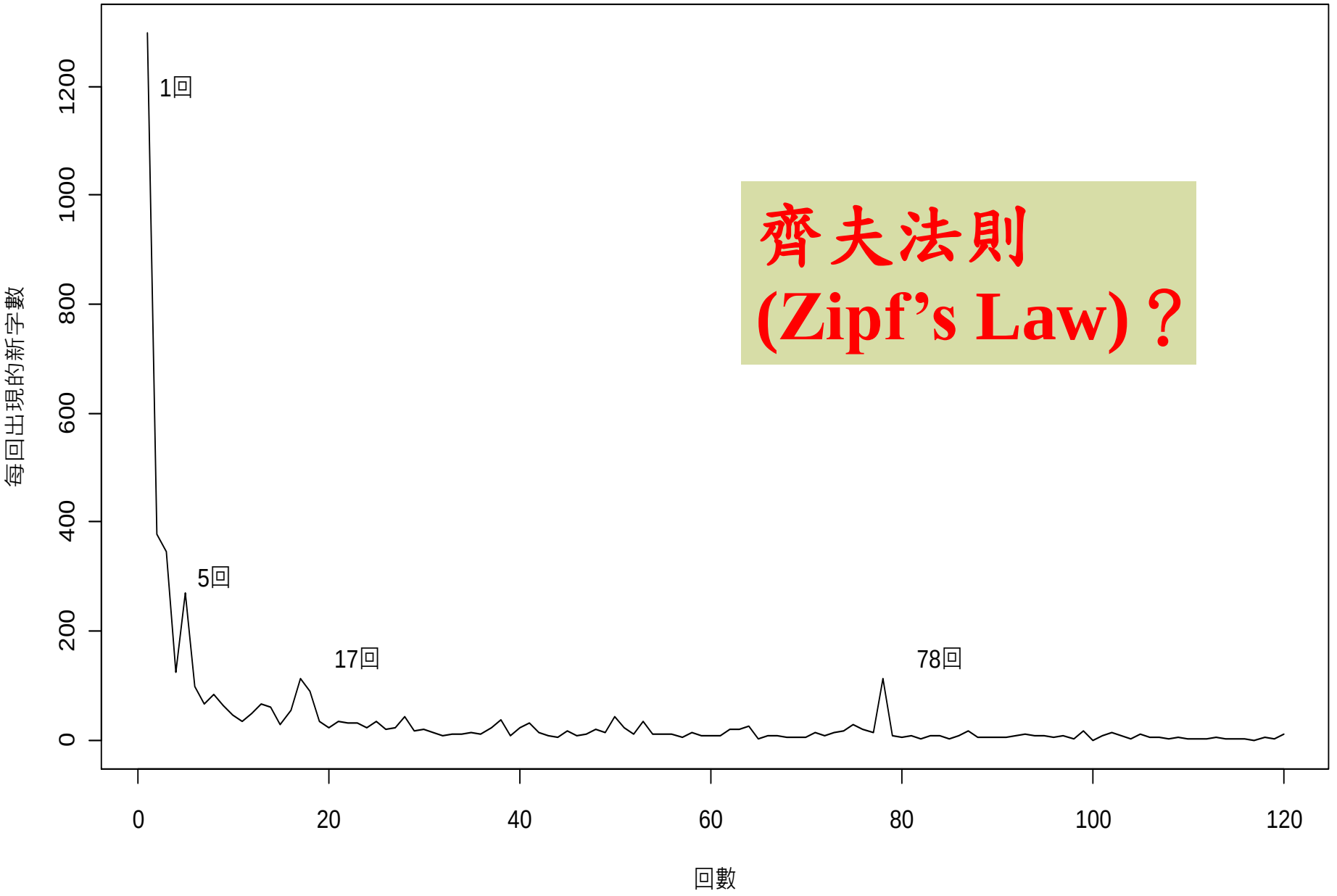
→透過圖表及模型的整理結果，可結合統計思維與方法（例如：離群值），進一步整理出文字背後的資訊。

註：上述分析尚未加入字詞文意及詞性、句子結構及長度，單純將每個字彙視為不同類別。

維基百科的字數統計(November 27, 2006)



紅樓夢各回新出現字彙的個數



齊夫法則
(Zipf's Law)？

中央研究院「現代漢語平衡語料庫」

中文斷詞系統

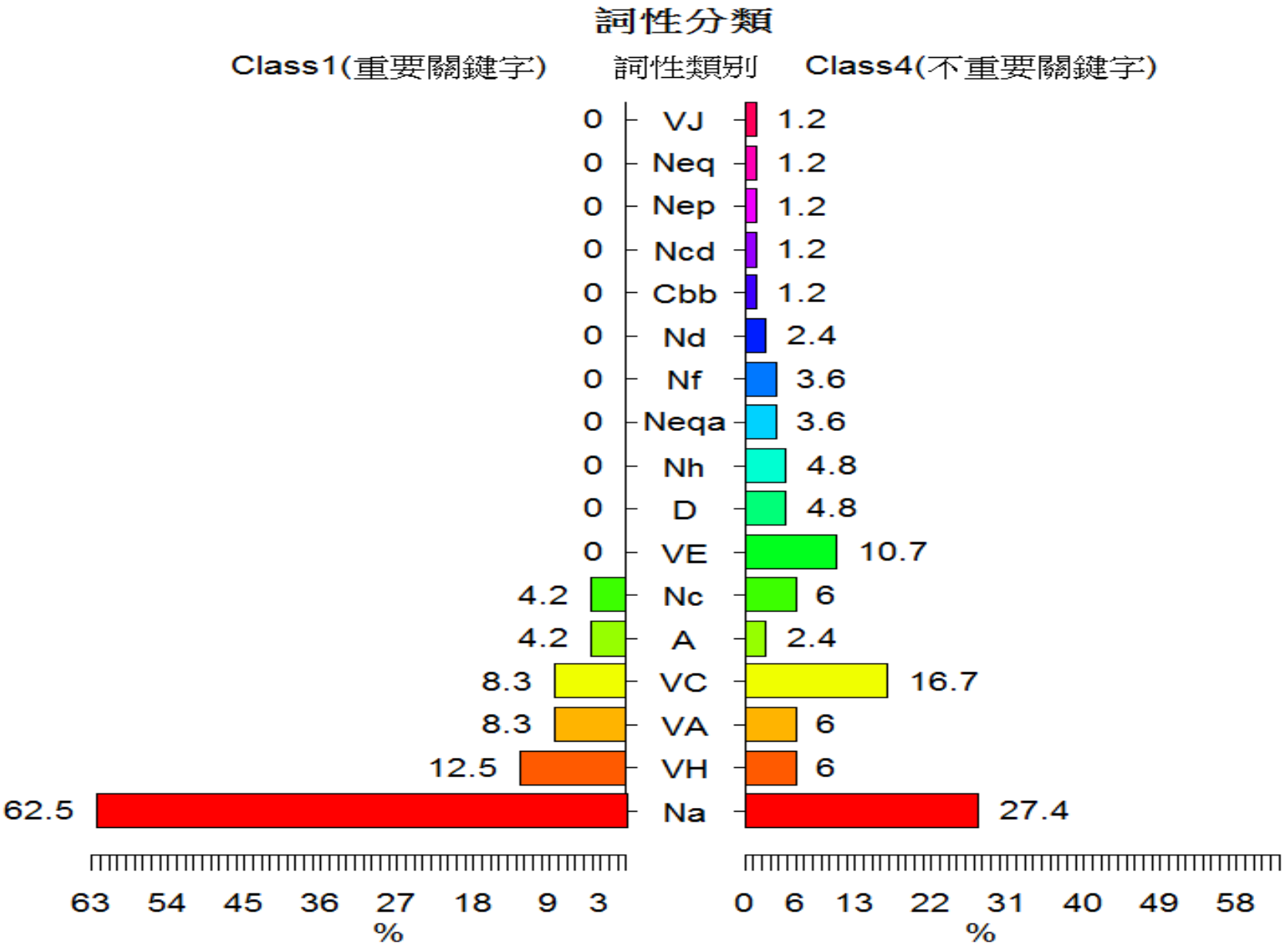
- 簡介
- 未知詞擷取做法
- 詞類標記列表
- 線上展示
- 線上服務申請
- 線上資源
- 公告
- 聯絡我們

線上展示使用簡化詞類進行斷詞標記，僅供參考並且系統不再進行更新。線上服務斷詞和授權mirror site僅提供精簡詞類，結果也與舊版的展示系統不同。

自 2014/01/06 起，本斷詞系統已經處理過 1848571 篇文章

送出 清除

由於中文詞集是一個開放集合，不存在任何一個詞典或方法可以盡列所有的中文詞。當處理不同領域的文件時，領域相關的特殊詞彙或專有名詞，常常造成分詞系統因為參考詞彙的不足而產生錯誤的切分。為了解決這個問題，最有效的方法是補充領域詞典加強詞彙的搜集。因此新的詞彙或關鍵詞的自動抽取成為分詞的先期準備步驟。領域關鍵詞彙多出現在該領域的文件中而少出現在其它領域，因此抽取關鍵詞時多利用此特性。



註：Class1→重要關鍵字； Class 4→非重要關鍵字

結巴斷詞與TFIDF

- 結巴(Jieba)是最近幾年最常用的斷詞演算法，據說是以《人民日報》為訓練資料。

→運用自然語言處理去針對文字資料進行斷詞，R及Python皆有支援套件。

- TFIDF (Term Frequency–Inverse Document Frequency)為文字探勘的常用方法，可用於決定關鍵詞。

→根據文本特性、個人經驗等，使用者挑選篩選門檻。

TFIDF的想法及特性

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

- TFIDF為採納統計思維的文字探勘方法，分為兩個步驟分析：TF、TDF。

→TF（詞頻）：類似「期望值」的概念，出現次數愈多、是關鍵詞機會愈高（至少不應太少）。

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

→IDF（逆向檔案頻率）：類似「變異數」，某個詞彙出現愈有規律（前後一致），可能是常用而非關鍵詞彙的機會愈大。

$$\mathrm{idf}_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$$

噁心餿水油 683噸銷全台

自由時報 自由時報－2014年9月5日 上午6:10

味全中鏢 12產品下架

〔本報記者／全台連線報導〕國內再爆黑心餿水油風暴，流入小吃、烘焙業的劣質食用油超過六百八十三公噸，食品大廠味全也中彈！

刑事局南部打擊犯罪中心昨逮捕涉案的郭姓業者等五男一女，南打偵三隊長李宏倫表示，主嫌郭烈成（卅二歲）在屏東縣竹田鄉經營地下油行，去年初開始從胡姓環保回收油業者處蒐購餿水油及皮革廢油等，過濾煮摻加工，降低油品中的「酸價值」到二．五以下，符合食用油檢測標準後，蒙混成豬油轉售給高雄強冠公司。

主嫌郭烈成坦承犯行。據了解，他以每公斤十二元買進餿水油，再以每公斤約廿二元到卅元轉售強冠，一年下來獲利達四百多萬元。

- 核心概念是什麼？（關鍵詞？）

→標題、大綱（第一段）的關連？

標點符號與斷句

- 中文標點符號在1919年五四運動後大致確定，古文的斷句會因解讀而產生差異，研究文言文時需特別謹慎。

→Q：白話文的斷句可依據哪些標點符號？

→現代網路文章的標點使用也頗為雜亂！

- 白話文、文言文的主要差異包括雙字詞，若將關鍵詞視為比較依據，根據研究議題確認關鍵詞為兩字、三字或以上。例如：食安風暴是一個關鍵詞、或是兩個關鍵詞。

關連性分析

- 關連性分析可提供詞彙之間的關係，輔助辨別文章特性和文章間的關係。

→列連表、用有字詞的交集與聯集等；

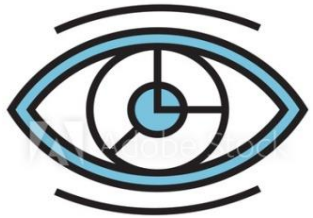
→相關係數、相似指數（若為連續資料）。

- 關連性分析可加上文意，或是單純評估字詞間的連結程度。

→文言文、白話文的虛詞(Function Words)；

→根據專家意見選擇重要關鍵詞，並計算這些詞彙間的關連，以及與研究目標間的關連。

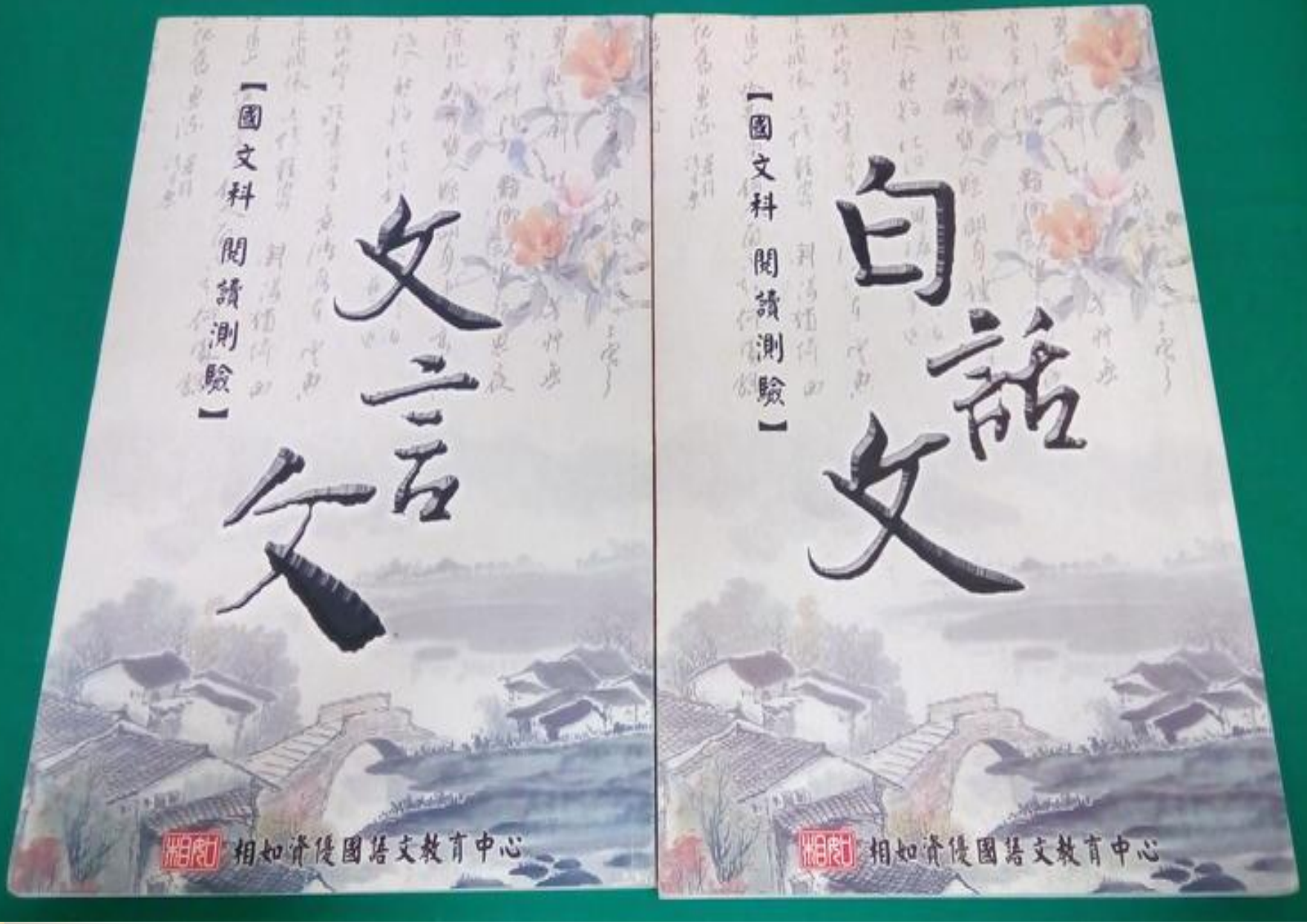
https://as1.ftcdn.net/jpg/02/12/48/84/500_F_212488495_u7CidAoTfSgxCCuliJ2rft28RhI7z9Uj.jpg



文字分析：觀念的視覺化

為什麼需要文學？了解文學、接近文學，對我們形成價值判斷有什麼關係？如果說，文學有一百種所謂「功能」而我必須選擇一種最重要的，我的答案是：德文有一個很精確的說法「macht sichtbar」，意思是「**使看不見的東西被看見**」。在我自己的體認中，這就是文學跟藝術最重要、最實質、最核心的一個作用。

——龍應台



https://img.ruten.com.tw/s2/4/a3/14/21912409931540_709.jpg

從《新青年》看文言文、白話文

- 問題：文言文的特色在於「**言簡意賅**」，如何將這個想法量化？

→圖表等量化工具可根據需求量身訂製。

- 文言文、白話文的比較盡量不牽涉內容（專家意見）判讀，可分為三個角度：

→方塊字(Word unit)

→標點符號(Punctuations)

→虛字(Function words)

新青年

《**新青年**》（法語：La Jeunesse）是新文化運動的核心雜志，後來成爲中國共產黨機關刊物。1915年9月陳獨秀在上海創辦，面嚮青年，宣揚民主與科學[1][2]。開始名爲《青年雜誌》，一年後改爲《新青年》，1917年搬到北京，1918年編輯部進一步擴大，包括錢玄同、劉半農、陶孟和、沈尹默、胡適、高一涵、魯迅、李大釗。這一時期中國的著名思想家、文學家們紛紛在該刊物上發表文章，《新青年》和北京大學成爲新文化運動的主要陣地。從1920年9月第8卷第1號開始，成爲中國共產黨機關刊物。1926年7月終刊，共出9卷54號。

方塊字分析與生物多樣性

大致可由以下三個角度切入：

- 豐富度(Richness)：

→總字（詞）數、字（詞）彙數、Type-Token-Ratio (相異字比例；TTR)。

- 不均度：

→字詞分布、前五十大或百大字詞的佔比、Simpson & Shannon Indices (Entropy)

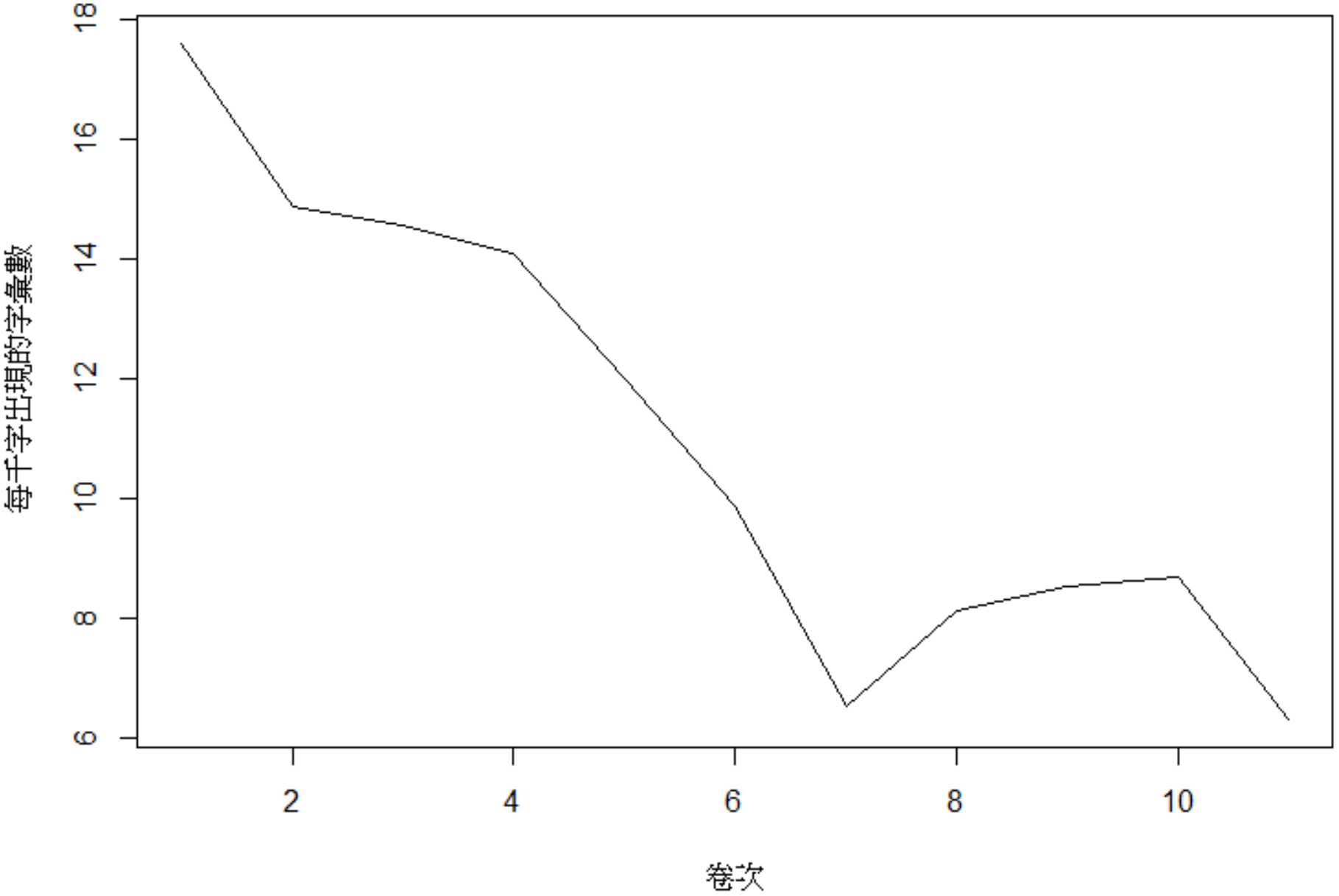
- 趨勢變化：

→共同字彙、新生及滅絕字彙

《新青年》前七卷字彙豐富度

	總字數	不同字數	前500字 字數比例	前500雙字詞 字數比例
第 1 卷	**261,905**	**4,403**	**195,584 (76.3%)**	**46,904 (18.3%)**
第 2 卷	298,279	4,344	223,940 (76.4%)	52,224 (17.8%)
第 3 卷	296,564	4,227	229,460 (78.8%)	58,604 (20.1%)
第 4 卷	309,914	4,298	243,288 (79.5%)	65,178 (21.3%)
第 5 卷	350,698	4,125	279,448 (81.0%)	88,136 (25.6%)
第 6 卷	397,012	3,848	324,976 (83.1%)	112,394 (28.7%)
第 7 卷	**596,713**	**3,850**	**493,201 (83.3%)**	**182,992 (30.9%)**

《新青年》各卷每千字出現的新字彙數



Standardized Type/Token Ratio

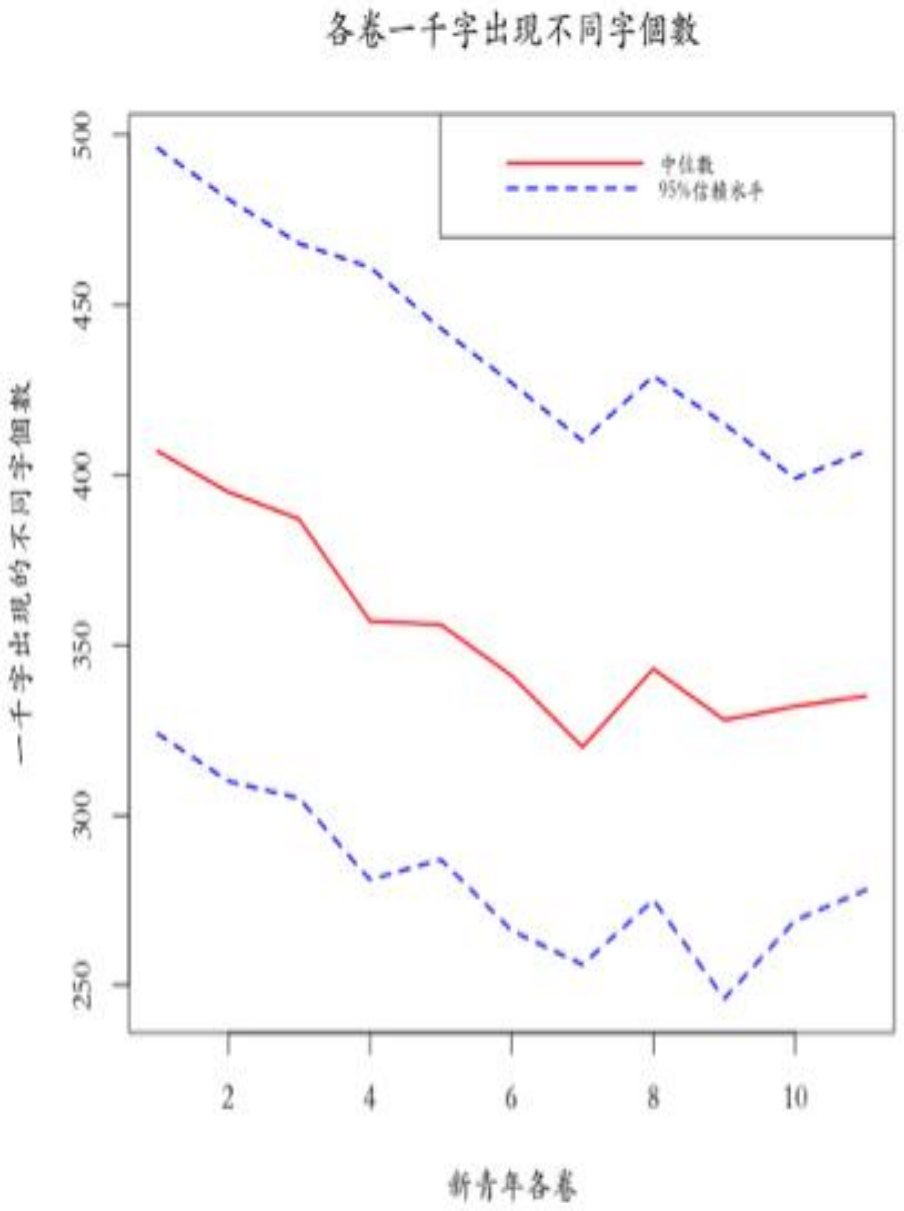
rank	word	freq	rank	word	freq	rank	word	freq	rank	word	freq
1	we	6	17	asleep	1	33	impressions	1	49	seem	1
2	and	5	18	at	1	34	instant	1	50	so	1
3	them	5	19	beliefs	1	35	is	1	51	the	1
4	are	3	20	brief	1	36	just	1	52	then	1
5	can	3	21	but	1	37	later	1	53	things	1
6	they	3	22	call	1	38	metaphor	1	54	thinking	1
7	to	3	23	coming	1	39	mull	1	55	this	1
8	again	2	24	concepts	1	40	notions	1	56	thoughts	1
9	as	2	25	described	1	41	occasions	1	57	times	1
10	in	2	26	endure	1	42	opinions	1	58	values	1
11	on	2	27	fall	1	43	other	1	59	variously	1
12	a	1	28	going	1	44	over	1	60	views	1
13	act	1	29	handle	1	45	perceptions	1	61	well	1
14	all	1	30	have	1	46	put	1	62	what	1
15	an	1	31	however	1	47	refer	1	**TOTAL**		**87**
16	aside	1	32	ideas	1	48	return	1			

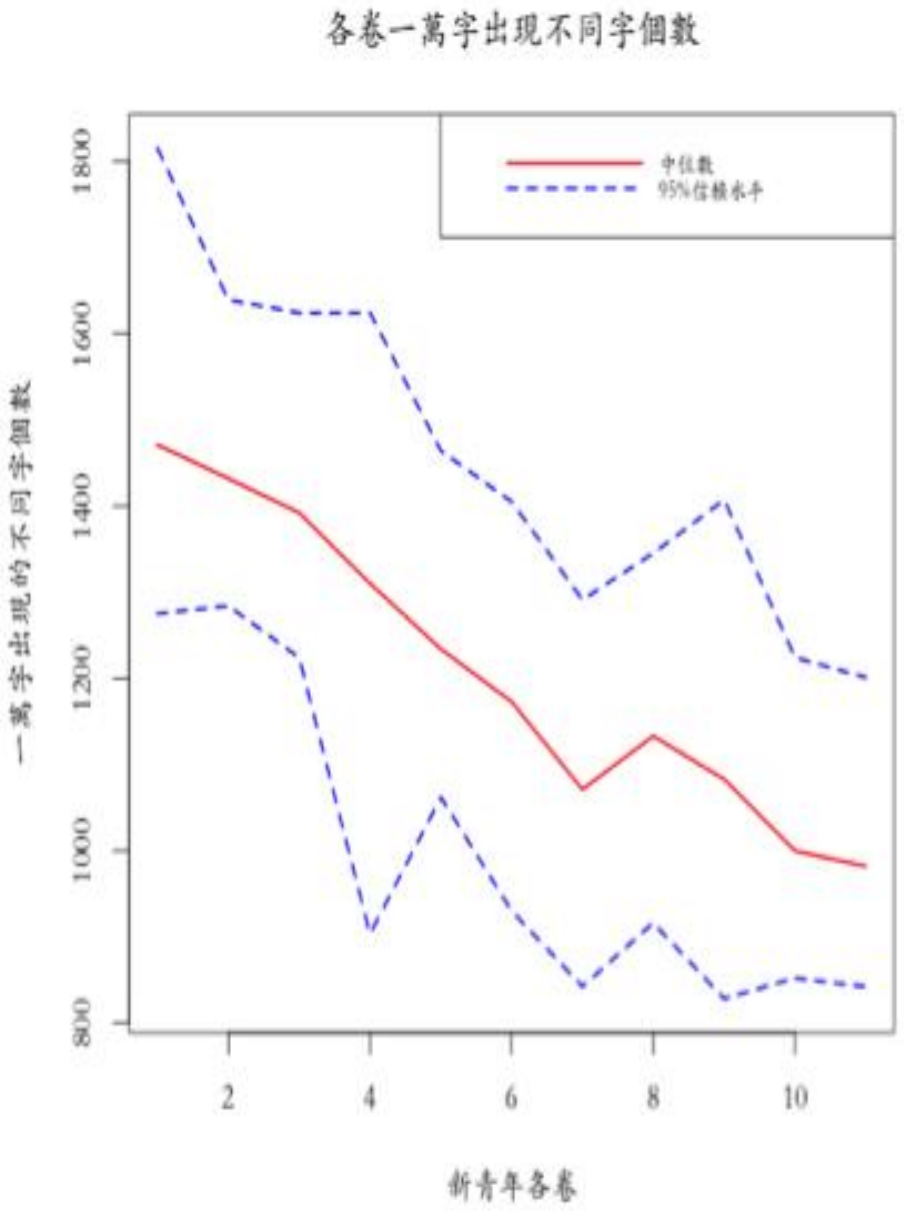
We see, then, that of the total of 87 tokens in this text there are 62 so-called *types*. The relationship between the number of types and the number of tokens is known as the *type-token ratio (TTR)*. For Text 1 above we can now calculate this as follows:

type-token ratio = (number of types/number of tokens) * 100

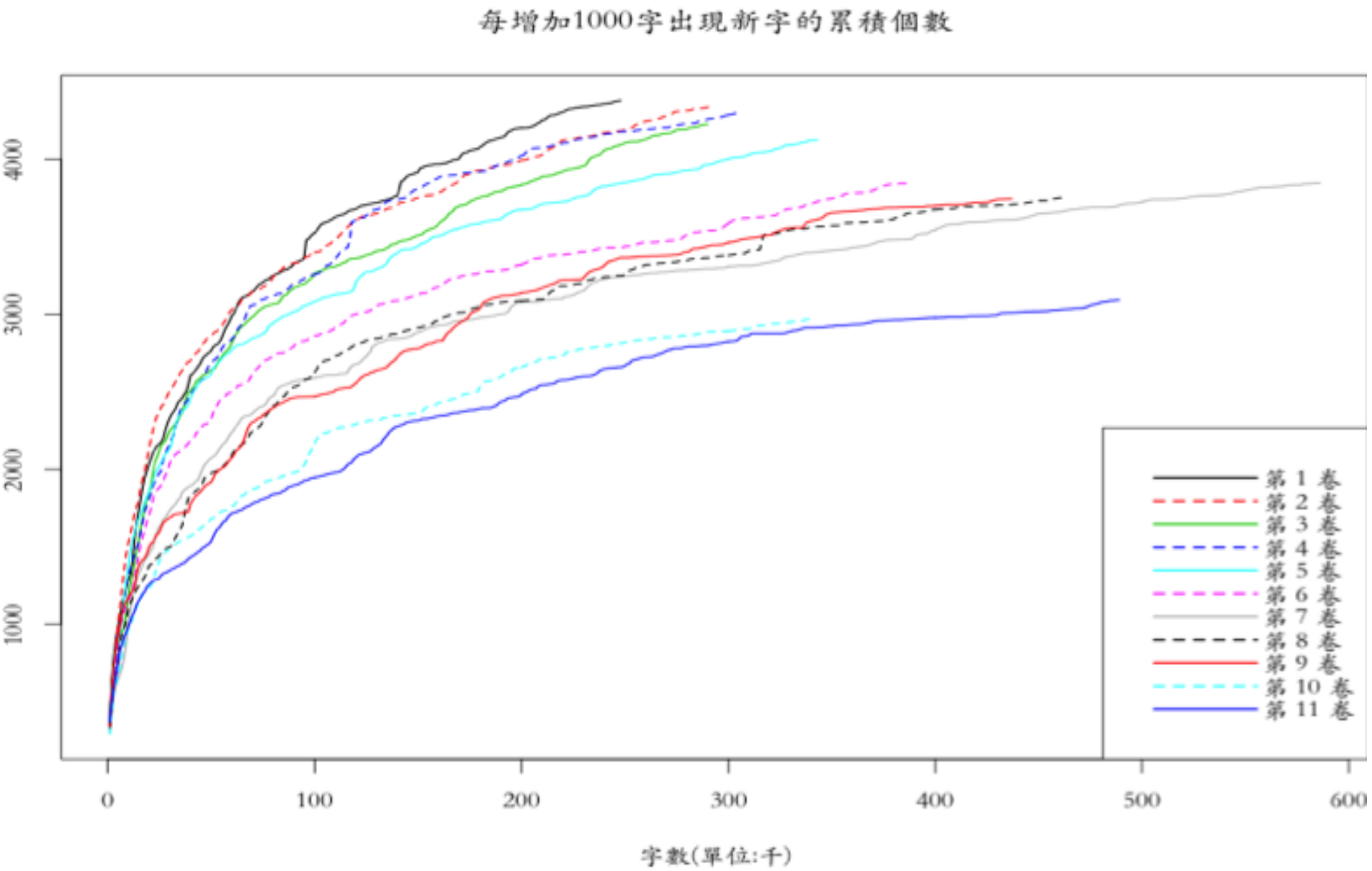
= (62/87) * 100 = **71.3%**

《新青年》各卷字彙豐富度(TTR)變化





《新青年》各卷新字出現頻率



字詞分布與趨勢分析

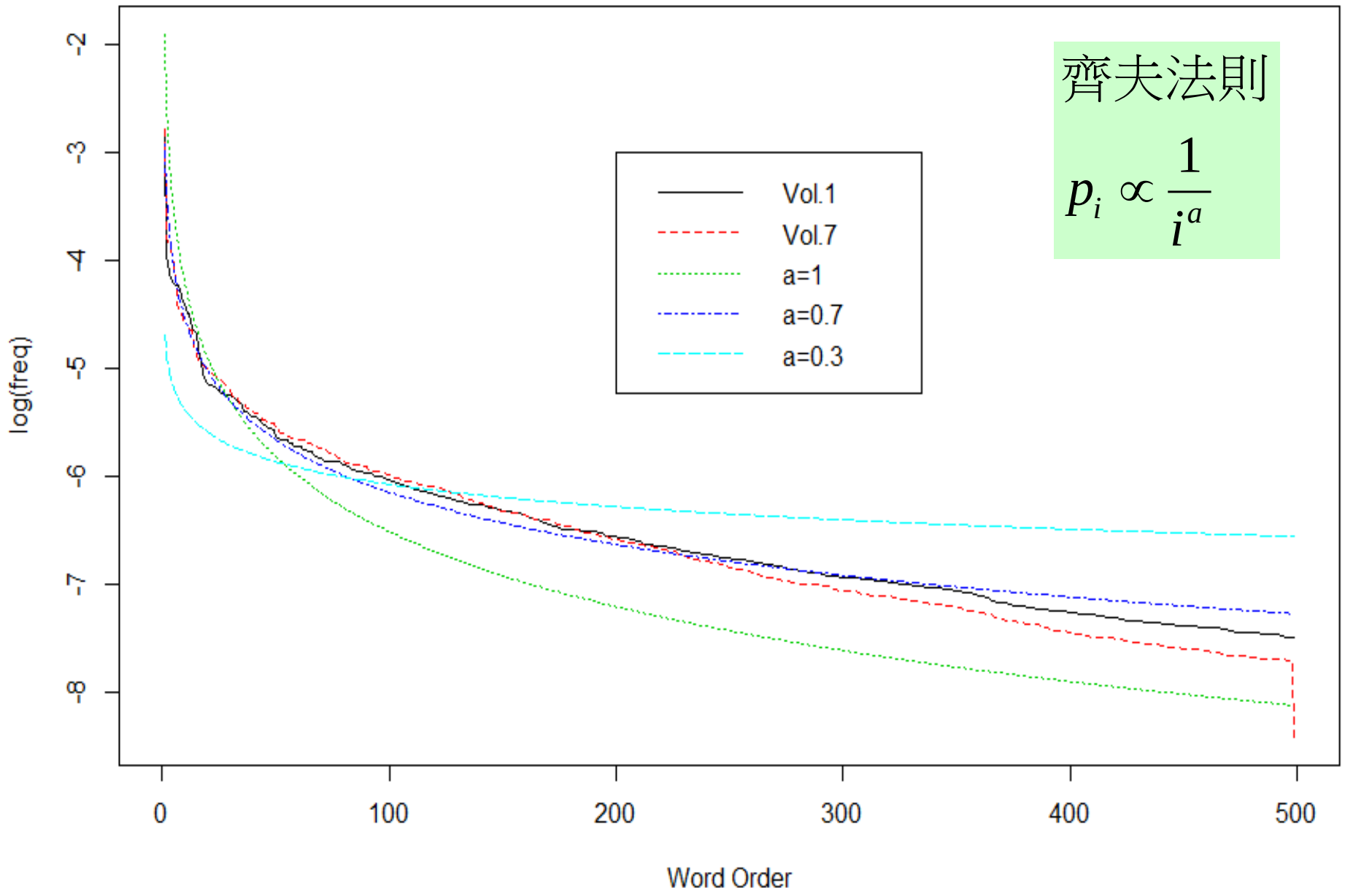
- 除了字彙豐富度，也可藉由分布、齊夫法則(Zipf's law)及Simpson指標及Entropy（熵）描述多樣性，其中：

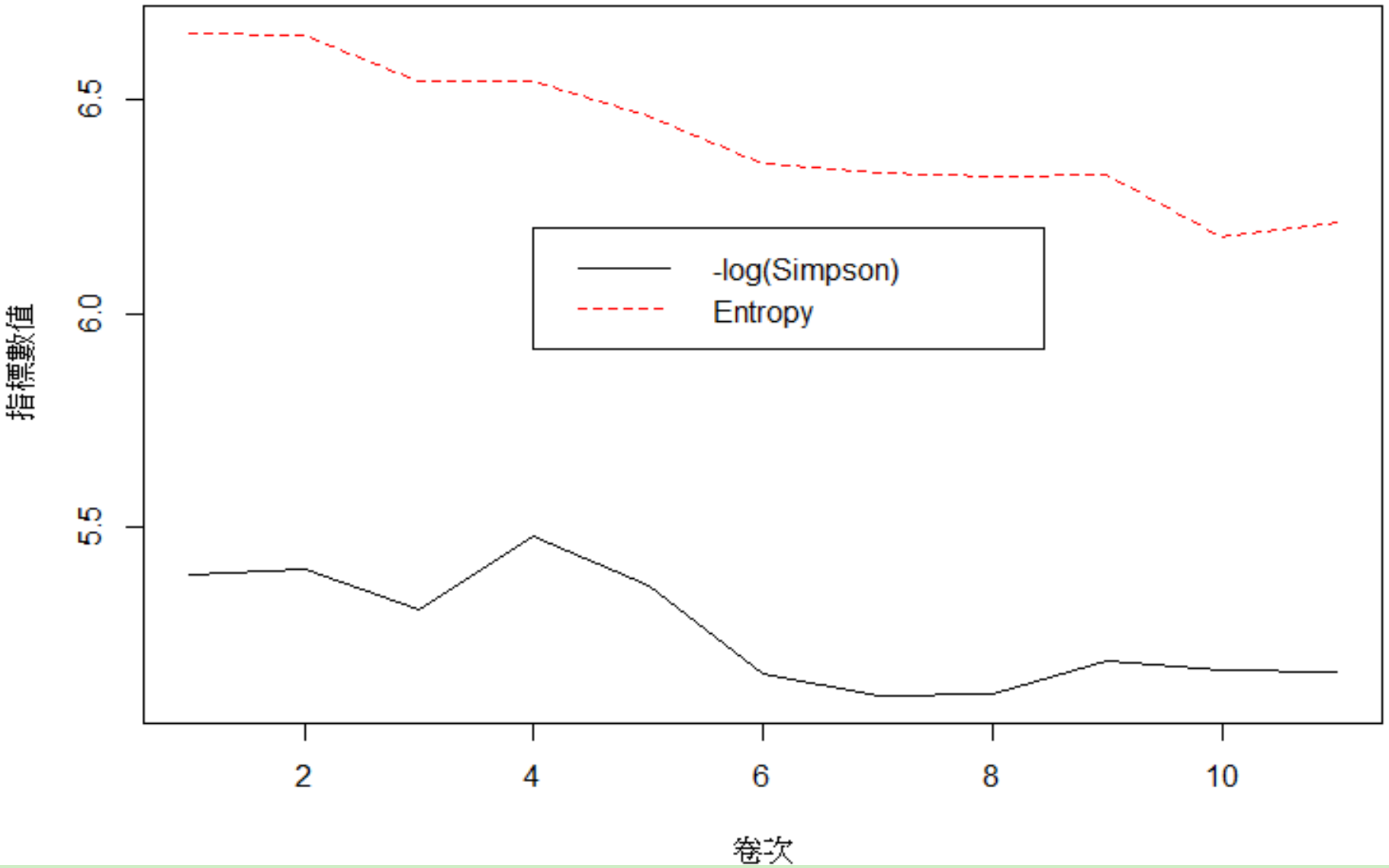
$$\theta_S = \sum_i p_i^2 \quad \text{及} \quad \theta_E = -\sum_i p_i \log(p_i)$$

- 趨勢分析則可透過**重複程度**，像是常見字詞的出現個數，與詞彙的滅絕及新生。

→相似指標：Jaccard Index及Yue Index。

《新青年》第一、七卷單字分布(Zipf's Law)





《新青年》各卷多樣性趨勢圖

Top 10 Words of *New Youth Magazine*

Rank	Vol.1	Vol. 2	Vol. 3	Vol. 4	Vol. 5	Vol. 6	Vol. 7
1	之	之	之	之	的	的	的
2	人	不	不	的	不	是	是
3	不	人	以	不	之	不	一
4	以	以	人	一	是	一	不
5	為	一	為	人	一	人	人
6	其	為	一	是	人	有	有
7	國	國	其	我	有	他	這
8	一	其	而	有	我	了	了
9	於	者	者	以	以	之	工
10	者	有	國	為	為	我	我

Top 10 2-Words of *New Youth Magazine*

Rank	Vol.1	Vol. 2	Vol. 3	Vol. 4	Vol. 5	Vol. 6	Vol. 7
1	國家	青年	社會	先生	中國	社會	我們
2	政府	社會	文學	文學	我們	我們	他們
3	自由	世界	中國	中國	他們	他們	社會
4	社會	國家	吾人	我們	現在	現在	勞動
5	薩稜	人類	世界	社會	先生	主義	工人
6	夫人	政府	先生	他們	文字	文學	現在
7	吾人	政治	吾國	現在	世界	中國	生活
8	政治	今日	政府	文字	社會	思想	時候
9	人民	主義	道德	學生	文學	先生	人口
10	青年	國民	教育	白話	主義	經濟	問題

第一~三卷

第七~九卷

文字雲：常見44個詞彙的比較

文言文、白話文的標點符號

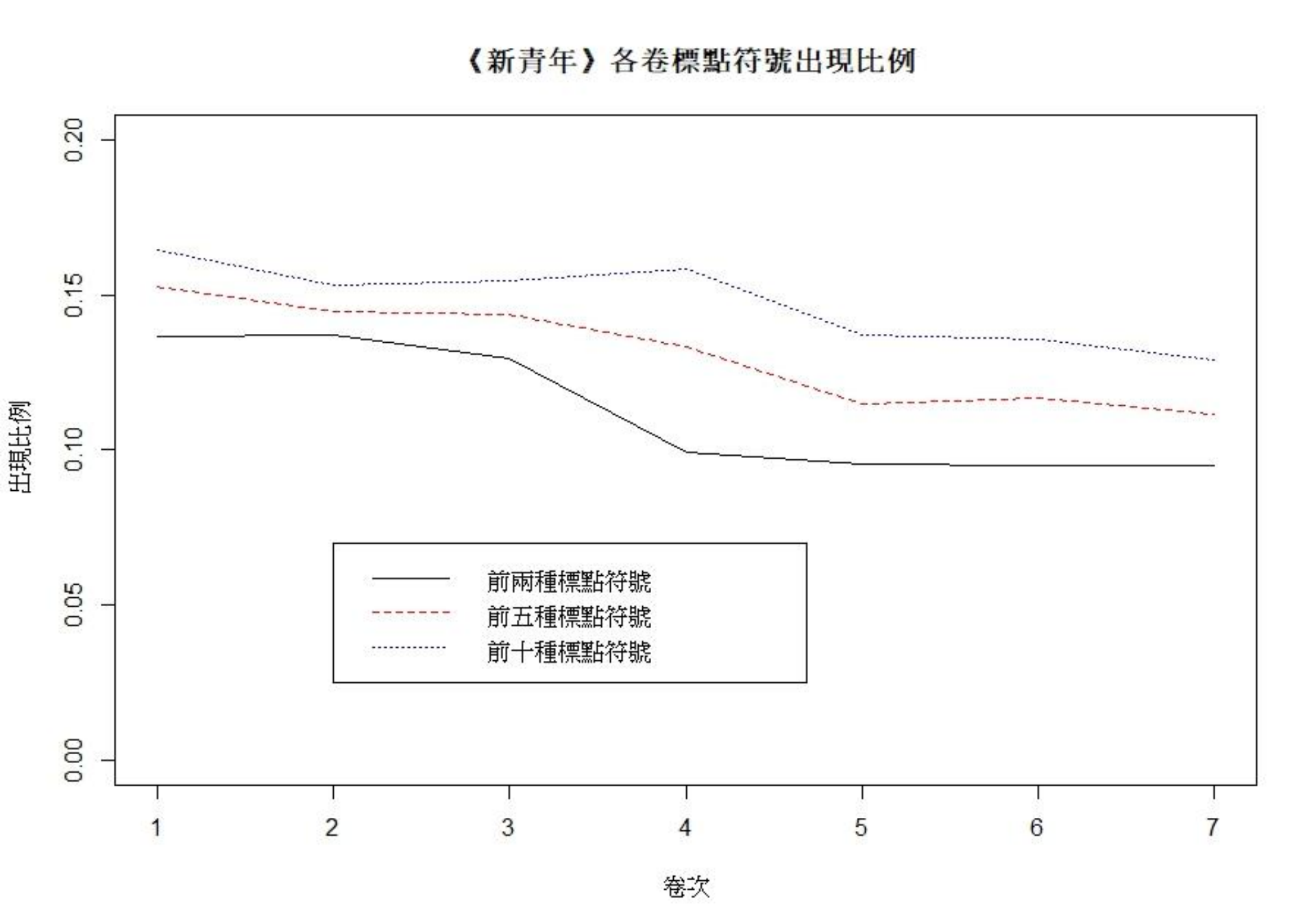
- 中文標點符號可追溯智商周時代，但文言文通常不加或不常使用標點符號，因而常有不同解釋。

→1920年教育部頒佈法令後逐漸一致。

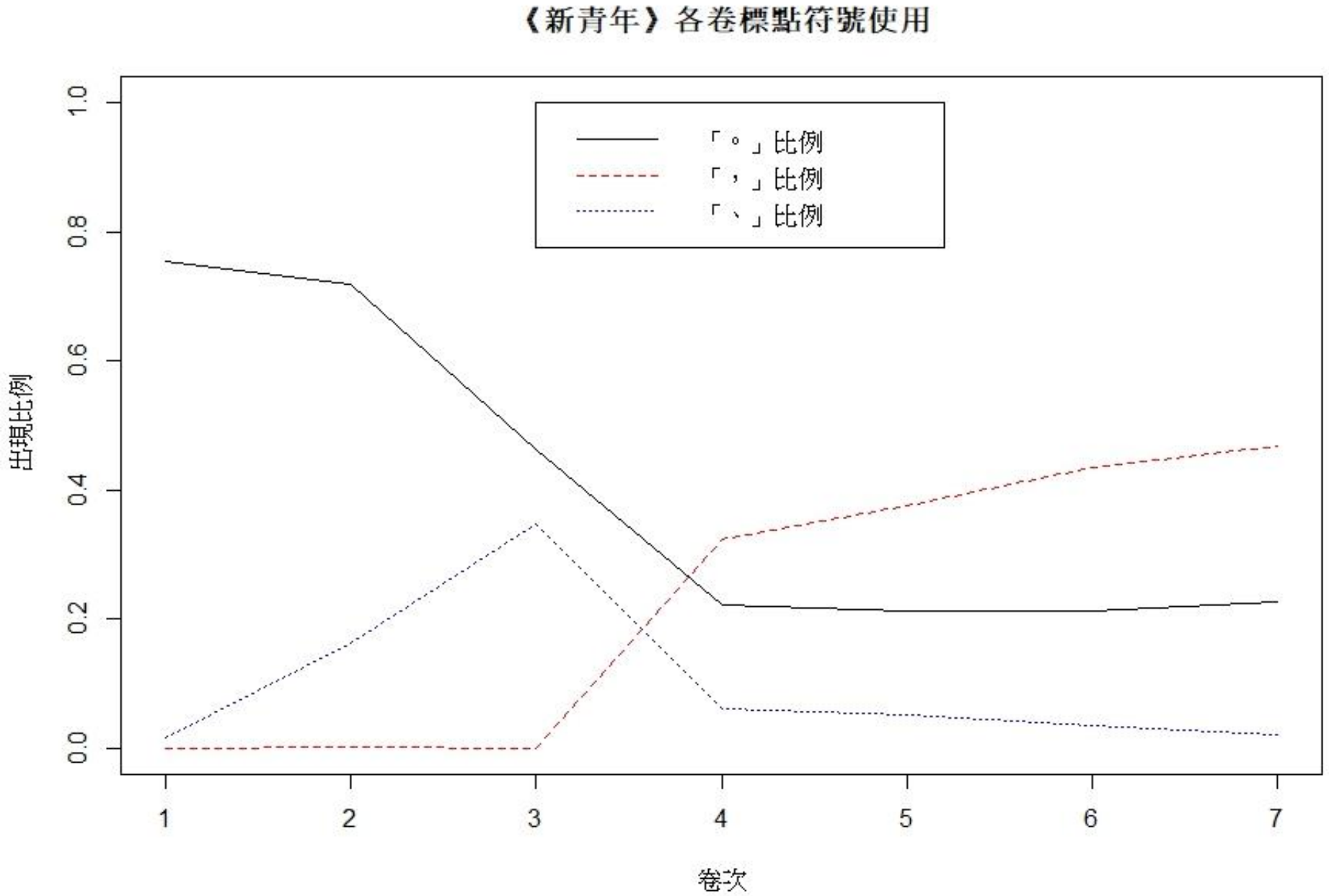
- 在此仿造英文的句法，採用「斷句」的標點符號計算各卷的句子長度。

→「。」、「，」、「；」、「！」、「？」

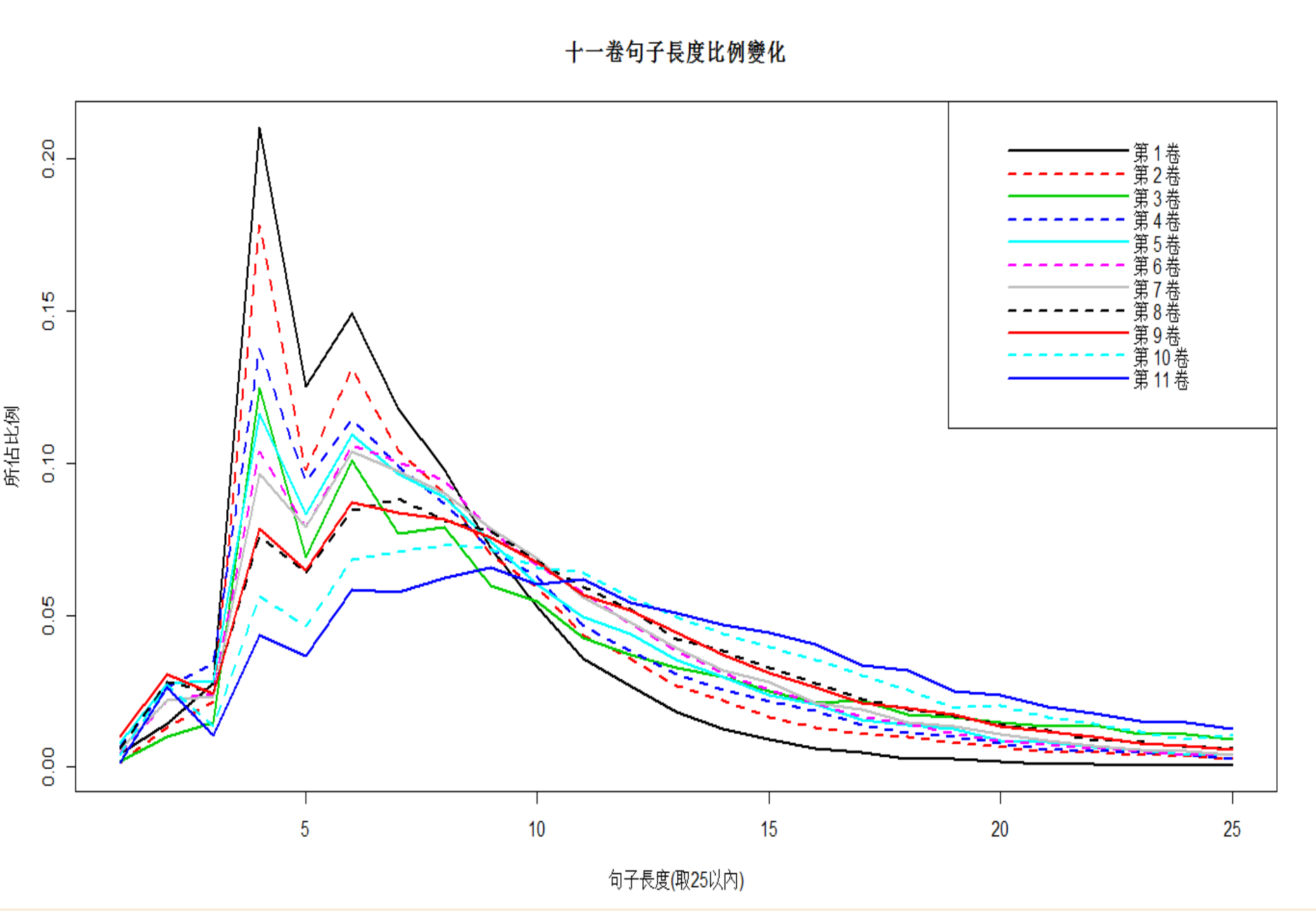
→標點符號的出現比例（加上「、」）



《新青年》各卷標點符號的出現比例趨勢



《新青年》各卷句號、逗號、頓號變化趨勢



《新青年》各卷每句字數分佈圖

文言文與白話文虛字

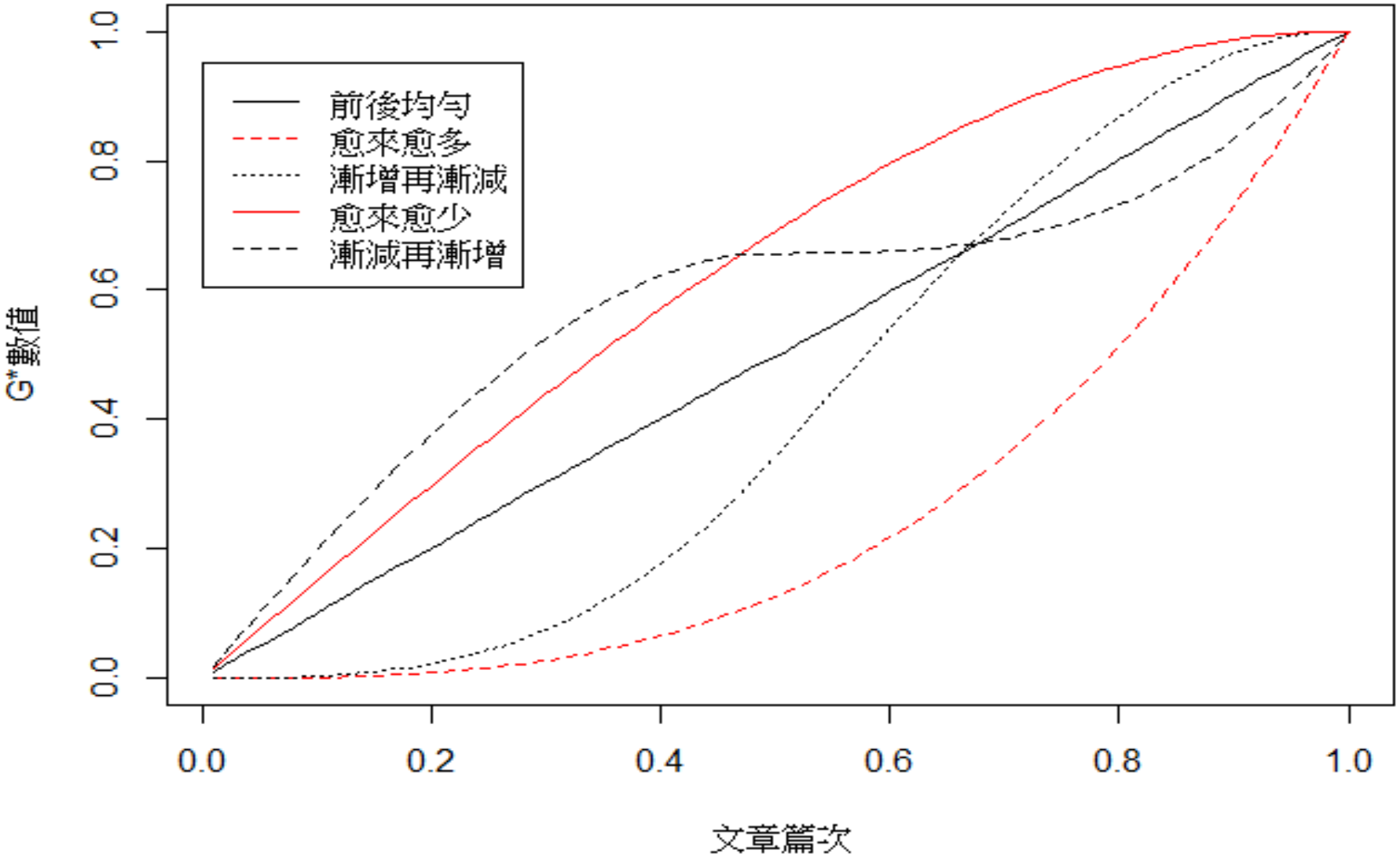
- 虛字（或虛詞）也稱為功能詞(Function words)，虛字無法獨立成句，但可用於理解表達的情緒或事物狀態。

→包括副詞、介詞、連接詞、助詞、歎詞等。

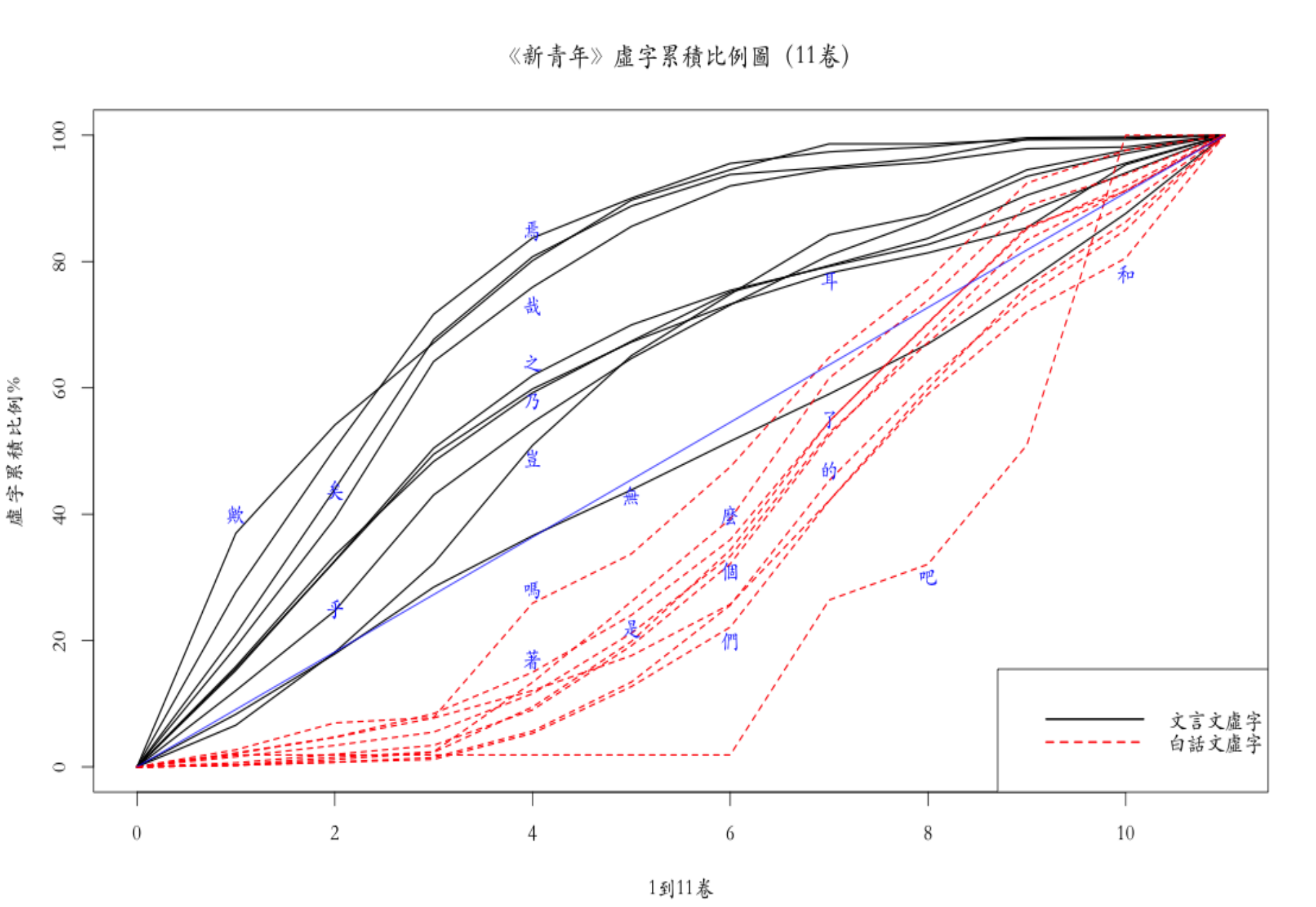
- 採用文言文、白話文各十個常用虛字：

→矣、乎、焉、歟、哉、耳、豈、之、乃、無（文言文）；

→的、是、們、個、了、和、麼、著、嗎、吧（白話文）。



不均度曲線類型代表的意義



文言、白話虛詞的趨勢變化（不均度）

虛字在各卷的吉尼係數變化趨勢

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