

CRL Asian Exchange Programme
Japan-China Natural Language Processing
Joint Research Promotion Conference
30/10 - 1/11 2001
Osaka



漢語共時語料庫：發展與應用

Development and Applications of A Synchronous Corpus of Chinese

鄒嘉彥 *Benjamin K Tsou*

香港城市大學 語言資訊科學研究中心
Language Information Sciences Research Centre
City University of Hong Kong

Email: rlbtsou@uxmail.cityu.edu.hk

摘要

LIVAC 共時語料庫於 1995 年由香港城市大學語言資訊科學研究中心建立。語料取自北京、香港、澳門、上海、新加坡和台灣六地有代表性的中文報紙，以及電子媒介上的新聞報道，每隔四日，同地抽取同日新聞。共時語料讓我們能以嚴謹科學方法和前所未有的「視窗」方式來研究各地華語社區漢語語用的情況。此外，語料中有關各地漢語使用的重要差異亦逐漸受資訊檢索(IR)和自然語言處理(NLP)等研究範疇的重視。直至 2001 年 6 月為止，語料庫已收集了 8000 多萬字、超過 49 萬個詞條，並在不斷擴充中。本課主要討論這共時語料庫的概念和設計，以及一些分析結果。此外，亦會談及和語料庫相關的計算機技術，如移動計算(mobile computing)、搜尋器(search engine)、分詞(segmentation)和句法分析(parsing)。

Abstract

Since 1995, the Language Information Sciences Research Centre of the City University of Hong Kong has developed a synchronous corpus of Chinese which draws textual data every 4 days at the same time from the news media of Beijing, Hong Kong, Macau, Shanghai, Singapore and Taiwan. It provides a scientific and unique WINDOW approach regarding the Chinese language as used by different Chinese speech communities, and reveals interesting and critical variations in the Chinese language that are increasingly important for IR and NLP. As of June 2001, the database has 80 million characters and a dictionary of over 49 thousand basic word types. This presentation discusses the conception and design of this unique database and the results obtained, as well as the many applications arising from it in, for example, mobile computing, search engine development, segmentation and parsing.



1. 平衡語料庫
(Balanced Corpus)

2. 共時語料庫
(Synchronous Corpus)

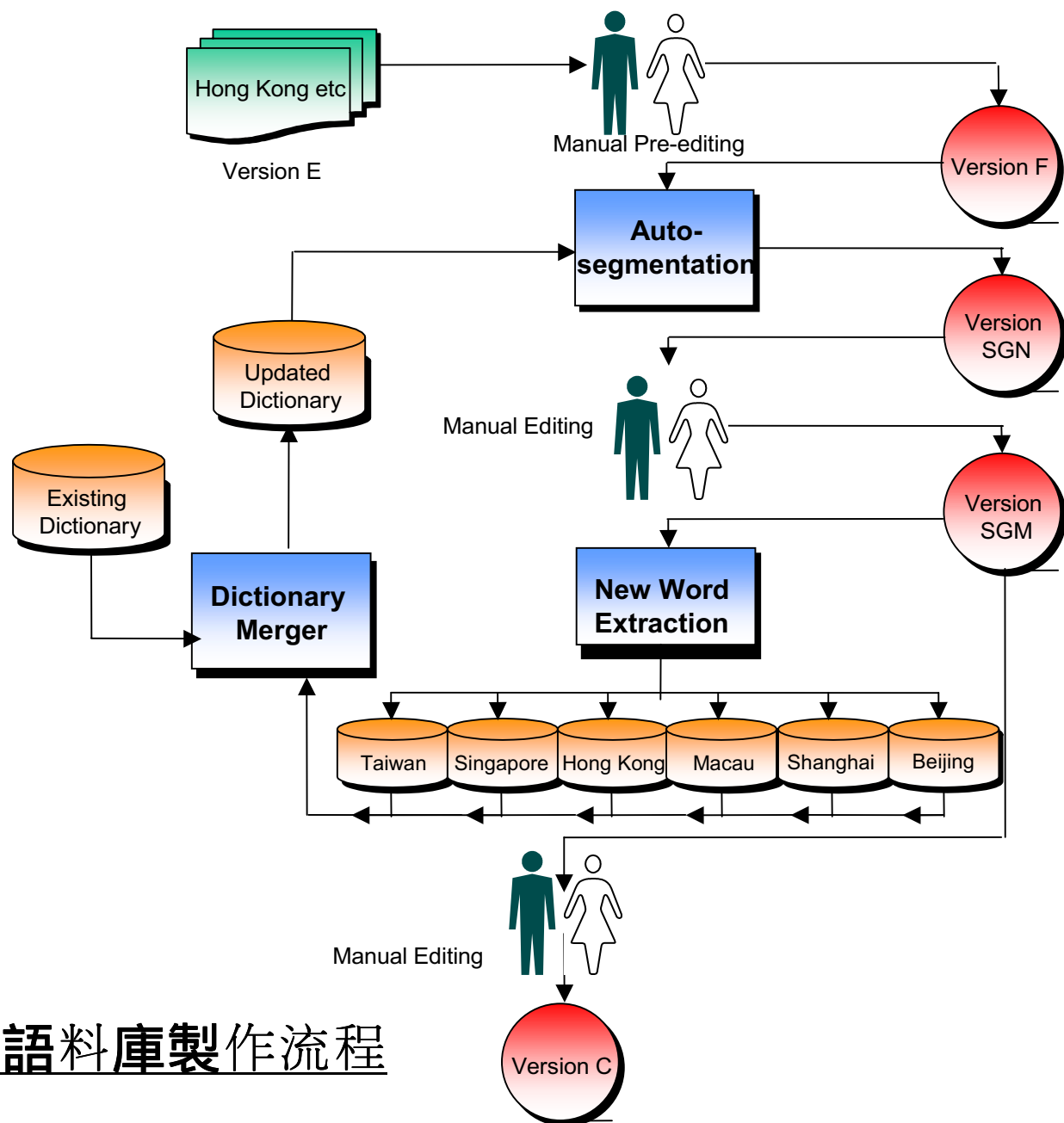
a. 生語料庫
(Raw Corpus)

b. 熟語料庫
(Processed Corpus)

LIVAC（漢語共時語料庫）

Linguistic Variation in Chinese Speech Communities

- 來自 香港城市大學 語言資訊科學研究中心
- 地區：涵蓋世界各主要華語地區
 - 香港 台北 澳門 新加坡 上海 北京 （廣州）（日本）
- 對象：各地主要中文報章
- 方式：每四天一次，各地同日抽取語料
- 時間：自1995年7月 至 2005年6月
- 內容：本地要聞、兩岸、國際、綜合、經濟、體育、文娛
- 語料處理：先以電腦自動分詞，再做人工校對
- 庫料：（2001年） 字：80,000,000+ [熟語料]
詞：490,000+



LIVAC 語料庫製作流程

LIVAC

Year 年份	Character 漢字	Word 詞
1	10.1M	0.12M
2	11.0M	0.13M
3	11.1M	0.13M
4	11.5M	0.13M
5	20.7M	0.23M
6 (up to Jun 2001)	13.5M	0.18M
Total 總計	80.2M	0.49M
LIVAC-Others		
Total 總計	6.5M	0.08M

LIVAC Synchronous Corpus 共時語料庫

<http://www.rcl.cityu.edu.hk/livac/index.htm>



<http://www.rcl.cityu.edu.hk/livac>

香港城市大學 語言資訊科學研究中心  LIVAC 共時語料庫																	
● 檢索文本 詞語檢索 原文檢索 ● 更改密碼 ● 更新使用者資料 ● 登出 ● 回到主頁  要得到最佳的視覺效果， 請將顯示器的解析度調至 800x600或以上	檢 索 條 件 <table border="1"><tr><td>詞 語：</td><td>台灣</td></tr><tr><td>漢語拼音：</td><td><不指定></td></tr><tr><td>粵語拼音：</td><td><不指定></td></tr><tr><td>詞 類：</td><td>疊詞 簡縮詞 人物名詞 地方名詞 專用名詞 原文詞語 其他名詞 無分類</td></tr><tr><td>地 區：</td><td>香港 新加坡 台灣 上海 澳門 新聞稿 北京</td></tr><tr><td>日 期：</td><td><不指定></td></tr><tr><td>出現頻率：</td><td><不指定></td></tr><tr><td>檢索範圍：</td><td>全部</td></tr></table> 檢 索 結 果 共找到 9774 組句子 ■ [香港 7/4/95] <冰島lg> <女> <總統> <語> <台> <傳媒> <歡迎> <李登輝ln> <私人> <到訪> @ 1. <本> <報> <專訊> <冰島lg> <總統> <芬恩博加多蒂爾ln> <上週六!*> <接受> <台灣lg> <傳媒> <專訪> <時> <指出>， <如果> <台灣lg> <總統> <李登輝ln> <要> <到> <冰島lg> <做> <私人> <訪問>， <冰島lg> <將> <非常> <歡迎>， <她> <也> <很> <樂意> <和> <李登輝ln> <見面>。 [*詳情*]	詞 語：	台灣	漢語拼音：	<不指定>	粵語拼音：	<不指定>	詞 類：	疊詞 簡縮詞 人物名詞 地方名詞 專用名詞 原文詞語 其他名詞 無分類	地 區：	香港 新加坡 台灣 上海 澳門 新聞稿 北京	日 期：	<不指定>	出現頻率：	<不指定>	檢索範圍：	全部
詞 語：	台灣																
漢語拼音：	<不指定>																
粵語拼音：	<不指定>																
詞 類：	疊詞 簡縮詞 人物名詞 地方名詞 專用名詞 原文詞語 其他名詞 無分類																
地 區：	香港 新加坡 台灣 上海 澳門 新聞稿 北京																
日 期：	<不指定>																
出現頻率：	<不指定>																
檢索範圍：	全部																

詞語檢索結果 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

網址(1) [N&type6=ON&type7=ON&type3=ON&type9=ON&re1=ON&re2=ON&re3=ON&re4=ON&re5=ON&re6=ON&re7=ON&fy=1995&fx=7&dl=1&ty=1996&tm=6&tl=30&et=ON&rt=ON&sym=ON](#)

- 詞語：中巴

漢語拼音：	zhongl bal
粵語拼音：	zungl baal
屬性：	其他詞類
英語：	1.China Motor Bus

地區：	香港	新加坡	台灣	上海	澳門	新聞稿	北京
出現次數：	8.8%	0.0%	0.0%	32.4%	50.0%	0.0%	8.8%
同類詞：	<無>	<無>	<無>	<無>	<無>	<無>	<無>
- 詞語：中巴

漢語拼音：	zhongl bal
粵語拼音：	zungl baal
屬性：	其他專名
英語：	1.China Motor Bus

地區：	香港	新加坡	台灣	上海	澳門	新聞稿	北京
出現次數：	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
同類詞：	<無>	<無>	<無>	<無>	<無>	<無>	<無>
- 詞語：中巴

漢語拼音：	zhongl bal
粵語拼音：	zungl baal
屬性：	地名
英語：	1.China Motor Bus

地區：	香港	新加坡	台灣	上海	澳門	新聞稿	北京
出現次數：	0.0%	0.0%	6.8%	0.0%	16.9%	0.0%	76.3%
同類詞：	<無>	<無>	<無>	<無>	<無>	<無>	<無>

Local Time: 2001年10月15日 AM 11:41:25

新華網絡

原文檢索結果 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

網址(1) [http://www.netatutor.net/...&ON&type7=ON&type3=ON&type9=ON&re1=ON&re2=ON&re3=ON&re4=ON&re5=ON&re6=ON&re7=ON&fy=&fm=&fd=&ty=&tm=&ti=&ff=&fb=&sec=all&num=5&out=all&display=segment](#)

原文檢索結果

共 193 組句子符合檢索條件，以下顯示前 193 組例子
193 entries found, first 193 entries displayed

- [9/14/95] 索 索 錄 錄 網 任 巧 快 老 台南 ①
 1. <不久> <前> <的> <一> <天>，<歸心似箭> <的> <老> <台南> <張繼全> <在> <四川> <峨嵋山市> <下> <了> <火車> <後>，<搭> <上> <一> <輛> <開往> <樂山> <的> <中巴>。
[* 註值]
- [10/4/95] 八 轉 三 上班族 蜂擁 出 紅綠 海陸 巴士站 亂 ①
 2. <中巴> <及> <九巴> <的> <發言人> <均> <指出>，<公司> <在> <接獲> <運輸署> <有關> <改> <掛> <三> <號> <風球> <的> <通知> <後>，<已> <盡快> <安排> <車輛> <接載> <乘客>，<而> <過> <海> <隧巴> <方面>，<兩> <巴> <各自> <派> <出> <三十> <部> <空> <車> <到> <隧道口> <載> <客>。
[* 註值]
- [12/11/95] 海口 費打 車匪路霸 破案 廿 宗 向 緝 犯 ①
 3. <十一月> <二十七> <日>，<市> <防暴> <二> <大> <隊> <隊長> <黃存> <帶領> <四> <名> <幹警> <在> <海秀路> <中巴> <上> <抓獲> <四> <名> <正在> <扒竊> <作案> <犯罪分子>。
[* 註值]
- [12/11/95] 九巴 新 車 未 滿月 起 十 漸 蝶蜂 搶 犯 轎 ①
 4. <這> <次> <已經> <是> <最近> <五> <個> <月> <來> <的> <第> <四> <宗> <同> <類> <事件>，<九巴>、<中巴> <及> <城巴>，<均> <曾> <先後> <發生> 「<甩> <轎> <事件>」，<引起> <市民> <對> <巴士> <的> <安全> <問題> <的> <關注>。
[* 註值]
- [1/16/96] 奇招 搶 多 忽 如 決河 ①
 5. <的士> <龍駛> <之時>，<珠海> <市面> <大致> <平靜>，<各> <種> <大>、<中巴> <也> <不見> <特別> <擠逼>。
[* 註值]
- [2/1/96] 中巴 加價 有 激 ①
 6. <中巴> <加價> <有> <激> ① [* 註值]
 7. <爲了> <審議> <中巴> <加價>，<運輸科> <和> <運輸署> <跟> <交通> <諮詢> <委員會>

Local Time: 2001年10月15日 AM 11:49:00

網際網路



Abbreviation 簡縮詞

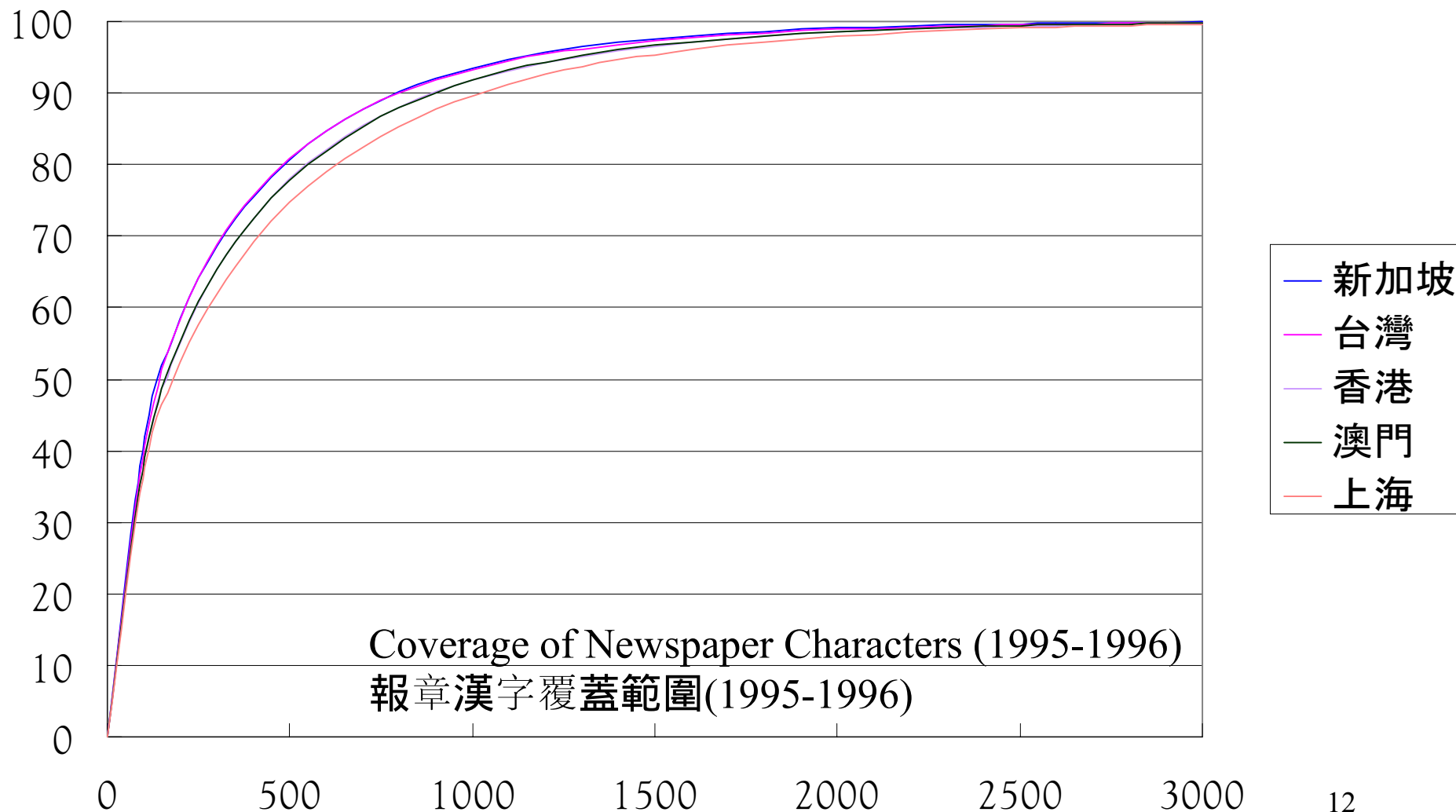
中巴 (Central - “BA”)

- 中型巴士 “midi”-bus
- 中國 - 巴基斯坦 China - Pakistan
- 台灣 - 巴拿馬 Taiwan - Panama
- 中華巴士公司 China Motor Bus

Character Distribution

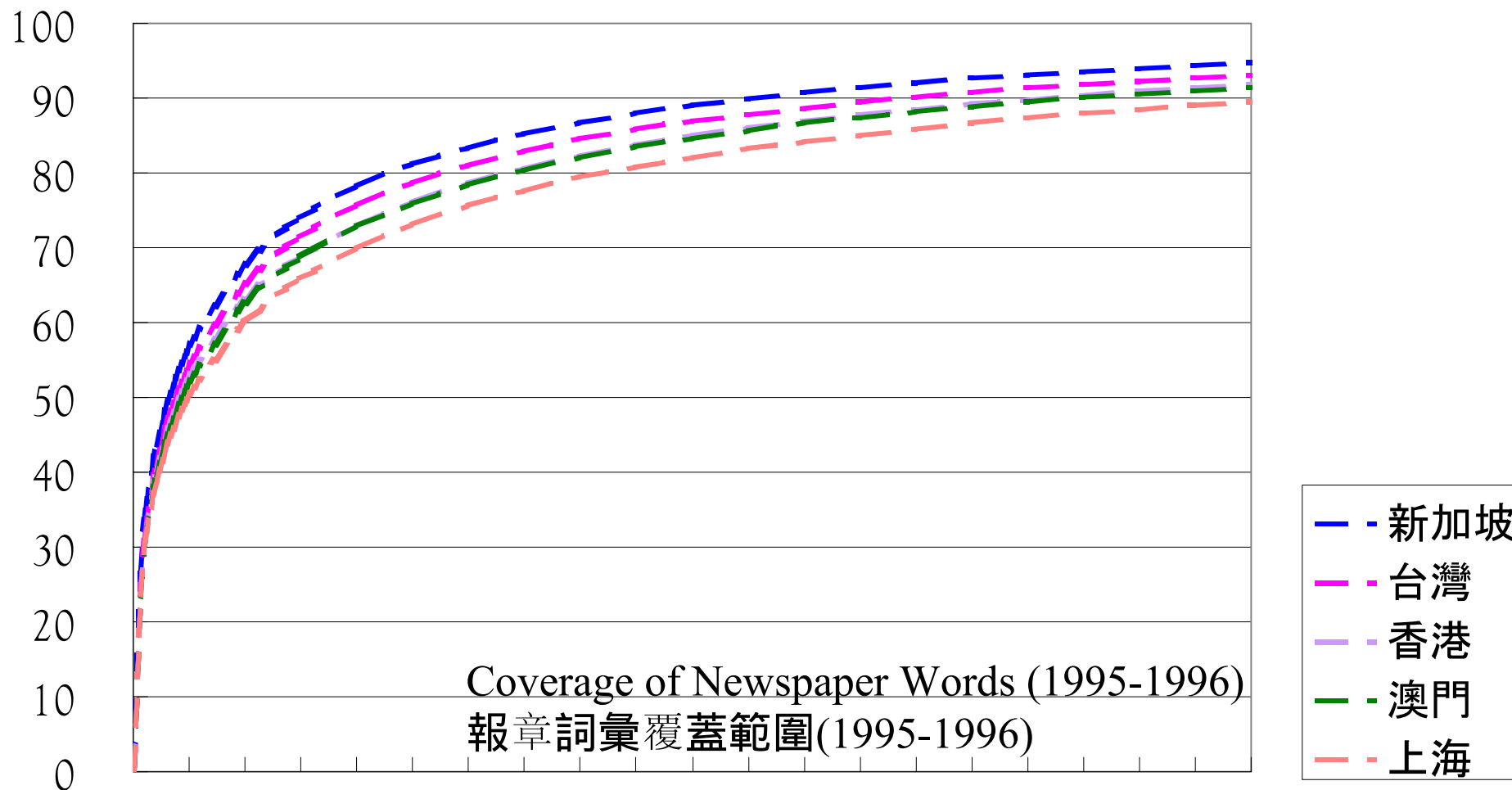


漢字分





Word Distribution 詞語分



Lexical Entries in Major Collections of Chinese Words

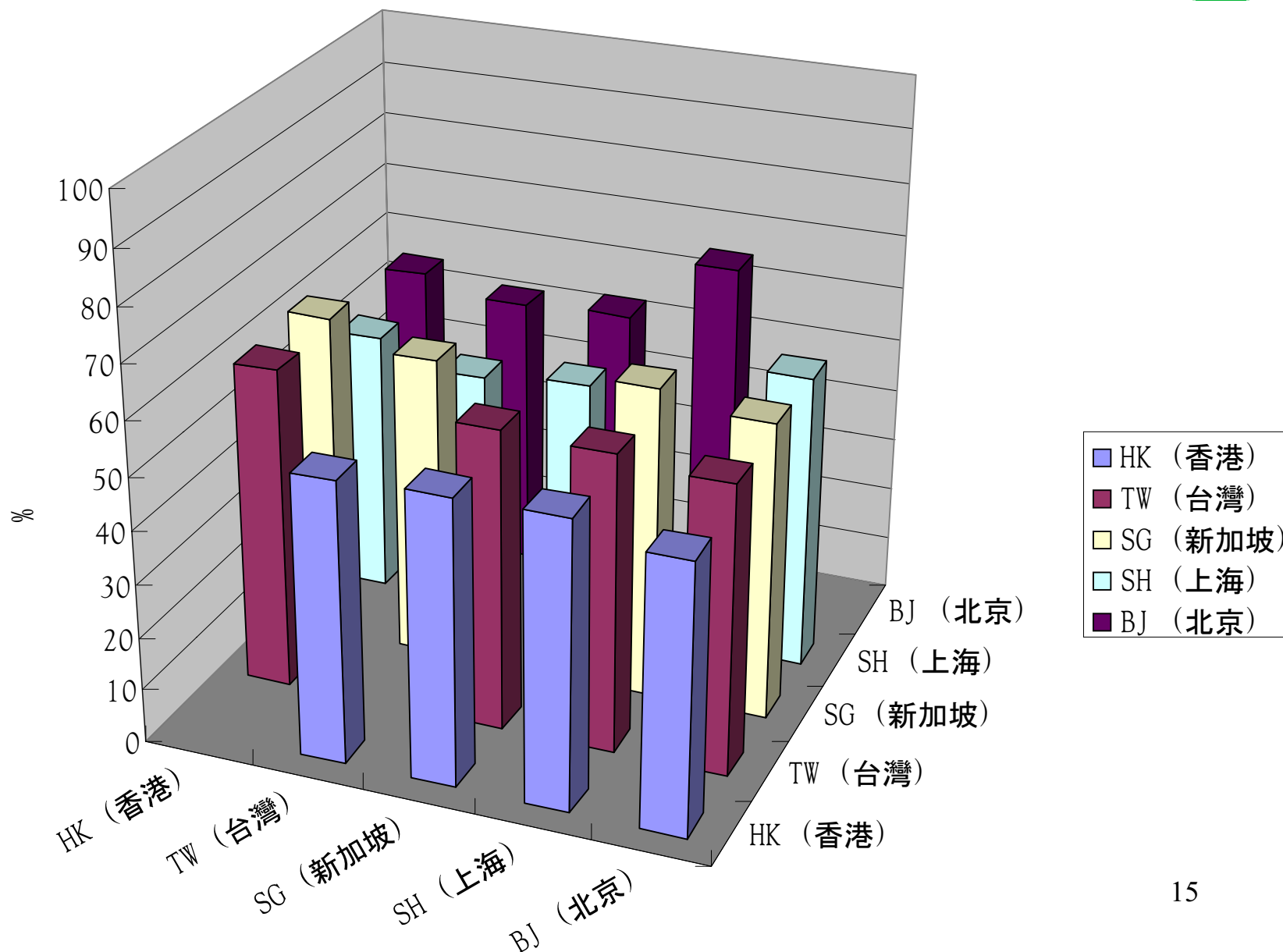


各主要漢語詞典詞條數量

	Total 總計
辭源 (CIYUAN) 1931	65,555
辭源 (CIYUAN) 1939	88,074
現漢 (XIANHAN)	50,333
辭海 (CIHAI)	127,841
LIVAC 共時語料庫 (1995-2001)	490,000+

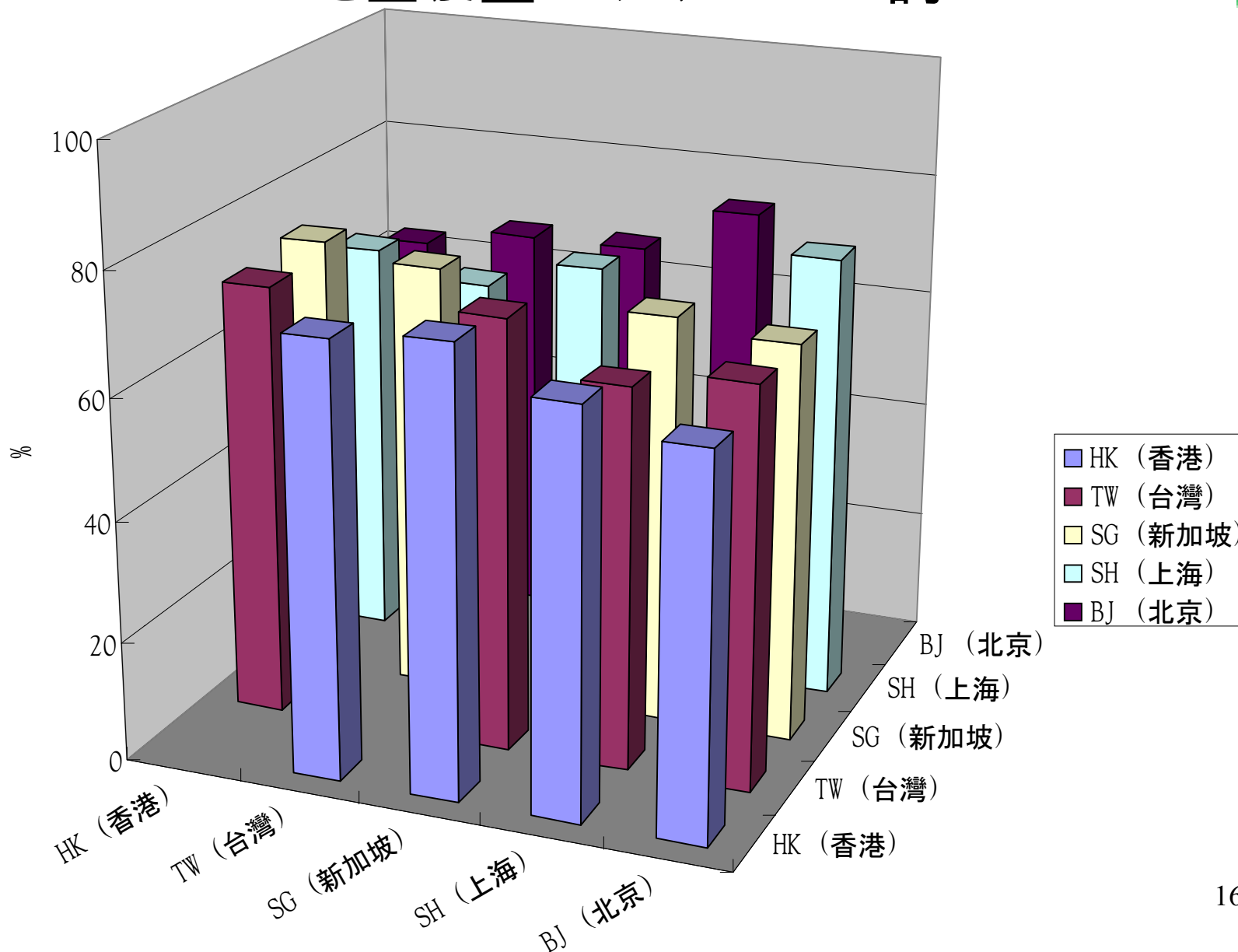
Overlap: All Words

地區覆蓋：全部詞語



Overlap: Top 5000 Words

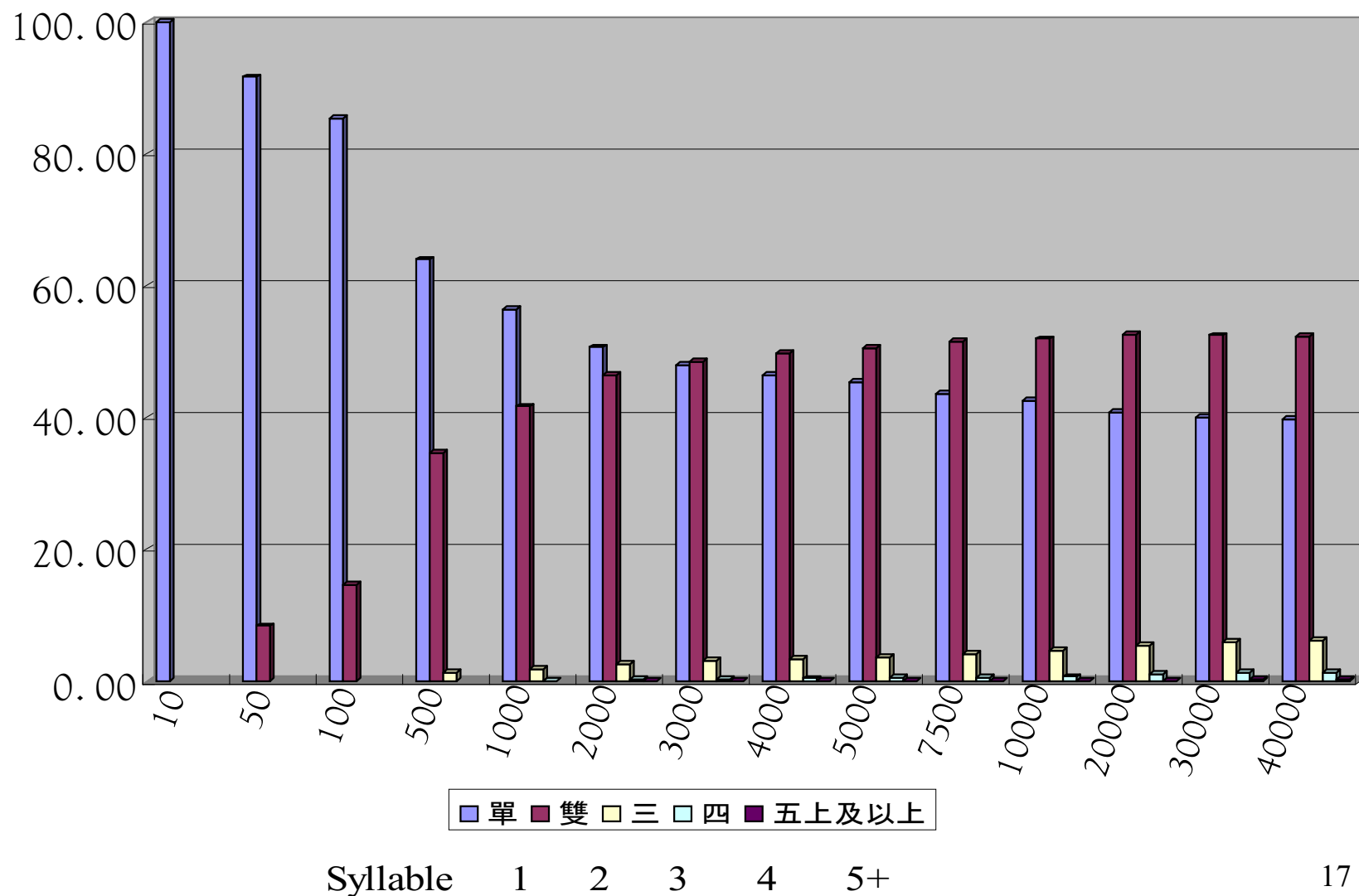
地區覆蓋：最常用5000詞



Token Distribution of Top 40,000 Words by Word Length: HK (1 yr)



詞標與音節分：香港常用40,000詞





LIVAC 共時語料庫 1.0 版 (1995--1996)

	HK 香港	MC 澳門	TW 台灣	SG 新加坡	SH 上海	Total 總計
(1) Character Tokens 字標	1.60 M	1.81 M	1.40 M	0.82 M	1.15 M	6.78 M
(2) Character Types 字種	4.4 k	4.4 k	4.2 k	3.6 k	4.3 k	5.5 k
(3) Word Tokens 詞標	0.95 M	1.06 M	0.82 M	0.48 M	0.69 M	3.99 M
(4) Word Types 詞種	42.9 k	48.7 k	37.5 k	25.6 k	45.0 k	107.8 k
(5) Ave. char. per word 平均? 詞字數	1. 69	1. 72	1. 72	1. 70	1. 67	1. 70

Top 10 Four-Character Words (四字格)

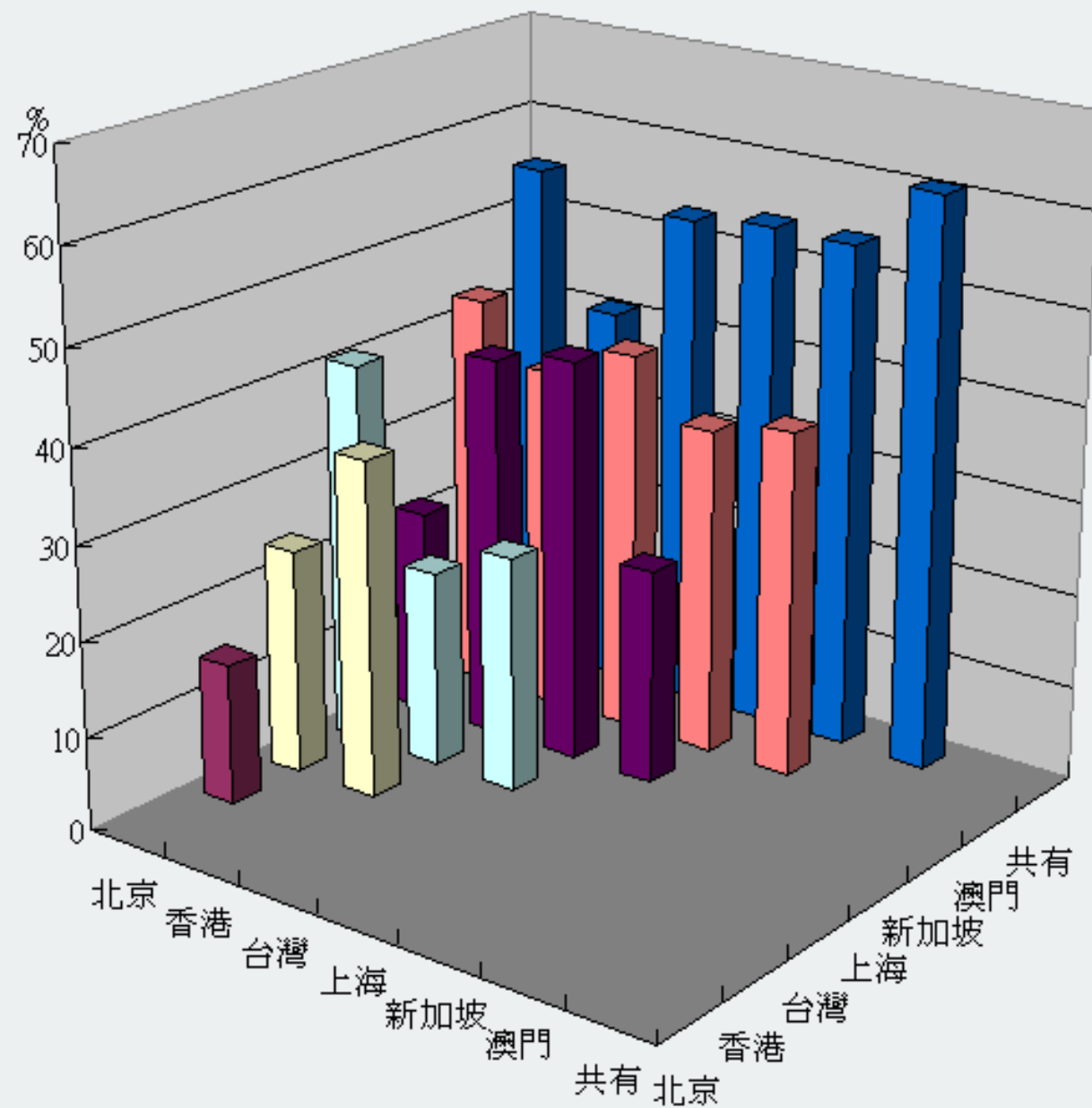
- 實事求是
- 堅定不移
- 艱苦奮鬥
- 全力以赴
- 前所未有
- 千方百計
- 全心全意
- 引人注目
- 一如既往
- 各行各業

Four-Character Words in Japanese & Chinese

中日共用四字格

1	小心翼翼	11	意味深長
2	一觸即發	12	共存共榮
3	一舉一動	13	天衣無縫
4	一網打盡	14	四面楚歌
5	一朝一夕	15	喜怒哀樂
6	自給自足	16	臥薪嘗膽
7	誠心誠意	17	有名無實
8	不眠不休	18	自暴自棄
9	自由自在	19	一石二鳥
10	半信半疑	20	神出鬼沒

各地區相同 四字格詞比例



- 北京
- 香港
- 台灣
- 上海
- 新加坡
- 澳門
- 共有

* 資料以各地首100個
高詞頻四字格詞為基礎

Distribution of Chinese Lexicon



漢語詞彙分

1995 -- 1998

%	HK	MC	TW	SG	SH	BJ	Total
	香港	澳門	台灣	新加坡	上海	北京	總計
Personal (人名)	12.8	16.7	14.1	15.4	13.7	16.2	23.2
Geog. (地名)	8.4	10.1	7.4	7.5	10.4	10.5	11.9
Others (其他專名)	4.2	5.6	4.5	5.1	4.9	3.9	6.4
Sub-total	25.4	32.4	26.0	28.0	29.0	30.6	41.5
Non-Proper Nouns (非專名)	74.6	67.6	74.0	72.0	71.0	69.5	58.5
Total (總計)							0.24M

Locus Classicus (Chinese) 典故



- 雷峰精神
- 武則天
- 事後孔明
- 鐵面關公/包公
- 蘆溝橋事變
- 五四運動
- 有眼不識泰山
- 西安事變
- 洛陽紙貴
- 空城計
- 劉備荊州
- 買長城, IBM

Locus Classicus 典故(I)



- Chaucerian
- Shakespearean
- Ocam's razor
- Sophie's choice
- King of France problem
- Shanghai-ed
- China's New York
- His Falklands
- His Malvenas
- Parisian
- Bostonian
- Bohemian
- Venice of the East
- Gettysberg address
- (Remember the) Alamo
- Malthusian
- Monroe doctrine
- Trojan horse
- Oedipal complex

名人榜網址: <http://www.rcl.cityu.edu.hk/livac/roster>

中港台三地2000年新聞名人榜 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

網址(AD) <http://www.rcl.cityu.edu.hk/livac/roster/>

香港城市大學 語言資訊科學研究中心
Language Information Sciences Research Centre

 **LIVAC** 共時語料庫
Synchronous Corpus

Chinese News Media Celebrity Roster 2000

「中港台三地2000年新聞名人榜」

1. BACKGROUND

Since July 1995, the Language Information Sciences Research Centre of the City University of Hong Kong has been working continuously on the LIVAC (Linguistics Variation in Chinese Speech Communities) synchronous corpus. The corpus is based on news media materials collected regularly from six Chinese-speaking communities, namely Hong Kong, Singapore, Taiwan, Beijing, Shanghai, and Macau. Up to now, the corpus contains over 70 million Chinese characters and over 400,000 words, and is still expanding. Interested parties are welcome to browse our LIVAC homepage at <http://www.rcl.cityu.edu.hk/livac>.

2. BI-WEEKLY CELEBRITY ROSTER

The bi-weekly "Celebrity Roster", first launched in January 2001, is a product of the LIVAC project. It is an index of the top 25 celebrities according to media exposure in the Chinese newspapers of Beijing, Hong Kong and Taiwan, compiled every two weeks and updated on the last Saturdays. The index reflects the communal differences in the orientation of news reports and the hottest topics in politics, economy, sports, and entertainment in the three major Chinese communities. It should be a useful reference source for anyone, from specialists to students, interested in Chinese current affairs and in the cultural and linguistic development of

1. 依據

由香港城市大學語言資訊科學研究中心主持的「中港台三地2000年新聞名人榜」是以本中心語料語言學實驗室開發的「LIVAC共時語料庫」為依據，該語料庫自1995年7月起至今五年多以來，定期收集香港、新加坡、台灣、北京、上海、澳門等六地主要報刊及電子媒介的語料，初步建立了以詞匯為基礎的語料庫，供從事語言學研究的學者和對語言及有關社會、文化有意探討者使用，語料庫至今累計已達7,000多萬字、40多萬個詞語，並仍在不斷增加中。點擊網址 <http://www.rcl.cityu.edu.hk/livac> 可進入「LIVAC共時語料庫」，歡迎參閱瀏覽。

2. 雙周榜

本中心自2001年一月開始推出本項目，每兩星期對香港、台灣、北京三地的主要傳媒統計一次，分別排列出各地見報率最高的前25位新聞人物，並於該兩星期的最後一個星期六上午更新上網。這個「雙周榜」緊貼三地的新聞報道，反映了兩周來各地區的重要新聞取向。從三地名人的上榜與排序變化，可觀察到各地的政治、經濟、體育及娛樂等方面的動態，供關心華人時事，與華人語言現象、社會、文化與動態發展有意探討者，或其他社會人士及學生參考。

新華網

各地名人榜 (2001/09/25 - 2001/10/08)

	香港	台灣	北京
1	<u>拉丹</u> Osama bin Laden	<u>賓拉登</u> Osama bin Laden	<u>布什</u> George W. Bush
2	<u>小布殊</u> George W. Bush	<u>小布希</u> George W. Bush	<u>江澤民</u> Jiang Zemin
3	<u>佐敦</u> Michael Jordan *	<u>李登輝</u> Lee Teng-hui	<u>米盧</u> Bora Milutinovic
4	<u>董建華</u> Tung Chee-hwa	<u>連戰</u> Lien Chan	<u>朱鎔基</u> Zhu Rongji
5	<u>范志毅</u> Fan Zhiyi *	<u>謝淑薇</u> Hsieh Su-wei	<u>李堅</u> Li Jian *
6	<u>汪明荃</u> Lisa Wong *	<u>倫斯斐</u> Donald H. Rumsfeld	<u>李嵐清</u> Li Lanqing *
7	<u>張柏芝</u> Cecilia Cheung *	<u>鮑爾</u> Colin Powell *	<u>李白</u> Li Bai *
8	<u>奧雲</u> Michael Owen *	<u>珍妮佛羅培茲</u> Jennifer Lopez *	<u>胡錦濤</u> Hu Jintao
9	<u>孫繼海</u> Sun Jihai *	<u>陳水扁</u> Chen Shui-bian	<u>錢其琛</u> Qian Qichen *
10	<u>曾蔭培</u> Tsang Yam-pui *	<u>林信義</u> Lin Hsin-i *	<u>安南</u> Kofi Annan

註：姓名後有「*」標記表示為本期新上榜者。

各地名人榜 (2001/09/25 - 2001/10/08)

11	<u>普京</u> Vladimir Putin *	<u>馬英九</u> Ma Ying-jeou	<u>李鵬</u> Li Peng
12	<u>甄妮</u> Jenny *	<u>瑞普肯</u> Carl Ripken Jr. *	<u>孟慶餘</u> Meng Qingyu *
13	<u>寶詠琴</u> Po Wing Kam, Theresa *	<u>艾許克羅夫特</u> John Ashcroft	<u>孫中山</u> Sun Yat-sen *
14	<u>米路天奴域</u> Bora Milutinovic	<u>王金平</u> Wang Jin-pyng *	<u>拉登</u> Osama bin Laden *
15	<u>安南</u> Kofi Annan *	<u>邦茲</u> Barry Bonds	<u>馬華</u> Ma Hua *
16	<u>艾歷臣</u> Sven Goran Eriksson *	<u>林豐正</u> Lin Feng-cheng *	<u>扎義夫</u> Mullah Abdul Salam Zaeef *
17	<u>李鐵</u> Li Tie *	<u>郭家宏</u> Guo Jia-hung *	<u>王兆國</u> Wang Zhaoguo *
18	<u>鄧小平</u> Deng Xiaoping *	<u>阿布杜拉</u> Ali Abdullah Saleh *	<u>李瑞環</u> Li Ruihuan
19	<u>陳慧琳</u> Kelly Chen *	<u>張俊雄</u> Chang Chun-hsiung	<u>尉健行</u> Wei Jianxing
20	<u>阿塔</u> Mohammed Atta *	<u>陳文茜</u> Sisy Chen *	<u>羅幹</u> Luo Gan
21	<u>科拿</u> Robbie Fowler *	<u>喬丹</u> Michael Jordan *	<u>蘇家富</u> Su Jiafu *
22	<u>施文</u> David Seaman *	<u>瑪丹娜</u> Madonna *	<u>丁關根</u> Ding Guangen
23	<u>于根偉</u> Yu Genwei *	<u>鍾欣凌</u> Lynn Jung *	<u>王英凡</u> Wang Yingfan *
24	<u>王菲</u> Faye Wong *	<u>林瑞圖</u> Lin Ruey-tou *	<u>布赫</u> Bu He *
25	<u>朱邦造</u> Zhu Bangzao *	<u>施明德</u> Shih Ming-teh *	<u>平傑三</u> Ping Jiesan *

中港台三地2000年新聞名人榜

香港、台灣、北京三地全年都曾上榜的19位名人

	香港	台灣	北京
1	李登輝	李登輝	江澤民
2	江澤民	柯林頓	朱鎔基
3	克林頓	江澤民	克林頓
4	董建華	呂秀蓮	普京
5	朱鎔基	朱鎔基	葉利欽
6	呂秀蓮	小淵惠三	埃連
7	埃利安	蒲亨	李登輝
8	普京	李文和	呂秀蓮
9	小淵惠三	艾里安	金正日
10	小布殊	董建華	阿薩德
11	金正日	金正日	戈爾
12	李文和	成克杰	金大中
13	成克杰	小布希	董建華
14	李平	葉爾勤	成克杰
15	葉利欽	金大中	布什
16	金大中	李平	李平
17	戈爾	高爾	小淵惠三
18	阿薩德	阿塞德	魯賓
19	魯賓	魯賓	李文和

Celebrity Roster 名人榜



Common for all 3 places

三地均曾上榜名人

USA 美國

1. 克林頓/柯林頓(Clinton)

2. 戈爾/高爾 (Gore)

3. 小布殊/小布希/布什
(G. W. Bush)

4. 埃利安/艾里安/埃連
(Elia)

5. 魯賓(Rubin)

6. 李文和(Lee Wen-ho)

Mainland China 中國

7. 江澤民(Jiang Zemin)

8. 朱鎔基(Zhu Rongji)

9. 成克杰(Cheng Kejie)

10. 李平 (Li Ping)

Celebrity Roster 名人榜



Common for all 3 places
三地均曾上榜名人

Korea 韓國

11. 金正日(Kim Jong-il) [KOR]

12. 金大中(Kim Dae-jung) [KOR]

Russia 俄羅斯

13. 普京/蒲亭 (Putin) [Russia]

14. 葉利欽/葉爾欽 (Yeltsin)
[Russia]

Taiwan 台灣

15. 李登輝(Lee Teng-hui) [TW]

16. 呂秀蓮(Annette Lu Hsiu-lien)
[TW]

Japan 日本

17. 小淵惠三(Keizo Obuchi) [JP]

Hong Kong 香港

18. 董建華(C H Tung) [HK]

各地地名錄(2001/09/25 - 2001/10/08)

	香港	台灣	北京
1	美國	美國	中國
2	香港	阿富汗	美國
3	中國	台灣	北京
4	阿富汗	中國	阿富汗
5	沙特阿拉伯	巴基斯坦	俄羅斯
6	巴基斯坦	大陸	巴勒斯坦
7	英國	紐約	香港
8	日本	華盛頓	波蘭
9	台灣	臺北	台灣
10	北京	俄羅斯	以色列
11	亞洲	巴勒斯坦	英國
12	大陸	以色列	上海
13	歐洲	英國	埃及
14	俄羅斯	沙烏地阿拉伯	華
15	烏茲別克	日本	德國
16	德國	中華	北京市
17	英格蘭	中東	法國
18	阿曼	歐洲	兩岸
19	深圳	巴國	韓國
20	喀布爾	法國	沙特阿拉伯
21	港澳	印度	中俄
22	瀋陽	德國	亞洲區
23	上海	白宮	浙江
24	中亞	埃及	雲南
25	卡塔爾	加拿大	四川

各地最常用詞語排行榜(2001/09/25 - 2001/10/08)

	香港	台灣	北京
1	的	的	的
2	在	在	和
3	是	美國*	在
4	有	是	了
5	美國*	與	是
6	他	他	中國*
7	會	說	年
8	香港*	不	個
9	及	日	國
10	但	將	要
11	和	也	黨
12	不	表示	爲
13	人	有	各
14	了	和	等
15	與	這	全
16	已	以	對
17	對	人	中
18	爲	對	不
19	個	爲	上
20	多	及	發展
21	後	中	與
22	將	攻擊	將
23	年	並	日
24	說	已	建設
25	表示	阿富汗*	這



Overall Lexical Growth

詞彙整體增長

(for 5 major locations)

(五個主要地區)

Comparison of Proper Nouns (3 yrs)

專有名詞比較 (三年)



Location 地區	%
Hong Kong 香港	25.4
Macau 澳門	32.4
Taiwan 台灣	26.0
Singapore 新加坡	28.0
Shanghai 上海	29.0
Beijing 北京	30.5
TV (HK) 電視新聞 (香港)	17.5

Examples of Proper Nouns 專有名詞包括

- (1) Personal names 人名;
- (2) Geographical names 地名;
- (3) Organizational names 機構名, brands 牌子, ...

Attrition/Growth of Proper Nouns



專用名詞增長/消減

1. Average 平均

%	Attrition 消失 Y1-Y2	Growth 增長 Y1-Y2	Attrition 消失 Y2-Y3	Growth 增長 Y2-Y3
Pers. 人名	76	79.2	76.2	77.5
Geog. 地名	58.2	61.5	60.3	59.8
Org. 機構名	60.5	66.9	66.1	66.3
Others 其他	36	36.5	55.9	36.7
Total 總計	100	100	100	100



Attrition/Growth of Proper Nouns

專用名詞增長/消減

2. Hong Kong 香港

	Attrition	Growth	Attrition	Growth
%	消失	增長	消失	增長
	Y1-Y2	Y1-Y2	Y2-Y3	Y2-Y3
Pers. 人名	77.5	77.1	77.1	78
Geog. 地名	59.5	59.3	60.6	59
Org. 機構名	51.1	61	58.8	61
Others 其他	34.6	35.2	34.7	35.1



Attrition/Growth of Proper Nouns

專用名詞增長/消失

3. Shanghai 上海

%	Attrition 消失 Y1-Y2	Growth 增長 Y1-Y2	Attrition 消失 Y2-Y3	Growth 增長 Y2-Y3
Pers. 人名	86.2	89.8	84.6	85.7
Geog. 地名	56.8	62.9	60.9	59.1
Org. 機構名	68.8	70.6	68.4	66.3
Others 其他	34.3	40.8	35.1	37.2



Shanghai and HK Compared—Who Leads?

上海與香港—誰勝？

	Attrition 消失	Growth 增長
Pers. N 人名	上海	上海
Geog. N 地名	香港	上海
Org. N 機構名	上海	上海



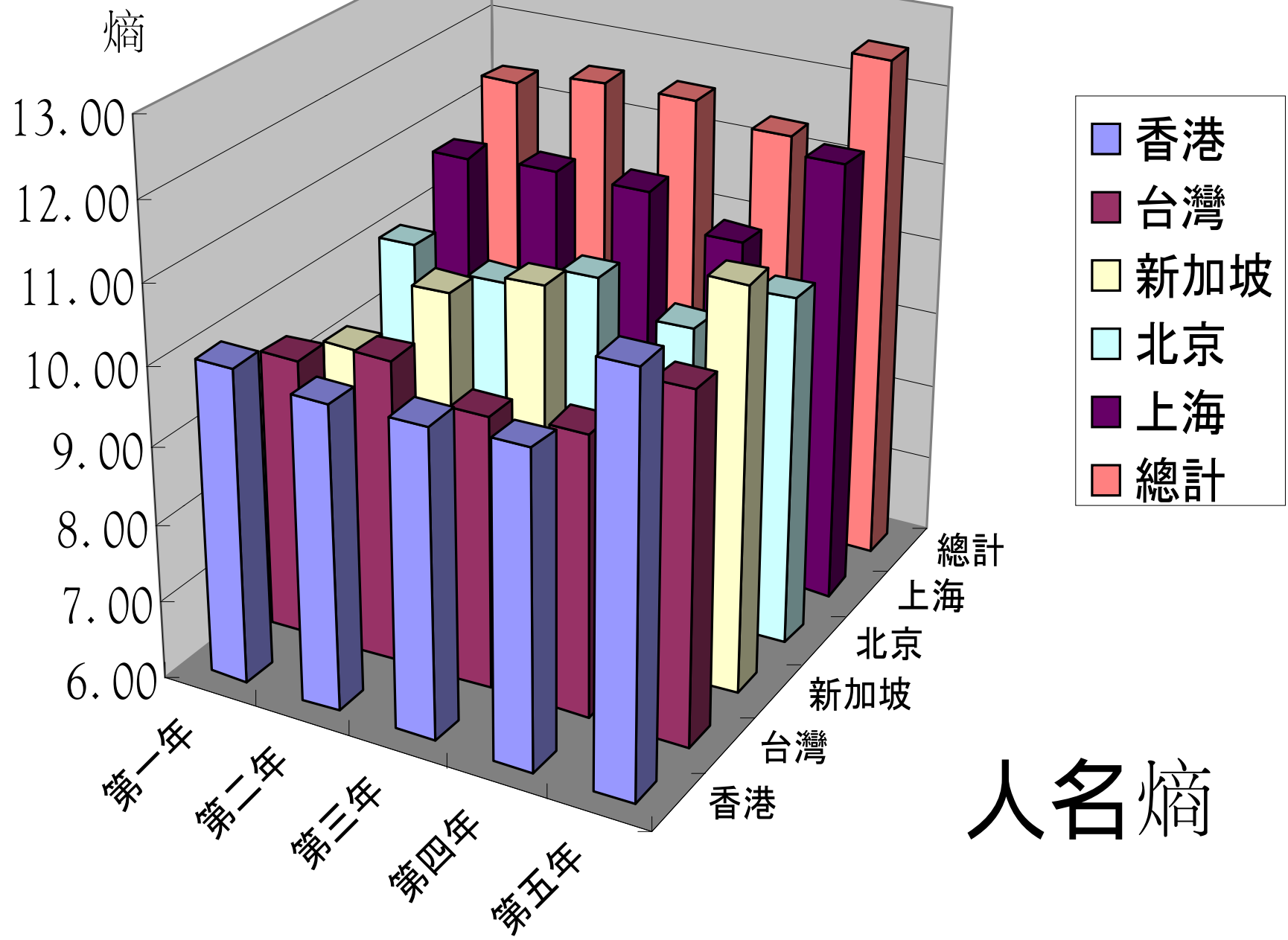
Entropy (熵)

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x)$$

where x_i = symbol type, character type, or word type
(字、詞或任何字符)

X = the set of x_i
(所有 x_i 的集)

$V(X)$ = the set of all x_i in the corpus
(語料庫中所有 x_i 的集)



人名熵

Summary of New Developments by Rank

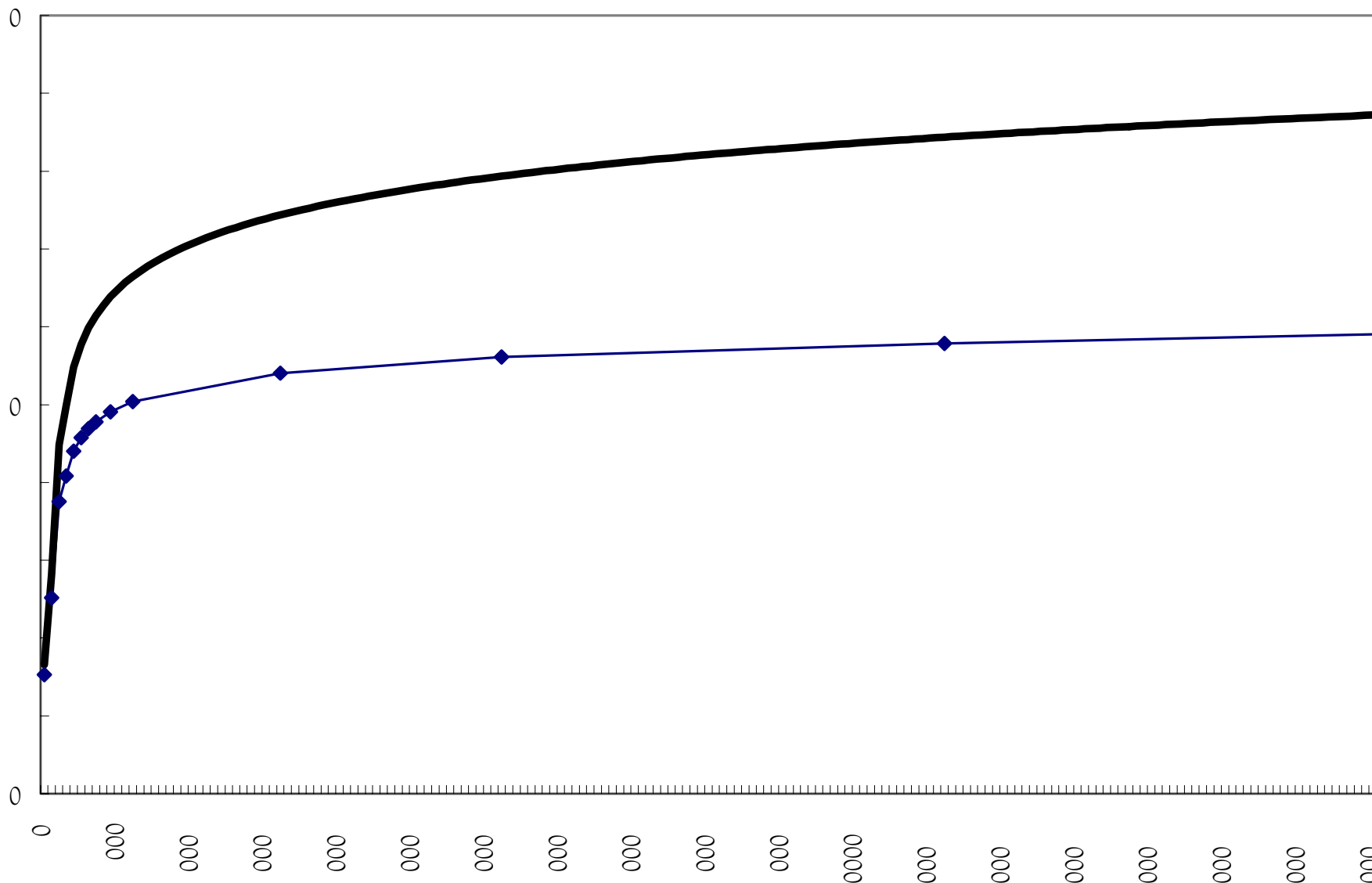
Pers. N	SH > HK > SG > BJ > TW
Geog. N	SH > BJ > HK > TW > SG
Org. N	SH > HK > SG > BJ > TW



零價熵 (H_0) 及 最大熵 ($H_{\max}$)

Entropy Graph of Proper Nouns 專用名詞熵

熵值



◆ Entropy of Proper Names — a Entropy

專用名詞詞種數目

Graphemic Entropy of Alphabets and Characters

字母與漢字的熵

Language 語言	a e n 字符熵
French 法文	3.98
Italian 意大利文	4.00
Spanish 西班牙文	4.01
English 英文	4.03
Russian 俄文	4.35
Chinese 漢語	9.65 (10.25 – 9.59)

Source: Feng (1989) and Liu et al (1987)



Maximum Entropy 最大熵

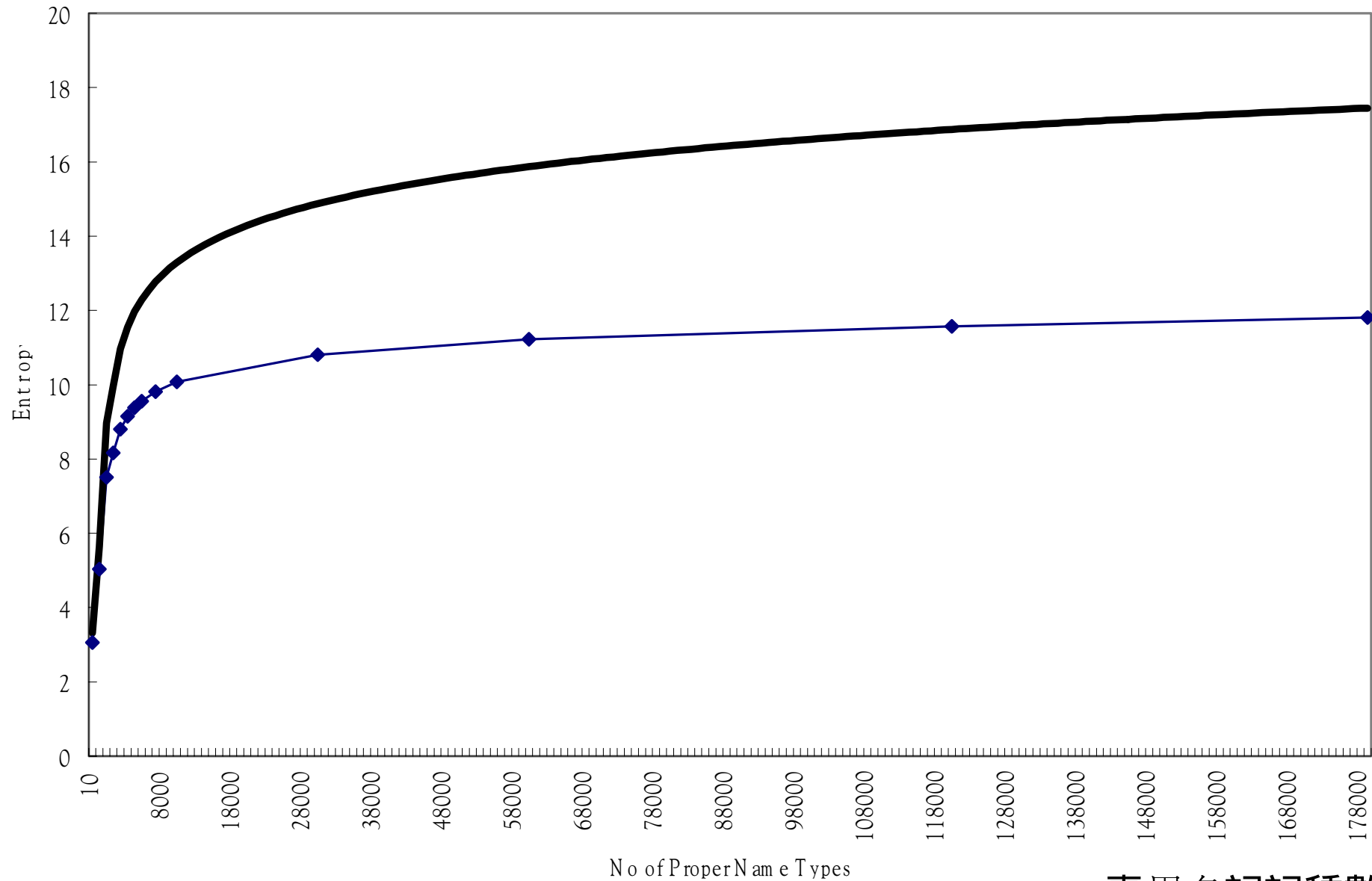
$H_{\max}$, maximum entropy, is defined as:

$$H_{\max} = \log_2 n$$

where n = number of all elements (symbols, character types, or word types) in the system

Entropy Graph of Proper Nouns 專用名詞熵

熵值



◆ Entropy of Proper Names — Max Entropy

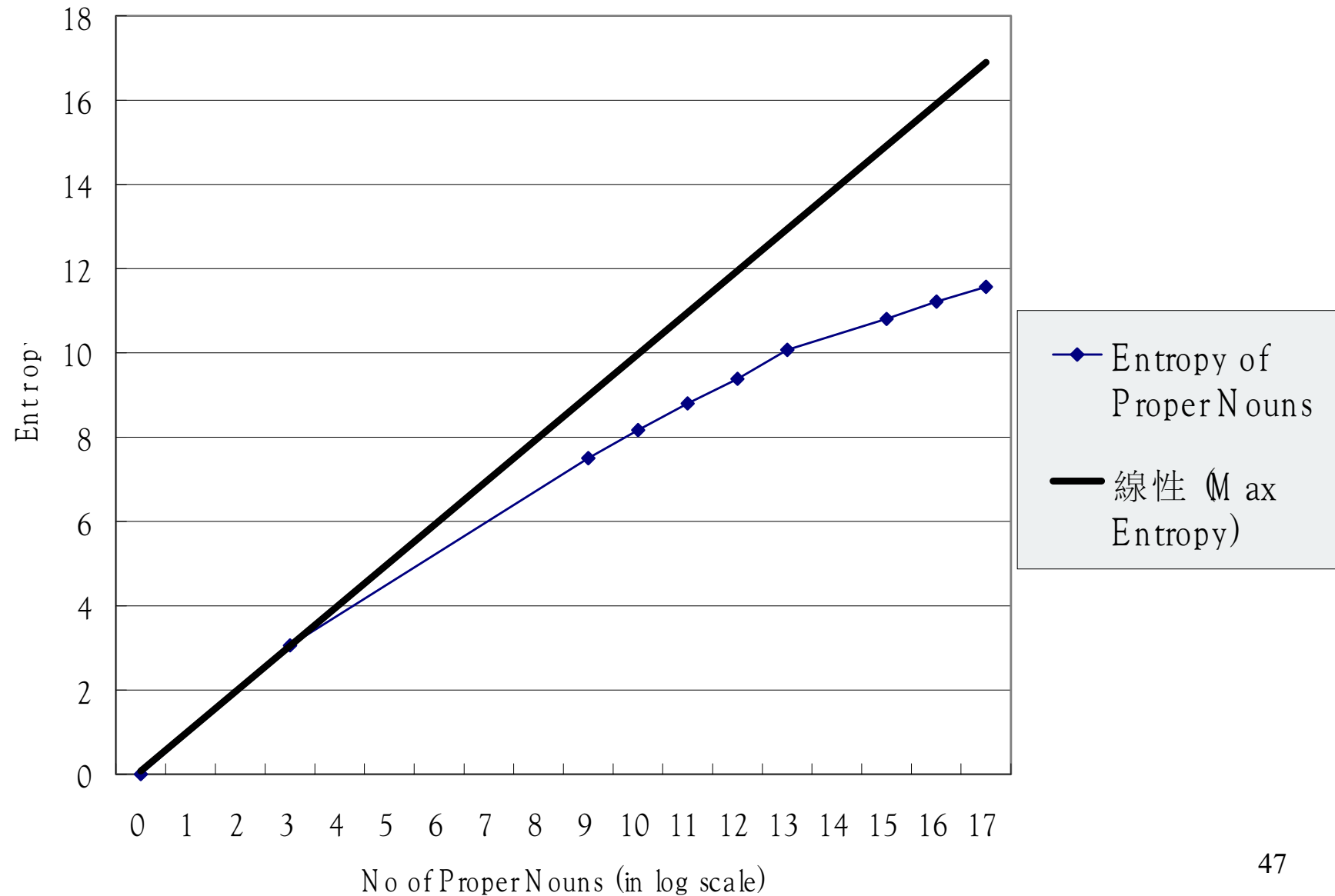
專用名詞詞種數目

Replot of the Entropy Graph in $\log_2$ scale

人名熵圖表



熵值

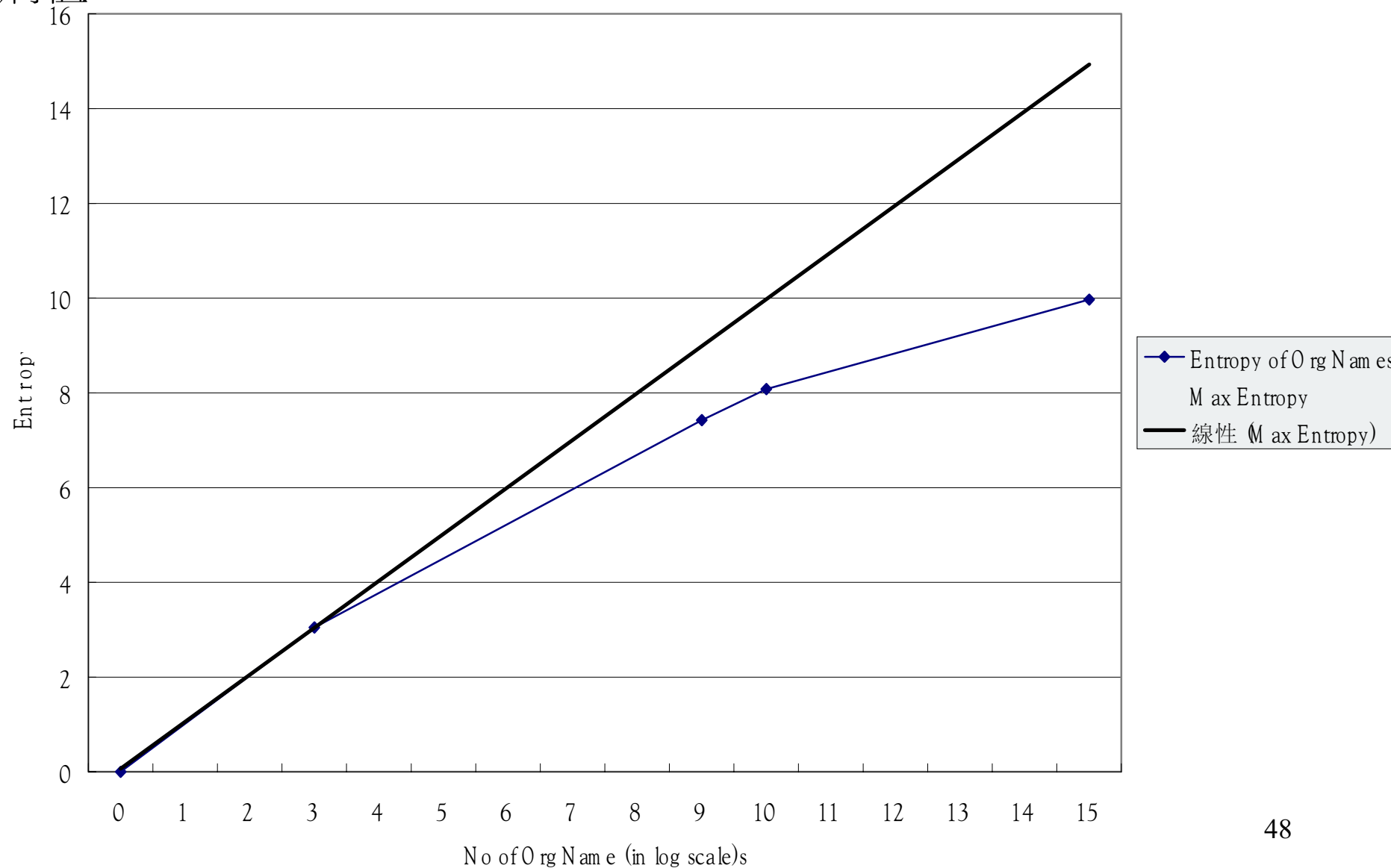


Entropy Graph for Organizational Names (in $\log_2$ scale)

機構名熵圖表



熵值



Mobile Computing 移動計算

為手提電話中文輸入提供字頻、詞頻排序



結語 Concluding Remarks

1. 各地漢語有別、需要分辨
 2. 漢語資訊發展迅速、引發新挑戰...
 3. 生語料有不足之處
 4. 平衡語料庫有限制
 5. 共時語料庫可提供新用途，開拓新領域。
1. Chinese language shows important local characteristics
 2. Rapid IT development brings new challenges to Chinese NLP...
 3. Simple raw text corpus shows inadequacies
 4. Balanced corpus has limitations
 5. Synchronous corpus can provide extra crucial benefits