

語料統計分析法在警察通識教育之運用及實例

曾榮汾 Tzeng Rung-Fen

中央警察大學通識教育中心教授

摘要：語料統計分析是一種利用語言資料為樣本，進行對語言環境綜合或專題觀察的研究法。警察學術範疇涵括所有的社會領域，語言的使用也正反映著社會環境的情況，所以利用語料統計來協助警學研究當是可行的方法。嚴格來說，警察的通識教育和其他專業知識一樣，都是警察養成教育中不可或缺的部分，它有如八面玲瓏的窗口一般，可讓警察認識全方位的社會，為達此目的，透過語言使用情形的觀察，進而了解到社會變遷的實況，是相當具有意義的。本文試舉三個實例來說明此法的運用模式，並期盼警大通識中心能積極擔負此長期的任務，既能協助警學的研究，也能成為警大通識教育的一種特色。

關鍵詞：語料／詞頻／警察 language data / word frequency /police

綱目：

- 一、前言
- 二、綜合語料的統計分析
- 三、特殊語料的統計分析
- 四、專題語料的統計分析
- 五、結語

一、前言

語料統計分析法是指將語料樣本作字詞出現頻次的統計，並進行各種語言屬性的分析，藉以了解語料中所具涵的語言資訊。這種方法常用在對語言環境的了解上。所謂的「語言環境」，從廣義來說，事實上就等於「社會環境」。社會環境的變遷，往往會反映在該時空中人們所使用的語言上，於是透過語料樣本的適當收集和統計，分析的結果也就呈現了社會環境的變異情形。

警察的學術，從廣義來說，就是整體的社會學問。在這一點上，警察學術除某些特殊的專業知識和技能外，幾皆為「通識」。在學術領域中大概再也無法找出第二種學術類似警學這麼廣泛的。甚至可以這麼說：

社會有什麼，警察就該學什麼。

一般總以為警察最重視「犯罪」問題，但就以「犯罪」而論，犯罪對象，士、農、工、商無所不包；犯罪手法，金、木、水、火、土概皆有之，幾可謂犯罪萬象有如「社會萬象」。更何況，警察工作實以服務為宗旨，但服務對象為何？社會各角落皆是。這也就是為什麼稱「警政為內政基礎」，警察為「國之楨榦」的原因了。

警學既然如此廣博，則警察養成教育中，「通識教育」必為重要。藉著通識教育伸入各領域，從修身至安民，由個人至群體，整個社會知識盡可括之。由此可知，警察不能不了解社會，因為他的職場就在社會。也因為如此，警察教育不能滿足於少數的「專業知識」，如從以上分析來看，「通識教育」不也是警察的專業教育？或者可說：因為有了通識的結合，讓警學包容了全社會。但是，「社會」

到底是什麼？若只憑個人感受，恰如摸象；如求宏觀，則語料統計應當可以幫上一點忙。本文將試舉幾個統計實例來說明此點。

二、綜合語料的統計分析

筆者曾幫教育部進行過民國八十四年至八十七年的常用語詞調查，這四年的調查工作主要就是在藉語言來記錄社會。因此所採用的樣本是社會環境全方位的考量，試舉八十七年調查報告書的樣本為例：¹

本次調查的資料來源有五：

- (一)八十七年度出版之雜誌。
- (二)八十七年度暢銷之書籍。
- (三)八十七年度印行之報紙。
- (四)八十七年奇摩站分類索引中之各類網站。
- (五)八十七年口語調查之資料。

資料蒐輯細目如下：

(一)八十七年度出版之雜誌

雜 誌 名 稱	雜 誌 月 份
新新聞	1-12
財訊月刊	1-12
錢雜誌	1-12
廣告雜誌	1-12
天下雜誌	1-12
遠見雜誌	1-12
汽車購買指南	1-12
柯夢波丹	1-12
解讀時代美語	1-12
博覽家	1-12
張老師月刊	1-12
科學月刊	1-12
跳蚤月刊	1-12
新女性	1-12
當代設計	1-12
家庭月刊	1-12
城市情報	1-12
風尚	1-12
GQ 瀟灑	1-12
VOGUE 時尚	1-12
大地	1-12
吃在中國	1-12
茉莉雜誌	1-12
哈潑時尚中文版	1-12
就業情報	1-12
健康世界	1-12
育兒生活	1-12
嬰兒與母親	1-12
美食天下	1-12
牛頓雜誌	1-12
尖端科技雜誌	1-12
光碟月刊	1-12

¹ 參曾榮汾，《民國八十七年常用語詞調查報告析介》，載《民國八十七年常用語詞調查報告》，教育部，國語文教育叢書第 38 種，2000-6。

雜誌名稱	雜誌月份
電腦繪圖與設計	1-12
新遊戲時代雜誌	1-12
PC HOME 電腦家庭	1-12
管理雜誌	1-12
讀者文摘	1-12
講義	1-12
傳記文學	1-12
藝術家雜誌	1-12
職業棒球	1-12
NBA 美國職籃	1-12
釣魚人	1-12
D&D 裝飾設計	1-12
雅砌 ARCH	1-12
媽媽寶寶	1-12
國際網球雜誌	1-12
中央月刊	1-12
世界電影	1-12
二手車訊	1-12
PC WORD 電腦世界	1-12
一手車訊	1-12
超越車訊	1-12
汽車百科	1-12
0 與 1 BYTE 科技	1-12
網路生活雜誌	1-12
麥客情報雜誌	1-12
電腦玩家	1-12
新電子科技	1-12
普門	1-12
聯合文學	1-12
光華雜誌	1-12
推理	1-12
皇冠	1-12
高手模型雜誌	1-12
今周刊	1-12
儂儂	1-12
電台雜誌	1-12
景秋高爾夫雜誌	1-12
中古車購買指南	1-12
車主汽車專業雜誌	1-12
世界民航	1-12
薇薇雜誌	1-12

(二)八十七年度暢銷之書籍

書名	出版社	作者
IC 教父張忠謀的策略傳奇	天下雜誌	楊艾俐
Go,Go,Go 健康飲食	天下遠見出版有限公司	郝龍斌
傷心咖啡店之歌	九歌出版社	朱少麟
善待自己 II	精美出版有限公司	賴志松
投資股票的第一本書	商業周刊出版有限公司	黃培源、楊偉凱
痛快日記	皇冠文化出版有限公司	蔡康永
鐵達尼號沉沒記	九歌出版社	華特勞德原著 黃文範譯
溫柔雙城記	大田出版有限公司	張曼娟

書 名	出 版 社	作 者
第三謊言	小知堂文化事業有限公司	雅歌塔·克里斯多著 簡伊玲譯
全球資本主義危機	聯經出版事業公司	索羅斯著 聯合報編譯組譯
鹹魚的滋味	圓神出版社有限公司	陳昇
攀上心中的巔峰	水雲齋文化事業有限公司	劉墉
觀念許文龍和他的奇美王國	商業周刊出版股份有限公司	黃越宏
品牌行銷法則	商業周刊出版股份有限公司	大衛·愛格著 沈雲驄、湯宗勳譯
幽默高手	時報文化出版企業股份有限公司	戴晨志
失樂園(上、下)	麥田出版股份有限公司	渡邊淳一著 譚玲譯
夏先生的故事	小知堂文化事業有限公司	徐四金著 彭意如譯
別跟豬打架	臉譜文化事業股份有限公司	BOB WOUNDED 著 沈雲驄譯
人子	遠景出版事業公司	鹿橋
心靈會說話	商業周刊出版股份有限公司	南西·奧哈拉著 邱紫穎譯
桃子罐頭	尖端出版有限公司	櫻桃子著 銀花譯
101 次求婚	台灣東販股份有限公司	野島伸司著 孫慧敏譯
醫生也瘋狂	麥田出版股份有限公司	歐陽林著
第一次的親密接觸	紅色文化事股份有限公司	蔡智恆
對錯都是為了愛	水雲齋文化事業有限公司	劉墉
李光耀回憶錄	世界書局	李光耀
綠化心靈	高寶國際有限公司	潘台成
昨日的叛逆	圓神出版社股份有限公司	光禹
總裁獅子心	平安文化有限公司	嚴長壽
呆伯特法則	經典傳訊文化股份有限公司	史考特·亞當斯著 張定綺譯
我坐在琵琶河畔，哭泣	天下遠見出版股份有限公司	保羅·科賀著 許耀雲譯
張忠謀自傳(上冊)	天下遠見出版股份有限公司	張忠謀
海德堡之吻	圓神出版社有限公司	鄭華娟
最後 14 堂星期二的課	大塊文化出版股份有限公司	Mitch Albom 著 白裕承譯
失戀雜誌	元尊文化企業股份有限公司	水瓶鯨魚
暗夜倖存者	平安文化有限公司	徐璐
湖濱散記	高寶國際有限公司	亨利·梭羅著 陳柏蒼譯
突圍	商智文化事業股份有限公司	沈慶京
蒼天有淚人間有天堂	皇冠文化出版有限公司	瓊瑤
Easy 談效率	精美出版股份有限公司	朱凱蕾
鮮活思維	聯經出版事業公司	施振榮
控股公司與股權規劃	知道出版有限公司	楊聰權
別為小事抓狂	時報文化出版企業股份有限公司	理察·卡爾森著 朱衣譯
非常誠實有點毒	方智出版社股份有限公司	吳淡如
香水婚紀念日	圓神出版社有限公司	鄭華娟
讓我躺在你身邊	麥田出版股份有限公司	小野
快樂是自找的	水晶圖書股份有限公司	朱凱蕾

書 名	出 版 社	作 者
還珠格格水深火熱	皇冠文化出版有限公司	瓊瑤
我不是教你詐 3	水雲齋文化事業有限公司	劉墉
轉個念頭，人生會更好	高寶國際有限公司	王尚智
讓自己變生活高手	圓神出版社有限公司	曹又方
別為小事抓狂得意人生 100 招	時報文化出版企業股份有限公司	理察·卡爾森著 朱恩伶譯
潛水鐘與蝴蝶	大塊文化出版股份有限公司	Jean-Dominique Ba 邱瑞璽譯
自戀總比自卑好	方智出版社股份有限公司	吳淡如
你是 EQ 高手嗎	時報文化出版企業股份有限公司	戴晨志
陌生人	聯經出版事業公司	李家同
橄欖樹	聯經出版事業公司	蔡素芬
二人證據	小知堂文化事業有限公司	雅歌塔·克里斯多 著 簡伊玲譯
還珠格格陰錯陽差	皇冠文化出版有限公司	瓊瑤
唐安麒瘦身健美湯方	圓神出版社有限公司	唐安麒
在勇氣邊緣	圓神出版社有限公司	光禹
共同基金聖經	商業周刊出版股份有限公司	蔣國樑、林一銘、 文琪、陳蟬虹 著
男女溝通高手	時報文化出版企業股份有限公司	戴晨志
活得更聰明	方智出版社股份有限公司	吳淡如
你只要負責笑就好	皇冠文化出版有限公司	小野
牧羊少年奇幻之旅	時報文化出版企業股份有限公司	保羅·科爾賀著 周惠玲譯
連戰風雲	時報文化出版企業股份有限公司	李建榮
李敖快意恩仇錄	李敖出版社	李敖
台灣念真情之這些地方這些人	麥田出版股份有限公司	吳念真

(三)八十七年度印行之報紙

報 紙 名 稱	收 錄 月 份
聯合報	1-12
民生報	1-12
經濟日報	1-12
聯合晚報	1-12
中央日報	1-12

(四) 八十七年奇摩站分類索引中之各類網站：參考奇摩站之分類索引，共蒐錄一百六十個各類型網站的文字用語。

(五)八十七年口語調查之資料，所包含的調查項目如下：

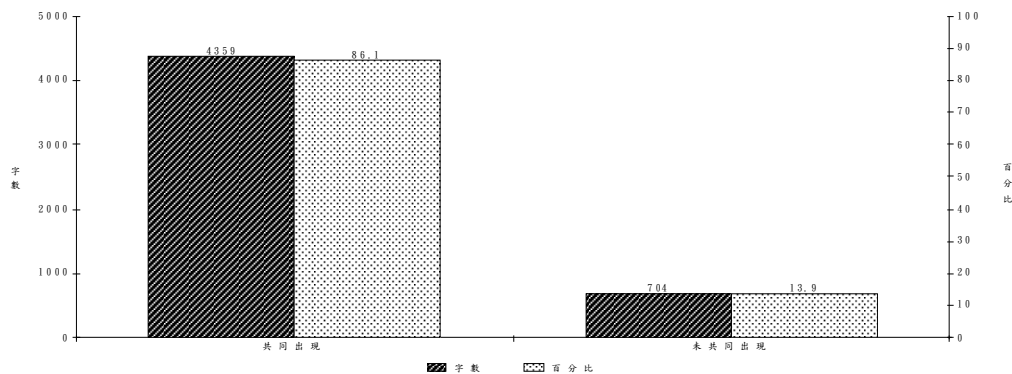
- 1.書面資料：包括演講稿、新聞稿、劇本、廣告詞等。
- 2.錄音資料：從各電視、電臺節目對話直接錄音取樣的資料。
- 3.口語問卷：透過問卷設計，廣收全國各級學校學生口頭用語資料。
- 4.BBS 討論站文章：以中華電信 Hinet 討論區為主要採集對象，選取口語程度較高、看版人數較多、討論公共事務及流行話題之資料。

從樣本採錄來看，可以知道這份報告所要追求的是社會環境的整體觀察，所得到的結果，無論用字的頻次或用詞的頻次資料，必能說明當年社會活動的概況。如果能持續逐年的觀察，當也可以了解到社會環境變遷的過程。例如八十七年統計樣本為 1,579,771 字，單字數為 5,063 字。統計分析後的數據如下：

- 1.出現頻次最高者為「的」字，頻次為 5,4861，佔總數 3.47%。

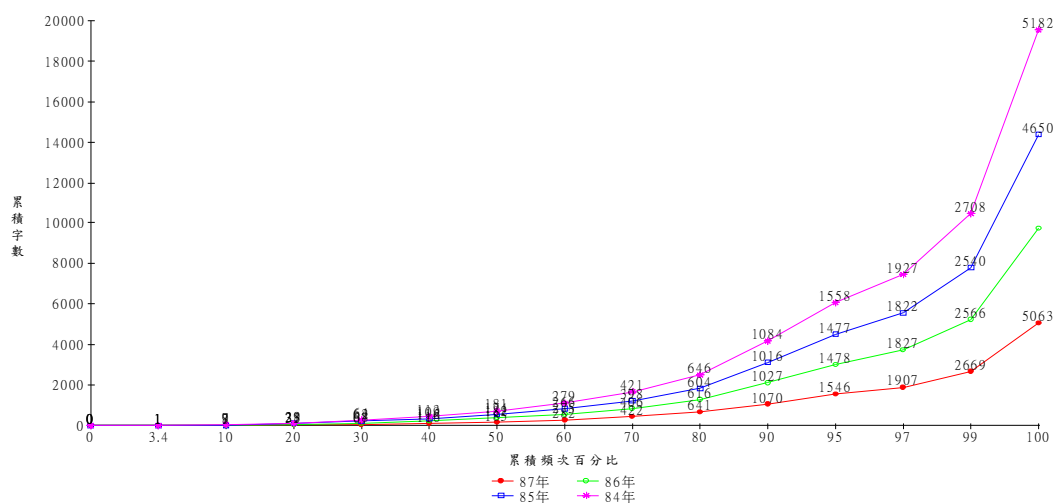
2. 累積百分比 90% 時，累積字數為 1,070 字。
3. 累積百分比 99% 時，累積字數為 2,669 字。
4. 單音節詞與多音節詞混編，總詞數為 62,058，純以多音節詞為準，總詞數為 58,810。
5. 將八十七及八十六年兩個統計結果作一比較，兩年共有字為 4,359 字，未共有字為 704 字。二者共有比率達 86%，正表示語言本身的穩定性質。

圖 13: 本字彙表與《八十六年常用語詞字頻表》共同出現的字數及其百分比



6. 將八十七與八十四至八十六四年字頻表作一比較，其分布曲線如下：

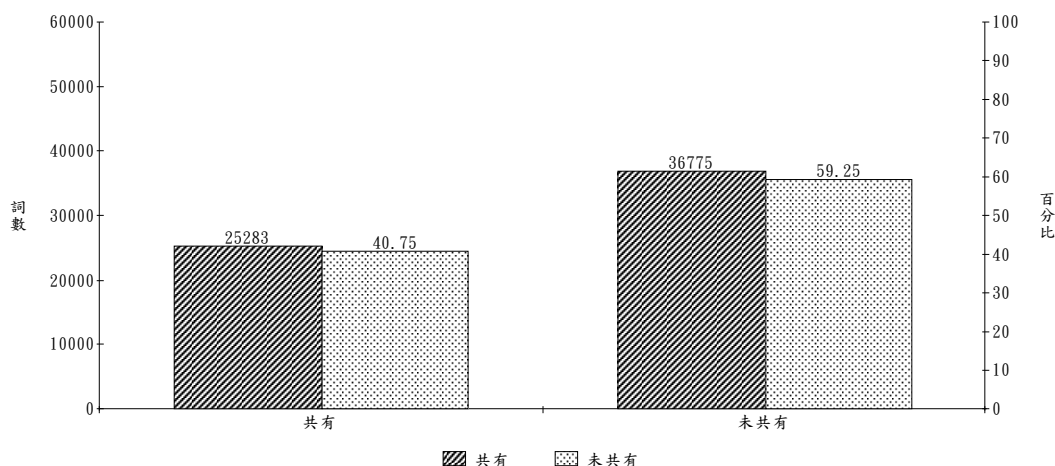
圖 3: 87年86年85年84年累積字數頻次百分比比較圖



由上圖可知四個年度的分布曲線十分相似。

7. 八十七年詞頻表共收 62,061 詞，八十六年共收 47,909 詞。以八十七年為主，二者共有與未共有情形如下：

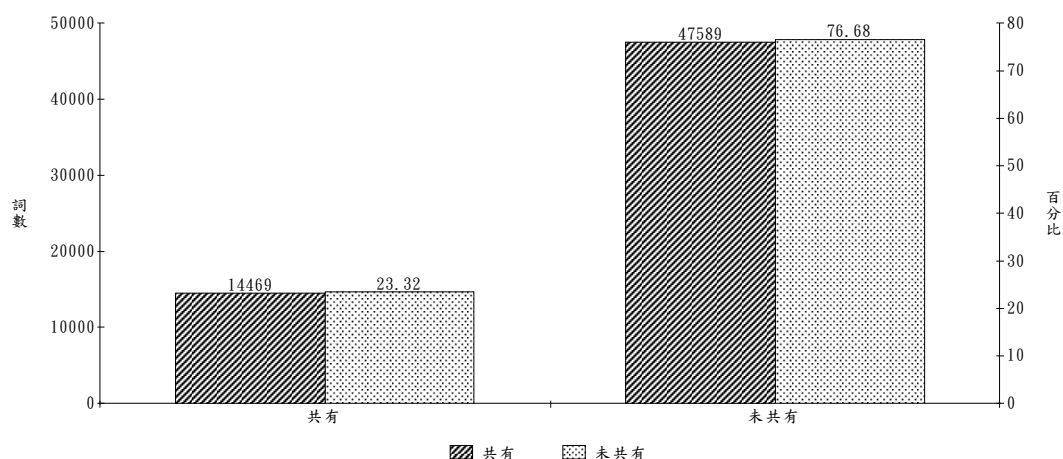
圖15:87年與86年詞頻表共同出現的詞數及其百分比



二者共有在 40%左右。

8. 八十七年與八十四至八十六年收詞比較，以八十七年為主，四者共有與未共有情形如下：

圖17:87年與84-86年詞頻表共有詞數及其百分比



四年共有詞為 14,469 詞，百分比為 23.32%。這一部分當是國內語言結構相當穩定的部分。未共有部分的比例越大，當表示出國內語詞周邊部分新陳代謝的快速現象。

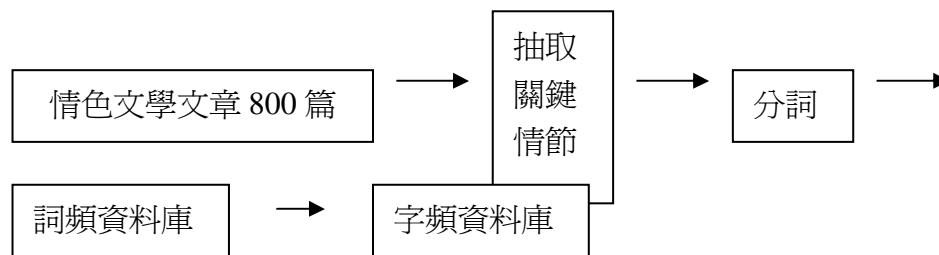
這種統計所得到的結果因為包括社會各階層，從四個年度統計的出現字數情形來看，呈穩定狀態，所以這種綜合觀察可以求得社會常態的語言使用情形。筆者曾藉用此常態標準來比較分析特殊階層的語料。

三、特殊語料的統計分析

筆者在民國八十九年嘗以網路情色文學為分析對象，求其使用語料之特色，²並以民國八十六年常用語詞調查報告為比較用的常態標準。³

² 《刑事案件資料庫系統之建立與綜合應用—犯罪語料庫建立之研究》，曾榮汾，2000-9，國科會專題計畫報告。

在八百篇網路情色文學作品的分析中，求得所出現的字詞頻次。分析的方法如下：⁴



這種分析法的特色在第二步驟，先將非情色情節的文字刪汰，只留下關鍵情節，如此可以避免統計結果在高頻次部分與一般文學作品相似。統計結果如下：

1. 收詞總樣本 41,428 詞，累計後為 6,938 詞。
2. 收字總樣本 185,074 字，累計後為 2,877 字。

頻次最高的前 100 個字詞如下，表中並列教育部《八十六年常用語詞調查報告書》以資比較，常用語詞未收者，以「*」表示：

1. 情色文學語料頻次最高前 100 個詞表：

詞目	詞頻	累積百分比	86 年常用語詞調查	
			詞序	詞頻
1 自己	414	0.99932413	6	960
2 男人	399	1.96244086	665	76
3 起來	289	2.66003669	142	233
4 已經	270	3.31176982	46	416
5 出來	261	3.94177851	198	185
6 開始	243	4.52833832	30	504
7 乳房	242	5.11248431	29581	1
8 陰戶	242	5.69663030	36903	1
9 太太	233	6.25905185	1101	48
10 陽具	218	6.78526600	9112	5
11 肉棒	214	7.30182485	*	*
12 身體	207	7.80148692	428	108
13 屁股	201	8.28666602	6407	7
14 女人	191	8.74770686	346	128
15 可以	183	9.18943710	1	1227
16 感覺	182	9.62875350	279	150
17 陰莖	181	10.06565608	7730	6
18 媽媽	171	10.47842039	292	144
19 沒有	163	10.87187409	4	1032
20 陰道	163	11.26532780	4504	11
21 雙手	156	11.64188472	3728	14
22 先生	153	12.01120015	220	172

³ 《八十六年常用語詞調查報告書》，教育部，1999-6。

⁴ 情色文學樣本取自 [www.members.easyspace.com/sexcap/story.](http://www.members.easyspace.com/sexcap/story/)，目前已關站。

23	我們	152	12.37810177	2	1169
24	手指	148	12.73534807	1621	32
25	快感	144	13.08293908	4743	10
26	雞巴	144	13.43053008	*	*
27	覺得	144	13.77812108	233	168
28	知道	143	14.12329825	106	287
29	龜頭	140	14.46123395	14785	3
30	舌頭	139	14.79675582	12527	3
31	乳頭	138	15.12986386	*	*
32	精液	135	15.45573042	40902	1
33	慢慢	134	15.77918316	1222	44
34	內褲	132	16.09780825	11973	3
35	時候	132	16.41643333	104	301
36	用力	131	16.73264459	6332	7
37	大腿	130	17.04644202	7083	6
38	進去	130	17.36023945	2090	25
39	一樣	128	17.66920923	140	235
40	現在	127	17.97576518	55	393
41	然後	124	18.27507966	244	164
42	同時	121	18.56715265	77	348
43	插入	121	18.85922564	6738	7
44	舒服	119	19.14647099	3672	14
45	撫摸	117	19.42888867	7933	6
46	什麼	116	19.70889254	40	446
47	輕輕	116	19.98889640	2365	22
48	緊緊	114	20.26407261	6089	8
49	呻吟	112	20.53442116	17473	2
50	高潮	106	20.79028676	1867	28
51	下來	105	21.04373853	568	87
52	但是	105	21.29719031	34	482
53	興奮	105	21.55064208	1757	30
54	雖然	104	21.80168002	59	384
55	抽送	103	22.05030414	*	*
56	東西	103	22.29892826	302	141
57	刺激	102	22.54513855	692	74
58	動作	101	22.78893502	417	110
59	而且	100	23.03031766	63	375
60	還是	100	23.27170030	103	304
61	衣服	98	23.50825529	859	61
62	裡面	97	23.74239645	1169	46
63	突然	96	23.97412378	746	70

64	於是	93	24.19860964	262	157
65	雙腿	89	24.41344019	*	*
66	小穴	88	24.62585691	*	*
67	聲音	88	24.83827363	517	93
68	因為	84	25.04103505	9	858
69	她們	84	25.24379647	716	72
70	所以	84	25.44655788	25	523
71	淫水	84	25.64931930	*	*
72	一陣	83	25.84966689	2908	17
73	臀部	82	26.04760066	14846	3
74	下體	81	26.24312059	23138	1
75	抽插	81	26.43864053	*	*
76	水管	78	26.62691899	8264	5
77	微微	78	26.81519745	20607	2
78	他們	77	27.00106208	5	967
79	第一	77	27.18692672	43	427
80	就是	76	27.37037752	12	723
81	剛才	74	27.54900068	5331	9
82	不要	73	27.72521000	151	222
83	完全	73	27.90141933	155	221
84	這麼	73	28.07762866	307	139
85	不斷	72	28.25142416	235	166
86	胸部	72	28.42521966	13453	3
87	一起	69	28.59177368	176	205
88	不是	69	28.75832770	28	508
89	扭動	69	28.92488172	17150	2
90	真是	69	29.09143574	680	75
91	馬上	69	29.25798977	897	59
92	婦人	69	29.42454379	2840	18
93	眼睛	69	29.59109781	731	71
94	好像	68	29.75523800	560	88
95	享受	68	29.91937820	795	66
96	強烈	68	30.08351839	657	77
97	陽物	68	30.24765859	*	*
98	繼續	68	30.41179878	242	165
99	地方	67	30.57352515	114	270
100	肉體	67	30.73525152	3551	14

2. 情色文學語料頻次最高前 100 個字表：

字	字頻	累積百分比	86 年常用語詞調查	
			字序	字頻
1 的	10469	5.65665626	1	36053
2 我	3547	7.57318694	16	4777
3 一	3465	9.44541102	2	12680
4 了	3015	11.07448912	15	4896
5 她	2807	12.59117974	217	1192
6 不	2561	13.97495056	7	8058
7 在	2197	15.16204329	5	9011
8 著	2120	16.30753104	136	1652
9 是	1802	17.28119563	3	10897
10 來	1638	18.16624701	19	4504
11 上	1634	19.04913710	14	5037
12 人	1323	19.76398630	6	8499
13 到	1302	20.46748868	25	3978
14 他	1271	21.15424101	21	4359
15 下	1267	21.83883203	60	2709
16 這	1199	22.48668100	20	4437
17 有	1170	23.11886056	4	9424
18 手	1168	23.74995948	172	1399
19 出	1124	24.35728411	22	4260
20 小	1108	24.95596356	114	1942
21 身	1044	25.52006225	241	1054
22 大	1014	26.06795120	8	6019
23 那	990	26.60287236	203	1251
24 時	952	27.11726120	18	4543
25 得	932	27.62084355	59	2712
26 也	929	28.12280493	28	3748
27 陰	917	28.61828242	1242	115
28 地	901	29.10511471	36	3428
29 起	897	29.58978571	150	1580
30 好	881	30.06581151	118	1894
31 你	880	30.54129699	183	1351
32 頭	874	31.01354053	298	887
33 子	849	31.47227595	68	2568
34 開	787	31.89751127	88	2232
35 把	786	32.32220625	36	711
36 個	782	32.74473994	23	4068
37 道	762	33.15646714	202	1253
38 中	749	33.56117013	9	6011
39 動	742	33.96209084	63	2629
40 說	740	34.36193090	66	2611
41 然	711	34.74610156	106	2014
42 看	709	35.12919157	149	1586
43 要	697	35.50579768	29	3723
44 就	687	35.87700055	41	3226
45 又	678	36.24334050	295	893
46 用	656	36.59779332	64	2620
47 去	653	36.95062516	147	1611
48 感	649	37.30129570	269	969
49 住	645	37.64980494	429	611
50 後	644	37.99777386	44	3076
51 女	633	38.33979922	198	1276
52 肉	625	38.67750197	799	256
53 快	610	39.00709986	471	533
54 自	602	39.33237516	50	2921
55 太	589	39.65062624	319	851

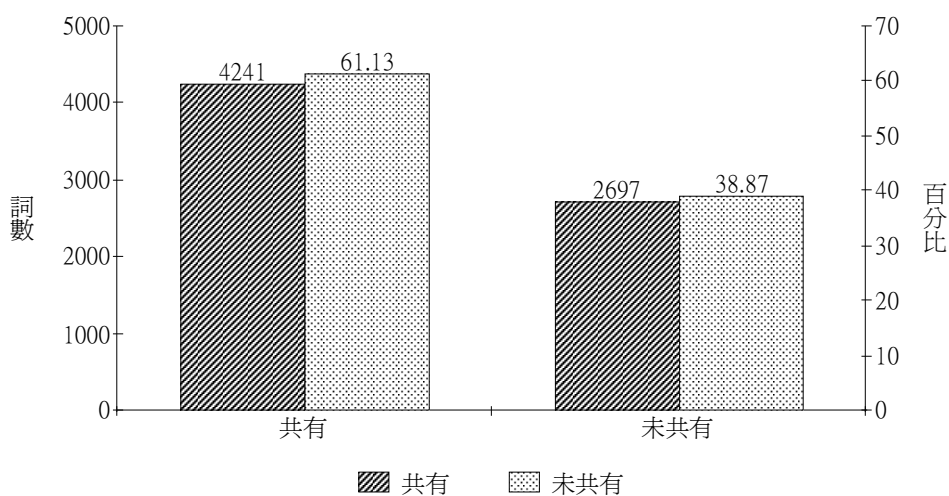
56 過	584	39.96617569	58	2726
57 男	582	40.28064450	549	454
58 樣	578	40.59295201	247	1049
59 已	576	40.90417887	144	1622
60 裡	565	41.20946216	393	680
61 面	557	41.51042286	101	2050
62 體	555	41.81030291	111	1983
63 將	548	42.10640068	112	1974
64 進	541	42.39871619	115	1934
65 啊	532	42.68616878	1189	126
66 而	526	42.97037942	34	3509
67 乳	526	43.25459006	1722	53
68 心	519	43.53501843	71	2524
69 插	501	43.80572095	1457	83
70 口	498	44.07480251	299	885
71 沒	486	44.33740018	168	1417
72 覺	486	44.59999784	438	596
73 媽	485	44.86205518	619	389
74 聲	483	45.12303187	552	452
75 緊	482	45.38346823	864	224
76 兩	475	45.64012233	186	1333
77 入	472	45.89515545	193	1298
78 還	472	46.15018857	164	1452
79 以	459	46.39819748	11	5793
80 想	459	46.64620638	199	1272
81 雙	452	46.89043302	637	375
82 和	451	47.13411933	53	2888
83 可	445	47.37456369	24	4054
84 會	445	47.61500805	13	5551
85 己	423	47.84356528	263	997
86 很	420	48.07050153	194	1292
87 次	415	48.29473616	201	1259
88 腿	411	48.51680949	1763	50
89 力	408	48.73726185	105	2018
90 們	406	48.95663356	56	2784
91 被	401	49.17330365	235	1093
92 讓	401	49.38997374	268	970
93 麼	388	49.59961961	215	1202
94 部	387	49.80872516	108	2009
95 對	384	50.01620973	42	3221
96 內	381	50.22207333	129	1783
97 兒	380	50.42739661	395	673
98 但	379	50.63217956	91	2194
99 之	377	50.83588186	31	3559
100 輕	377	51.03958417	545	456

表中所列雖僅一百個字詞，已可看出情色文學語料與取自一般文獻樣本的《民國八十六年常用語詞調查報告》的差異情形，其全面性的比較情形如下：

1. 詞頻表

共 有：4,241 詞 佔百分比 61.13%
未共有：2,697 詞 佔百分比 38.87%

圖1:情色詞頻與86年詞頻共有及未共有情形

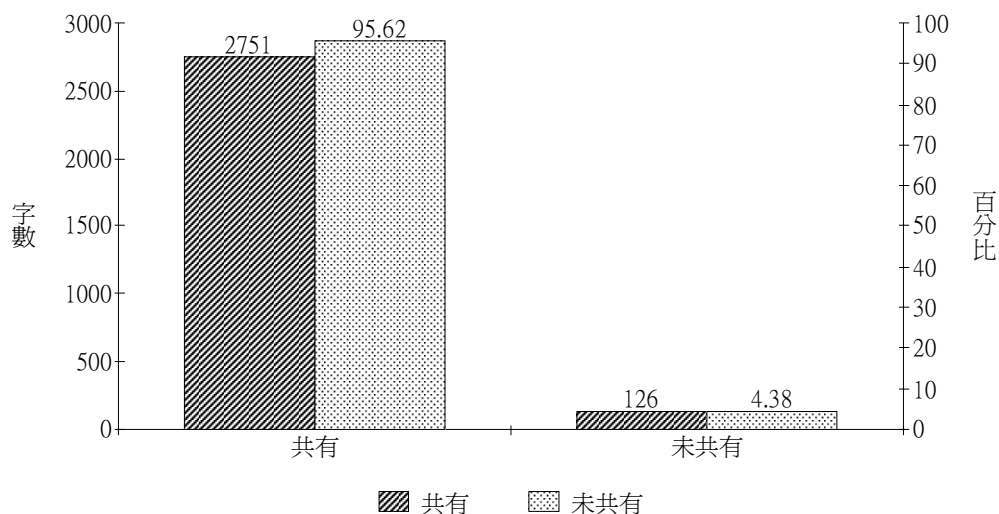


2. 字頻表

共有：2751 字 佔百分比 95.62%

未共有：126 字 佔百分比 4.38%

圖2:情色字頻與86年字頻共有及未共有情形



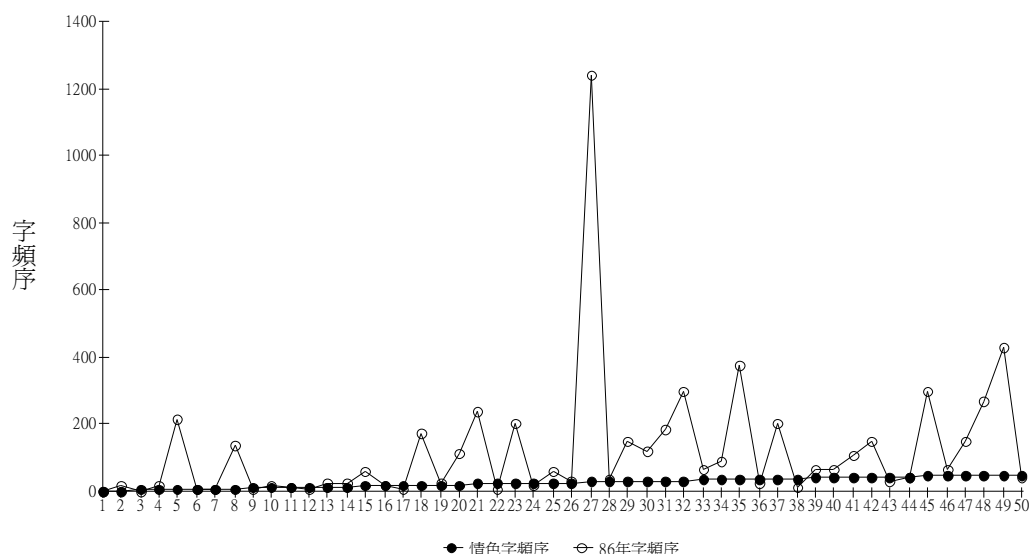
從百分比來看，似乎詞的觀察較字重要，因為單字的使用，無論何種性質的文章，都有其固定性⁵。尤其在詞的未共有部分，以最高頻次前 100 詞為例，未見於 86 年詞頻的詞有 9 個：

肉棒／雞巴／乳頭／抽送／雙腿／小穴／淫水／抽插／陽物

顯然都是情色作品中的特殊用詞，這是字頻差異較難看出的。不過字頻的部分，若進一步從排序先後來看，也可以看出所存在的差異。試以最高頻次的前五十字為例，作一比較，參看〔圖3〕：

⁵ 此點從教育部 84 年、85 年、86 年三年統計的字頻表可以看出。

圖3:情色字頻序與86年字頻序比較
以最高頻次前50字為例



二者使用文字的頻次高低存有相當差異。在情色文學出現最高頻次的 100 個字中，86 年的字頻超出此範圍的用字計 57 字，列之如下：

她／著／手／小／身／那／陰／起／好／你／頭／把／道／然／看／又／去／
感／住／女／肉／快／太／男／樣／已／裡／面／體／將／進／啊／乳／插／
口／沒／覺／媽／聲／緊／兩／入／還／想／雙／己／很／次／腿／力／被／
讓／麼／部／內／兒／輕

許多的用字所以出現頻次異於一般用法，也是因為情色文學的特殊行文需求所造成。所以這種統計的結果是可以用來檢視一篇文章是否為情色文學。凡使用這些字詞較為頻繁的，都有可能是屬於這類的文學。

四、專題語料的統計分析

在專題研究上，筆者也嘗試將此法運用在文學作品的欣賞上面，⁶晉朝王羲之寫過一篇〈蘭亭集序〉，原文如下：

永和九年，歲在癸丑，暮春之初，會於會稽之蘭亭，修禊事也。群賢畢至，少長咸集。此地有崇山峻嶺，茂林修竹，又有清流激湍，映帶左右。引以為曲水流觴，列坐其次，雖無絲竹管絃之盛，一觴一詠，亦足以暢敘幽情。是日也，天朗氣清，惠風和暢，仰觀宇宙之大，俯察品類之盛，所以遊目騁懷，足以極視聽之娛，信可樂也。

夫人之相與，俯仰一世，或取諸懷抱，晤言一室之內；或因寄所託，放浪形骸之外。雖趣舍萬殊，靜躁不同，當其欣於所遇，暫得於己，快然自足，不知老之將至。及其所之既倦，情隨事遷，感慨係之矣。向之所欣，俯仰之間，已為陳跡，猶不能不以之興懷；況修短隨化，終期於盡。古人云：「死生亦大矣！」豈不痛哉！

⁶ 同注 2，頁 10。

每覽昔人興感之由，若合一契，未嘗不臨文嗟悼，不能喻之於懷。固知一死生為虛誕，齊彭殤為妄作，後之視今，亦猶今之視昔，悲夫！故列序時人，錄其所述，雖世殊事異，所以興懷，其致一也。後之覽者，亦將有感於斯文。

這篇文章記述了東晉穆帝永和九年（353AD），王羲之、謝安等人聚集於蘭亭，舉行修禊，飲酒吟詩一事。唐代李白也有一篇〈春夜宴桃李園序〉，內容如下：

夫天地者，萬物之逆旅；光陰者，百代之過客。而浮生若夢，為歡幾何？古人秉燭夜遊，良有以也。況陽春召我以煙景，大塊假我以文章。會桃李之芳園，序天倫之樂事。群季俊秀，皆為惠連；吾人詠歌，獨慚康樂。幽賞未已，高談轉清。開瓊筵以坐花，飛羽觴而醉月。不有佳作，何伸雅懷？如詩不成，罰依金谷酒數。

李白於春夜宴請從兄弟於桃李園，也是飲酒作詩。兩篇文章都具感懷人生短暫，當及時行樂的意念。《古文鑑賞大辭典》評《蘭亭集序》說：⁷

玩此文中段，因樂極生悲，感生死事大，見不可不隨時行樂之意，乃曠達一流。

清人林雲銘評《春夜宴桃李園序》說：

大意謂人生短景，行樂猶恐不及，況值佳辰，豈容錯過？⁸

也因兩篇文字有相似之處，所以或以為二文「若出一人之手」。⁹為了解二文相似之處，下文試著利用語料分析法來觀察這兩篇文章的用字。《蘭亭集序》全文計 322 字，累計後共 206 個字。《春夜宴桃李園序》全文計 117 字，累計後共 101 個字。二文共有字為 35 字。列之如下：

之以懷也為人事有會生夫文大觴清群詠坐幽天地樂惠萬春遊已況古若未作序者

共有字佔《蘭亭集序》用字只六分之一不到，佔《春夜宴桃李園序》三分之一不到。好像從用字來看，二文似無相同之理。但是進一步分析，則共有字中有一些正是二文精神關鍵所在。例舉析述如下：¹⁰

1. 懷

《蘭》：遊目騁懷、猶不能不以之興懷、不能喻之於懷、所以興懷

《春》：何伸雅懷

2. 人

《蘭》：人之相與、古人云、每覽昔人興感之由、列序時人

《春》：古人秉燭夜遊、吾人詠歌

3. 生

《蘭》：死生亦大矣、一死生為虛誕

《春》：浮生若夢

4. 觴

《蘭》：引以為曲水流觴、一觴一詠

⁷ 503 頁。

⁸ 清林雲銘《古文析義·卷十》。

⁹ 警大《國文·蘭亭集序·題解》引過商侯之說。

¹⁰ 為便於比較，二文簡稱為《蘭》、《春》。

《春》：飛羽觴而醉月

5. 詠：

《蘭》：一觴一詠

《春》：吾人詠歌

6. 幽

《蘭》：暢敘幽情

《春》：幽賞未已

7. 樂

《蘭》：足以極視聽之娛，信可樂也

《春》：序天倫之樂事

8. 春

《蘭》：暮春之初

《春》：況陽春召我以煙景

這幾個字於二文中的用法確有相似之處。如果進一步從傷時光之流逝，嘆人生的短暫的意念來看，則《蘭》文有「向之所欣，俯仰之間，以爲陳跡」、「修短隨化，終期於盡」等說法，《春》文則有「浮生如夢，爲懽幾何」、「古人秉燭夜遊，良有以也」的說法。意思也是相近的。

所以，從以上遣詞用字的分析來看，這兩篇文章在創作意念確有相近之處，但是這是否就能判斷這二文「如出一人之手」？創作者的個性是否極其相近？筆者曾利用此點強調了「形上證物」觀念的重要。筆者的看法是：¹¹

如果說語言可爲「形上證物」，分析「證物」所要追求的目的當就是創作者個性。這點似乎很難單憑實體的語料分析得到答案，也正是本文所要特別強調的一點。事實上，以這兩篇文章而論，寫作《蘭》文時，王羲之的感情是痛苦的，而寫《春》文的李白則是曠達快樂的。這裡的痛苦和快樂已難從字面上了解，而必須從字裡行間「蘊涵」的感情去「品味」。

《蘭》文寫古今詩人創作靈感若合一契，皆出於靈感中哀喜的情感。樂的時候不知老之將至，哀的時候，情隨事遷，向之所欣，已爲陳跡。於是，人就這樣浮浮沉沉，來到了生命的盡頭。古人如此，今人如此，想來後人也將如此。作者從這角度看事情，情感怎會是快樂？反觀李白則不同。正因浮生若夢，所以希望學古人秉燭以夜遊。「群季俊秀，皆爲惠連；吾人詠歌，獨慚康樂」，說自己的兄弟都像謝惠連一般出色，所詠的詩只輸給謝靈運。二謝是南朝大文豪，這正顯現李白的狂傲。「開瓊筵以坐花，飛羽觴而醉月」，「詩若不成，罰依金谷酒數」，都可以品味出李白浪漫的情懷。所以兩篇文章縱然意念有相似之處，但風格不同，作者創作心態不同，個性也不同。不過，這並非單憑語料分析所能得到的結論。

類似的情形在蘇洵的《六國》和蘇轍的《六國論》中也可看到。父子兩人同論六國所以覆亡之因，但蘇洵藉六國「賂秦而亡」來諷刺當時朝廷對外族的軟弱態度，是借古傷心，痛快淋漓之作。蘇子由的作品，客觀的分析六國敗亡是因爲當時諸侯、謀士不知韓、魏二國形勢的重要所致。文氣固盛，但情感冷靜，不似蘇洵有感而發的作品。這種看似傳統文學欣賞的觀念，卻正是「形上證物」所以立論的基礎。劉彥和在《文心雕龍·知音篇》說：「夫綴文者情動而辭發，觀文者披文以入情，沿波討源，雖幽必顯。世遠莫見其面，覘文輒見其心。」是有道理的。

¹¹ 同注 2，15 頁。

五、結語

利用語料統計法來觀察社會，除了作整體的觀察，也可以較深入地用在社會族群區別度的觀察上，也就是利用此法去建立語言層次的特色。筆者曾替教育部調查過國內四級學校學生使用口語詞彙的情形、國小學童常用字詞、大陸小學教科書所收字詞等，都是運用此法去進行。¹²上文所舉的三個實例，剛好說明了這種統計分析法在不同研究工作上的運用情形。

在這三個實例中，筆者首先建立了綜合觀察的標準，再以此標準去探尋特殊語料的特徵，並利用此法嘗試去建立「形上證物」的觀念。第一個實例是基礎，有此基礎後，可去多向觀察社會，進行不同的研究。事實上，除上文所舉例子外，他如性別、職業別、學歷層次、地域區別、方言區隔等等都可延伸利用，這就是此法可以協助警學研究的價值所在。

警大是所專門研究社會治安學問的學校，除了警學專門學術外，爲了要輔助了解整體社會變遷的實況，有許多方法可以利用，諸如問卷調查、實地訪談等，語料統計應當也是一種可加利用的方法。假設警大每年都能對社會環境提出一份變動的觀察報告，累積多年下來，相信無論對警學專門學術的研究或實務政策的制訂，必具有參考指引的意義。而在警大中，最適合擔負此統計報告責任的單位，豈非就是通識教育中心？所以本文謹就個人過去的一點經驗提出呼籲：警大要成爲治安學問的知識庫，有些工作是可以主動發軔的。在此，誠盼通識教育在達成此任務指標上有所貢獻，也希望個人的淺見能有些許的助益。

¹² 《八十七年口語問卷調查報告書》，教育部，1999-10；《八十七年口語語料調查報告書》，教育部，1999-8；《國小學童常用字詞調查報告書》，教育部，2000-12；《大陸小學教科書字詞調查報告》，教育部，2000-9。

參考書目

- 1.曾榮汾，1997，〈試論語言學與犯罪學結合的可能發展〉，警察資訊學術研討會
- 2.教育部，1999-6，《八十六年常用語詞調查報告書》
- 3.教育部，1999-8，《八十七年口語語料調查報告書》
- 4.教育部，1999-10，《八十七年口語問卷調查報告書》
- 5.教育部，2000-6，《民國八十七年常用語詞調查報告》
- 6.教育部，2000-9，《大陸小學教科書字詞調查報告》
- 7.曾榮汾，2000-9，《刑事案件資料庫系統之建立與綜合應用—犯罪語料庫建立之研究》，國科會專題計畫報告