統計偏離值分析於人文研究上的應用 ---以《新青年》為例

金觀濤　梁穎誼
姚育松　劉昭麟

前言:

- 藉由數位方法進行人文研究，已經是人文研究的共識，也是未來的趨勢。
- 其中，“關鍵詞研究”是研究主題之一。
- 關鍵詞能幫助學者，理解史料背後的意義，進一步了解其思想，或解釋一些現象。

相關的關鍵詞研究:

- 《清季外交史料》
  - 近現代中國華人觀念，與保護華工的外交糾紛有關。

〈共現」詞頻分析及其運用－以「華人」觀念起源為例〉, 金觀濤、邱偉雲及劉昭麟 (2011)

- 《清末籌備立憲檔案史料》
  -清末籌備立憲失敗的背景，與行政機關不斷遭受諮議局挑戰有關。

〈社會行動的數位人文研究：以清末預備立憲為例〉, 金觀濤、姚育松及劉昭麟 (2011)

- 若未細讀某文本之前，就能找出其關鍵詞，能大量提升研究效率，(例 : 時間成本，人為疏失)

- 然而，未細讀文本時，找出關鍵詞是一個極具挑戰性的工作。

- 過去的相關研究，只根據詞頻高低判斷關鍵詞，並無考慮用詞模式，而且相當主觀，分析結果較不精確。

- 另一個缺點是，關鍵詞一般來說數量不會很多，而過去的研究因無有效方法擷取，使得詞數太多。

- 本研究提供了一個方法，幫助學者更客觀的擷取關鍵詞。

- 想法:
  考慮典型用詞分佈，若某詞詞頻出現太高，與典型用詞偏離太遠，我們可視為關鍵詞。

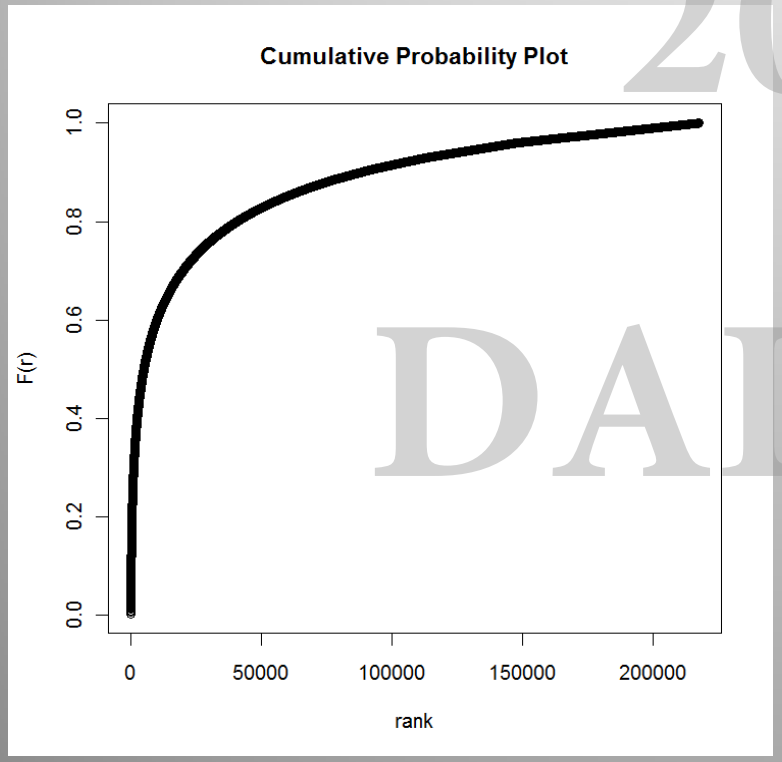
- 典型用詞分佈與「齊夫定律」(Zipf’s law)有關

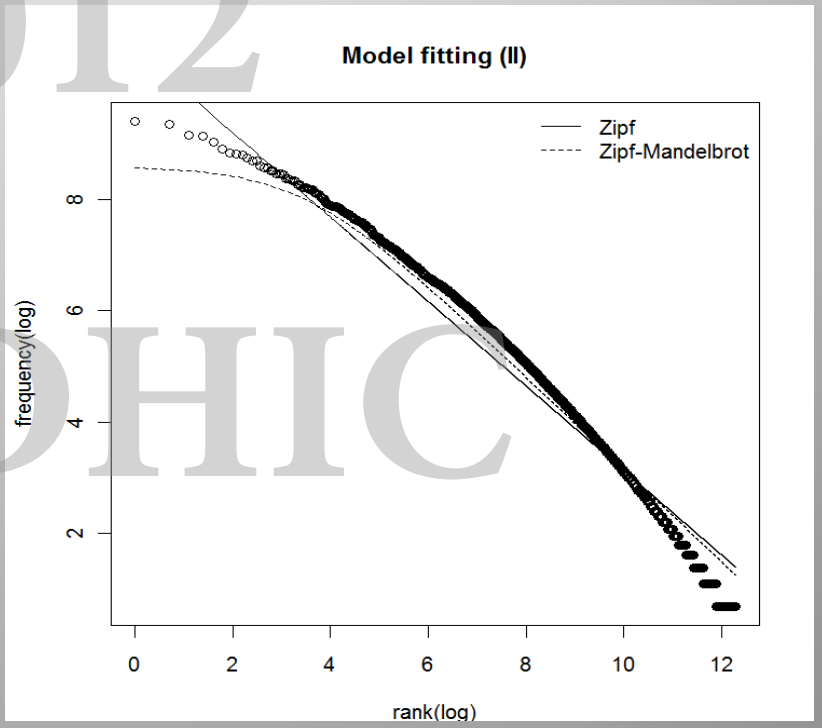
方法論:

- 「齊夫定律」為一經驗法則。
- 詞(字)數的頻數，以“冪定律”(Power law)遞減。
  - 如: 1/2 ， 1/4 ， 1/9, …
- 後來，有學者提出了「 Zipf-Mandelbrot 」模型，修正了某些線段的準確性。

Zip-Mandelbrot :

$$p(R=r)=\frac{1}{c(r+b)^{a}},\ a,b>0,$$





- a , b 參數估計： (最小平方法)

$$\min_{a,b}\{\sum_{r=1}^{k}[\log(y_r / \sum_{r=1}^{k} y_r) - (\log c + a\log(r-b))]^2\}$$

subject to : $\{a\ ,\ b\} > 0$

- 使用R，Matlab軟體即可方便求解。

- 定義 $R^*=(R_1,...R_i,...,R_k)$ 為一隨機向量，也為多項分配。若只看單項 R_i ，該隨機變數的分配為二項分配。 $R_i \sim Bin(n, p(i))$

- 二項分配:
  - 期望值： $\mathrm{E}_{\mathrm{R}_i} = np(i) = \dfrac{n}{c(i+b)^a}$
  - 變異數： $\mathrm{Var}(\mathrm{R}_i) = np(i)[1-p(i)]$

- 期望值可視為詞頻分佈的 “理論值”，若觀察到的詞頻偏離“理論值”太遠，就可視為可能的關鍵詞。

- 偏離值： $e_i = \dfrac{\mathrm{R}_i - \hat{\mathrm{E}}_{\mathrm{R}_i}}{\hat{\mathrm{Var}}(\mathrm{R}_i)}$

- 二項分配:
  - 期望值： $\mathrm{E}_{\mathrm{R}_i} = np(i) = \dfrac{n}{c(i+b)^a}$
  - 變異數： $\mathrm{Var}(\mathrm{R}_i) = np(i)[1-p(i)]$

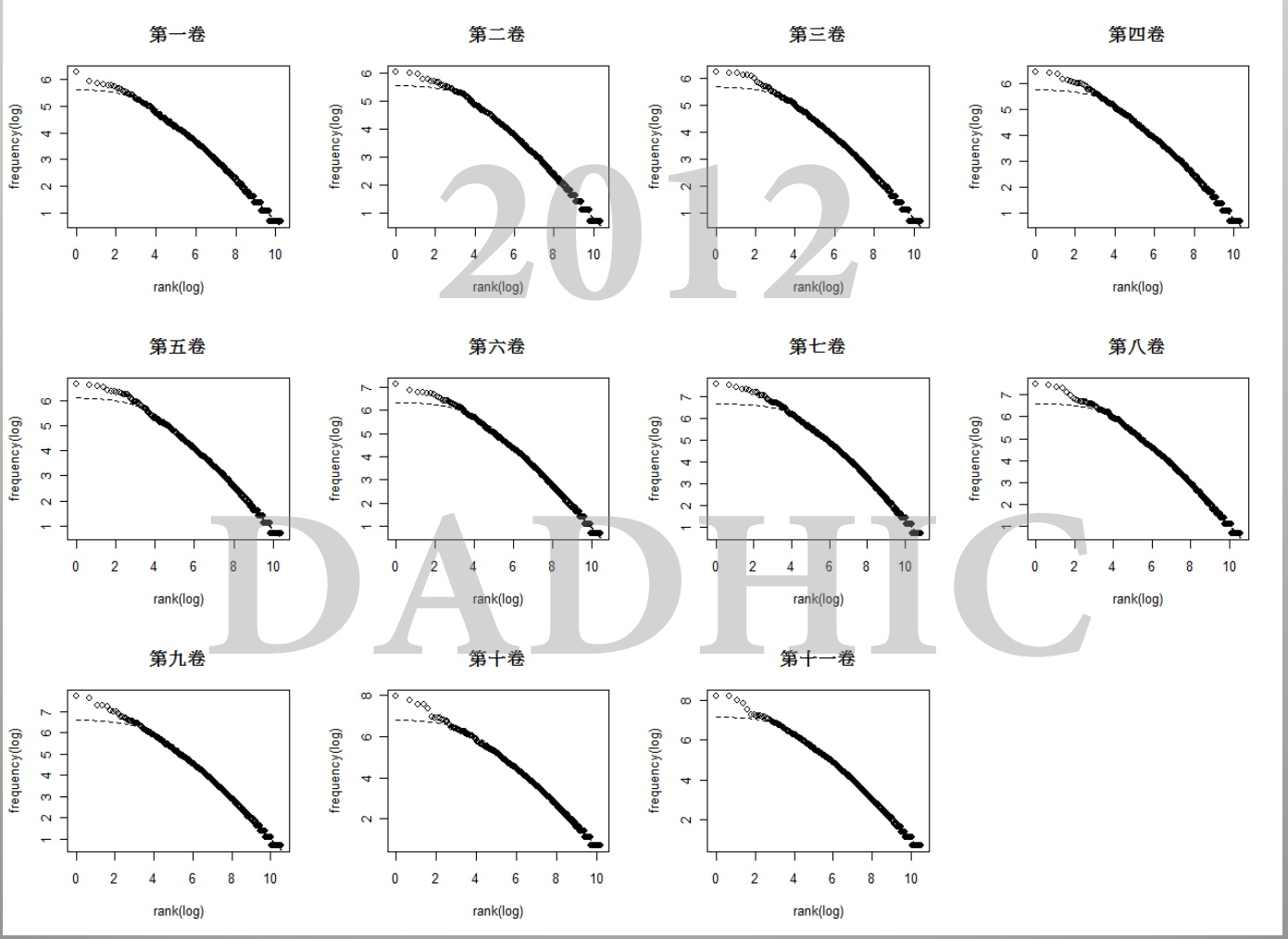
- 期望值可視為詞頻分佈的 “理論值”，若觀察到的詞頻偏離“理論值”太遠，就可視為可能的關鍵詞。

- 偏離值： $e_i = \dfrac{\mathrm{R}_i - \hat{\mathrm{E}}_{\mathrm{R}_i}}{\hat{\mathrm{Var}}(\mathrm{R}_i)}$

《新青年》資料分析

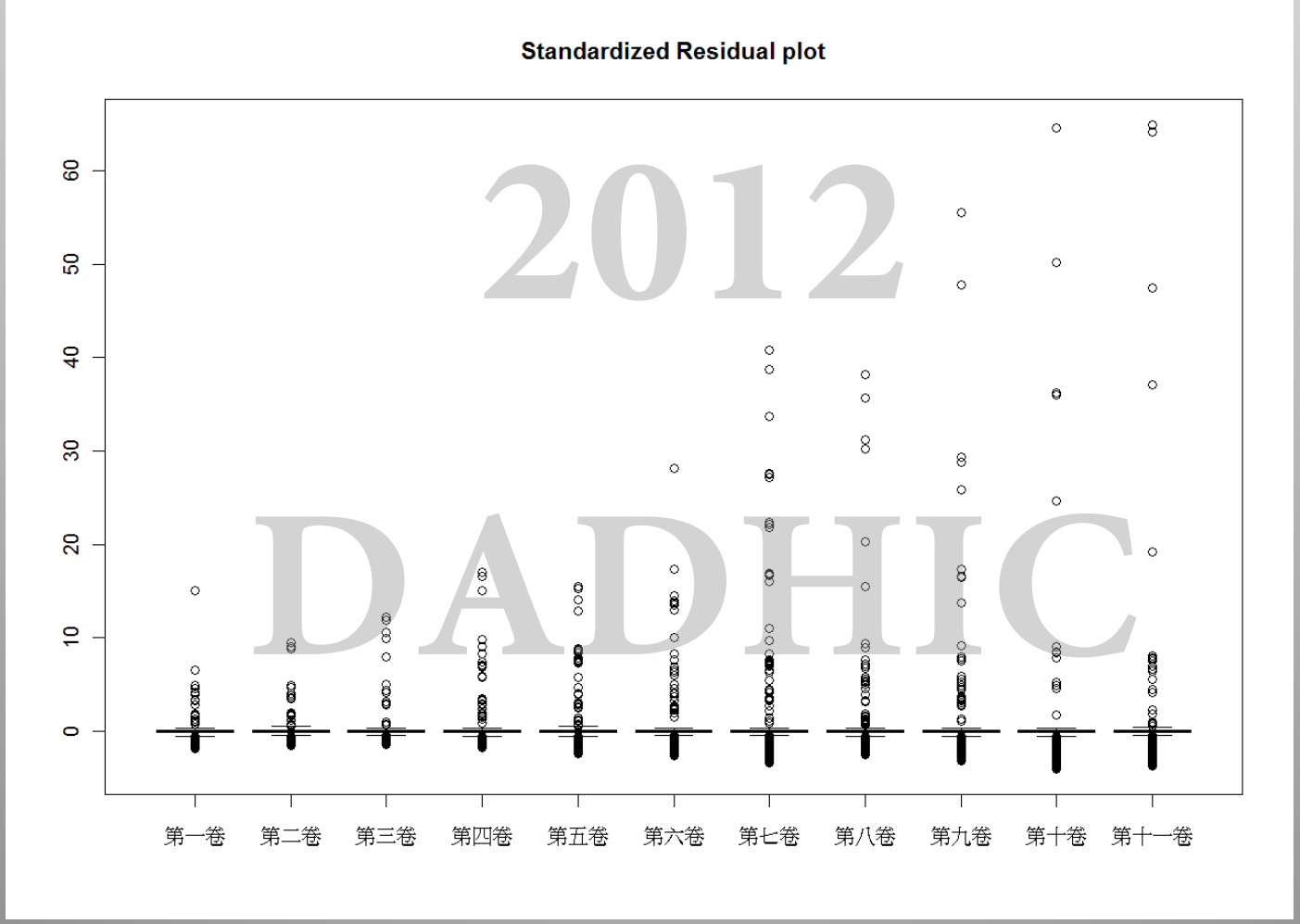
- 《新青年》共有11 卷 ，每卷總詞頻大約為20多萬。
- 待分析的詞之定義:
  - 該卷出現兩次以上的詞
- 由於文本龐大，我們使用Pat Tree 技術提取所需要的詞。

模型配適：



- 直接看配適曲線，對選取關鍵字的幫助不大。
- 可將偏離值畫成盒裝圖(Boxplot)，盒裝圖讓我們一目了然的了解偏離現象。
- 一般上，偏離值大於三或四以上，就可視為統計上的顯著，研究者可依據情況，給予更嚴苛的門檻。

盒裝圖 (Boxplot):


Standardized Residual plot
2012
DADHIC
0
10
20
30
40
50
60
第一卷
第二卷
第三卷
第四卷
第五卷
第六卷
第七卷
第八卷
第九卷
第十卷
第十一卷

關鍵詞與觀念變遷的關係

一、問題意識一的提出：

《新青年》的創刊目的即是要教育青年，那麼為何「青年」一詞要到第二卷才成為關鍵詞呢？ 而第一卷關鍵詞則是「國家」與「政府」？

二、歷史議題再陌生化：

數位人文學提供歷史學者一個透過數位技術協助的宏觀圖像，可幫助研究者對於歷史發展產生再陌生化的功能，幫助研究者重新審視與思考歷史。

三、當研究者面對上述再陌生化下的問題意識時，就必須回到文獻中去閱讀、分析，尋找可能的歷史解釋，完成數位與人文的協同研究。

卷數	一	二	三	四	五	六	七	八	九	十	十一
國家	•										
政府	•										
青年		•									
娜拉				•							
文學			•	•							
中國			•	•	•		•				
社會			•	•		•	•	•	•	•	•
我們				•	•	•	•	•	•	•	•
他們					•	•	•	•	•		
主義						•		•	•	•	•
資本								•	•	•	•
階級								•	•	•	•
革命									•	•	•
產階									•	•	•

關鍵詞與觀念變遷的關係

再前述再陌生化的問題意識下，本文進一步觀察新青年文獻得知兩點：

一、歐戰之前：國家之強大有賴於國民之自強，而能夠期待之國民，也只有青年。如〈青年與國家之前途〉：「蓋民為國之根本。而青年又民之中堅也。欲國之強。強吾民其可也。欲民之強。強吾青年其可也。」

二、歐戰之後：歐戰的慘烈使人發現到西方國家理論所可能帶來的危害。加上戰後談判中國作為戰勝國竟無法取回山東半島的權益，更深深刺激了時人的神經，由此西方國家理論是否能夠作為一普遍原則便得到質疑。

卷數	一	二	三	四	五	六	七	八	九	十	十一
國家	•										
政府	•										
青年		•									
娜拉				•							
文學			•	•							
中國			•	•	•		•				
社會			•	•		•	•	•	•	•	•
我們				•	•	•	•	•	•	•	•
他們					•	•	•	•	•		
主義						•		•	•	•	•
資本								•	•	•	•
階級								•	•	•	•
革命									•	•	•
產階									•	•	•

關鍵詞與觀念變遷的關係

一、問題意識二的提出：為何第三卷開始出現「中國」、「文學」、「娜拉」等等 。

二、經本文觀察得出兩點：

（一）從第三卷到第五卷（1917年3月1至1918年12月15），學習西方理論的潮流，在此之後，被新文學運動所取代 。

（二）而文學運動的成功使白話文取代文言文，於是「我們」、「他們」便被經常使用，而成為關鍵詞。

卷數	一	二	三	四	五	六	七	八	九	十	十一
國家	•										
政府	•										
青年		•									
娜拉				•							
文學			•	•							
中國			•	•	•		•				
社會			•	•		•	•	•	•	•	•
我們				•	•	•	•	•	•	•	•
他們					•	•	•	•	•		
主義						•		•	•	•	•
資本								•	•	•	•
階級								•	•	•	•
革命									•	•	•
產階									•	•	•

關鍵詞與觀念變遷的關係

一、問題意識三的提出：

「我們」與「他們」是作為一種區別我者與他者的常用詞。但從圖可發現，「我們」與「他們」隨著時間展延，其偏離值越趨下降，取而代之的是「主義」、「階級」、「社會」，「資本」。這是為什麼呢？

二、經過本文觀察得知：因為在「意識型態化」下，原屬於此意識型態下用於區別我者與他者的概念用語，亦即特有的分別主詞，會取代常用分別主詞。如：「我以為共產黨底基礎建築在無產階級上面，在理論上，自然要好過基礎建築在有產階級上面用金力造成的政黨」從這變化證明了「意識型態化」（主義化）的形成軌跡。

卷數	一	二	三	四	五	六	七	八	九	十	十一
國家	•										
政府	•										
青年		•									
娜拉				•							
文學			•	•							
中國			•	•	•		•				
社會			•	•		•	•	•	•	•	•
我們				•	•	•	•	•	•	•	•
他們					•	•	•	•	•		
主義						•		•	•	•	•
資本								•	•	•	•
階級								•	•	•	•
革命									•	•	•
產階									•	•	•

偏離值與觀念確立的關係

一、問題的提出：偏離值即為樣本的實際曲線與理論曲線在座標軸上的差距，越偏離即代表樣本與文本按照齊夫定律應有的分佈的比較上，顯得越特殊，我們說這即代表作為樣本的關鍵詞越加關鍵。這樣的假定是否能夠成立呢？

偏離值與觀念確立的關係

- 經過本文觀察根據樣本的最大偏離值來劃分三個時期：第一卷到第五卷樣本的最大偏離值平均是13.84，第六到第八卷是35.7，第九到第十一卷是61.44。我們認為，這正好反映了《新青年》的整個思想變化。
- （一）偏離值越大即表示關鍵詞越關鍵，第一卷到第五卷是社會主義尚未引入起得反響的時期，故此可知當時尚未有具體的、共識的觀念出現，因此出現的關鍵詞，自然偏離值偏低。

偏離值與觀念確立的關係

- （二）第六卷到第八卷之後，也就是1919年到1921年間，是社會主義引入取得反響的時期，而具有最大偏離值的關鍵詞，在第六卷到第八卷分別是「社會」、「他們」、「他們」，還有大量的屬於無產階級革命意識型態的關鍵詞，如「勞動」、「工人」、「資本」。這樣的落差可以得到這樣的解釋，此時期尚未出現觀念共識，因此《新青年》的作者們仍然沒有非常一致地共同使用屬於同一意識型態的概念語言，因此才用常用語言，即「他們」、「我們」來討論問題，故此最大偏離值的關鍵詞是「他們」。

偏離值與觀念確立的關係

- （三）到了第九卷到第十一卷，我們知道中國共產黨成立後，標示著意識型態的確立，對照最大偏離值的關鍵詞正好能夠證明。第九卷到第十一卷具有最大偏離值的關鍵詞分別是「主義」、「階級」、「主義」，都是意識型態用語，並且偏離值極高，表示「非常關鍵的關鍵詞」的出現，那正好應證了具體的、共識的觀念的確立。

偏離值與觀念確立的關係

- 可以這麼歸結，以偏離值來觀察關鍵詞的關鍵性是可以成立的，偏低的偏離值即表示越可能無具體、共識的觀念出現，偏高的偏離值即表示越可能具體、共識的觀念出現。而從偏離值的觀察會發現第一卷到第五卷，是「無共識觀念期」，第六卷到第八卷，是「思想討論期」，第九卷到第十一卷，是「意識型態確立期」。這樣的分期與觀察關鍵詞詞數的變化，也是相互應證的。

關鍵詞詞數與觀念確立的關係

我們歸結出《新青年》的思想分期，對照關鍵詞詞數，可發現第一卷到第五卷的「無共識觀念期」的關鍵詞偏少，平均為8.2個；第六卷到第八卷的「思想討論期」的關鍵詞偏多，平均為21個；第九卷到第十一卷的「意識型態確立期」的關鍵詞偏少，平均為13個。

關鍵詞詞數與觀念確立的關係

「無共識觀念期」的關鍵詞偏少，是因為當時人們找不到核心價值來代表當時的觀念；「思想討論期」的關鍵詞偏多，是因為思想激盪，人們各執一言，因此討論東西多，可以視為明顯的觀念便多；「意識型態確立期」的關鍵詞偏少，是因為人們有具體、共識的觀念，觀念集中濃縮於某關鍵詞，故此關鍵詞就偏少。

餘論

- 總結而言，利用齊夫定律偏離值計算應用於人文研究上，以《新青年》為例，除了證明了可行性之外，還發現取得三大成果。
- 第一，能夠更加準確地將關鍵詞篩選出來，避免了主觀判斷的爭議，以及人工篩選的負擔，從而更加細緻地觀察思想變化，以及反映在文本上的語言現象，前者透過「國家」與「青年」這兩個關鍵詞來說明，後者透過「他們」、「我們」、「社會」、「主義」、「階級」、「產階」來說明；
- 第二，應證了偏離值大小能夠代表詞的關鍵性，從而揭示出《新青年》的思想變化，從無共識觀念，到思想討論，最後到意識型態確立的過程；
- 第三，應證了關鍵詞詞數能夠部份反映文本的思想變化，由於觀念濃縮於關鍵詞，關鍵詞的多寡能夠反映明顯觀念的多寡。