TR-92-005

廿五史的文字統計與分析

作者：謝清俊*、林晰**、許金定**、傅武坡***、張翠玲***

* 中央研究院資訊科學研究所 研究員，元智工學院客座教授

** 中央研究院計算中心 程式設計師

*** 中央研究院資訊科學研究院 研究助理

中華民國八十一年二月廿日

廿五史的文字統計與分析

作者：謝清俊*、林晰**、許金定**、傅武堉***、張翠玲***

* 中央研究院資訊科學研究所 研究員，元智工學院客座教授

** 中央研究院計算中心 程式設計師

*** 中央研究院資訊科學研究院 研究助理

中華民國八十一年二月廿日

目　　錄

壹、語文處理的基礎工程

語文的計量統計是呈現一種語言文字整體性質的有效方法。早期的語文統計工作包含字彙的蒐集、常用字彙的選取等工作，其統計規模都不大，數據亦不易確實，這情形大致是由於語料蒐集不易，以及其工作量依語料量之增加而大幅提升的緣故（非線性關係）；在沒有適當的工具以前，似乎不可能從事大規模、高品質的語文統計工作。然而，由於一些應用須要有些關於字彙、詞彙的數據，如編字書、編啓蒙教材、準備活版鉛字等等，於是便有些語文統計工作之肇始。

自從1949年，仙農（C. E. Shannon）的消息理論（Information Theory）發表以後，賦予語文統計全新的方法和詮釋能力，奠定了語文統計的理論基礎和地位，於是語文統計邁進了一個新天地 [1]。仙農的理論是從工程的立場出發，譬如要設計一個高效能的語文傳輸系統，在設計之初我們並不能限制（或明白）這個系統將來要傳送什麼樣的文句（指內容而言）；於是只能依語文整體的統計性質作爲設計機體的準繩，並研擬出量測資訊量的方法以爲工程設計之基礎。據此，發展出編碼的理論（Coding Theory），加密的技術（Encryption），資料壓縮技術，通信通道容量的估算和對雜訊之處理方法等等。甚至有將之應用於拼字遊戲的設計和紅樓夢作者的鑑別等特殊應用之上 [2]。而事實上，其理論發展之概念影響深遠，在傳播、教育、人工智慧、電子通信、電腦設計等等領域中皆有不可或缺的貢獻[3]，而這種種成就的基礎功夫即語文統計。

自從1960年代計算語言學（Computional Linguistics）興起之後，語文統計更成爲用計算機分析語言文字最基本的工具。此時，語文統計更與計算機內語法（grammars）之表達與分析密切結合 [4]；由另一方面來看，由於計算機的普遍使用，也使得語文統計工作有了強有力的工具而如虎添翼，在以前許多繁鎖量大極耗人力的工作便由計算機承擔了，因而改變了人們對語文整理的觀念和態度，譬如文獻語言學（Corpus Linguistics）的誕生就是很好的例子[5.6]。

文獻語言學的基本觀點是要從實際大量的語料中去分析語文的性質，因此它能夠發前人之未見，且對語言文字做較徹底、較精確的分析，並藉此導出新的理論 [7]。此外，它也使我們對於語言應用的態度起了改變，譬如給一本書編個字典、引得、索引、字彙表等等並非難事（利用電腦來做）。

以上所談的都是由語文統計爲基礎而孳生出的應用領域，是故稱呼語文統計爲「漢語處理的基礎工程」並不爲過[8]。

貳、漢字語文統計之回顧

我國文字歷史悠久，就現在可見的殷墟甲骨文迄今，也有三千多年。語言文字是時代的公器，隨社會之演進而有生滅，亦而生生不息。我國文字歷代遞增的情形可依據艾偉 [9]及黃得時 [10]之資料匯整並增補 [11.12.13]如表一所示。

在表一中蒐集的字彙，大體上均依印刷的字形爲對象，是故原則上不收錄書法上或字體上之字形變化，然而在字形的認定上卻不能夠那麼清楚地劃分。是故這些統計在異體字的認別上並不一致，在參閱時宜加注意。這些蒐集亦因異體字形處理的方式不同而影響到其精確的程度。大體而言，做漢字統計的工作最不易做好的就是對相關字形的認定：若從字的外觀來劃分，則細部差異者甚衆而難以取捨[14]；若依字義而分，則涉及文字之孳乳考據，以致於牽涉太廣而難以定論。其實，這個問題就是字的定義的問題，在沒有釐清字的定義之前，所有的文字統計工作將缺少一個共通的衡量參考。

根據表一的字彙，我們並不能夠知道歷來或現在一共有多少漢字。表一中最大的字彙大致告訴我們一個已知的上限。其實，還有比表一中更大的字彙。例如中文資訊交換碼中收集到75684字 [12]，又如全漢字庫號稱將蒐集十餘萬字[15]。然而，這些大字彙集的字還須仔細核對，不像列在表一中的字彙已經經過嚴謹的校對。

表一中所列的字彙並沒有每個字使用頻度的資料。對於漢字使用頻度的統計工作，可溯至1856年上海長老會印刷局William Gamble主持的《文言字彙》編纂工作[16]。在此之前，是有常用字彙之編輯記錄，如《千字文》，然而卻無每個字的頻率。若能將古時漢語文字的頻度統計出來，相信對歷史語言學的研究和對古文獻處理工程均將有所助益。

我國對漢語字、詞之使用頻度的研究工作起步甚早，甚至先於英、美、法、西班牙等各國著名的字詞彙統計。從民國十年至抗日戰爭以前的十幾年是一段蓬勃發展的時期，此期間對中文字詞之統計和在教育心理學上的應用甚有成就。詳請參閱艾偉的《漢字問題》 [9]。在此期間，工作的內容大體上均描準一般性的常用字彙，目的在解決語言在應用和教育上的一些基本問題；統計的母體均不大，在數十萬字左右，而字形之差異問題並未受到重視。

表一　歷代漢字字數遞增情形統計表

時代	西元	書名	作者	所收單字	遞增字數
殷		甲骨文		約 3000字	
秦		倉頡篇	李斯	3300字	300字
漢	1-5	訓纂篇	揚雄	5340字	2040字
漢	60-70	續訓纂篇	班固	6120字	780字
漢	100	說文解字	許愼	9353字	3233字
魏	227-239	聲類	李登	11520字	2167字
晉		字林	呂忱	12824字	1304字
後魏		字統	楊承慶	13734字	910字
後魏	480	廣雅	張揖	18150字	4416字
梁	543	玉篇	顧野王	22726字	4576字
唐	751	唐韻	孫愐	26194字	3468字
唐	753	韻海鏡源	顏真卿	26911字	717字
宋	1037-67	集韻	丁度	約 30000字	約 3081字
宋	1066	類篇	王洙等	31319字	1319字
明	1615	字彙	梅膺祚	33179字	1860字
明	1675	正字通	張自烈	33440字	261字
清	1716	康熙字典	張玉書等	42174字	8734字
民國	1915	中華大字典	中華書局	44908字	2734字
民國		大漢和辭典	諸橋轍次	48902字	3994字
民國	1969	中文大辭典	張其昀等	49888字	986字
民國	1986	中文資訊交換碼第三冊	國字整理小組	53940字	4052字
民國	1990	漢語大字典	徐中舒等	54678字	738字

抗戰勝利之後大規模的文字統計工作消沈了一段時期，在方言調查方面卻引用了抽樣和統計的方法[17]。在台灣，早期有民國四十六年國立編譯館的《國民小學常用字彙表》，之後直到民國五十九年底才有新聞界中文報業協會的《新聞常用字彙》和民國六十一年林樹的《中文電腦基本用字的研究》；這些都是字彙及其頻度的統計工作。在詞彙方面，則以民國六十四年劉英茂等《常用中文詞的出現次數》爲主要的數據依據[18]。至於此後發表之字彙，如教育部陸續公佈之常用字，次常用字,和罕用字表,中文資訊交換碼的各個字集,國家標準CNS 11643的 13051字集等等，則因缺少字頻資料而無法做分析之用。

在大陸方面，文字和詞彙的統計做得比台灣多，而且成就比較大。由於缺乏文獻引徵，早期的工作無法描述，而近期的成就，擇其要者則有：(1) 1986年出版的《現代漢語頻率詞典》 [19]，語料母體約200萬字次，並分別作了字與詞的統計；(2)北京航空學院等十個單位共同完成的漢語詞頻統計工程[8]，語料母體約三億字次，經抽樣選擇了約2500萬字次做統計工作，並將抽樣之材料分爲四個時期:1919-1949,1950-1965，1966-1976，1977-1982分別統計；在內容上亦分爲社會科學和自然科學兩大類，每大類再分爲五個子類分別統計。這種大規模分年代，別類型的統計工作應屬首創。

在統計內容方面，除了有字詞的各別使用頻率外，對其累積頻率、字與詞之覆蓋率、各種資訊量之量測（各種情況之機率probability，各種熵entropy，重覆率等）以及對於 Zipf分佈經驗公式之驗證等，皆有很值得參考的成就 [8.19.20.21]。要言之，大陸上對語文統計工作已進入了仙農開創的對資訊量的量測以及計量語言量測的領域，這是台灣沒有能做好的。

在國外方面,應用馬可夫模式 (Markoff Model) 以N-連(N-gram)方法斷詞則做得很成功[22]，這方面的研究，在海峽兩岸也都做了一些，成績都很好。

誠然大陸上的語文統計比此地做得多、做得好，然而其統計成果主要的缺點仍在於對字和詞的定義沒有釐清，以致於兩岸在文字統計方面的成就，舉例而言，均不足以支持編好一個漢字用的交換碼，在其他語文應用方面的影響力也有待彰顯。

參、研究計劃簡介

本計劃是利用業已完成的廿五史全文資料庫所做的基礎文字統計工作，其目的是多重的，希望藉此工作能夠：

- 多了解一些國語文的性質
- 發展出全文資料庫能共用的統計軟體
- 建立語文統計分析之能力以支援各種語文之應用

上述的目的與廿五史文字資料的性質，以及其計算環境都有關係。茲將各點分述如下：

一、資料背景

廿五史記載的是中國官方核定的歷史資料；其中最早的一部史是《史記》，成書於公元前93年；最晚的是清史稿，成書於民國16年(1927)；以成書年代來算，它跨越了2020個年頭。關於文體方面則都是文言文，而且在體裁上都遵循《史記》的形式撰寫。關於廿五史的作者、成書年代等一些資料請參閱表二。

統計廿五史文字有下列的意義：(1).它跨越了很長的年代，因此可觀察隨年代而產生的文字變遷狀況；(2).它的體裁一致，記載的事務有明顯的相關性，較有相同的比較基礎；(3).古代的文字沒有頻度統計資料，希藉此產生一些頻度數據以爲參考。雖然對古文之頻度計量若只採廿五史的樣本將較偏狹，但這只是起步，以後尚可增補。譬如十三經或漢以前文獻若能變成機讀形式，就可添加漢以前文字統計之頻度數據；(4).白話文的文字統計資料已經有一些了，我們可以把文言文的資料和白話文的加以比對，從其異同中更可獲得我們漢語的一些性質[16.18.19.20]。

在我們這個統計工作中，可將資料依年代、作者、文體、記事對象等分別作統計。廿五史全文資料庫裡是採用鼎文版，經標點和分段落之後的廿五史，它包含有三家註的資料。因此，我們可以以全部書爲單位，或抽取其原作者所撰的部份，或選原作者所著之某一類文體，如列傳，來分別統計。

在文字的正確性方面，廿五史全文資料庫的每一段文章及註釋都經過四至五次詳細的校對，錯誤率可以說已很低，是故原始機讀資料的文字應有相當高之正確性。

表二　二十五史相關資料

史別	卷數	撰者	成書年代	記載年代（範圍）
a 史記	130 卷	[西漢]司馬遷	漢武帝太始四年(93B.C.)	黃帝－漢武帝
b 漢書	120 卷	[東漢]班固	漢章帝建初八年(83A.C.)	漢高帝元年(206A.D.)－王莽地皇四年(23A.D.)
d 後漢書	120 卷	[南朝宋]范曄	宋文帝文嘉二十二年(445A.D.)	東漢
e 三國志	65 卷	[西晉]陳壽	約晉武帝太康十年(289A.D.)	魏、蜀、吳
g 晉書	130 卷	[唐]房玄齡等	唐太宗貞觀二十二年(648A.D.)	宣帝司馬懿(179A.D.)－宋武帝劉裕取代東晉(420A.D.)
f 宋書	100 卷	[梁]沈約	齊武帝永明六年(488A.D.)	東晉安帝義熙元年(405A.D.)－宋順帝昇明三年(479A.D.)
i 南齊書	59 卷	[梁]蕭子顯	梁武帝天監十三年(514A.D.)	南朝齊(479-502A.D.)
m 梁書	56 卷	[唐]姚思廉等	唐太宗貞觀十年(636A.D.)	南朝梁(502-557A.D.)
n 陳書	36 卷	[唐]姚思廉等	唐太宗貞觀十年(636A.D.)	南朝陳(557-589A.D.)
h 魏書	130 卷	[北齊]魏收	北齊文宣帝天保五年(554A.D)三月紀、傳先奏上，十一月復上十志	北魏
k 北齊書	50 卷	[唐]李百藥	唐太宗貞觀十年(636A.D.)	東魏建立(534A.D.)－北齊亡(577A.D.)
l 周書	50 卷	[唐]令狐德棻等	唐太宗貞觀十年(636A.D.)	北周
o 南史	80 卷	[唐]李延壽	唐高宗顯慶四年(659A.D.)	宋永初元年(420A.D.)－陳禎明三年(589A.D.)
j 北史	100 卷	[唐]李延壽	唐高宗顯慶四年(659A.D.)	魏登國元年(386A.D.)－隋義寧二年(618A.D.)
s 隋書	85 卷	[唐]魏徵等	紀傳成於唐太宗貞觀十年(636A.D.) 十志成於唐高宗顯慶元年(656A.D.)	
y 舊唐書	200 卷	[後晉]劉昫等	後晉出帝開運二年(945A.D.)	唐(618-907A.D.)
x 新唐書	225 卷	[宋]歐陽修 宋祁	宋仁宗嘉祐五年(1060A.D.)	
t 舊五代史	150 卷	[宋]薛居正等	唐太祖開寶七年(974A.D.)	朱溫稱帝(907A.D.)－北宋建立(960A.D.)
w 新五代史	74 卷	[宋]歐陽修	宋神宗熙寧五年(1072A.D.)	後梁開平元年(907A.D.)－後周顯德七年(960A.D.)
p 宋史	496 卷	[元]脫脫等	元順帝至正五年(1345A.D.)	宋建隆元年(960A.D.)－祥興二年(宋亡,1279A.D.)
q 遼史	116 卷	[元]脫脫等	元順帝至正四年(1344A.D.)	遼(907-1125A.D.)
r 金史	135 卷	[元]脫脫等	元順帝至正四年(1344A.D.)	金
z 元史	210 卷	[明]宋濂等	明太祖洪武三年(1370A.D.)	元太祖稱成吉思汗(1206A.D.)－元順帝至正二十八年(1368A.D.)
u 明史	332 卷	[清]張廷玉等	清高宗乾隆四年(1739A.D.)刊行	明太祖洪武元年(1368A.D.)－崇禎十七年(1644A.D.)
v 清史稿	529 卷	[民國]趙爾巽等	民國十六年(1927A.D.)	

二、廿五史全文資料庫的環境及統計功能

廿五史全文資料庫有好多個版本，最近使用的是在AT&T公司出產的3B系列迷你計算機上建立[23]。它的操作系統是UNIX System V，中文字部份的擴充是由宏碁公司出版，其基本字彙是BIG-5的13051個字再加上3923個補充字彙，是故總共約有16974字的支援能力，用的字碼是SHIFTED BIG-5。

廿五史全文資料庫是以自己發展的一個資料管理系統(DBMS)，稱爲「中文文獻處理系統」（簡稱爲CTP，即Chinese Text Processor），而製作的 [24]。CTP 是一個通用的全文資料庫DBMS，不僅廿五史的每一個史的全文資料庫皆用它製作，我們還用它做了些其它的全文資料庫，諸如：示範用的《大藏經全文資料庫》，其中包括三論宗主要的三部經典，即《中論》，《十二門論》，及《百論》，總共七卷，約十萬字次，其中包括巴利文的註釋；一個廿五萬字的《報紙新聞資料庫》[25]；以及正在製作中的《十三經全文資料庫》和《台灣地方誌全文資料庫》等等。

在 CTP中，以樹狀結構來表達文章的內容結構，譬如各卷、篇、章、回、節和段落等；亦用另一樹狀結構來記載原書排版的結構，如頁次、行數等。在 CTP中，上述的兩種結構劃分得很清楚；文章內容結構不容更易，而排版結構可視需要修改。文章經上述方式分割後，其最小也是最基本的結構單位稱爲「段落」或「段」。其實，「段」不易清晰地界定；在本系統中所稱的段，原則上，就內容而言求其所記載完整，就長度而言希望在200至500字以內，這是爲了便於查詢、顯示，作索引和節省計算資源、增加效率等等的緣故。

我們軟體統計程式的設計是針對著上述的文章資料結構而設計的。換言之，它是一個附屬在CTP之下的統計軟體：任何用CTP所製作的全文資料庫都可以用這個統計軟體來做以「段」或「段」以上結構的文字統計 [26.27]。譬如：我們可以指定統計某一段的文字，也可以選一些段來做統計；可統計某一卷書的文字，一些卷的文字，或者是選某一個史爲單位，甚至是全部的廿五史。當然在以上選擇時，還可以有選正文（作者撰的）、或校勘（三家註或後人加的），或標題、或內文等等的自由。

在統計功能方面，目前已有的主要功能是：每個字使用的次數和頻度，字彙的累積頻度，頻譜之繪製，依成書年代表現某些字的使用頻度變化，和與內建的一些字彙集合作比對的工作等。正在設計中的則包括各種熵及機率的量測軟體，馬可夫程序相關的機率及熵的計算的軟體等。關於目前已用內建的字彙集合清單則請參閱表三。

三、異體字與符號之整理

在前面已經提到：作文字統計時必須先對文字的認定方法（即文字的定義）先加界定，否則統計的結果將產生誤差，並且極易造成錯誤之詮釋。在這節裡我們特別報告這方面的經驗以供各方參考。

製作電子化的文獻，首先要決定的是：文件中那些訊息應該保留，那些應該割捨；其次才是設計用什麼方法來表達應該保留的訊息。以一本書中包含的訊息爲例，則不難發現其種類繁多且式樣各異；僅就文字有關者而言，就已經涉及甚廣了。以下則就文字及符號部份所涉及者，包括字體、字形變異之處理方式，特殊符號之表達方式，標點符號之混淆如何解決，格式控制之表達方式，以及錯誤之檢查與校正等等來做說明。

古書中之字體、字形和今日計算機中使用之字形有極多的差異。例如『者』與『者』這類相異的狀態如何處置？它不僅影響到電子文件和原文件差異的程度，更嚴重地影響到文件之檢索；當然，也影響到文字統計的結果。像這樣的字，算一個字呢還是兩個字？如果從文件內容而言，應算是同樣的字；但是從文字學研究的立場，卻寧可保留原文件中原來之字形。因此，在本統計工作中，遇到這樣的字則以兩者皆可的方式任由使用者選擇；使用者若建立一個〈異體字對照表〉，則統計時就將之合併計算，若在表中所無者，則分別計算。至於文件檢索方面，則建議在CTP的下一個版本中，加入異體字對照表的功能，以增加檢索的彈性。

其次是沒有的字如何處理的問題。沒有的字必須造字，而造字之規則和工作程序必須事先妥爲規劃。在廿五史中一共造了3016個字。新造的字一多，如何檢查有無重複造字？則又是一個要面對問題。在打字時，若發現沒有該字形，當然要依原書中字形造字；但是在字造成以前，如何臨時表示該處有字待造？是什麼字？在原來那裡？以及造好後如何將之補回等等，都是要預先仔細規範之事。通常這類的事，若無事先妥爲規範，將造成工作之中斷，或造成各個打字人員以不同的方法處理它的混亂情形，而待產生混亂後再事後補救則其工作十分艱鉅，且所費龐大。

關於特殊符號方面，讓我們舉個例子說明。鼎文版廿五史中有所謂「補文」與「贅文」。補文是原文未見而被考據後認爲應該補入才妥當者；贅文反之，乃公認重贅該刪節之文字。在鼎文版中，補文以『〔』及『〕』表示，贅文則以『（』及『）』標示。然而〔、〕、（、）等亦作其他用途，這種情形將會使程式不易認別補文贅文。處此情境，可能在打字時另選補文與贅文所用的標示符號是較好的解決方式。像這類之規範，必須在做全文資料庫以前，將文件之表達方

表三　內建的字彙集合

序號	字彙集合名稱	字數	說明
1	國民學校常用字彙表	2738	1.內建的字彙集合可以依需要增加，字彙的數目並無限制
2	教育部標準國字常用字彙表	5343	
3	聯合報自動排鑄機常用字彙表	2345	2.此表中前七個字彙取自馬立君論文[25]
4	聯合報電腦排版常用字彙表	2358	
5	中國時報鉛字常用字彙表	1987	
6	中國時報電腦排版常用字彙表	2578	
7	新聞常用字彙表	2962	
8	CNS11643	13051	

表四　二十五史資料庫符號頻率表

	符號	頻率	符號	頻率	符號	頻率	符號	頻率
正常符號	(空白)	846294	0	33052				
	，	3268909	、	441867	？	30454	！	18658
	．	1756601	；	59447	：	219310		
	〔	179955	〕	179956	（	15525	）	15527
	「	13722	』	13721	『	290196	」	289896
	『	14977	」	14983	★	77991	■	79068
	▲	1225	□	565	●	6983	＝	229
異常符號	．	329	．	16	：	197	，	23
	、	21700	；	1	：	3	？	351
	ヽ	1	…	58	○	94		
	」〔	1	」〕	1	」	42	」	46
	（．	1	）	1	（	45	）	45
	／	3	，	10	．	1	．	4
	」	1						
	一	32	—	8	—	16	—	1
	～	4	|	6				
	厶	1	x	1	Y	5	X	2

說明：1.表中正常符號部份是廿五史中常態使用且經校核無誤者，其中

○表示全形零

□表示原書中缺漏字

●表示待補（等待造字）的字，依表中數目顯示，製此表時尚有6983字次（約1200不同的字）待補

2.正常符號中有六對成雙使用的符號，即〔 〕、（ ）、「」、『』、★ ■與▲ ■，最後兩對符號用以標示夾在正文中的小字注文以及大字。

3.異常符號部份是指在文中查出有誤而待改正者，例如『。』有 329個『．』有16 個;這些都應該是在正常符號中的『。』（1756601）。這些符號形成的原因 之一是在字碼中有這些符號（譬如好幾種句點），經打字人員誤用， 亦有因不熟習規則而錯用。也有事先未曾規定而隨便用者，如短橫符號 事先未發覺，而使用的結果有『－』(32個)，『—』(8個)，『-』(一個)，『 』(4個)等 。這些均有待劃一。

式詳爲規劃，並依之編成輸入規則。此外，校注文字之標示亦可能有類似的情形，也有待規範。

在標點符號方面，有許多小問題須事先釐清。像書名號、私名號、斷節號等與文字重複的符號如何處理，便是一問題。此外，有些中文的標點符號與 ASCII 中符號極易混淆，如『・』與『．』，『：』則有兩個，諸如此類，亦待訂定規範方不致於表達不一。

在格式控制方面，則包括空白、空行、空頁之表示方式（目前字集中無此類控制符號），頁碼之標示方法，行位之標示方法，以及大小字之標示方法等等。

以上這些處理細節，若做得好，均可提供檢索和統計上許多不同之選擇；對系統之適用範圍功能和品質方面均有極大助益。廿五史全文資料庫在當初製作時，這些事沒有規劃得很好，是一邊做一邊訂正，因此至目前仍有一些問題正在修正中。關於符號之混淆情形及其頻度之統計請參閱表四。

對於以上種種可能出錯的問題，可以設計一些程式來協助校對，也可以利用程式將其一部份錯誤加以更正。例如『（』和『）』之匹配，或任何成對符號之匹配，以及表四中短橫的五種不同打法(表四中倒數第三行至倒數第二行處)等等，目前都是以程式處理的。其實表四中大部份異常符號之偵測亦是由程式尋得。

至於廿五史中之造字表和異體字表，則由於篇幅過大，在此謹以表五顯示一其部份之情況。要言之，字形極類似者，即音義皆同而僅些許字形差異者共 183 字，這些字在統計時合併計算；字集中已有該字，而字形明顯地不同，建議建立異體字表處理者共計1442字形，合併後爲 687字。至於造字時不慎而重複者，則有四十字共92形，目前這些重複造的字已經更正。關於異體字之類別，則依字書之稱呼計有：俗字、亦作、同、本字、古字、別體字、互通字、或作、譌字、簡字、即、一作、省作、之誤、本作等。目前，這些建議以異體字表合併之字僅限於已發現者，以後很可能陸續將會再添加。

表五　新造字及異體字表之部份例

鰽：鰽（FCFB）
蟁：蚊（FCFA）同
：滬（FCF9）
珌：璋（FCF8）古
：𣪞（FCF7）
暢：暘（FCF6）同
：𪗂（FCF5）
：譪（FCF1）
樗：樢（FCF0）同
鶴：鸖（FCEF）同
冉：冄（FCEE）同
：縐（FCED）
：騟（FCE9）
篡：簒（FCE8）俗
：餁（FCE7）
：䵷（FCE6）
雷：靁（FCE5）古　䨻（FCDF）古　𩆦（C865）同
棋：棊（FCE4）本
顰：顰（FCE3）同　𩕄（FCCD）同　𩕊（C8BA）同
：騐（FCE2）
留　　䚧：駵（FCE1）同　騮（C786）同
譽：佫（FCE0）同
：冣（FCDD）
：叜（FCDC）同　叟（FCDB）
：陁（FCDA）
：㚣（FCD9）
：偲（FCD7）
：㪯（FCD5）　傲（A0FE）
浸：㴆（FCD3）同
額：頟（FCD2）同
：埭（FCD1）
：尸（FCD0）
：鐈（FCCF）
舞：儛（FCCE）同
：䑛（FCCC）
聃：聃（FCCB）同
：鴹（FCC9）
：楖（FCC8）
蛙：鼃（FCC7）同
鴃：鴂（FCC6）同　鴂（FCAF）同
慫：慫（FCC5）俗　慫（C89C）俗
：劮（FCC4）
：䱷（FCC3）
灘：歎（FCC2）俗
蛇：虵（FCC1）俗
：詷（FCC0）
：𡾋（FCBE）
：眭（FCBD）
：䁃（FCBB）
屑：屑（FCBA）本
：撥（FCB9）
嵕：㚇（FCB7）同
：縆（FCB6）
：抖（FCB4）
：鎇（FCB3）
：㕊（FCB2）
：徐（FCB0）

說明：表中在冒號『：』左側是電腦中文字集中已存在的字，右側是新造的異體字之字形，字碼及異體性質的描述（如同、古、俗等）。在一行裡有兩個或兩個以上者，爲建議列入異體字對照表中之字，如：蟁、蚊，又如：雷、靁、䨻、𩆦等四字。在一行中僅一字者爲新造之字。

肆、統計舉要

以人工估算，廿五史及三家註總共約有5650萬餘字次，然而經計算機逐字計數，含標點符號但不算空白格，共計爲4061萬0593字次，其中漢字共3140萬9450字次。標點符號共計920萬1143字次；佔總字次之 22.25%，與漢字總字次之比爲28.62%。這些數據對於古書不分段亦不加標點的理由作了一個很好的解釋。對於全部書的統計請參考表六。

二十五史總共用了13966 個不同的字，此即稱爲廿五史的字集字數或字彙字數；其中在BIG-5碼或 CNS11643中已有的字共9951個，而不在其中的共有4015個字。這4015個字在罕用字中的有 999個（即在前文所述的附加3923餘字字集中），以及新造字共3016個。如此看來，CNS11643或 BIG-5 所選的13051字彙並不適合用於處理文言文資料。

廿五史共計3778卷，共用了27個不同的標點及符號（當直向顯示文件時，程式自動將所有標點及符號轉換爲直向之形式，這些直向顯示用的標點和符號並未計入在內）。在全文資料庫中總共的段落數目爲22萬1736段，總共頁次計爲 9萬4545頁（其中部份表格未建在全文資料庫中，是故所有表格均未列入統計）。關於其餘統計之要項則分述如下。

一、依體裁之統計

在表七所列爲正文部份的統計資料。所謂正文，即爲作者所撰之文章，不含序、跋、校勘、註釋等等。在表八中顯示的是各史列傳部份的統計，其中總字次爲2052萬6253；而漢字計1663萬2388字次，標點爲 389萬3865次。表中之比值爲總字數除以字集字數；平均句長爲在二個標點符號中間字串之平均長度；各節平均值爲每個段落平均之字數；各節字集數爲各平均使用了多少個不同字的數目。也許有讀者會認爲：爲什麼只做兩個標點符號之間字串長度的統計，而不做句子長度的統計呢？坦白說這是因爲有些句子實在不易界定的緣故，若有可行的句子定義供計算機使用，則可做句子長度的統計了。

由於本統計軟體可以相當自由的選定統計範圍，例如表九則是以史記中之本紀爲範圍之統計，這樣依體裁不同而設計的統計有利於文學方面的分析。

表六　二十五史文字統計表——全部書

製表日期：80/10/08

正史內文	分項	字數 a 中文字	字數 b 標點	字數 c 小計	中文字集數 d	平均每字使用次數 e
史記	1本 文	533505	162655	696160	5122	104.160
	2校勘及注釋	451981	189382	641363	5825	77.593
	3小計	985486	352037	1337523	6276	157.025
漢書	4本 文	742298	231257	973555	5833	127.258
	5校勘及注釋	516374	286959	803333	6507	79.357
	6小計	1258672	518216	1776888	6716	187.414
後漢書	7本 文	894020	292347	1186367	6161	145.110
	8校勘及注釋	435865	183335	619200	6161	70.746
	9小計	1329885	475682	1805567	6859	193.889
三國志	10本 文	377807	91762	469569	4388	86.100
	11校勘及注釋	321518	104626	426144	4556	70.570
	12小計	699325	196388	895713	5125	136.454
晉書	13本 文	1158126	281792	1439918	5997	193.118
	14校勘及注釋	71085	32713	103798	2521	28.197
	15小計	1229211	314505	1543716	6055	203.008
宋書	16本 文	811893	211640	1023533	5842	138.975
	17校勘及注釋	102567	39612	142179	2894	35.441
	18小計	914460	251252	1165712	5940	153.949
南齊書	19本 文	299257	80115	379372	4962	60.310
	20校勘及注釋	53666	24950	78616	2628	20.421
	21小計	352923	105065	457988	5064	69.693
梁書	22本 文	294438	71010	365448	4973	59.207
	23校勘及注釋	27613	11355	38968	1866	14.798
	24小計	322051	82365	404416	5012	64.256
陳書	25本 文	163382	38644	202026	4033	40.511
	26校勘及注釋	17435	7623	25058	1541	11.314
	27小計	180817	46267	227084	4088	44.231
魏書	28本 文	998329	251382	1249711	5417	184.296
	29校勘及注釋	133719	56150	189869	2958	45.206
	30小計	1132048	307532	1439580	5601	202.115
北齊書	31本 文	212506	49825	262331	4032	52.705
	32校勘及注釋	54501	20963	75464	2127	25.623
	33小計	267007	70788	337795	4145	64.417
周書	34本 文	262659	65296	327955	4161	63.124
	35校勘及注釋	87763	37566	125329	2486	35.303
	36小計	350422	102862	453284	4328	80.966

正史內文	分項	字數 a 中文字	字數 b 標點	字數 c 小計	中文字集數 d	平均每字使用次數 e
南史	37本 文	677624	160463	838087	5376	126.046
	38校勘及注釋	61365	25887	87252	2681	22.889
	39小計	738989	186350	925339	5442	135.794
北史	40本 文	1106543	267954	1374497	5572	198.590
	41校勘及注釋	163759	66282	230041	3115	52.571
	42小計	1270302	334236	1604538	5666	224.197
隋書	43本 文	701698	180673	882371	5592	125.482
	44校勘及注釋	19646	9594	29240	1754	11.201
	45小計	721344	190267	911611	5624	128.262
舊唐書	46本 文	2002600	469113	2471713	6346	315.569
	47校勘及注釋	88855	35497	124352	2603	34.136
	48小計	2091455	504610	2596065	6373	328.174
新唐書	49本 文	1694794	448400	2143194	6771	250.302
	50校勘及注釋	33492	12087	45579	1881	17.805
	51小計	1728286	460487	2188773	6795	254.347
舊五代史	52本 文	790879	177447	968326	5109	154.801
	53校勘及注釋	52652	23591	76243	2183	24.119
	54小計	843531	201038	1044569	5143	164.015
新五代史	55本 文	291476	79182	370658	3909	74.565
	56校勘及注釋	17970	6174	24144	1309	13.728
	57小計	309446	85356	394802	3937	78.599
宋史	58本 文	3980123	1007878	4988001	7389	538.655
	59校勘及注釋	190879	70260	261139	3312	57.633
	60小計	4171002	1078138	5249140	7428	561.524
遼史	61本 文	296254	89675	385929	4071	72.772
	62校勘及注釋	38077	12435	50512	1847	20.616
	63小計	334331	102110	436441	4120	81.148
金史	64本 文	931070	225142	1156212	5264	176.875
	65校勘及注釋	98102	39184	137286	2557	38.366
	66小計	1029172	264326	1293498	5316	193.599
元史	67本 文	1611849	418384	2030233	5854	275.341
	68校勘及注釋	116797	47511	164308	2617	44.630
	69小計	1728646	465895	2194541	5880	293.987
明史	70本 文	2802544	719292	3521836	7124	393.395
	71校勘及注釋	102168	32339	134507	2799	36.502
	72小計	2904712	751631	3656343	7157	405.856
清史稿	73本 文	4514567	1547619	6062186	8080	558.734
	74校勘及注釋	1360	277	1637	571	2.382
	75小計	4515927	1547896	6063823	8081	558.833

表七　二十五史文字統計表——正文部份（二之一）

製表日期：80/10/14

正文內史	分項	字數 a 中文字	b 標點	c 小計	中文字集數 d	平均每字使用次數 e
史記	1本紀	77233	25159	102392	2804	27.544
	2表	4053	1432	6385	1000	4.953
	3書	44699	14773	59472	2393	18.679
	4世家	133839	42309	176148	2998	44.643
	5列傳	257560	74069	331629	4254	60.545
	6小計	518284	157742	676026	4987	103.927
漢書	7本紀	54341	18366	72707	2346	23.163
	8表	7420	2368	9788	1149	6.458
	9志	163156	55046	218202	3796	42.981
	10列傳	515642	155008	670650	5254	98.143
	11小計	740559	230788	971347	5815	127.353
後漢書	12本紀	94492	32424	126916	3110	30.383
	13列傳	642250	195910	838160	5658	113.512
	14志	157278	64013	221291	3464	45.404
	15小計	894020	292347	1186367	6161	145.110
三國志	16列傳	368842	89160	458002	4343	84.928
	17小計	368842	89160	458002	4343	84.928
晉書	18紀	79326	19635	98961	3014	26.319
	19志	227225	61718	288043	4057	56.008
	20列傳	851575	200439	1052014	5646	150.828
	21小計	1158126	281792	1439918	5997	193.118
宋書	22本紀	65369	15043	80412	2914	22.433
	23志	325047	95595	420642	4512	72.041
	24列傳	421477	101002	522479	5112	82.449
	25小計	811893	211640	1023533	5842	138.975
南齊書	26本紀	35745	8476	44221	2640	13.540
	27志	65372	20249	85621	3030	21.575
	28列傳	198140	51390	249530	4482	44.208
	29小計	299257	80115	379372	4962	60.310

正文內史	分項	字數 a 中文字	b 標點	c 小計	中文字集數 d	平均每字使用次數 e
梁書	30紀	50746	12439	63185	3100	16.370
	31列傳	243692	58571	302263	4752	51.282
	32小計	294438	71010	365448	4973	59.207
陳書	33紀	40033	9625	49658	2861	13.993
	34列傳	123349	29019	152368	3653	33.766
	35小計	163382	38644	202026	4033	40.511
魏書	36紀	107597	23750	131347	3088	34.844
	37列傳	703623	165671	869294	5077	138.590
	38志	187109	61961	249070	3535	52.930
	39小計	998329	251382	1249711	5417	184.296
北齊書	40紀	36352	8555	44907	2600	13.982
	41列傳	176154	41270	217424	3833	45.957
	42小計	212506	49825	262331	4032	52.705
周書	43帝紀	39418	9829	49245	2533	15.561
	44列傳	223243	55467	278710	3991	55.937
	45小計	262659	65296	327955	4161	63.124
南史	46本紀	99579	24225	123804	3517	28.314
	47列傳	578045	136238	714283	5195	111.269
	48小計	677624	160463	838087	5376	126.046
北史	49本紀	147645	35458	183103	3604	40.967
	50列傳	958898	232496	1191394	5442	176.203
	51小計	1106543	267954	1374497	5572	198.590
隋書	52紀	35034	8901	43935	2659	13.176
	53志	369199	100674	469873	4717	78.270
	54列傳	296831	70979	367810	4601	64.514
	55小計	701064	180554	881618	5589	125.436
舊唐書	56本紀	315342	63388	378730	4218	74.761
	57志	435832	126662	562494	4621	94.316
	58列傳	1249520	278694	1528214	5824	214.547
	59小計	2000694	468744	2469438	6340	315.567

表七　二十五史文字統計表——正文部份（二之二）

正史內文	分項	字數 a 中文字	字數 b 標點	字數 c 小計	中文字集數 d	平均每字使用次數 e
新唐書	60本紀	88229	22224	110453	2605	33.869
	61志	449908	125772	575680	4903	91.762
	62表	32991	28251	61242	2272	14.521
	63列傳	1123145	272034	1395179	6181	181.709
	64小計	1694273	448281	2142554	6771	250.225
舊五代史	65本紀	346820	72942	419762	4144	83.692
	66列傳	384502	87645	472147	4467	86.076
	67志	54028	15024	69052	2551	21.179
	68小計	785350	1756118	960961	5078	154.657
新五代史	69本紀	32386	7754	40140	1955	16.566
	70傳	177847	39406	217253	3316	53.633
	71考	18504	18474	36978	1211	15.280
	72世家	46909	10282	57191	2502	18.749
	73小計	275646	75916	351562	3775	73.019
宋史	74本紀	339369	73555	412924	3879	87.489
	75志	1538012	454428	1992440	6267	245.414
	76表	1772	441	2213	460	3.852
	77列傳	2098170	478811	2576981	6474	324.092
	78小計	3977323	1007235	4984558	7383	538.714
遼史	79本紀	96943	23387	120330	2745	35.316
	80志	98940	40120	139060	2964	33.381
	81表	3806	1189	4995	883	4.310
	82列傳	91062	23001	114063	3030	30.053
	83小計	290751	87697	378448	4020	72.326
金史	84本紀	148394	31842	180236	3262	45.492
	85志	285445	84529	369974	4024	70.936
	86表	659	151	810	299	2.204
	87列傳	493492	107691	601183	4478	110.204
	88小計	927990	224213	1152203	5243	176.996
元史	89本紀	386266	77958	464224	3855	100.199
	90志	561594	185867	747461	4585	122.485
	91表	3815	3196	7011	561	6.800
	92列傳	658602	151037	809639	4946	133.159
	93小計	1610277	418058	2028335	5853	275.120

正史內文	分項	字數 a 中文字	字數 b 標點	字數 c 小計	中文字集數 d	平均每字使用次數 e
明史	94本紀	104716	25360	130076	2890	36.234
	95志	786537	263940	1050477	5464	143.949
	96表	3412	1371	4783	1101	3.099
	97列傳	1904803	425220	2330023	6496	293.227
	98小計	2799468	715891	3515359	7118	393.294
清史稿	99本紀	411833	80797	492630	4105	100.325
	100志	1602704	612137	2214841	6938	231.004
	101表	114020	357814	471840	1857	61.403
	102列傳	2383325	496092	2879417	6867	347.069
	103小計	4511888	1546840	6058728	8079	558.471

表八　二十五史文字統計——各史<列傳>部份（不含校勘記及注釋）

史　別	總字數	中　文	標　點	總節數	總字集數	比　值	每字使用次數	句長平均	各節平均值	各節字集數
2 史　記（列傳）	337948	263755	74193	2067	4265	79.238	61.842	5.126	163.497	76.691
3 漢　書（列傳）	689423	533865	155558	3485	5256	131.169	101.572	5.178	197.826	99.260
4 後漢書（列傳）	849005	652896	196109	4115	5665	149.868	115.251	4.951	206.320	101.950
5 三國志（無）	459708	370546	89162	2143	4348	105.729	85.222	5.059	214.516	115.373
6 晉　書（列傳）	791778	638271	153507	3596	5441	145.521	117.308	5.084	220.183	117.879
7 宋　書（列傳）	524670	423561	101109	2053	5113	102.615	82.840	5.077	255.563	130.450
8 南齊書（列傳）	251666	200170	51496	1407	4485	56.113	44.631	4.958	178.867	99.306
9 梁　書（列傳）	303101	244489	58612	1505	4754	63.757	51.428	4.998	201.396	109.389
10 陳　書（列傳）	152864	123824	29040	890	3654	41.835	33.887	5.104	171.757	97.492
11 魏　書（列傳）	870748	705012	165736	5231	5080	171.407	138.782	5.042	166.459	92.263
12 北齊書（列傳）	218208	176899	41309	1064	3834	56.914	46.140	5.134	205.083	115.907
13 周　書（列傳）	280933	225353	55580	1630	3992	70.374	56.451	4.972	172.352	98.107
14 南　史（列傳）	716963	580590	136373	5232	5199	137.904	111.673	5.242	137.034	82.953
15 北　史（列傳）	1196330	963586	232744	7643	5443	219.591	176.870	4.989	156.526	91.192
16 隋　書（列傳）	368010	297026	70984	1517	4607	79.881	64.473	4.968	242.591	131.430
17 舊唐書（列傳）	1529096	1250389	278707	6696	5830	262.281	214.475	5.329	228.360	124.162
18 新唐書（列傳）	1395374	1123340	272034	8255	6191	225.387	181.447	4.966	169.034	98.077
19 舊五代史（列傳）	56144	45972	10172	153	2802	20.037	16.407	5.522	366.954	169.346
20 新五代史（傳）	217344	177938	39406	1480	3318	65.505	53.628	5.492	146.854	80.657
21 宋　史（列傳）	2578924	2100063	478861	16094	6483	397.798	323.934	5.277	160.241	96.459
22 遼　史（列傳）	114297	91291	23006	1221	3031	37.709	30.119	4.957	93.609	59.283
23 金　史（列傳）	602098	494365	107733	3628	4485	134.247	110.226	5.556	165.959	96.237
24 元　史（列傳）	810809	659724	151085	4768	4948	163.866	133.331	5.298	170.052	97.885
25 明　史（列傳）	2331385	1906128	425257	13759	6501	358.619	293.205	5.247	169.444	100.310
26 清史稿（列傳）	2879427	2383335	496092	13699	6879	418.582	346.465	5.393	210.192	118.906

*備註：<<三國志>>因無列傳部份，故仍以統計其魏書、蜀書、吳書、的數字為準。

表九　二十五史文字統計表——史記本記

製表日期 80.08.10

分項		字數 a中文字	字數 b標點	字數 c小計	中文字集數 d	平均每字使用次數 e
正史內文						
卷一	1本文 2校勘及注釋	4030 17740	1739 6975	5769 24715	897 1693	4.493 10.478
	3小計	21770	8714	30484	1765	12.334
卷二	4本文 5校勘及注釋	3291 14887	1580 5810	4871 20697	852 1477	3.863 10.079
	6小計	18178	7390	25568	1615	11.256
卷三	7本文 8校勘及注釋	2959 4561	1029 1937	3988 6498	704 827	4.203 5.515
	9小計	7520	2966	10486	1050	7.162
卷四	10本文 11校勘及注釋	8482 16246	2982 6493	11464 22739	1189 1539	7.134 10.556
	12小計	24728	9475	34203	1768	13.986
卷五	13本文 14校勘及注釋	8547 10722	2794 4138	11341 14860	1093 1180	7.820 9.086
	15小計	19269	6932	26201	1514	12.727
卷六	16本文 17校勘及注釋	13177 13763	4357 5707	17534 19470	1567 1595	8.409 8.629
	18小計	26940	10064	37004	2006	13.430
卷七	19本文 20校勘及注釋	8893 9510	2593 4032	11486 13542	1061 1168	8.382 8.142
	21小計	18403	6625	25028	1501	12.260
卷八	22本文 23校勘及注釋	9560 11248	2929 4869	12489 16117	1135 1358	8.423 8.283
	24小計	20808	7798	28606	1649	12.619
卷九	25本文 26校勘及注釋	4503 2169	1156 981	5659 3150	765 557	5.886 3.894
	27小計	6672	2137	8809	964	6.921
卷十	28本文 29校勘及注釋	5634 4643	1467 1952	7101 6595	972 935	5.796 4.966
	30小計	10277	3419	13696	1298	7.918
卷十一	31本文 32校勘及注釋	1464 2085	521 862	1985 2947	450 501	3.253 4.162
	33小計	3549	1383	4932	691	5.136
卷十二	34本文 35校勘及注釋	6693 7199	2012 3166	8705 10365	1048 1164	6.386 6.185
	36小計	13892	5178	19070	1480	9.386

二、字頻統計

表十中呈現的是史記前 100個最常用字在25史中出現之次序。類似這類的表，亦可選定範圍製作，其排序可依25史中任一史爲基準。

根據表十中的資料，我們可以了解一些文字變遷的狀況，例如表中所列每個字在2020年間使用頻度之變化。再者像是：「之不以爲」這四個字在每一史的使用頻度都名列前矛，約在前十名之內。若擴大至每個史中都能保留在前50個字，則增加了：「一子其二人三十年有大」等十個字，而「而曰於太軍中亭（月、書、州）」等十字則庶幾乎均排名在50之內。這些字合計起來，共有24字；也就是前五十字中約有近一半的字，在二千多年間其使用頻度始終維持在同一個水平上。若是從這些詞使用的方法，如語意、語法分類、或構詞等方向來分析，或可有所發現。

若以前100個字的範圍來討論，則其累積頻度將近每一史的40%（請參閱表十二，下文將有說明），此時，再增加下列表十中之11字：「王者下四與五天將所至南」，而幾乎在 100名之內的尚有：「也使是上國無自後時東及百令文平帝出從」等18字。其合計共53字，亦約近半數。

在計算機裡，表十是一個完全的表，亦即25史全部的字共計 13966個都列在其中。我們把它轉到PC上並且做了個軟體程式來方便使用者查此表，圖一與圖二則是用此軟體分別查詢字組「之不以爲而」和「余我吾予」的例子。其實，用此軟體可以觀察許多有趣的統計，諸如依詞性質，如「東南西北」和「一二三四五」或依語意，如「余我吾予」，或依語法分類，如「之不以爲而」等介詞；可隨使用者之想像和興趣，提供25史之數據。

此外，我們發現另一個值得探討現象，那就是各個史的文字若照使用頻度排列時，在170字至200字之間，彼此交叉，聚成一塊。此數據請參考表十一。其原因有待解釋。

三、累頻與頻譜

廿五史全文的各史累頻曲線如圖三所示。其中最上一條爲史記者，最下一條爲清史稿者，若此二曲線單獨繪出則如圖四，至於累積頻率達 60%以前之資料，則在表十二中。換言之，常用的前100個字的累積頻率均在40%上下。

由圖三看來，這廿五條曲線有一致性。而事實上不僅限於文言文有這種現象，根據林樹 [16] 的資料所繪之曲線正好落入這群曲線之中間。此外，依據《現代漢語頻率詞典》[19]的資料，以及新加坡在1988年所做的聯合早報和國小教科書字彙統計的資料 [28.29]，亦有類似的結果。除廿五史外，這些都是白話文的統計。從這些現象看來，無論文言文或白話文，當文獻字數甚大時，其累頻譜都有類似的性質。這可以解釋爲文言文體或白話文體都頗一致的漢語統計結構上的特徵。究竟爲何存此現象？則有待日後之詮釋了。

表十　依史記前100常用字排序（四之一）

	史記	漢書	後漢書	三國志	晉書	宋書	南齊書	梁書	陳書	魏書	北齊書	周書	南史	北史	隋書	舊唐書	新唐書	舊五代史	新五代史	宋史	遼史	金史	元史	明史	清史稿
1) 之	25.530	21.368	17.302	22.316	27.346	20.363	16.193	18.801	15.574	21.103	18.598	20.529	19.590	19.888	19.781	19.505	14.605	16.852	19.283	17.683	13.634	17.202	15.817	14.223	10.803
2) 王	15.461	9.203	4.565	4.070	6.485	5.838	7.418	7.568	8.116	5.829	7.141	4.004	8.014	6.182	4.493	4.640	5.115	5.080	6.371	2.886	6.331	2.754	3.477	4.108	2.600
3) 不	14.907	13.415	10.221	13.499	12.597	9.731	9.603	9.419	6.361	8.993	9.789	7.532	11.282	9.984	8.025	8.649	10.635	6.796	10.308	9.222	6.700	9.152	7.406	9.058	5.528
4) 以	14.162	14.321	11.143	14.770	14.314	11.422	9.940	10.776	10.206	11.243	11.640	12.269	11.705	12.736	10.664	12.384	13.418	13.046	18.021	14.128	12.488	13.932	12.344	10.000	9.183
5) 為	13.984	14.103	10.788	13.437	12.171	10.306	12.393	13.030	11.949	10.523	12.788	11.648	13.945	12.893	10.519	12.672	11.357	11.882	16.647	10.389	9.871	9.283	8.819	7.787	7.523
6) 二	13.072	14.287	17.129	5.211	4.445	6.243	6.806	4.559	4.034	5.316	4.442	5.483	4.687	4.346	8.801	5.708	5.632	5.490	4.210	7.371	5.602	7.248	9.334	6.516	7.402
7) 子	12.526	7.210	4.898	5.133	5.411	5.166	5.650	7.006	6.599	6.182	6.790	4.691	8.302	7.095	5.452	5.152	6.143	3.447	5.062	3.741	5.183	3.533	4.012	4.194	3.323
8) 而	12.259	8.112	6.117	6.898	8.786	4.987	3.919	5.659	3.985	5.447	4.892	5.280	4.773	5.153	5.728	4.858	5.300	4.141	10.599	5.291	2.832	4.193	3.593	4.944	2.694
9) 曰	11.440	8.609	5.036	7.060	7.426	5.332	4.952	5.040	3.802	5.606	6.115	4.853	7.143	6.389	6.830	4.270	6.799	3.849	7.074	4.162	4.937	6.068	4.287	3.554	2.445
10) 其	10.465	8.727	6.630	8.210	8.137	5.304	4.009	6.552	5.959	6.758	6.495	7.687	5.834	7.888	8.139	7.519	6.920	6.525	11.204	7.215	5.009	6.945	7.858	6.551	5.030
11) 三	9.835	10.587	12.359	4.313	4.367	7.702	6.938	4.542	4.533	6.094	3.697	5.483	4.529	4.221	7.631	6.160	7.015	4.470	3.753	6.566	6.783	7.300	9.821	7.101	8.099
12) 人	9.109	7.595	7.372	6.938	6.827	5.593	6.220	6.854	5.051	5.700	7.849	6.031	8.555	8.410	7.014	7.246	8.295	6.398	8.039	7.096	6.237	9.042	7.219	7.009	5.732
13) 也	8.897	7.396	4.840	7.044	6.523	4.435	3.595	3.066	2.748	3.360	3.205	3.144	3.297	3.626	4.380	3.393	3.109	3.315	5.297	2.504	1.744	2.784	2.280	2.621	1.291
14) 公	8.874	3.623	2.379	3.317	2.564	1.960	2.006	1.848	2.023	2.597	2.362	6.585	2.399	4.062	2.974	2.382	2.504	1.394	1.447	1.374	0.986	1.325	0.968	0.950	0.950
15) 三	8.245	8.902	11.735	3.373	4.210	5.714	5.071	3.890	4.503	5.246	3.763	6.325	3.714	4.083	6.509	5.480	5.821	4.510	3.604	6.207	5.107	5.709	7.041	5.850	7.210
16) 者	8.090	7.218	5.223	4.987	4.693	3.610	2.483	3.117	2.553	3.778	2.797	3.374	3.053	3.971	4.984	3.781	5.468	2.909	4.975	5.372	3.211	5.800	6.295	4.929	2.747
17) 於	8.083	5.532	4.910	5.681	6.515	4.734	3.992	5.825	5.222	6.716	6.977	7.095	5.016	6.785	5.865	6.397	3.422	6.777	4.166	3.440	1.990	3.764	3.295	4.483	3.655
18) 十	6.660	6.688	6.381	3.730	4.504	7.029	4.141	4.745	4.229	6.271	3.280	3.868	4.352	3.870	10.419	7.667	8.965	5.589	5.047	8.198	7.765	7.846	12.254	9.952	12.550
19) 年	6.420	4.496	5.985	4.941	4.276	8.532	7.577	6.414	7.440	6.209	4.391	6.657	5.692	4.184	4.812	7.573	6.501	4.992	4.427	7.387	6.916	7.419	8.476	10.026	11.128
20) 有	6.218	6.997	6.783	6.979	7.295	6.805	5.905	6.265	5.277	7.354	6.293	6.317	6.723	6.871	8.957	5.073	6.274	4.836	6.204	5.811	4.845	5.975	5.133	6.971	4.408
21) 大	6.059	7.190	5.534	5.897	5.675	4.978	3.373	5.168	5.380	6.380	6.602	11.344	4.788	7.281	7.395	6.057	5.752	5.357	4.583	5.308	5.923	6.277	6.503	6.358	6.134
22) 系	5.854	1.330	0.413	0.297	0.669	0.664	0.470	0.437	0.341	0.704	0.398	0.623	0.259	0.550	0.431	0.362	0.429	0.430	0.595	0.432	0.336	0.220		0.212	
23) 下	5.423	5.929	4.418	4.291	3.716	2.801	2.562	2.329	2.102	2.470	2.596	2.404	2.608	2.626	3.529	4.101	5.296	3.352	3.756	3.641	2.611	2.936	3.306	3.703	1.986
24) 四	5.280	6.486	7.026	1.943	2.720	4.330	3.572	2.511	2.376	3.428	1.977	2.864	2.055	2.205	4.627	3.737	3.784	2.557	2.248	4.017	3.539	3.858	5.149	4.577	5.279
25) 侯	5.274	4.590	2.363	2.944	1.217	0.924	0.877	1.703	2.986	1.083	1.031	1.472	1.870	1.112	0.757	0.287	0.292	0.312	0.352	0.206				0.532	

單位：千分之一

表十　依史記前100常用字排序（四之二）

	史記	漢書	後漢書	三國志	晉書	宋書	南齊書	梁書	陳書	魏書	北齊書	周書	南史	北史	隋書	舊唐書	新唐書	舊五代史	新五代史	宋史	遼史	金史	元史	明史	清史稿
26) 與	5.045	3.847	3.283	4.914	2.931	2.346	2.463	2.674	2.602	2.711	3.608	3.483	3.683	3.646	2.329	2.783	3.199	3.227	4.155	2.458	2.254	3.161	2.300	2.789	2.649
27) 五	5.036	6.449	6.940	2.032	2.999	4.267	3.466	2.792	2.705	3.688	2.034	3.178	2.205	2.403	5.393	3.848	4.030	5.027	2.542	4.200	4.085	4.293	5.630	3.939	4.430
28) 天	4.741	5.070	2.853	2.793	3.497	3.269	2.539	2.901	2.449	2.861	3.271	2.827	2.025	2.469	3.940	3.681	3.686	2.962	3.579	3.071	3.160	2.692	2.502	2.954	1.854
29) 使	4.625	3.332	2.074	3.603	2.358	2.010	2.072	2.373	2.376	2.544	2.741	2.476	2.426	2.500	1.430	5.225	4.914	10.418	9.326	5.831	11.518	7.076	4.104	2.562	1.745
30) 將	4.572	4.605	4.249	8.220	6.280	5.885	6.925	6.918	10.194	8.052	5.065	8.178	4.588	5.077	3.489	3.345	3.555	4.381	4.935	2.329	2.196	2.464	2.057	2.548	2.765
31) 是	4.470	3.432	2.908	3.060	3.044	2.696	2.059	2.785	2.376	2.629	2.877	2.676	2.736	2.789	2.579	2.423	2.138	2.901	2.476	2.100	2.059	2.252	1.674	1.998	1.574
32) 君	4.370	2.039	1.225	2.067	1.340	0.858	0.871	0.833	0.920	1.118	0.900	0.717	0.751	0.815	0.925	0.887	0.976	0.543	0.761	0.552	1.033	0.433	0.333	0.392	0.214
33) 齊	4.317	1.331	0.393	0.505	0.617	0.631	1.102	1.618	1.602	1.029	1.645	2.551	1.632	1.673	1.855	0.491	0.479	0.359	0.297	0.287	0.318	0.340	0.273		0.583
34) 太	4.268	4.388	4.380	6.531	5.183	7.027	7.180	5.192	5.880	6.149	5.262	5.993	4.985	4.954	4.276	4.686	4.147	5.707	6.617	4.175	6.877	4.004	5.080	3.119	2.074
35) 上	4.236	5.946	3.164	2.297	2.564	3.045	3.767	1.499	1.310	2.230	1.902	2.434	2.910	2.959	4.985	3.852	3.338	2.843	1.893	4.059	5.020	6.102	3.305	2.625	4.107
36) 臣	4.198	3.785	2.063	2.634	2.688	1.904	1.907	1.293	1.395	2.370	1.256	1.260	1.084	1.724	1.796	2.846	2.192	2.161	2.049	2.777	3.019	2.299	2.265	3.307	2.448
37) 乃	4.154	2.643	2.878	2.537	2.363	1.546	1.705	2.261	2.468	2.313	2.432	3.834	2.974	2.766	1.288	1.807	2.667	1.810	4.101	1.392	0.885	1.791	1.387	1.952	0.914
38) 四	4.041	4.279	3.134	3.160	2.978	2.880	2.645	3.297	2.066	3.302	1.804	4.687	2.127	3.065	3.316	3.023	2.457	2.437	2.342	2.490	5.244	2.859	2.810	2.375	2.729
39) 楚	3.928	1.205	0.275		0.308	0.242		0.223		0.202		0.215			0.229	0.304	0.319		0.591		0.220				0.232
40) 相	3.924	3.541	1.900	2.666	2.003	2.120	1.923	1.916	1.542	1.923	2.685	1.868	1.992	2.200	2.224	2.197	2.337	1.651	2.219	1.873	2.033	1.939	2.099	1.228	0.938
41) 所	3.907	4.628	3.773	5.254	5.053	4.399	3.750	4.522	4.064	4.134	4.170	3.930	4.532	4.436	4.109	4.108	3.409	3.317	3.477	3.602	2.550	3.371	3.864	3.617	2.647
42) 立	3.861	2.377	1.543	1.325	1.388	2.451	1.615	1.202	1.377	1.336	1.162	0.962	1.563	1.340	1.238	1.119	1.484	1.043	2.324	1.208	1.741	1.534	2.373	1.222	0.861
43) 至	3.810	4.193	2.675	3.519	2.936	3.158	2.698	3.520	3.467	3.147	3.304	3.619	3.883	3.500	3.409	3.612	3.360	3.705	3.872	3.910	2.976	3.442	5.338	3.602	3.437
44) 軍	3.763	3.536	2.759	8.514	6.661	9.029	11.731	11.314	14.800	8.819	6.228	10.823	7.268	5.380	3.910	4.270	4.434	8.786	7.846	4.454	7.057	6.598	5.028	2.956	4.340
45) 得	3.697	3.142	2.198	2.753	2.060	1.901	1.817	1.838	1.243	1.696	2.188	1.721	2.374	2.060	1.650	1.703	2.328	1.581	2.672	2.146	1.799	1.954	1.574	2.079	1.826
46) 無	3.695	2.496	3.308	3.713	3.805	3.791	3.618	3.527	3.004	2.889	3.041	2.468	3.398	2.982	2.789	2.937	2.606	2.293	2.284	2.433	1.531	2.406	1.669	2.422	1.874
47) 兵	3.648	2.615	2.189	3.589	2.005	1.780	1.456	1.858	2.681	1.558	1.645	2.627	1.791	1.743	2.223	2.943	3.378	3.262	7.179	3.151	3.417	4.225	2.904	4.240	4.200
48) 中	3.642	4.366	4.504	4.666	4.763	4.627	6.455	6.803	7.214	6.326	6.195	5.068	5.841	5.209	3.747	6.256	6.421	5.238	5.015	5.097	3.879	4.732	5.269	5.051	2.671
49) 六	3.534	4.663	5.519	1.182	1.927	2.898	2.158	1.577	1.560	2.417	1.467	2.411	1.399	1.625	3.647	2.538	2.616	2.047	1.653	2.852	3.196	2.990	3.699	3.017	3.440
50) 故	3.529	3.764	2.382	2.164	2.484	2.509	2.820	2.207	2.224	1.676	1.621	1.000	1.806	1.452	1.367	1.454	1.889	1.729	1.715	1.190	1.098	1.315	0.880	1.506	0.775

表十　依史記前100常用字排序（四之三）

	史記	漢書	後漢書	三國志	晉書	宋書	南齊書	梁書	陳書	魏書	北齊書	周書	南史	北史	隋書	舊唐書	新唐書	舊五代史	新五代史	宋史	遼史	金史	元史	明史	清史稿
										0.393	0.858	0.683		0.613	0.332	0.376	0.452	0.956	1.091	0.614	0.643	0.275	0.360	0.483	0.288
51) 趙	3.338	1.100	0.474	0.289	0.539	0.233																			
52) 誰	3.164	2.384	1.817	3.314	2.135	2.017	1.675	1.895	1.938	2.197	2.202	2.494	1.731	2.012	1.924	1.610	1.734	2.072	1.987	2.099	2.781	2.003	3.615	2.700	1.982
53) 夫	3.101	2.885	1.427	1.684	1.267	1.258	1.106	1.381	1.176	1.313	0.947	2.087	1.152	1.350	1.448	1.587	1.066	0.946	0.830	1.021	0.339	0.962	1.309	0.645	0.295
54) 言	3.060	3.266	2.251	2.418	2.236	1.647	1.605	1.797	1.438	1.817	2.540	1.400	2.089	2.354	1.833	2.035	2.262	1.885	2.317	3.566	1.029	2.407	2.443	3.882	1.704
55) 自	3.029	3.348	3.089	3.800	3.298	3.116	2.827	3.080	3.351	3.154	3.632	2.883	3.683	3.498	2.492	3.118	3.321	3.308	4.482	2.819	1.782	2.856	2.087	2.933	3.335
56) 帝	3.009	4.242	4.223	2.871	4.741	4.919	2.847	1.415	2.431	2.342	4.775	4.925	9.208	6.539	4.797	1.645	4.449	4.172	3.339	2.039	3.485	1.132	2.490	4.092	0.503
57) 事	2.929	3.269	2.949	3.730	3.767	4.017	4.975	4.870	4.692	3.792	4.494	3.649	4.485	3.843	3.461	4.602	3.732	4.398	4.558	4.699	5.147	6.232	5.520	5.253	3.014
58) 欽	2.869	2.248	1.404	2.518	1.423	1.141	1.155	1.032	0.731	1.151	1.546	1.310	1.585	1.321	0.716	0.924	1.003	0.883	1.523	1.068	0.709	1.110	0.623	0.934	0.418
59) 行	2.853	2.789	8.845	2.342	2.044	2.380	2.949	2.423	1.913	2.659	3.168	2.706	2.271	2.817	2.841	2.670	2.790	3.307	3.557	3.000	2.962	3.445	4.529	2.548	2.551
60) 可	2.717	2.348	1.667	2.739	2.439	2.224	2.185	2.210	1.590	2.075	1.753	1.789	1.806	1.930	1.677	2.191	2.458	1.507	2.799	2.181	1.604	2.720	1.511	2.053	1.288
61) 漢	2.706	3.288	2.163	1.978	1.222	2.792	1.066	0.721	0.463	1.389	0.445	0.559	0.444	0.582	1.340	1.202	0.717	1.208	2.248	0.589	2.055	0.502	0.703	0.466	1.177
62) 後	2.697	3.352	2.838	3.589	2.581	2.479	2.241	1.872	2.852	3.572	3.627	2.849	3.251	3.735	4.714	2.413	2.482	2.516	3.006	2.192	1.712	1.880	1.684	2.223	1.805
63) 卒	2.681	1.054	1.133	1.619	1.018	0.746	0.874	1.499	1.517	1.579	1.392	1.208	1.613	1.634	0.690	1.347	1.329	1.140	2.128	1.242	1.116	0.720	0.901	1.487	1.138
64) 皆	2.680	3.205	2.705	2.987	2.169	1.482	1.625	1.503	1.359	1.477	1.589	1.525	1.775	1.893	2.453	2.083	2.910	1.492	3.789	2.151	1.705	2.797	1.932	2.194	1.516
65) 陽	2.585	2.702	3.178	2.210	2.684	3.242	3.393	2.823	3.394	2.936	4.119	2.876	2.837	2.749	2.164	1.674	1.673	2.487	1.494	1.015	0.567	0.956	1.422	1.425	1.369
66) 刑	2.562	2.282	1.962	2.105	2.245	2.202	2.195	1.631	0.756	1.373	0.947	1.464	1.321	1.282	2.685	2.245	1.908	1.046	1.084	2.316	0.907	1.647	1.059	1.519	1.149
67) 比	2.558	2.202	2.073	2.785	2.449	2.701	2.685	2.525	1.743	1.873	2.643	2.140	2.987	2.330	1.765	1.803	0.917	1.417	1.769	1.386	1.091	1.709	0.873	1.153	0.944
68) 七	2.538	3.463	4.539	0.812	1.552	2.576	2.013	1.472	1.280	2.207	1.153	1.593	1.325	1.308	2.559	2.267	2.243	1.619	1.371	2.340	2.488	2.397	3.593	2.473	2.639
69) 見	2.532	2.286	1.918	2.291	1.813	2.458	2.291	1.889	1.359	1.615	2.193	1.528	2.744	2.106	1.705	1.240	1.492	1.245	2.487	1.330	1.441	1.486	1.042	0.979	0.737
70) 死	2.473	1.937	1.347	1.652	1.510	1.033	0.940	1.020	0.615	1.447	1.528	0.838	1.377	1.509	0.953	1.039	2.007	0.833	2.038	1.150	0.899	1.186	1.229	2.141	1.077
71) 出	2.468	2.397	1.662	2.237	1.787	1.854	2.516	2.620	2.066	2.113	2.601	2.079	2.287	2.096	1.865	1.754	2.087	1.690	3.031	2.554	2.030	2.271	1.973	2.142	2.338
72) 矣	2.439	1.476	0.848	1.185	1.224	0.797	0.728	0.800	0.780	0.796	0.469	0.766	0.698	0.730	0.714	0.783	0.943	0.758	1.400	0.950	0.693	1.002	0.541	0.797	0.370
73) 今	2.428	2.170	3.168	2.842	1.970	1.957	1.893	1.435	1.188	1.696	1.223	1.298	1.421	1.369	1.403	1.715	1.348	3.196	1.023	1.640	0.654	1.716	1.300	1.092	0.625
74) 人	2.405	2.331	1.414	1.579	1.488	1.670	1.827	1.604	2.340	1.955	2.390	2.076	2.057	2.008	2.240	2.243	2.667	2.147	3.089	2.845	1.835	2.157	2.310	3.153	3.627
75) 時	2.392	3.087	3.034	3.301	3.465	3.008	2.181	3.493	3.412	3.016	3.243	3.298	3.608	3.610	2.935	3.060	2.566	2.975	3.049	2.180	1.566	1.904	1.713	2.891	2.154

表十　依史記前100常用字排序（四之四）

	史記	漢書	後漢書	三國志	晉書	宋書	南齊書	梁書	陳書	魏書	北齊書	周書	南史	北史	隋書	舊唐書	新唐書	五代史	新五代史	宋史	遼史	金史	元史	明史	清
76) 給	2.383	1.948	1.984	1.684	1.726	1.656	1.175	1.218	1.304	2.003	1.246	1.242	1.081	1.472	1.835	1.626	1.447	1.191	1.204	1.676	1.524	1.697	1.814	1.782	1.971
77) 施	2.350	1.875	1.539	2.183	1.652	1.157	0.990	1.611	1.115	1.088	1.326	1.302	1.441	1.350	1.019	1.177	1.403	0.895	1.860	1.157	0.907	1.579	0.882	1.041	0.903
78) 從	2.345	1.952	2.054	2.588	1.835	1.697	2.208	1.364	1.420	2.106	2.671	2.962	1.991	2.612	2.237	2.146	1.878	3.991	3.916	1.813	1.972	2.819	3.342	2.191	1.895
79) 東	2.307	2.505	2.172	2.553	2.405	2.298	3.188	2.921	3.437	2.902	2.048	3.302	2.630	2.584	2.585	2.401	2.741	1.509	1.646	2.306	3.070	2.831	2.599	3.771	4.618
80) 如	2.290	1.908	1.290	1.702	2.150	2.071	2.761	2.369	2.254	1.739	1.818	1.242	1.779	1.575	1.885	1.416	1.464	1.447	1.229	1.983	2.257	2.162	1.548	1.775	1.441
81) 何	2.284	1.591	1.268	1.646	1.722	1.917	1.572	1.652	0.865	1.223	1.729	1.011	1.748	1.441	1.030	1.127	1.108	1.046	1.508	0.840	0.733	1.100	0.638	1.013	0.539
82) 前	2.273	1.813	1.304	1.830	1.226	1.208	1.168	1.537	1.085	1.261	1.303	1.102	1.409	1.386	0.876	1.241	0.970	1.372	1.465	1.087	0.798	1.148	0.907	1.108	0.731
83) 儀	2.273	0.456	0.439	1.328	1.127	1.047	0.526	1.960	0.335	1.055	2.905	3.332	1.472	1.633	1.765	0.801	0.688	0.967	1.646	0.202	0.426			0.286	
84) 生	2.265	1.678	1.368	1.079	1.552	1.857	1.430	1.757	1.377	1.496	1.232	1.147	1.632	1.435	1.071	1.011	1.073	0.687	0.805	1.058	1.073	1.026	0.882	1.199	1.073
85) 六	2.201	3.070	4.406	0.783	1.463	2.236	1.824	1.147	1.164	1.905	0.970	1.396	1.109	1.237	2.701	2.128	2.184	1.695	1.135	2.121	2.420	2.394	3.520	2.465	2.959
86) 及	2.194	2.438	2.693	2.213	3.014	2.102	1.738	2.643	3.187	2.490	2.797	3.619	4.015	3.662	3.348	2.934	2.413	2.875	2.085	2.572	2.644	3.177	2.684	2.749	2.040
87) 百	2.188	3.486	2.945	1.997	2.687	3.069	1.890	2.010	1.712	3.222	1.846	2.683	1.753	1.972	3.810	3.620	3.537	3.096	1.751	2.981	2.225	3.090	4.458	2.610	2.828
88) 地	2.163	1.912	1.206	0.896	1.283	1.384	0.887	0.866	0.719	0.874	0.787	0.951	0.711	0.847	1.332	1.083	1.160	0.820	0.671	1.260	1.925	1.302	1.565	1.489	2.143
89) 周	2.158	1.413	0.842	0.983	1.062	1.046	0.738	0.748	2.108	0.849	1.790	0.989	0.858	1.974	2.762	0.911	0.822	1.402	2.034	0.799	0.470	0.252	0.342	0.864	0.594
90) 九	2.146	2.850	4.084	0.926	1.509	2.182	1.725	1.171	1.158	1.779	0.890	1.260	1.080	1.138	2.447	2.111	1.897	1.706	1.193	1.800	2.091	1.804	2.574	2.216	2.434
91) 令	2.127	2.381	2.169	2.245	2.003	3.280	2.383	2.044	2.358	2.210	3.332	2.547	2.528	3.013	2.622	3.145	1.925	2.718	1.008	1.954	1.257	2.726	1.761	1.623	1.280
92) 長	2.107	2.586	2.071	2.065	2.039	2.122	2.406	2.630	2.291	2.495	2.872	2.125	2.387	2.548	2.321	1.936	1.938	1.344	0.943	0.999	1.470	1.090	1.148	1.440	1.472
93) 段	2.088	1.615	1.242	1.282	1.081	0.927	0.824	0.457	0.238	1.125	0.993	0.551	1.000	1.028	0.625	0.891	1.316	1.010	2.959	0.525	1.015	1.132	0.888	0.980	0.403
94) 文	2.073	1.766	2.443	2.011	1.952	2.390	2.701	2.992	4.174	2.283	4.156	3.106	4.405	5.314	2.804	2.488	2.140	2.482	1.719	2.206	1.307	1.403	1.887	2.701	2.224
95) 雨	2.006	2.157	2.301	2.423	2.473	3.890	5.101	3.544	4.588	3.578	2.127	2.027	3.244	2.525	2.449	2.778	3.164	2.209	3.332	2.665	5.797	2.721	4.005	5.053	5.730
96) 高	2.003	1.771	1.171	1.139	1.154	1.997	1.347	4.136	5.618	3.221	4.508	2.159	1.715	2.021	2.811	1.692	1.618	1.647	2.719	0.928	1.820	1.359	1.015	1.062	1.266
97) 然	2.001	1.609	1.044	1.997	1.657	0.999	0.732	1.049	0.725	1.249	1.293	1.230	0.929	1.350	1.025	1.067	1.521	0.845	1.733	1.197	0.679	0.920	0.674	1.011	0.514
98) 皆	1.970	0.341	0.201	0.267	1.109	3.040	1.344	0.978	1.158	1.241	2.024	0.996	1.105	1.022	1.461	0.588	0.493	1.840	3.764	0.297	1.394		0.245		0.347
99) 平	1.959	2.111	2.239	2.297	2.380	2.374	1.946	1.858	2.462	3.662	3.514	2.676	1.932	2.886	2.392	2.248	2.194	2.220	1.875	2.079	2.200	2.257	2.987	2.059	2.027
100) 試	1.942	1.846	2.345	1.703	2.352	2.935	2.387	1.750	2.321	2.209	5.464	4.034	4.593	4.735	2.781	2.686	2.977	2.553	1.976	1.374	1.123	1.273	1.033	2.209	1.075

表十一　二十五史各史頻譜交會點

（表中之數字爲該字之千分之頻度）

常用字等第 史名	第170字	第200字	至第200字的累計頻率
史記	1.208	1.034	61.78
漢書	1.212	1.077	57.64
三國志	1.274	1.101	53.00
後漢書	1.225	1.069	55.35
宋書	1.256	1.056	52.62
南齊書	1.235	1.089	51.23
魏書	1.253	1.125	52.57
梁書	1.273	1.117	50.94
陳書	1.304	1.176	53.10
北齊書	1.303	1.157	53.12
周書	1.310	1.112	54.78
晉書	1.222	1.072	50.82
隋書	1.282	1.112	53.82
南史	1.291	1.084	51.80
北史	1.319	1.129	52.12
舊唐書	1.232	1.097	52.70
舊五代	1.417	1.199	54.66
新唐書	1.186	1.043	52.00
新五代	1.352	1.146	57.98
遼史	1.390	1.232	55.32
金史	1.384	1.179	53.59
宋史	1.292	1.112	50.97
元史	1.139	1.238	54.15
明史	1.328	1.111	51.50
清史	1.238	1.128	48.39

表十二

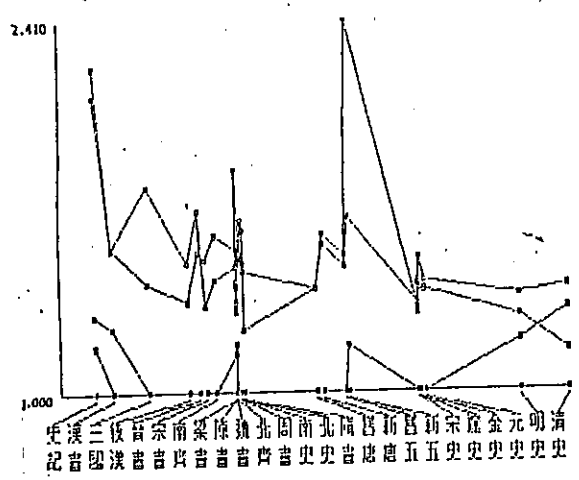
各史累積頻譜達60%前之分佈概要

成書年代 （時程）	史名 累積頻度	20%	40%	60%
88±5BC	史記 漢書	20 15	73 73	194 217
289	三國志	27	108	280
445	後漢書	27	99	261
501±13	宋書	31	103	260
	南齊書	32	113	286
554	魏書	30	107	269
648±12	梁書	30	116	293
	陳書	30	108	264
	北齊書	30	99	249
	周書	28	92	229
	晉書	29	116	303
	隋書	29	104	266
	南史	30	113	285
	北史	30	110	273
960±15	舊唐書	31	108	273
	舊五代	30	101	242
1006±16	新唐書	30	108	286
	新五代	25	87	219
1358±12	遼史	30	101	233
	金史	29	106	255
	宋史	31	115	289
	元史	28	100	252
1739	明史	32	110	281
1927	清史稿	34	127	321

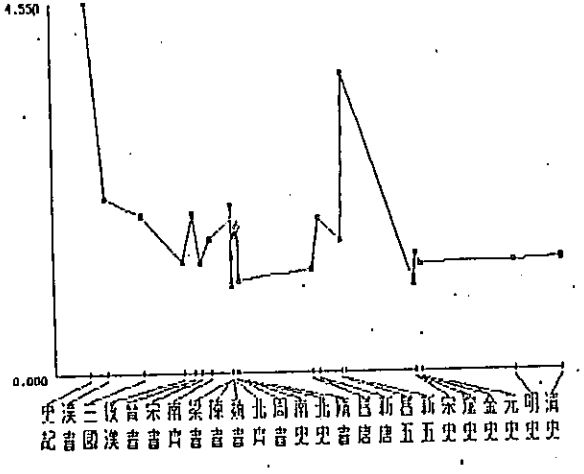
* 表中數字 n 字表示：超過漢書累積頻度（最左列）時，爲該書之 n 個常用字。

* 由表中，可知此分佈情形與該史文字之多寡關係甚微。

圖一　「余吾我予」之頻率變化圖



(a)「余我吾予」之個別曲線

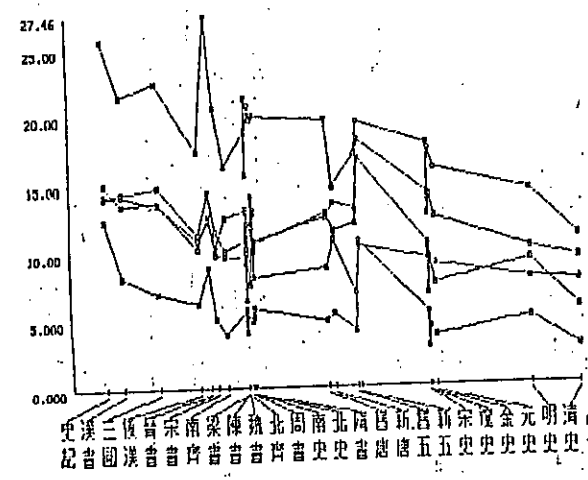


(b)「余我吾予」之合計之曲線

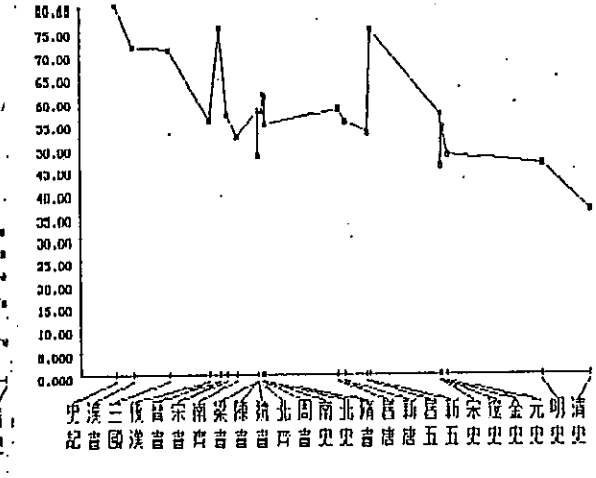
	史記	漢書	三國	後漢	晉書	宋書	南齊	梁書	陳書	魏書	北齊	周書
余	.4000	0.000	0.000	0.000	0.000	0.000	0.000	.0000	.5000	0.000	0.000	0.000
我	1.170	.8000	.8660	.0500	.4900	.1000	.5100	.7500	.6100	.0100	1.090	.6500
吾	1.760	.7600	1.230	.0200	.7000	.6400	.0200	.8000	.3300	.5200	.9500	.2100
予	.1400	.4900	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	4.550	1.010	1.110	1.070	1.270	1.020	1.530	1.550	1.440	.3600	1.040	1.760
	4.550	1.010	1.110	1.070	1.370	1.020	1.530	1.550	1.440	.3600	1.040	1.760

	南史	北史	隋書	舊唐	新唐	舊五	新五	宋史	遼史	金史	元史	明史	清史
余	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
我	.2300	.7300	.4300	.3000	.0100	.7800	.4600	.7600	.8100	.2000	.9300	.1700	.640
吾	.1400	.9500	.2500	.5200	.5200	.4300	2.410	.5900	.2300	.5200	.5000	.0100	.100
予	0.000	0.000	0.000	0.000	0.000	0.000	.4700	0.000	0.000	0.000	0.000	.6700	.390
	1.370	1.600	.8000	1.109	1.530	1.210	3.340	.3500	1.040	1.720	1.430	1.950	1.13
	1.370	1.600	.8000	1.100	1.530	1.210	3.340	.3500	1.040	1.720	1.430	1.950	1.13

圖二　「之不以爲而」之頻率變化圖



(a)「之不以爲而」之個別曲線

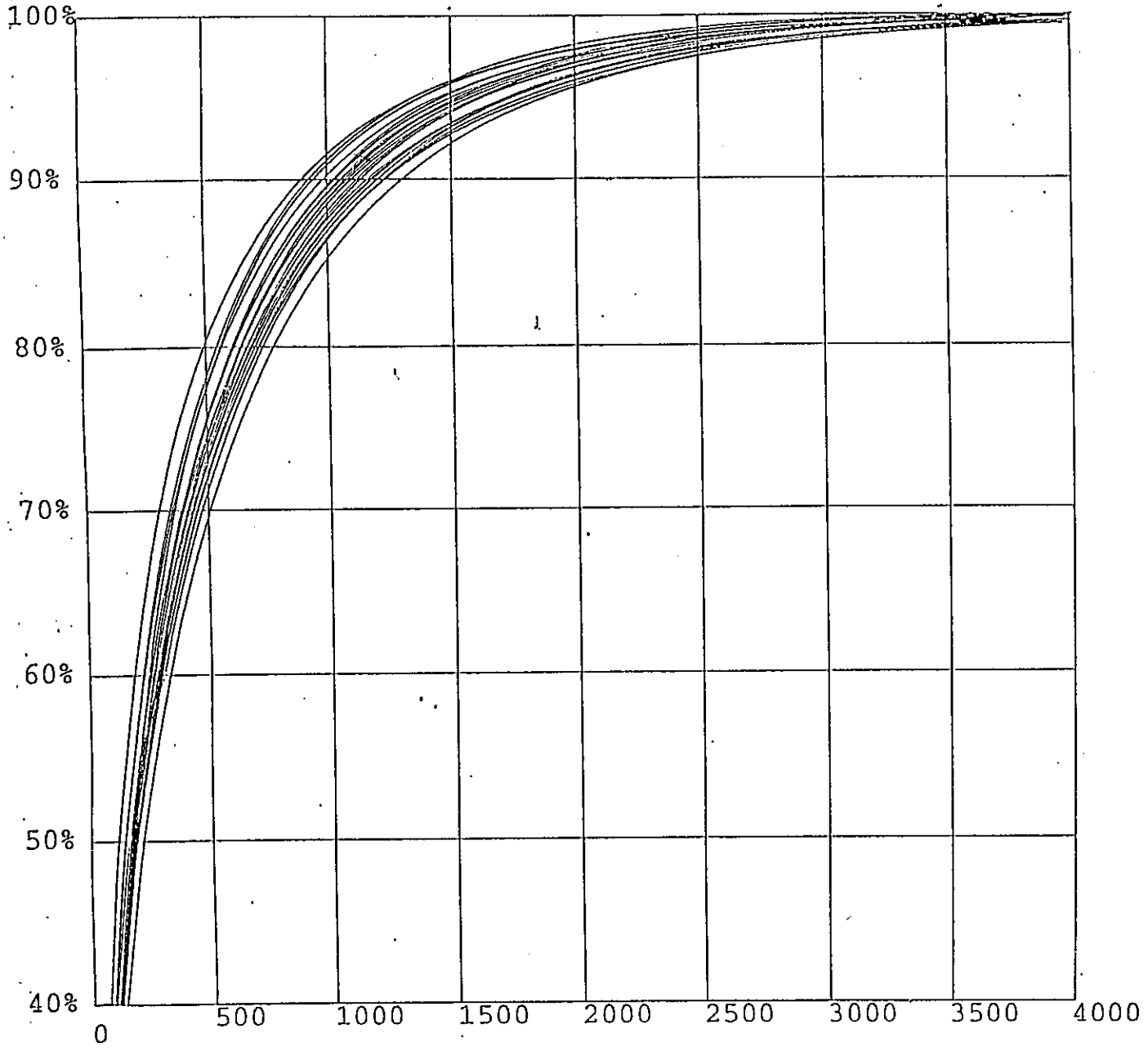


(b)「之不以爲而」之合計之曲線

	史記	漢書	三國	後漢	晉書	宋書	南齊	梁書	陳書	魏書	北齊	周書
之	25.30	21.68	22.16	17.02	27.46	20.63	16.93	18.01	15.74	21.03	18.98	20.29
不	14.07	13.19	13.39	10.21	12.97	9.310	9.030	9.150	6.610	8.930	9.890	7.320
以	14.62	14.21	14.70	11.43	14.14	11.22	9.400	10.75	10.06	11.43	11.40	12.69
爲	13.84	14.03	13.37	10.86	12.71	10.06	12.93	13.30	11.47	10.23	12.88	11.45
而	12.59	8.120	6.980	6.170	8.060	4.870	3.190	3.590	3.850	5.470	4.920	5.800
	80.42	71.19	70.20	55.69	75.14	56.09	52.48	57.83	48.73	57.09	57.07	57.56
	80.42	71.19	70.20	55.69	75.14	56.09	52.48	57.83	48.73	57.09	57.07	57.56

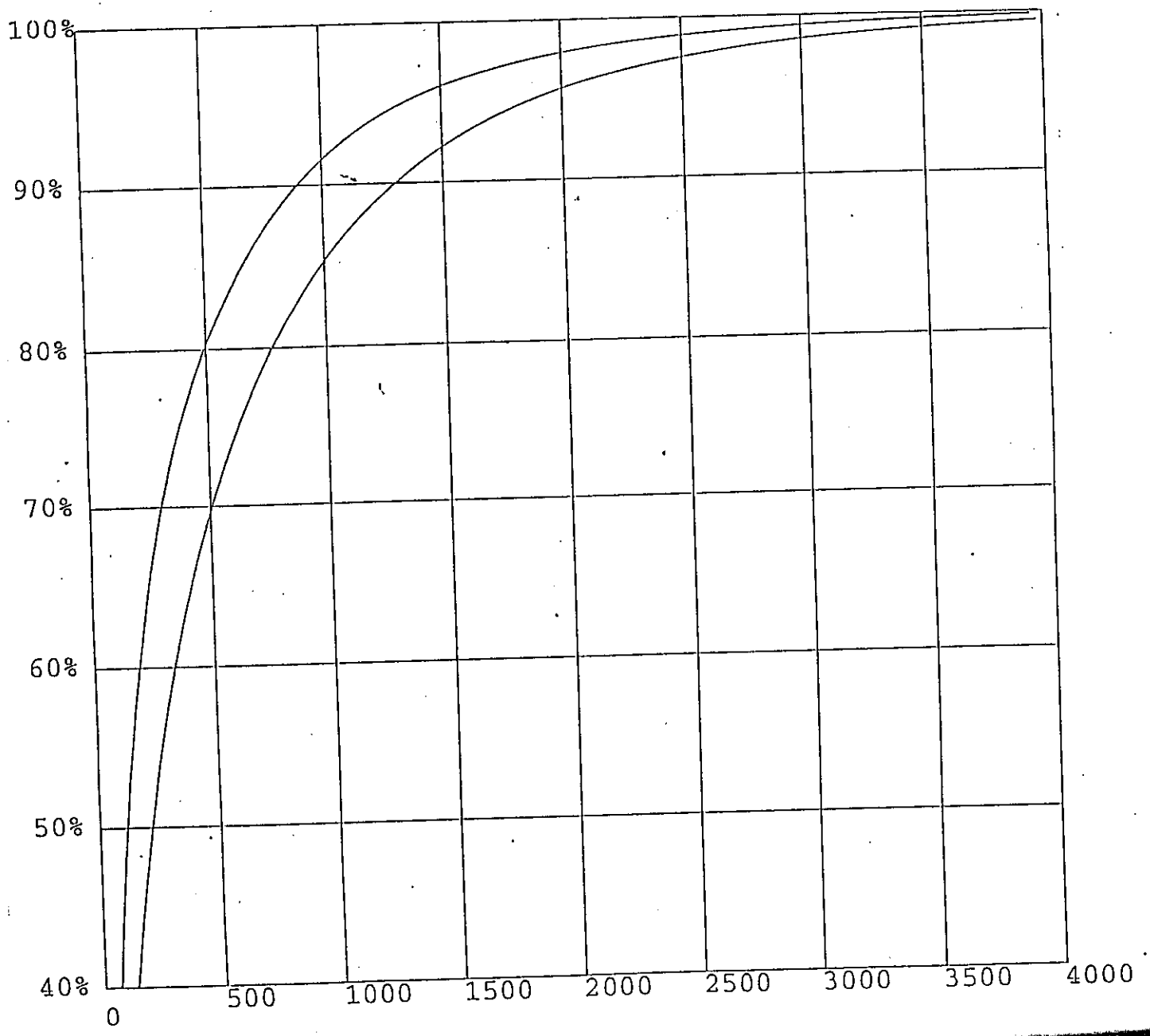
	南史	北史	隋書	舊唐	新唐	舊五	新五	宋史	遼史	金史	元史	明史	清史
之	19.80	19.88	19.21	19.05	14.05	16.52	19.83	17.91	13.34	17.02	15.17	14.23	10.0
不	11.82	9.840	8.250	8.490	10.35	6.960	10.08	9.220	6.000	9.320	7.060	9.500	5.28
以	11.05	12.36	10.64	12.84	12.10	13.46	18.21	14.20	12.86	13.32	12.44	10.60	9.87
爲	13.43	12.93	10.19	12.72	11.37	11.82	16.47	10.89	9.710	9.830	8.190	7.870	7.23
而	4.720	5.520	5.200	4.580	5.000	4.410	10.99	5.910	2.220	4.930	2.960	4.440	2.94
	61.83	60.34	54.17	58.68	55.15	52.17	74.58	56.13	43.23	53.62	47.82	46.12	35.3
	61.83	60.54	54.17	58.68	55.15	52.17	74.58	56.13	43.23	53.62	47.82	46.12	35.3

圖三　二十五史累頻譜



上線：史記（在一千三百字左右有交叉現象）
下線：清史稿
縱軸：字集累計使用頻率
橫軸：字集字數依使用頻率排序

圖四　史記累頻譜



上線：史記
下線：清史稿
縱軸：字集累積頻率
橫軸：字集字數依頻率排序

四、文件字數與其字集字數之關係

一個文件的中文字數越多，它的字彙就可能越大，也就是它的字集字數就可能越多。當我們把許多文件集中在一起觀察它們上述二數字的關係時，就會呈現出這許多文件此二者的共同統計性質；而這個共同的統計性質經常爲一單向增大的函線（monotonic function）。這條經驗曲線的用途之一就是我們不必經由統計便用它來估計這類新文件的性質。例如，在大略知道某文件有多少中文字數時，便可由上述經驗曲線概略地估計它字集字數的大小。

對於各史列傳部份中文字數（不含標點及其他符號）各它們字集字數的經驗曲線如圖五中所示。它是一個四次的回歸曲線。由於它是四次的高次方程，較不易使用，於是我們也可以稍作變化，以觀察文件的中文字數和它每個字平均使用的次數（即中文字數除以其字集字數），亦可得到相似的效果。和圖五相當的這條曲線如圖六所示。在圖六中，此曲線呈現出十分接近直線的面貌；若 Y表示文件的中文字數，以萬爲單位，而X表示此文件平均每個字使用之次數，則此直線之公式可略爲：

$$Y = 0.7X - 17$$

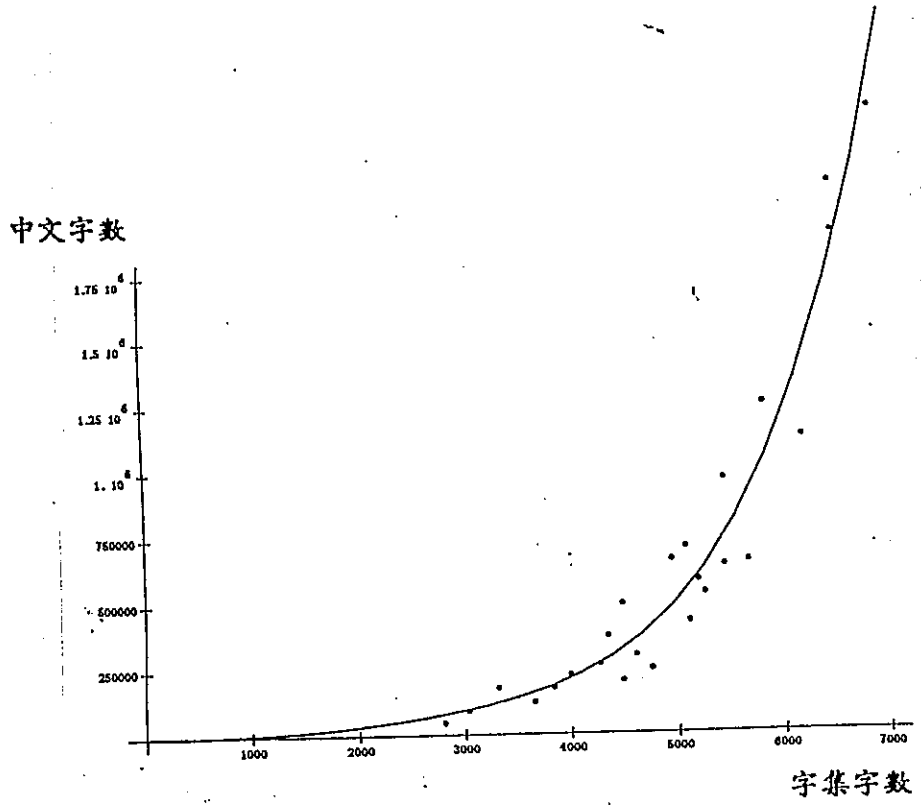
或 $$X = 1.5Y + 24$$

此兩公式在 $250 \geq Y \geq 10$之間可得到相當不錯的效果。若覺得上式的精確度不夠，那麼可以用二次曲線的回歸方程式來表示圖六中的關係則準確度可提至滿意的程度。

文件的中文字數與它的每個字平均使用次數的關係，亦可由最小的單位「段」來觀察起。若一個段的中文字數在200至500之間，則廿五史列傳中之每段每個字平均使用的次數約在2±0.5之範圍內。在這範圍左右的標準差較大是因爲文件字數較少而與內容關係密切的緣故。

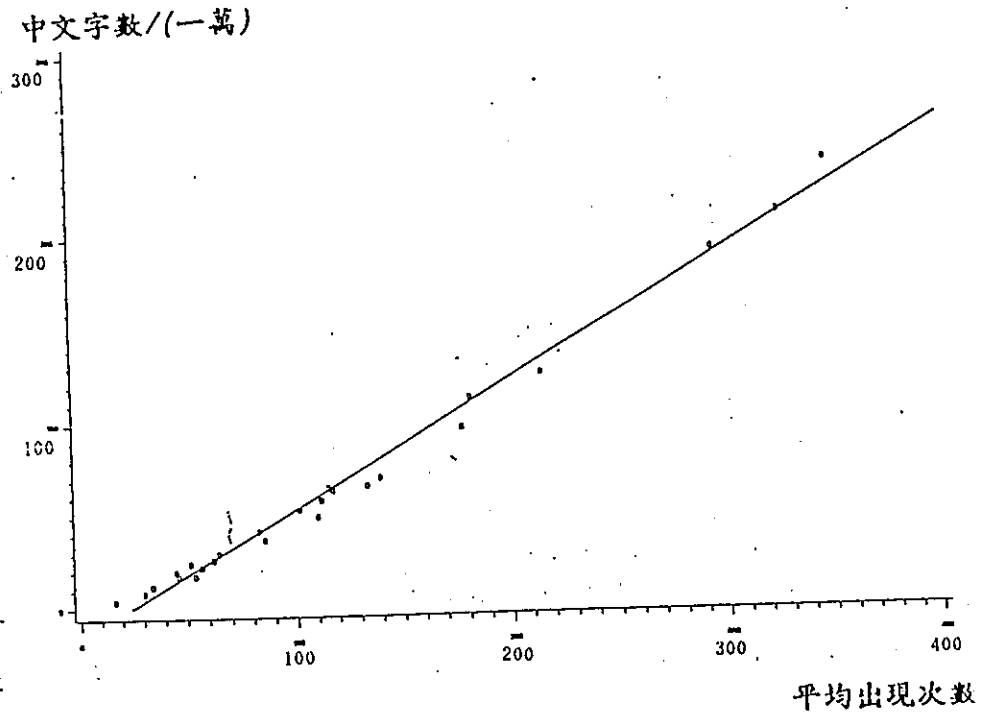
表十三中所示的是依三國志卷二中各階層之統計摘要（原表太長）。大體而言，此範圍內（500字至10000字）之關係亦可用直線作估算。圖十三甚具代表性，即廿五史中其他各列傳之統計亦有極類似之性質。

圖五　二十五史各史列傳之中文字數
(Y:以萬爲單位) 與其平均每個字使用次數(X)的關係



1.4950616107871 62*10^-8 + 0.00003451428253565 37*x + 0.0547880927557086*x^2 - 0.0000244816078874 5915*x^3 + 3.5029861771424 63*10^-9*x^4

圖六　各史列傳之中文字數
(Y:以萬爲單位) 與其平均每個字使用次數(X)的關係



$$Y = -16.769213 + 0.696104 \times X$$

表十三　以三國志卷二爲例，說明文件字數與其每字平均使用次數的關係(約從500字至10000字之範圍內)

文件大小		字集	(4) 比值	(5) 平均每字使用次數
(1) 總字數	(2) 中文字數	(3) 字數	〔(1)/(3)=(4)〕	〔(2)/(3)=(5)〕
619	505	256	2.42	1.97
634	501	258	2.46	1.94
610	504	272	2.24	1.85
692	571	305	2.27	1.87
755	612	343	2.20	1.78
756	617	335	2.26	1.84
890	714	349	2.25	2.04
898	723	369	2.43	1.96
995	813	373	2.67	2.18
1050	870	394	2.66	2.20
1106	885	398	2.78	2.22
1252	981	538	2.33	1.82
1201	966	446	2.69	2.17
1363	1008	455	3.00	2.22
1754	1437	623	2.82	2.31
2230	1829	630	3.54	2.90
2321	1834	690	3.36	2.66
2978	2387	721	4.13	3.31
3145	2522	765	4.11	3.30
4807	3918	1008	4.77	3.89
5139	3817	977	5.26	3.91
5343	4251	1156	4.62	3.68
6271	5101	1097	5.72	4.65
6413	5073	1140	5.63	4.45
7261	5526	1137	6.39	4.86
73151	58709	2752	26.58	21.33

表十四　史記列傳卷八十六 刺客列傳

(a) 與『國民小學常用字彙表』之比對

新校本史記三家注/新校本史記/列傳/卷八十六 刺客列傳第二十六

不屬該字表

路徑 2.1.5.26(卷標頭+段)，層數 $a，費時 6 秒，40 節
總計：字數 7373(中文 5709，註解碼 124)
　　　中文(符號)字集數 928(20)，中文字數與字集數比值 6.152
句長：平均值 5.622，標準差 2.700
各節平均值：字數 184.325，字集數 80.775，比值 1.574
各節標準差：字數 150.177，字集數 55.523，比值 0.371

比對「國民學校常用字彙表」(2732 字)
766 字屬該表，出現 5085 次；162 字不屬該表，出現 624 次
(不)屬該表的字集字數 / 該表字數：　0.280 (0.059)
(不)屬該表的字集字數 / 字集字數：　0.825 (0.175)
(不)屬該表的字之出現次數 / 中文總字數：0.891 (0.109)

字	節數	次數	字	節數	次數	字	節數	次數	字	節數	次數
眦	1	1	莒	2	5	恭	7	8	即	1	1
伉	1	1	乞	1	1	匕	5	11	尸	1	2
亢	2	2	兮	1	2	勿	1	2	曰	25	88
毋	2	3	叱	3	3	札	1	5	刎	2	2
妄	1	1	朴	1	1	劫	3	3	匣	1	1
矣	14	23	肖	1	1	邑	5	6	邪	3	4
卒	7	9	妾	1	4	忿	1	1	怡	1	1
歿	1	1	枉	1	1	沫	5	9	泄	2	4
泣	4	5	炙	2	2	臾	1	1	俠	5	8
哉	8	8	狎	1	1	殆	1	1	畏	2	2
否	1	3	扈	1	1	衍	1	1	倚	1	1
俱	2	3	冤	1	1	恙	1	1	恣	1	1
挾	1	1	晏	1	1	涕	5	5	逕	1	1
荊	16	51	豈	8	8	逆	2	2	陛	2	2
臣	2	2	庶	1	1	徙	2	2	斬	2	2
焉	4	4	祖	1	1	敖	2	3	頃	5	5
俾	2	4	咽	1	1	愕	1	1	悍	1	3
筑	3	11	果	1	1	詔	2	2	軻	15	46
酣	3	3	傷	1	1	既	5	5	嗜	1	1
嗣	2	2	嫁	1	2	惡	1	1	持	1	1
溢	2	2	宕	1	2	詣	1	1	誅	5	5
辟	1	1	遂	13	17	遁	1	1	敦	1	1
侯	4	11	候	1	1	誡	1	1	鄙	1	1
嘿	2	2	戮	1	1	暮	1	1	燕	5	16
款	1	1	發	1	1	頭	1	1	隙	1	1
蔽	1	1	踞	1	1	閭	1	2	篤	2	2
魄	2	2	彊	2	2	擅	2	2	孰	1	1
罪	1	1	謁	1	1	豫	5	17	鄣	1	1
殷	1	1	儒	1	1	擘	1	1	濡	2	2
濮	2	2	纏	1	1	剄	1	1	蹊	1	1
聶	8	22	讓	3	3	圖	1	2	寵	2	2
曠	1	1	轟	2	3	與	1	1	壇	1	2
忤	1	1	邯	1	1	倪	1	1	到	1	1
柰	3	6	睞	1	4	偃	1	1	掩	1	1
憯	1	1	搖	2	2	毀	1	1	盛	1	1
焠	1	1	馭	3	3	儇	3	3	措	1	1
睚	1	1	偕	1	1	弱	1	1	鄲	1	1
鎰	1	1	騏	1	1	曜	1	1	闢	1	1
腸	1	1	譬	1	1	驟	2	2	櫝	1	1
黨	1	1	臨	4	6						
污	1	2	潔	1	1						

五、字集比對

在本計劃發展的文字統計軟體中，將某文件與既定字彙比對亦是一項重要的功能。目前已建在系統內的字彙表如表三所示。事實上，表三中之字彙表可以隨意地增加，譬如可將廿五史每個史的字彙亦納入字彙集合之中以備後用。

茲舉一例以說明字集比對之功用。史記刺客列傳總共5709字次（不含標點）。此段文章與《國民小學常用字彙表》和《教育部標準國字常用字彙表》比對之結果如表十四(a)與(b)。由比對結果知：有43個字（共出現98字次）爲《教育部標準常用字彙表》中所無，約佔1.4%的出現機會；而對《國民小學常用字彙表》來說，則有162字不在表中（共計624字次），約佔使用機會之10.9%。

表十四

(b) 與『教育部標準國字常用字彙表』之比對

新校本史記三家注/新校本史記/列傳/卷八十六 刺客列傳第二十六

不屬該表

路徑 2.1.5.26(卷標頭+段), 層數 $a, 費時 6 秒, 40 節
總計: 字數 7373(中文 5709, 註解碼 124)
　　　中文(符號)字集數 928(20), 中文字數與字集數比值 6.152
句長: 平均值 5.622, 標準差 2.700
各節平均值: 字數 184.325, 字集數 80.775, 比值 1.574
各節標準差: 字數 150.177, 字集數 55.523, 比值 0.371

比對「教育部標準國字常用字彙表」(5343 字)
885 字屬該表, 出現 5631 次; 43 字不屬該表, 出現 78 次
(不)屬該表的字集字數 / 該表字數: 0.166 (0.008)
(不)屬該表的字集字數 / 字集字數: 0.954 (0.046)
(不)屬該表的字之出現次數 / 中文總字數: 0.986 (0.014)

【字/	節數/	次 數】	【字/	節數/	次 數】	【字/	節數/	次 數】	【字/	節數/	次 數】
徧	1	1	眦	1	1	堊	2	5	衆	7	8
卽	1	1	伉	1	1	泄	2	4	款	1	1
亶	1	2	忤	1	1	邯	1	1	俛	1	1
剄	1	1	柰	3	6	眛	1	4	偟	1	1
捥	1	1	愊	1	1	揕	2	2	戮	1	1
轟	1	1	焠	1	1	軹	3	3	僂	1	1
搡	1	1	踁	1	1	佶	1	1	聶	3	3
郰	1	1	諸	1	1	騏	1	1	餽	1	1
鬩	1	1	腸	1	1	醳	1	1	臞	1	1
檷	1	1	黛	1	1	韓	4	6	驟	2	2
贏	1	1	污	1	2	濮	1	1			

伍、檢討、未來的工作與結語

由於這個計劃並未申請經費補助，是故人力頗爲拮据，工作在時斷時續下進行了約有二年時光，到目前尚未完全完工。然而，工作人員均覺得這是值得做的事。目前已做的部份多屬於基本之文字統計工作，正在進行中尚未完工的有：

- 各種熵之量測
- n-連 (n-gram) 馬可夫程序相關的機率和熵之量測

若有經費支援，則擬做下列之統計分析

- 對於聲韻，如聲、韻、調之頻率與熵之分析
- 對於聲韻之 n-連量測
- 對於詞頻與相關熵之量測
- 對於分級、分業字彙及詞彙之建立
- 對於字、詞在語法分類上之統計與分析等

至於應用方面，尚可：

- 做字、詞之自動索引、自動分類之套裝軟體
- 做文件檢索 (document retrieval) 之研究
- 做文件認別 (document identifictaion) 之特徵表達 (signature) 研究等

在文件檢索方面已有一些初步之研究發表[30]，目前的這些統計結果顯示：中文文件檢索的方法上是和拼音文字不同的，我們有把握發適合中文文件的檢索方法。在文件認別和表達方面，不僅有助於文學之考據 [2]，更有助於歷史語言學方面的計量分析。

總之，語文之統計是了解漢語的有力工具與方法，希望本文能夠拋磚引玉以促使更成熟的語文統計早日展開。

誌　謝

本文承丁邦新先生和管東貴先生的鼓勵，才得以進行，在此深表感謝。又在工作期間，陳克健先生、何大安先生、曾士熊先生、丁之佩先生對本工作諸多建言，使得本研究獲益頗多。本文承林靜萍小姐細心的打字、校對和排版亦功不可沒。謹借此篇幅一並致謝。

參考資料

1. C. E. Shannon & W. Weaver, 《The Mathematical Theory of Communication》, Univ. of Illinois press; 1949。

2. 徐秉尊、蔡偉鴻，<從信息論角度探討《紅樓夢》的作者>；中文信息學報，第四卷第二期；1989。

3. Warren Weaver, <Recent Contributions to the Mathematical Theory of Communication>, 出處同1.; 1949。

4. Abraham Bookstein & Shmuel T. Klein, <Compression, Information Theory, and Grammars: A Unified Approach>; ACM Trans. on Information Systems, Vol. 8, No.1, pp 27-49; 1990年1月。

5. John M. Sinelair,《Looking Up: An account of the COBUILD Project》, Collins publishers and the Univ. of Birmingham press, London; 1987. ISBN:0-00 370256-1。

6. John M. Sinclair, 《Corpus, Concordance, Collocation》; Oxford Univ. Press; 1991。 ISBN:0-19-437144-1

7. 鄭錦全，<電腦在漢語音韻研究上的應用>，《思與言》雜誌；1972, 第9期, pp 344-348。

8. 劉源、梁南元，<漢語處理的基礎工程>，中國文信息學報，第一卷第一期,pp.17-25; 1986。

9. 艾偉，《漢字問題》，台灣中華書局印行；1965 台二版。

10. 黃得時，<歷代字書與常用字數>，圖書館學報第七期，東海大學印行；1965。

11. 徐中舒主編，《漢語大字典》，北京國家出版事業管理局規劃出版，1990。

12. 謝清俊、黃克東，《國字整理小組十年》，台北，國字整理小組出版；1989年12月。

13. 謝清俊，<談中國文字在電腦中的表達>，台北，海基會主辦：中國文字的未來研討會；1991年6月21日。

14. 謝清俊,<On the formalization of Glyphs> a contritution paper to the ISO/JTCI/SC18/WG 8 AFII SWG meeting at Kyoto, 1991年2月6日。

15. 黃賢，<全漢字庫及其編碼>，煙台市，中國文信息學會漢字編碼專業委員會第四屆年會論文；1988年10月。亦發表於《深圳大學學報（人文社會科學版）》，1989年。

16. 林樹，《中文電腦基本用字之研究》，交通大學計算機工程研究所技術報告 CS-001 號；1972年6月。

17. 王力，《中國語言學史》，台灣，駱駝出版社；1987年7月台一版。

18. 劉英茂、莊仲仁、王守珍，《常用中文詞的出現次數》，台北，六國出版社；1975。

19. 王還、常寶儒等，《現代漢語頻率詞典》，北京語言學院出版；1986年。

20. 王德造、張社英、劉源，<漢語言的幾個統計規律>，中文信息學報，第一卷第四期；1988。

21. 黃昌寧，<大陸計算語言學研究的回顧與展望>，台灣，計算語言學學會ROCLing IV年會論文；

1991年9月。（註：作者在北京清華大學計算機科學系述職）

22. 蘇克毅，<An Introduction to Corpus Based，Statistical Orieccted Techingues of Natual Language Processing>，台灣，計算語言學學會年會會前講習講義，中華民國計算語言學學會出版；1991年9月。

23. 謝清俊等，（On the Antomation of Chinese History Literatures），ROC-Japan Symposium on Information management and exchange，台北；1986年，11月。

24. 謝清俊等，《中文全文處理系統的設計與製作》史籍目動化計劃第二期研究報告，中央研究院計算中心；1986年9月。

25. 馬立君，《我國主要報紙使用字彙之內容分析——以新聞全文資料庫及新聞索引資料為研究工具》，輔仁大學傳播研究所碩士論文，1989。

26. 陸念慈，《中文文獻資料庫之文字統計系統套裝軟體使用手冊》，中央研究院計算中心，1989年 3月。

27. 楊志明，《廿五史全文資料庫使用手冊（修訂版）》，中央研究院計算中心，1990年 7月7日。

28. 聯合早報用字詞調查工作委員會，《南洋·星洲聯合早報用字用詞調查報告書》，新加坡；1988年5月。

29. 盧紹昌等，《小學華文教材字詞頻率詞典初稿(1)》，新加坡國立大學華語研究中心；1988。

30. 曾士熊、楊鍵樵、謝清俊，<An Experimental Model Of Chinese Textual Database>，Jounal Of The Chinese Institute Of Engineers. Vol. 13. No 6. , pp. 607-622; 1990年6月。