從［紅樓夢］論齊普夫律中文之適用性

何 光 國

Assistant Director, Howard University Libraries

【摘　要】

齊普夫律（Zipf Law）的主要特點有二：一爲文獻中的單字出現頻率與其遞減排名次序之積爲一常數。而且該常數等於全文字數的1/10。另一爲由排名與單字出現頻率之對數值所獲得的曲線是一條自左向右，斜率等於－1的直線。歷年來，經過各種不同的研究和試驗，證實英語確實具有這二個特點。齊普夫還根據不完全的資料統計所獲得結論，認爲該律也適用於中文。本文特以［紅樓夢］第四十回爲樣本，對齊普夫律的特點作一比較性的探討，以鑑定它對中文的適用性。研究發現，該律基本上並不適用於中文。

【ABSTRACT】

Zipf Law has two foundamental characteristics：（1）the product of the frequency of a word's appearance in a text and its relative ranking is a constant. It equals to 0.1 of the total number of words of the text；（2）the slope of its curve equals to -1. Many studies have proven that Zipf Law does fit English pattern. Using an imcomplete statistical data, Zipf concluded that the law is also applicable to Chinese.This study uses Chapter Forty of the Dream of the Red Chamber as a sample to test whether the law is truly applicable to Chinese language. Based on the findings of this study, it is found that Zipf Law is not applicable to Chinese language.

一、齊普夫律

齊普夫（George K. Zipf）在1949年出版了一本名爲［人類行爲與省力原理：人類生態學導論］（Human Behavior and the Principle of Least Effort：An Introduction to Human Ecology）（註1）的巨著。在那本書裡，他以James Joyce撰著的希臘神話小說［尤里西斯］（Ulysses）爲樣本，作了文字出現頻率的計量研究。［尤里西斯］全文共260,430字，單元字數爲29,899字。齊普夫將每字的出現次數（頻率）f，從多至少的依序排名（Rank），將出現次數最多的單元字排在第一，將出現次數最少者排在最後。他將排名序定爲r。他發現r與f之積爲一常數（Constant），C。

而且這一常數等於全文字數的0.1。這種關係可以下式表示：

$$rf=c \qquad ①$$

①式即爲齊普夫第一律，也就是一般通稱的齊普夫律。其實這個公式實在有些因陋就簡，馬馬虎虎，因而受到了不少學者的嚴厲批評（註2）。原因之一便是①式中隱略了一重要部份，那就是斜率。正確的公式應如下式：

$$g(f)=\frac{A}{r^{\beta}} \qquad ②$$

②式中的A與β皆爲常數。

將②式二邊各取對數，可得：

$$\ln g(f)=\ln A-\beta\ln(r) \qquad ③$$

顯然③式爲直線公式。因斜率β爲負值，所以該直線爲一自左向右下傾斜的直線。且因$\beta=1$，所以傾斜度恰等於45°。唯有$\beta=1$的時候，齊普夫的①式方能成立。

從③式，我們可以清晰地看出齊普夫特別重視文字分佈的線性和斜率等於－1的二個條件（註3）。因此，我們可以將齊普夫的特性歸納爲以下二點：

1. 排名，r，與相對單字出現頻率，f，之積等於一常數，c。且c值等於文獻全部字數的0.1。
2. 斜率等於－1。

若欲鑑定齊普夫律對中文是否適用，或者鑑定中文文字的分佈是否具備以上二種特點，我們就必須取樣做文字計量研究，查看中文文字出現頻率與其相對排名之積是否確實等於全文字數的1/10，而所獲直線，其斜率又是否確實等於－1。

二、研究動機：

㈠齊普夫在其所著的〔人類行爲與省力原理〕一書中，曾以下圖（見圖1）爲依據，認爲中文分佈也符合齊普夫律所定下的線性和向下傾斜的基本規律（註4）。換句話說，他認爲該律也適用於中文。不過，若我們仔細閱讀他的論述，實可發現很多的漏洞。譬如，圖中的中文分佈絕不像德文那樣有規律和呈線性。相反，中文分佈毫無確定線性可言。既使找到一條直線，它的標準差誤一定也會很大。雖然從圖形的整體看，中文分佈大致上確有自左向右下傾的勢態，但並不足以證明該線的斜率即等於－1。追根究底，齊普夫對中文分佈那種大而化之的結論，實由於他對中文文字未曾作過落實的計量研究的結果。因此，他那〝也適用於中文〞的推斷，不無可疑之處。

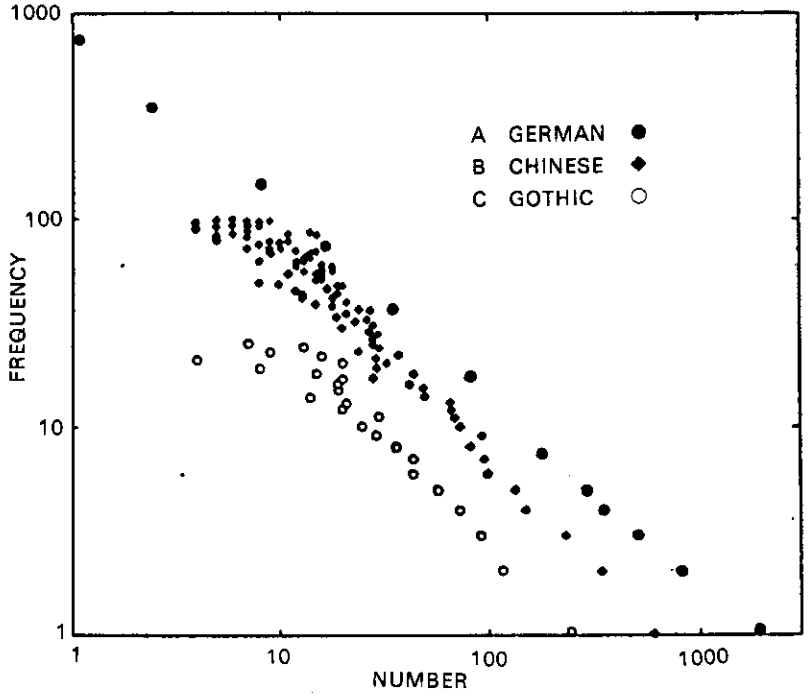

圖1 The number-frequency distribution of (A) German stem forms ; (B) Chinese characters ; (C) Gothic root morphemes.

㈡中國大陸學者研究發現中文詞類分類，在「基本上」也符合齊普夫律（註5）。「基本上」這三個字意義模糊。僅從字面上解，齊普夫律是否百分之百的適用於中文，還可以有商榷研究的餘地。

㈢筆者直覺的認爲中文與英文之間，無論是在文字的結構上，或語義和語法上，都有非常顯明的差別。英文行得通的，中文並不一定行得通。至於究竟行不行得通，只有經過一番研究試驗，才能獲得正確的答案。

㈣齊普夫律的應用面廣，諸如考證作者真僞，發掘作者寫作風格，編製各種字典、詞典，以及重要人物名錄等等，都可利用該律的研究方式，從文字計量的分析和研究上，獲得較爲正確的結論，樹立採信的價值。因此，該律的是否確能適用於中文，或在何種條件下才能適用等等，都值得我們去細心研究和探討，務必要求出一個比較明確的答案，否則即使我們挪用齊普夫律作各種有關的研究或應用，所得結論和結果，也必定會籠罩上一層躲閃不掉的陰影。

三、研究方法

㈠樣本的選擇：

在統計方法中，利用樣本作個案研究，似已成了不移公理。它的好處是只要樣本選擇妥貼適當，則自會收到見微知著之效，不必勞師動眾，大費週張。尤其，統計這玩意兒，講的就是機率，100%的機率究竟不常見，於是低於100%的情況，也可以接受。更何況齊普夫律之被稱的〝律〞（law），實有它四海皆準的含義和特性。換句話說，假如齊普夫認的該律也適用於中文，那麼採用任何中文文獻當作樣本來研究，能得到的結論，都應該完全一樣。否則，豈可當之爲〝律〞？所以筆者毫無保留的選擇了［紅樓夢］（註6）。唯個人能力有限，不得不又在全書中挑選了［紅樓夢］裡面的第四十回〝史太君兩宴大觀園，金鴛鴦三宣牙牌令〞，作爲樣本，進行研究。

筆者選用［紅樓夢］雖屬一番偏愛，老實說也不無其他原因。譬如：(1)［紅樓夢］爲中國家喻户曉的一部古典小說，被紅學家們譽爲「古今獨步」的奇書。在文體性質上，該書與［尤里西斯］很相近。假如採用一本

非小說類的作品作爲樣本，筆者會很不放心，恐怕造成選樣的偏差（Bias）。其實，這份考慮倒是多餘的。此點且留待〝討論〞一節再說。(2)以［紅樓夢］爲研究對象的文章非常多。這篇研究也希望湊個數，共襄盛舉。(3)筆者選上第四十回，除了也想旁看劉姥姥進到〝畫兒裡〞去逛上一逛之外，主要原因還是在這一回裡，不僅人物齊全，文字優美動人，文筆輕鬆風趣，而且作者的語言和意境，就像那高山頂上的一線溪流，格外地清新純淨，了無矯作之態。這一回可說是全書最令人喜愛的一章。人間的歡樂本不多，何不用它來稍稍沖淡一絲閒愁？(4)筆者獨力實無法將全書近三百萬字統統鍵入電腦，只有退而求其次的選擇其中一回，共計13,525字。只希望這是一個恰當的選擇。同時也希望利用這一回爲樣本，所得的結論能與全文研究所得相同。筆者這份大膽的希望，只有等待有興趣的學者專家印證指教。

㈡將第四十回全文共13,525字，去除標題及標點符號，全部鍵入微電腦。

㈢將13,525字劃分成三組：

1.單字組

共1,047字

2.字詞組

共1,466字

3.字詞組（不含人名與稱呼）

共1,427字

㈣分別計算各組字詞出現頻率，並依序排名。

㈤製作各種有關圖表，並根據資料，分別作了迴歸分析和最小平方值的計算。

㈥根據所獲結果，繪製齊普夫律曲線。

㈦分析研究各種資料及圖表。

㈧研究發現及結論。

四、研究發現

㈠單字組：

第四十回全文共13,525字，單元字爲1,047字。根據表1（見表1），發現平均每字被重複利用了13次，排名第一的〝了〞字與排名第二的〝道〞字，平均每一百字中便會出現三次，〝的〞和〝一〞字，大約平均每一百字中便會發現二次。再根據單字逐字出現頻率排名序（註7），編成〝單字出現頻率對數值〞（見表2）。從迴歸分析，得迴歸係數－1.14272，

常數4404.9。利用最小平方法，獲得係數－1.1894，常數5696.66（見表3）。

㈡字詞組：

假如我們將人名、稱呼、及各類習慣用語、口語、方言、成語等由單字組合而成的連綴詞，譬如賈母、鳳姐兒、我們、咱們、寶玉、商議、文官、小丫頭子們、一副兒、慌慌張張、不得了、毛毛蟲等等都當作一個字計算，那麼全回共有1,466個字（註8）。根據各個字組出現頻率製成〝字詞出現頻率對數值〞（見表4）。從迴歸分析，得迴歸係數－0.95004，常數714.9。利用最小平方法，獲得係數－0.95833，常數745.2（見表5）。

㈢字詞組（不包含人名及稱呼）：

［紅樓夢］第四十回，單個兒講起來，可就是全書笑得最多也最風趣的一章。這一回不僅情意特別，而且稱呼也格外的多。譬如劉姥姥、姥姥、劉親家、老太太、老祖宗、鳳姐兒、鳳丫頭等等。爲了避免文字統計上的偏誤，特別將這些人名及稱呼剔除，共得1,427字。比帶人名及稱呼的字詞組少了39字（見表6）。由迴歸法獲得迴歸係數－0.93622，常數601.57。由最小平方法得係數－0.95814，常數676.22（見表7）。比較（見表5），去掉人名及稱呼，並未能增加係數的準確性，反而偏低了－0.00019。這種結果，甚出意料之外。

㈣根據表2、表3、表4的估計字數及全文字數的1/10，繪製成圖2（見圖2）。從圖示，發現三種文字計算方法所得之積與1353字之間都有很大的距離。在程度上，以單字組距離最遠，字詞組不含人名及稱呼者次之，字詞組包含人名及稱呼者最近。

㈤根據表3、表5、表7中所列六條直線方程式，求得排名及出現頻率值（見表8）。再利用雙對數圖紙，繪製成單字、字詞、及字詞不含人名及稱呼等六條齊普夫律曲線（見圖3）。發現單字迴歸法求得的斜率角度最小，僅－35°，最佳的角度爲以最小平方法求得的字詞組包含人名及稱呼的斜率角度－43.8°。比齊普夫律的規定－45°，僅小－1.2°。

㈥研究結果：

1. 由各字組字數出現頻率與排名之積，所得之估計值均不等於全文的0.1，或1353字（見表2

表1　單元字每一百字出現頻率

排名 r	單元字	出現頻率 f	每一百字出現頻率	排名 r	單元字	出現頻率 f	每一百字出現頻率
1	了	347	2.6		過	62	0.5
2	道	346	2.6	43	忙	52	0.4
3	的	326	2.4	44	没	50	0.4
4	一	312	2.3		兩	50	0.4
5	說	230	1.7	45	樣	48	0.4
6	不	216	1.6	46	起	47	0.3
7	這	198	1.5	47	看	46	0.3
8	笑	190	1.4		小	46	0.3
9	來	179	1.3		呢	46	0.3
10	上	164	1.2		坐	46	0.3
11	也	159	1.2		邊	46	0.3
12	我	150	1.1	48	花	45	0.3
13	是	145	1.1		姨	45	0.3
14	賈	140	1.0		姐	45	0.3
	人	140	1.0		頭	45	0.3
15	姥	132	1.0	49	拿	44	0.3
16	劉	128	0.9		出	44	0.3
17	鴛	124	0.9		得	44	0.3
18	有	122	0.9	50	要	43	0.3
19	著	111	0.8	51	寶丫	42	0.3
20	個	110	0.8		叫	42	0.3
21	他	109	0.8	52	紗	40	0.3
22	那	107	0.8	53	如	39	0.3
23	大	105	0.8	54	玉	38	0.3
24	都	104	0.8	55	薛	36	0.3
25	子	95	0.7		等	36	0.3
26	下	93	0.7	56	王	35	0.3
27	去	90	0.7	57	別	34	0.3
28	只	88	0.7		李	34	0.3
	兒	88	0.7		罷	34	0.3
29	們	87	0.6		令	34	0.3
30	吃	86	0.6		二	34	0.3
31	你	84	0.6		回	34	0.3
	又	84	0.6	58	把	33	0.2
	鳳	84	0.6		家	33	0.2
32	裡	82	0.6		三	33	0.2
33	便	81	0.6	59	話	32	0.2
34	好	76	0.6		些	32	0.2
35	聽	75	0.6	60	桌	31	0.2
36	見	74	0.5		自	31	0.2
37	母	73	0.5	61	東	30	0.2
38	還	72	0.5		進	30	0.2
39	在	70	0.5		今	30	0.2
	就	70	0.5	62	帶	29	0.2
40	鴇	66	0.5		咱	29	0.2
41	老	65	0.5	63	幾	28	0.2
42	眾	62	0.5		與	28	0.2
	太	62	0.5				

續表1　單元字每一百字出現頻率

排名 r	單元字	出現頻率 f	每一百字出現頻率	排名 r	單元字	出現頻率 f	每一百字出現頻率
	成	28	0.2	72	夫	19	0.1
	到	28	0.2		雲	19	0.1
	比	28	0.2		當	19	0.1
	給	28	0.2		湘	19	0.1
64	中	27	0.2		春	19	0.1
	後	27	0.2	73	正	18	0.1
65	天	26	0.2		几	18	0.1
	黛	26	0.2		再	18	0.1
	放	26	0.2		什	18	0.1
	開	26	0.2		五	18	0.1
66	早	25	0.2		姑	18	0.1
	倒	25	0.2		送	18	0.1
67	擺	24	0.2		探	18	0.1
	可	24	0.2		西	18	0.1
	色	24	0.2	74	素	17	0.1
	窗	24	0.2		知	17	0.1
	張	24	0.2		用	17	0.1
68	手	23	0.2	75	席	16	0.1
	四	23	0.2		高	16	0.1
69	和	22	0.2		各	16	0.1
	因	22	0.2		預	16	0.1
	之	22	0.2		板	16	0.1
	婆	22	0.2		多	16	0.1
	茶	22	0.2		式	16	0.1
	拉	22	0.2		完	16	0.1
	走	22	0.2		打	16	0.1
	銀	22	0.2		內	16	0.1
	飯	22	0.2		怕	16	0.1
70	作	21	0.2		妹	16	0.1
	日	21	0.2		間	16	0.1
	才	21	0.2		杯	16	0.1
	已	21	0.2		紅	16	0.1
	地	21	0.2	76	至	15	0.1
	屋	21	0.2		盒	15	0.1
71	命	20	0.1		使	15	0.1
	明	20	0.1		瞧	15	0.1
	先	20	0.1		風	15	0.1
	奶	20	0.1		畫	15	0.1
	收	20	0.1		十	15	0.1
	想	20	0.1		設	15	0.1
	方	20	0.1		何	15	0.1
	房	20	0.1		時	15	0.1
	心	20	0.1		點	15	0.1
	左	20	0.1		聲	15	0.1
	問	20	0.1		站	15	0.1
	面	20	0.1	77	姊	14	0.1
	媽	20	0.1		揀	14	0.1
	麼	20	0.1		喜	14	0.1

續表1　單元字每一百字出現頻率

排名 r	單元字	出現頻率 f	每一百字出現頻率	排名 r	單元字	出現頻率 f	每一百字出現頻率
	原	14	0.1	80	數	11	0.1
	親	14	0.1		園	11	0.1
	往	14	0.1		逛	11	0.1
	迎	14	0.1		鬧	11	0.1
	閑	14	0.1		錦	11	0.1
	會	14	0.1	81	快	10	0.1
	劄	14	0.1		落	10	0.1
	若	14	0.1		該	10	0.1
	住	14	0.1		請	10	0.1
	案	14	0.1		抬	10	0.1
	行	14	0.1		雖	10	0.1
	娘	14	0.1		越	10	0.1
	右	14	0.1		答	10	0.1
	帳	14	0.1		楞	10	0.1
	長	14	0.1		荷	10	0.1
	愛	14	0.1		惜	10	0.1
	處	14	0.1		夾	10	0.1
	糊	14	0.1		石	10	0.1
	箸	14	0.1		舡	10	0.1
	紈	14	0.1		怪	10	0.1
	然	14	0.1		名	10	0.1
78	外	13	0.1		軟	10	0.1
	滿	13	0.1		找	10	0.1
	香	13	0.1		離	10	0.1
79	定	12	0.1		遠	10	0.1
	做	12	0.1		畢	10	0.1
	並	12	0.1		無	10	0.1
	酒	12	0.1		眼	10	0.1
	向	12	0.1		水	10	0.1
	梯	12	0.1		偏	10	0.1
	佛	12	0.1		翠	10	0.1
	讓	12	0.1		地	10	0.1
	年	12	0.1		洒	10	0.1
	怎	12	0.1		很	10	0.1
	竟	12	0.1		既	10	0.1
	卻	12	0.1		六	10	0.1
	書	12	0.1		么	10	0.1
	顏	12	0.1		管	10	0.1
	取	12	0.1		試	10	0.1
	身	12	0.1		盤	10	0.1
	端	12	0.1		誰	10	0.1
	菜	12	0.1		骨	10	0.1
	碗	12	0.1		換	10	0.1
	雙	12	0.1		煙	10	0.1
	椅	12	0.1	82	‧		
	湊	12	0.1		‧		
	飲	12	0.1	總計		13,525字	

表2　單字出現頻率對數值

［紅樓夢］第四十回

排名 r	出現頻率 f	估計字數 r×f	Ln(r)	Ln(f)	Ln(r)× Ln(f)	Ln(r)²
10	164	1640	2.3026	5.0999	11.7429	5.3019
20	111	2220	2.9957	4.7095	14.1085	8.9744
30	88	2640	3.4012	4.4773	15.2283	11.5681
40	74	2960	3.6889	4.3041	15.8772	13.6078
50	52	2600	3.9120	3.9512	15.4574	15.3039
60	45	2700	4.0943	3.8067	15.5858	16.7637
70	40	2800	4.2485	3.6889	15.6722	18.0497
80	34	2720	4.3820	3.5264	15.4526	19.2022
90	30	2700	4.4998	3.4012	15.3047	20.2483
100	28	2800	4.6052	3.3322	15.3454	21.2076
150	18	2700	5.0106	2.8904	14.4826	25.1065
200	14	2800	5.2983	2.6391	13.9826	28.0722
250	10	2500	5.5215	2.3026	12.7136	30.4865
300	8	2400	5.7038	2.0794	11.8607	32.5331
350	6	2100	5.8579	1.7918	10.4960	34.3154
400	6	2400	5.9915	1.7918	10.7353	35.8976
450	4	1800	6.1092	1.3863	8.4692	37.3229
500	4	2000	6.2146	1.3863	8.6153	38.6214
550	4	2200	6.3099	1.3863	8.7474	39.8151
600	3	1800	6.3969	1.0986	7.0277	40.9207
650	2	1300	6.4770	0.6931	4.4895	41.9512
700	2	1400	6.5511	0.6931	4.5409	42.9167
750	2	1500	6.6201	0.6931	4.5887	43.8254
800	2	1600	6.6846	0.6931	4.6334	44.6840
900	2	1800	6.8024	0.6931	4.7151	46.2726
1000	1	1000	6.9078	0.0000	0.0000	47.7171
1047	1	1047	6.9537	0.0000	0.0000	48.3537
		合計：	143.54	62.52	279.87	809.04
		平均：	5.32	2.32		

表3　字詞出現頻率對數值

［紅樓夢］第四十回

迴歸估計值		最小平方法
常數	8.390479	$n=\frac{-53.3748}{44.8752}=-1.1894$
Y估計標準誤	0.252891	
可決係數	0.974031	
觀察項	27	
自由度	25	$Ln(c)=8.6476$
		$c=5696.66$
迴歸係數	−1.14272	
係數估計標準誤	0.037316	
$r^{1.14272}$	$*f=4404.9$	$r^{1.1894}$ $*f=5696.66$

表4　字詞出現頻率對數值

［紅樓夢］第四十回

排名 r	出現頻率 f	估計字數 r×f	Ln(r)	Ln(f)	Ln(r)×Ln(f)	Ln(r)²
10	66	660	2.3026	4.1897	9.6470	5.3019
20	42	840	2.9957	3.7377	11.1971	8.9744
30	31	930	3.4012	3.4340	11.6979	11.5681
40	22	880	3.6889	3.0910	11.4025	13.6078
50	19	950	3.9120	2.9444	11.5187	15.3039
60	15	900	4.0943	2.7081	11.0877	16.7637
70	13	910	4.2485	2.5649	10.8972	18.0497
80	11	880	4.3820	2.3979	10.5076	19.2022
90	10	900	4.4998	2.3026	10.3612	20.2483
100	9	900	4.6052	2.1972	10.1186	21.2076
150	7	1050	5.0106	1.9459	9.7502	25.1065
200	5	1000	5.2983	1.6094	8.5273	28.0722
250	4	1000	5.5215	1.3863	7.6544	30.4865
300	3	900	5.7038	1.0986	6.2662	32.5331
350	3	1050	5.8579	1.0986	6.4356	34.3154
400	2	800	5.9915	0.6931	4.1530	35.8976
450	2	900	6.1092	0.6931	4.2346	37.3229
500	2	1000	6.2146	0.6931	4.3076	38.6214
600	2	1200	6.3969	0.6931	4.4340	40.9207
700	1	700	6.5511	0.0000	0.0000	42.9167
800	1	800	6.6846	0.0000	0.0000	44.6840
900	1	900	6.8024	0.0000	0.0000	46.2726
1000	1	1000	6.9078	0.0000	0.0000	47.7171
1466	1	1466	7.2903	0.0000	0.0000	53.1484
		合計：	124.4707	39.47895	164.1802	688.2426
		平均	5.19	1.64		

表5　字詞出現頻率對數值

[紅樓夢] 第四十回

迴歸估計值		最小平方法
常數	6.572146	
Y估計標準誤	0.148199	$n=\frac{-40.0358}{41.7762}=-0.95833$
可決係數	0.987618	
觀察項	24	$Ln(c)=6.613732$
自由度	22	$c=745.26$
迴歸係數	−0.95004	
係數估計標準誤	0.022678	
$r^{0.95004}$	$*f=714.9$	$r^{0.95833}$ $*f=745.2$

表6　字詞出現頻率對數值（不包含人名和稱呼）

［紅樓夢］第四十回

排名 r	出現頻率 f	估計字數 r×f	Ln(r)	Ln(f)	Ln(r)× Ln(f)	Ln(r)²
10	57	570	2.3026	4.0431	9.3095	5.3019
20	38	760	2.9957	3.6376	10.8972	8.9744
30	27	810	3.4012	3.2958	11.2098	11.5681
40	21	840	3.6889	3.0445	11.2309	13.6078
50	16	800	3.9120	2.7726	10.8464	15.3039
60	13	780	4.0943	2.5649	10.5018	16.7637
70	11	770	4.2485	2.3979	10.1874	18.0497
80	10	800	4.3820	2.3026	10.0900	19.2022
90	9	810	4.4998	2.1972	9.8871	20.2483
100	9	900	4.6052	2.1972	10.1186	21.2076
200	5	1000	5.2983	1.6094	8.5273	28.0722
300	3	900	5.7038	1.0986	6.2662	32.5331
400	2	800	5.9915	0.6931	4.1530	35.8976
500	2	1000	6.2146	0.6931	4.3076	38.6214
600	1	600	6.3969	0.0000	0.0000	40.9207
700	1	700	6.5511	0.0000	0.0000	42.9167
800	1	800	6.6846	0.0000	0.0000	44.6840
900	1	900	6.8024	0.0000	0.0000	46.2726
1000	1	1000	6.9078	0.0000	0.0000	47.7171
1427	1	1427	7.2633	0.0000	0.0000	52.7560
		合計：	101.9445	32.5478	127.5329	560.6189
		平均：	5.10	1.63		

表7　字詞出現頻率對數值（不包含人名和稱呼）

［紅樓夢］第四十回

迴歸估計值		最小平方法
常數	6.399539	$n=\frac{-38.7271}{40.4189}=-0.95814$
Y估計標準誤	0.174821	
可決係數	0.984917	
觀察項	20	
自由度	18	
迴歸係數	−0.93622	$Ln(c)=1.63+0.95814*5.1$
係數估計標準誤	0.027307	$=6.516514$
		$c=676.22$
$r^{0.93622}*f=601.57$		$r^{0.95814}*f=676.22$

表8　單字出現頻率曲線

	迴歸估計法			最小平方法		
排名(r)	1	3	10	1	3	10
出現頻率(f)	4404.9	1255.3	317.1	5969.7	1542.1	368.3
	斜率：−1.14272			斜率：−1.1894		

表8　字詞出現頻率曲線

	迴歸估計法			最小平方法		
排名(r)	1	3	10	1	3	10
出現頻率(f)	714.9	251.8	89.5	745.2	260.1	82.0
	斜率：−0.95004			斜率：−0.95833		

表8　字詞出現頻率曲線（不包含人名和稱呼）

	迴歸估計法			最小平方法		
排名(r)	1	3	10	1	3	10
出現頻率(f)	601.6	215.1	69.7	676.2	236.0	74.5
	斜率：−0.93622			斜率：−0.95814		

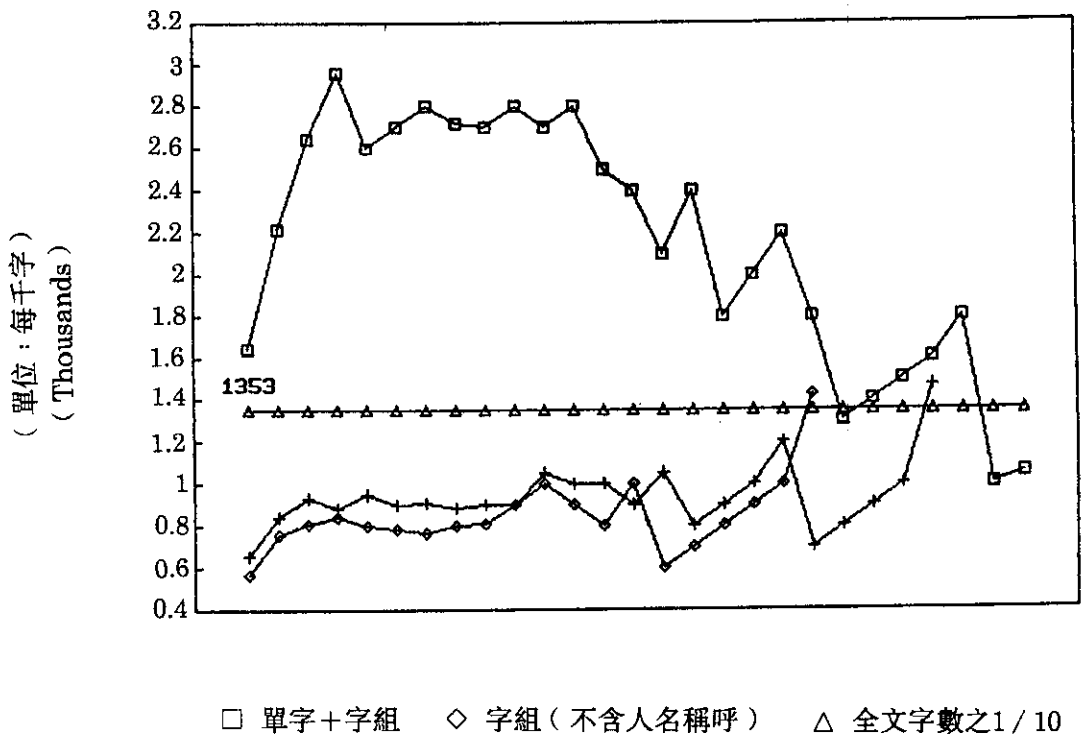

圖2　中文單字及字組分佈比較
［紅樓夢］第四十回

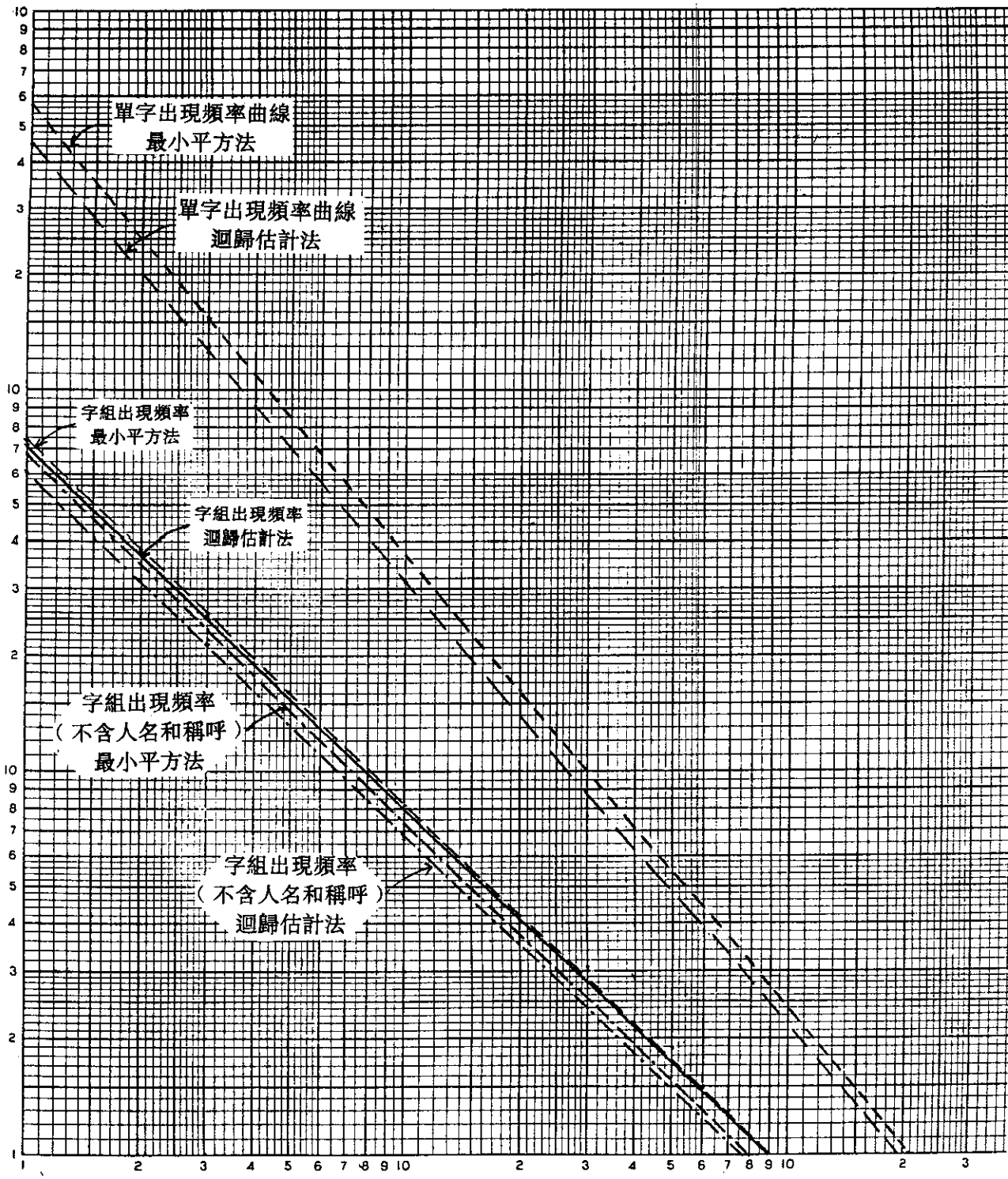

圖3　齊普夫律曲線
［紅樓夢］第四十回

表9

	迴歸估計法			最小平方法		
	估計係數值	理論係數值	差距	估計係數值	理論係數值	差距
單字	−1.14272	−1.0	0.14272	−1.18940	−1.0	0.18940
字詞	−0.95004	−1.0	−0.04996	−0.95833	−1.0	−0.04167
字詞（不含人名與稱呼）	−0.93622	−1.0	−0.06378	−0.95814	−1.0	−0.04186

、表4、表6）

2.所得斜率皆不等於－1（見表9）

(七)綜合結論：

齊普夫律基本上不適用於中文。

五、討論：

中文之所以被稱爲中文，就是中國文字有與他國文字不相同的文化特點。因此，凡以他國文字爲研究對象的題材或律例，都不一定能適用於中文。反之，亦然。

齊普夫律的研究對象，原本爲英文，才有所謂齊普夫律的發現。後來，他又借助他人不成熟的研究結果，認爲該律也適用於中文。在前一節中，筆者已確切證明齊普夫在45年前所作〝也適用於中文〞的推論，並不正確。

研究任何語言，必須研究它的基本音素與音節，同樣的道理，研究任何文字，也必須研究它的字、詞、句的結構。談到後者，英文字、詞、句的結構，與中文字、詞、句的結構，相去極遠。實不能一概而論。中國大陸學者曹聰孫教授在他最近的新著中曾這樣說（註9）：

〝無論那一種語言中的詞彙，都不可能反映整個世界的全部情況（事物、性狀、動作、變化），而只是反映使用這種語言的人所注意到的地方。人注意到的地方在不同的語言裡，當然是不同的。這樣，各種語言的詞彙，系統的內部結構與意義分配也就肯定是不一樣的。〞

齊普夫律是文字計量學（Quantita-

tive Linguistics）中的一種理論。我們談論文字計量，就必須先認清它的計量單位不可。

㈠文字計量與計量單位的問題：

齊普夫律、洛特卡律、甚至布萊德福律都是研究排名（Rank）和頻率（Frequency）之間的相對關係。齊普夫律講的是排名與文字的出現或使用頻率；洛特卡律講的是作者人數與其出版文獻數量，而布萊德福率講的則是區域與其涵蓋的期刊數量。在研討布萊德福律時，計量單位爲每一種期刊，很容易瞭解和接受。然而若仔細追究起來，問題仍然很多。譬如有關文獻的可能出現率與期刊的容量（篇幅）大有關係。一種月刊與一種年刊，對相關文獻的出現機率就會有很大的差別。這是題外的話，我們且別過不提。再就，洛特卡律中所使用的計量單位：作者人數（每人）與發表文章數量（每篇），真可就是世界性的單位，不會引起爭論。其實不然，一些不十分顯眼的問題，仍然存在。譬如，一篇連載性的文獻，究竟將它算成一篇呢？或是每期都算一篇？又譬如一篇文章爲x位作者共同撰寫，那麼究竟應該是每位作者都分配上1/x篇呢？或是將全篇都歸計到Xi作者一人名下？總而言之，不同的計量單位，必會獲得不同的結論。這似已成了不言可喻的真理。

齊普夫律是文字頻率的計算。乍看起來，也好像毫無爭論的餘地，一個字就是一個字！可是，若仔細推敲起來，仍不免有一大堆的問題。譬如英文與中文對下列各詞的用字便不相同：

一張桌子	4字	One table	2字
一 本 書	3字	One book	2字
一 條 魚	3字	One fish	2字
一雙鞋子	4字	One pair of shoes	4字
一 幅 畫	3字	One picture	2字
一件文件	4字	One piece of document（or one document 2字）	4字
一 碗 湯	3字	One bowl of soup	4字
合計	24字		20字

(1)若以單字計算，則分別獲得16字和12字：

一（7），張，桌，子（2）	one（7）,table,book
本　，書，條，魚	fish,pair,of（3）
雙，鞋，幅，畫	shoes, picture, piece,
件（2），文，碗，湯	document, bowl, soup

（括弧內之數字爲該字出現次數，無括弧者均只出現一次）

(2)若以字、詞計算，則分別獲得14字和12字

一張，桌子，一本，書，	英文字數不變
一條，魚，一雙，鞋子，	
一幅，畫，一件，文件，	
一碗，湯	

單字與字詞之間在數量上，雖然只有2個〝字〞的差別，然而每字的出現頻率和排名卻有了變化。在單字計算中，〝一〞字共出現7次，〝子〞和〝件〞字各出現2次。因此，〝一〞字排在第一，〝子〞字排在第二，〝件〞字排在第三，其他各字則依次排名4至16。相對的，在字詞的安排和計算之下，固無一重複，所以順著筆劃多寡，從1排到14。顯而易見，這二種計量方法所得的排名與出現頻率之積，就有了不同。筆者所以提出這些小問題，旨在說明①文字的計量研究，必須先確定計量單位；②文字是語言的化身。由於地區和國家的不同，語言和文字的組合與結構也都會有不同。論及文字，尤其是研究文字計量，我們既不能以偏蓋全的囫圇吞棗，也不必削足適履的投其方便。爲了避免這二種差誤，唯有先瞭解各種文字的特色，再從特色中，決定最適當的計量單位。文字的特色對文字的計量研究實有不可分割的親密關係。

㈡中國文字的字與詞：

中國有數千年文化，自從倉頡造字開始迄今也已經過了廿多個世紀。除去各地方言或有語言卻沒有文字的地方不談，我國現在究竟有多少個可以考據得出來的單字，好像還沒有十分可靠的答案。尤其是幾年前中國大陸簡體字風行了好一陣子，常常看到一

些像字又不像字的字，不知它們是新字，是簡體字，或是無中生有的別字、鬼字、訛字，真是亂糟一團。最近經過學者專家整理以後，情況已經好多了。根據在台北召開的第三屆中國文字學國際學術研討會的專題報告，歷代漢字的發展，從甲骨文到1990年，一共找出了54,678字。曹雪芹只用了其中之8.34%，或4,561個單字，便寫出了一本「千古奇書」。寫小說本就是作文字遊戲。就看作者怎麼樣將單字排列組合起來，將一塊「頑石」點化成樸玉。我們可以肯定的說，在文字遊戲上，曹雪芹確實是我國少見的天才。

中國文字蘊育在一個獨特的文化背景裡。它的保守性早在一百多年前就已爲一位美國傳教士所發現（註11）。幾千年下來，文字變化極少。直到民國八年的「五四運動」以後，中國的語言和文字才起了急劇的變化。宋元開始的〝白話〞，逐漸取代了文言；北京官話也逐漸變成了通行全國的普通話。再加上〝西化〞的結果，〝吸收了許多外來語和歐化的造句法，新的語言形式和新的思想內容互相隨伴著而來〞。（註12）〝五四運動〞前後，單字本身事實上並沒有甚麼新進展，倒是以單字組合成的〝詞〞和〝詞彙〞卻開始向前邁大步。

從計量上講起來，〝字〞是單個的。凡由二個以上的單字組合成的〝字〞才稱爲詞。有關〝詞〞的定義很多，筆者斗膽爲〝詞〞定下四個基本條件：①詞必須由二個以上的單字組合而成；②詞字組不可分割；③詞字組中之字序不可更易；④組成詞字組不可更改。現在我讓我們看看這些條件，對〝詞〞的形成有些甚麼樣的影響？對文字計量又會產生甚麼樣的結果。

(1)詞必須要由二個以上單字組合而成：單字不能稱爲〝詞〞，只能稱爲〝字〞。譬如〝一粒〞爲〝一〞和〝粒〞二個單字組成。分開來爲字，合組則成詞。分開來，〝一〞字爲基數，〝粒〞字爲圓形細小的東西。二字合在一齊成〝一粒〞，意指計算圓形小物的單位名稱。再譬如，〝話〞字指言語；〝梅〞字爲姓氏，也爲一種植物，早春開花、結子、生葉。二字合在一齊成〝話梅

〃，指酸梅，爲一解渴食品。由這二個例子，我們便可知道詞爲字之組合。若撤散開來，每個字都各有其義，與組合後之詞意不同。

(2)詞字組不可分割：單字組成詞以後，便不可再改變。否則原意盪然。譬如桌子，洗衣機，［紅樓夢］，這些都是詞，若將這些詞都分解成字，那就再也看不到原來詞意的影子。

(3)組成詞字組不可更改：詞的來源很多。但多由生活環境進化變遷和〝西語中化〃而來。譬如〝椅子〃，最初用〝倚〃字，後來才寫做〝椅〃（註13）。〝桌子〃最初稱爲〝卓子〃（〝卓〃意高而直，與几之矮小成對比）。後來才改爲〝桌子〃。此外如咖啡、可口可樂、雪茄、冰淇淋、白蘭地、芭蕾舞等中譯舶來品，由於日久〝約定成俗〃（註14）爲國人所接受。假如有人將咖啡寫成〝加非〃或將〝冰淇淋〃寫成〝冰其林〃，閱讀的人一定會丈二金鋼，摸不著頭腦。

(4)詞字組中之字序不可更易：詞中各單字，次序已定，不容隨意更動。否則原意會面目全非。譬如〝愚翁移山〃、〝投鞭斷流〃二句成語，不可改寫成〝移山愚翁〃和〝斷流投鞭〃。更改後的詞與原詞，在語義和語法上都產生了相當大的差異。

愚翁移山　　愚翁爲本體，移山爲客體

（笨老公公想搬移一座山）

移山愚翁　　移山爲本體，愚翁爲客體

（搬山的笨老公公）

投鞭斷流　　投鞭爲本體，斷流爲客體

（將馬鞭投入河裡，阻竭河水流動）

斷流投鞭　　斷流爲本體，投鞭爲客體

（爲了切斷河流，而將馬鞭投進河裡）

像這類本客體互換的結果，使得〝新〞詞的含意非常模糊。除了字組內之單字字序更改會影響原詞的語意以外，還會產生一些意想不到的後果。請看：

子椅、啡咖、樂口可可、茄雪、淋冰淇、地白蘭、蕾芭舞。

像這些詞彙有誰能懂！

〝詞〞可由單字組合而成，也可因語法的規定和需要而組成。譬如（註15）：

坐、坐了、坐著、坐坐

見、見了、見過、見見

看、看了、看見了、看見

過、看著

答、答應、答應了、答應著

稍前，我們提到過齊普夫將語法變換的單字如Give,Giving, Gave,Given都當作個別單字計算，那麼給，給著，給過，給過了，以及上述各詞也都應該算著一個字。這類由語法變化而成的〝字〞，實際上也都符合前面所提的四個條件。

㈢中國文字的計量：

中國文字有它與眾不同的特點。字與詞之間的分野有時非常模糊。因此，從事中國文字計量研究，只有二個途徑：一種從單字上著手，另一種則採用〝詞〞的方法，以〝詞〞爲計量單位。根據本文研究發現，以〝詞〞爲計量單位，顯然較以字的計量單位爲優越。然而，由於〝詞〞的組織層面極廣，涵蓋的獨特因素如方言、成語、口語、土話等等也特別多，因此，決定何者爲詞，何者又爲字，確實困難非常。它們的決定必須經過逐字分析推敲，否則中文文字計量的研究將很難達到理想的水平。

六、結論

齊普夫律是文字計量方法之一。嚴格的說起來，該律本身並無十分可取之處。該律的公式，漏洞也很多。因此，我們研究齊普夫律，應集中在它的研究方法上。利用同樣的方法，不僅可以做很多有關作品和作者的考證，而且還可以做很多資訊理論上的研究。

任何有關文字的計量研究，都必須先解決計量單位的問題。本文利用單字和詞彙二種不同的計量單位，進行了齊普夫律中文適用性的研究實驗，發現無一能夠肯定該律也能適用於中文的推論。不過，看我們細心比較二種計量結果，我們當會發現以〝詞〞爲單位的計

算結果，比較〝最接近〞齊普夫律的二個條件，也就是說：r與f之積約〝等於全文之0.1和斜率〝約〞等於－1（請參看圖3，表5及表9）。至於筆者的〝約等於〞是否就是中國大陸學者專家所謂的〝基本上〞，那就不得而之了。

註　釋

註 1：George Kingsley Zipf, Human Behavior and the Principle of Least Effort, An Introduction to Human Ecology（Cambridge, MA：Addison-Wesley Press, 1949）.

註 2：G. Herdan, Quantitative Linguistics（Washington D.C.：Butterworth, 1964）, 49.

註 3：在所有研究論之中，還未見有$\beta=1$或c值等於全文0.1的研究發現。這可是齊普夫律的根本弱點。

註 4：同註1，頁89－90。

註 5：王崇德，文獻計量學教程（天津，南開大學出版社，1990），頁179。

註 6：曹雪芹、高鶚，紅樓夢一二〇回（台北：聯經，民80年），共三册。

註 7：逐字排名序是每字依次排名，共1047名。單字排序，則爲相同出現頻率之單字同序，故共89名－－筆者註。

註 8：由於單元字的重複利用，因此，字詞組的〝字〞數較單字多出了419字－－筆者註。

註 9：曹聰孫，齊普夫律和語言的〝熵〞（天津：人民出版社，1994），頁63。

註10：謝清俊著，「二十五史的文字統計與分析」，第三屆中國文字學國際學術研討會，台北，民81年3月21日至22日。

註11：Rev.R.H. Graves, Forty Years in China（Baltimore：R.H. Woodward, 1895）, 54。（這是一本100年前的老書，從書裡可以見到百年前中國的模樣，值得一讀）

註12：黎錦熙，新著國語文法今序。

註13：呂叔湘，語文常談（香港：三聯書店，1982），頁67。

註14：語出荀子。

註15：詞例取自紅樓夢第四十回。