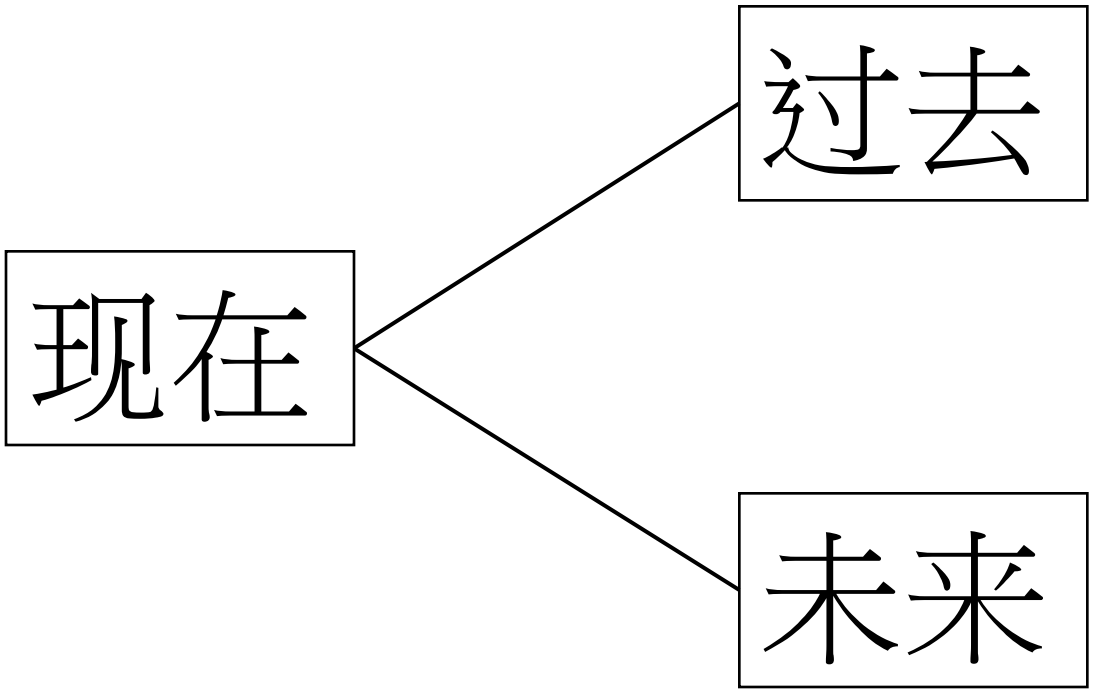
中国中文信息学会2012战略研讨会
江西婺源
2012.4.13-15

「中文信息处理」与「中国自然语言处理」发展新方向的初步探讨

邹嘉彦
香港教育学院　语言资讯科学研究中心
香港城市大学
btsou99@gmail.com


120°W
歐洲
北美
俄羅斯
哈薩克
蒙古
烏茲別克
吉爾吉斯
土庫曼
塔吉克
北韓
日本
阿富汗
中國
中東
不丹
巴基斯坦
尼泊爾
南韓
台灣
印度
緬甸
香港
泰國
孟加拉
菲律賓
越南
柬埔寨
10°N
10°N
斯里蘭卡
汶萊
馬來西亞
馬爾地夫
新加坡
印度尼西亞
大洋洲
東帝汶
120°W
N

汉语的世界地位



过去的困难

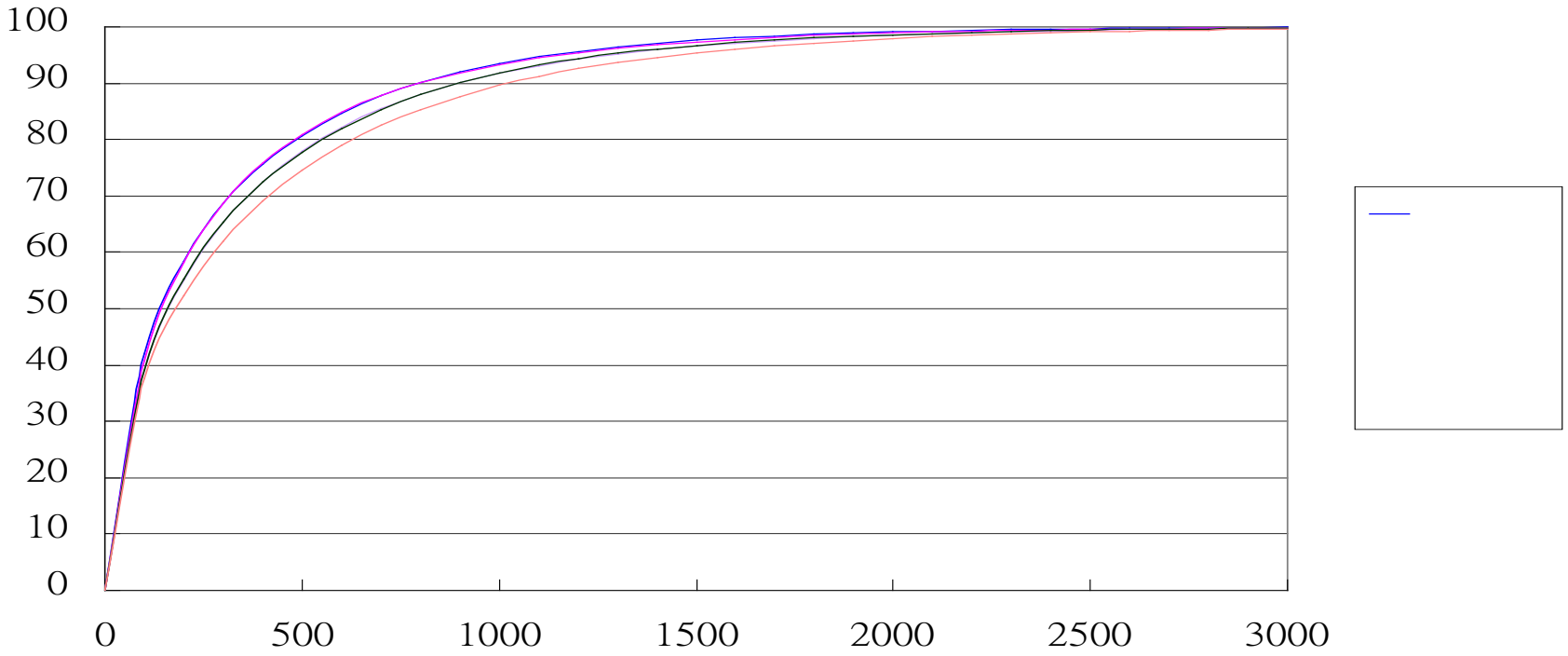
- 输入
- 硬件

- 现在:
  外来理论 --- 消费者

信息处理对象:「字」与「词」(以外?)

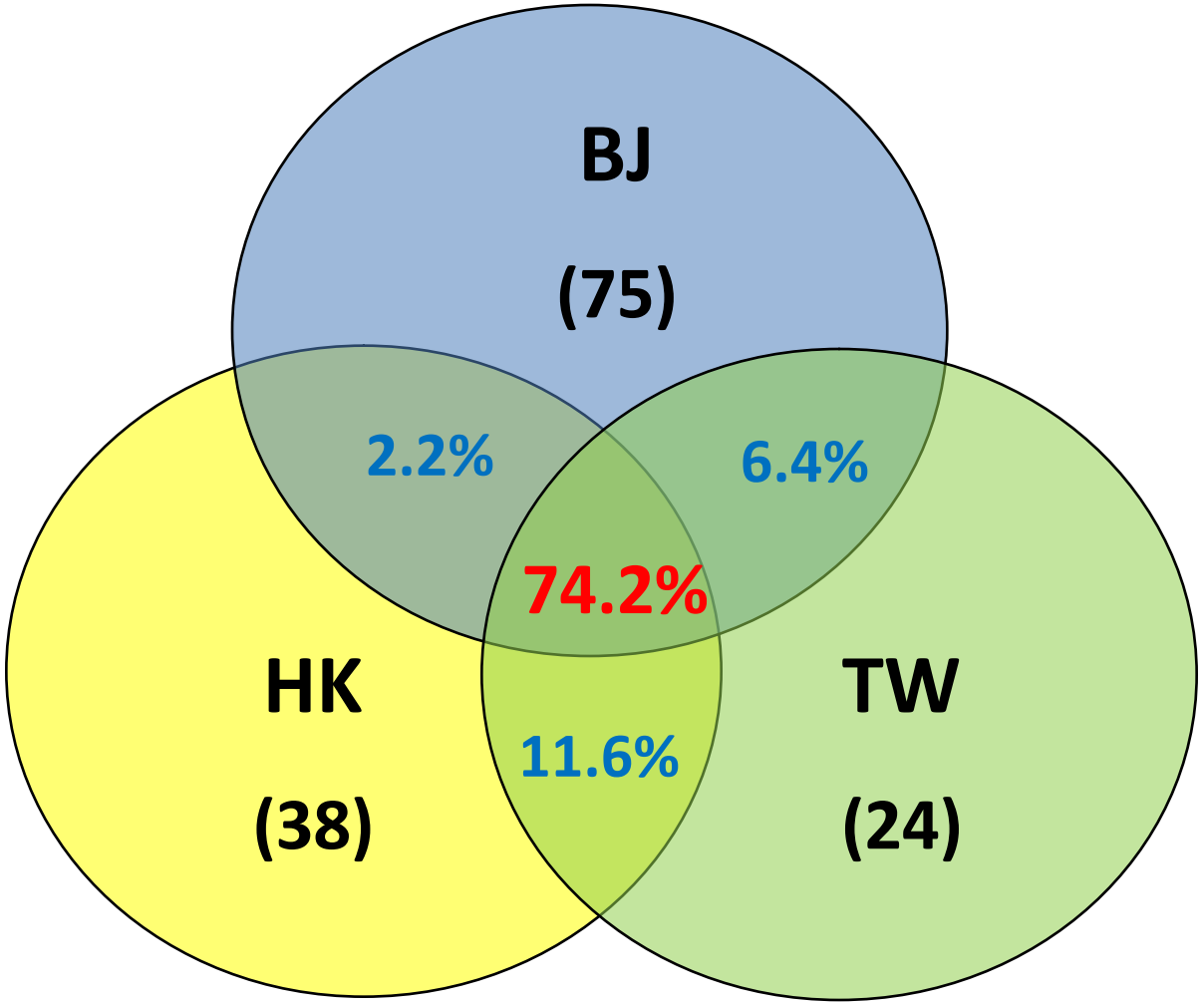
汉语一致性

各地区首三千常用字覆盖率



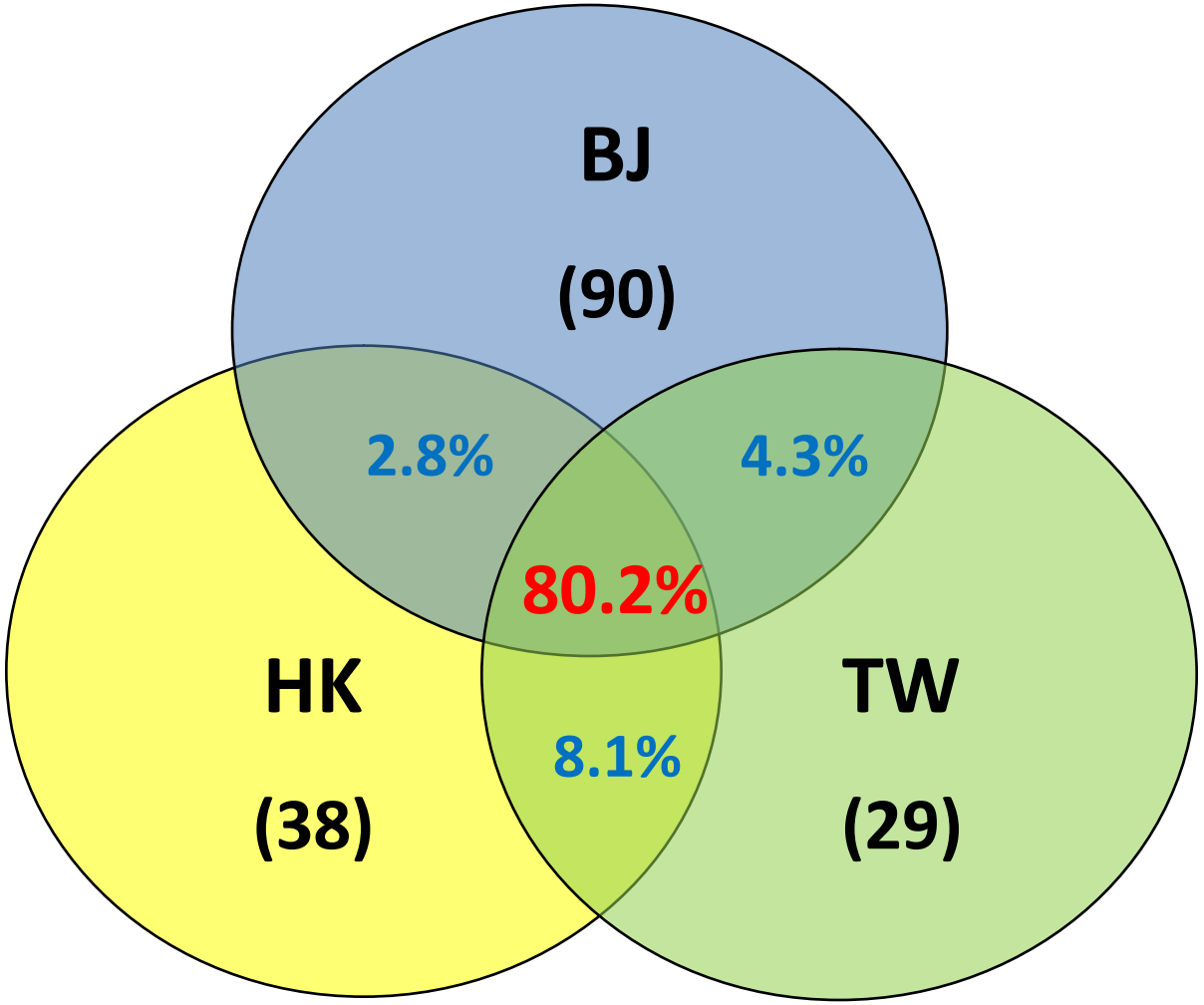
京，港，台三地首　常用单字比较

(1995 - 2009)



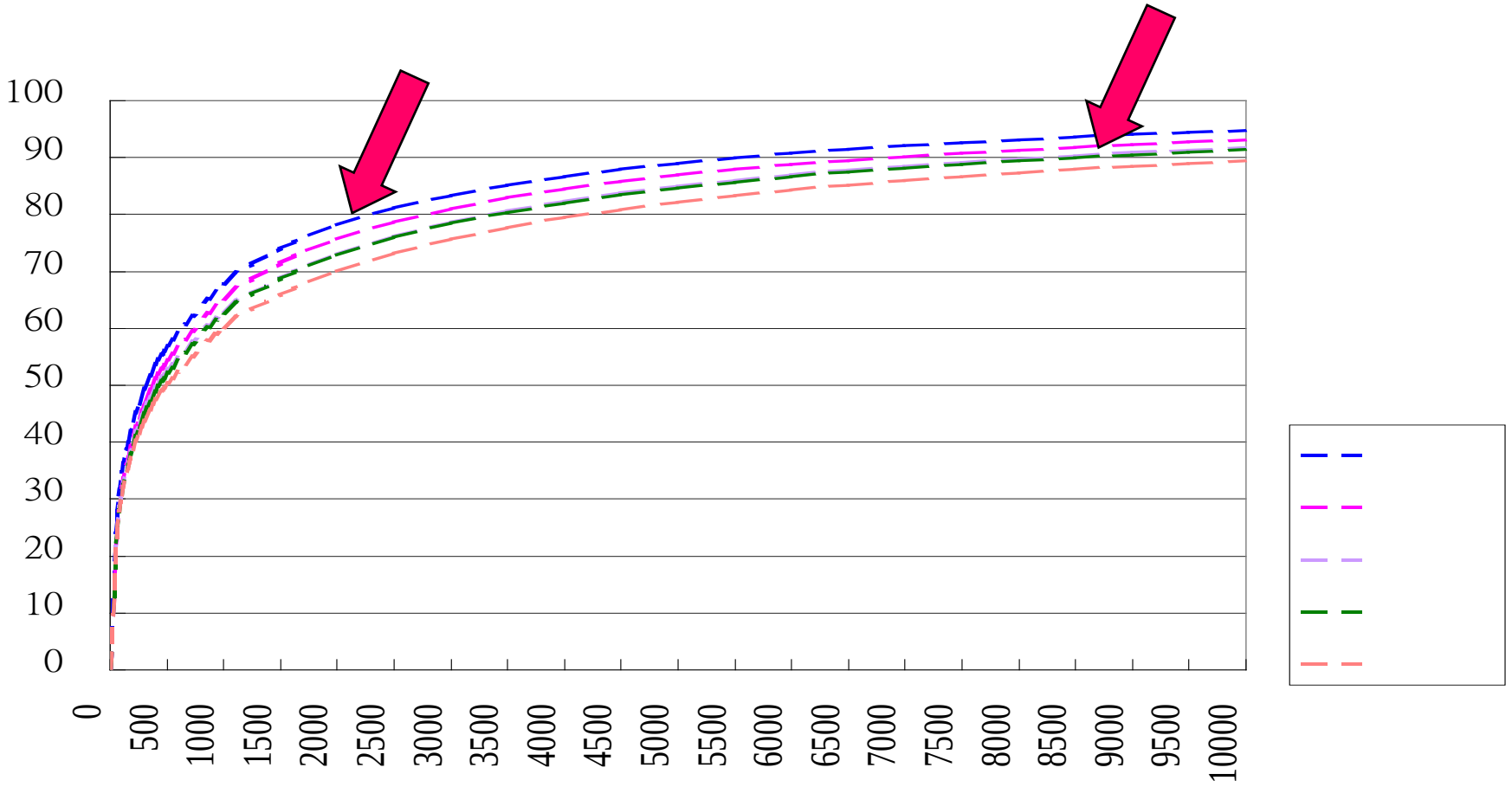
京，港，台三地首8　常用单字比较

(1995 - 2009)



TN#642

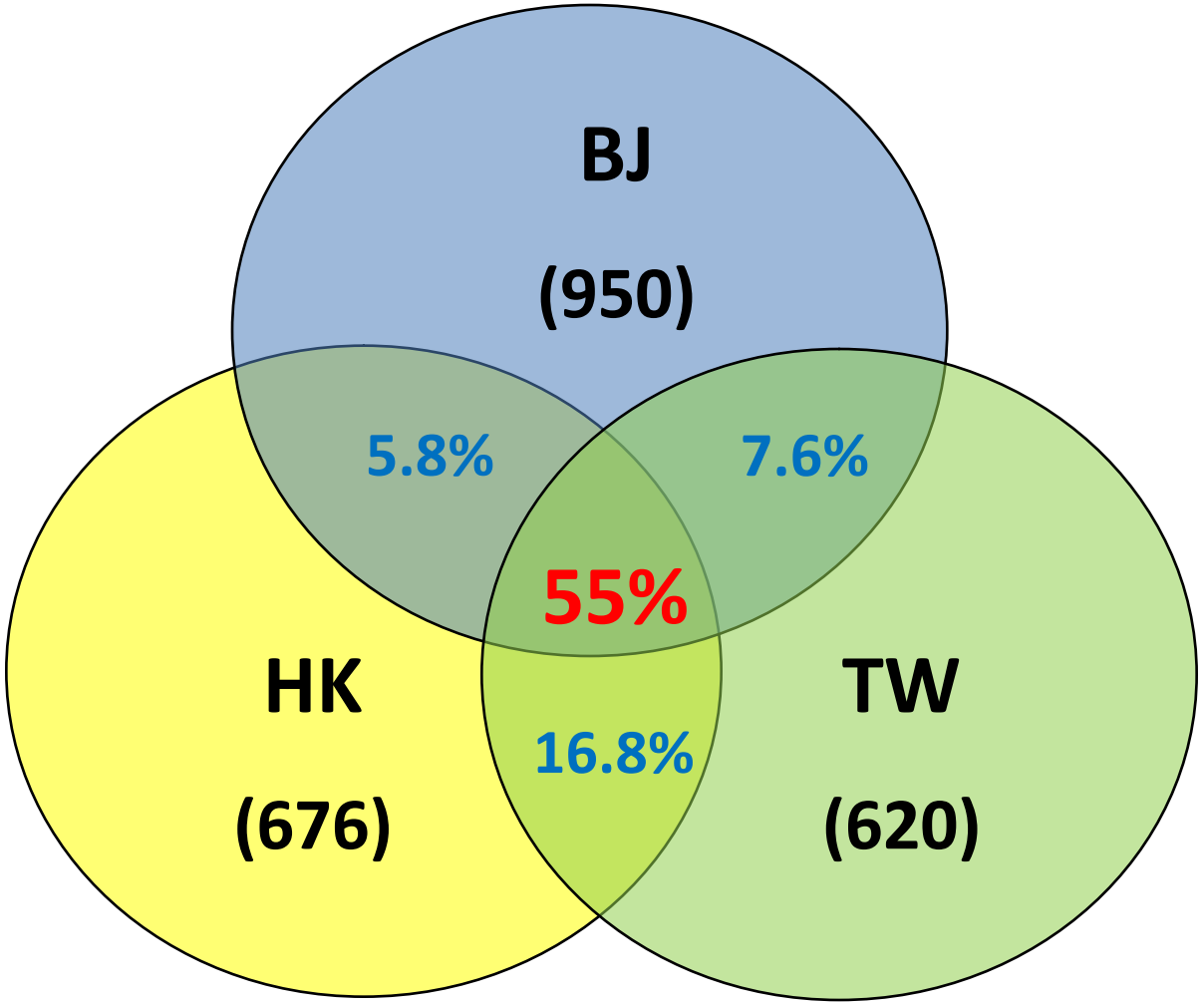
各地区首一万常用词语覆盖率



1. 2000 3500 80%
2. 10000 93%

京，港，台三地首　　常用詞比较

(1995 - 2009)



京，港，台三地最高频50词

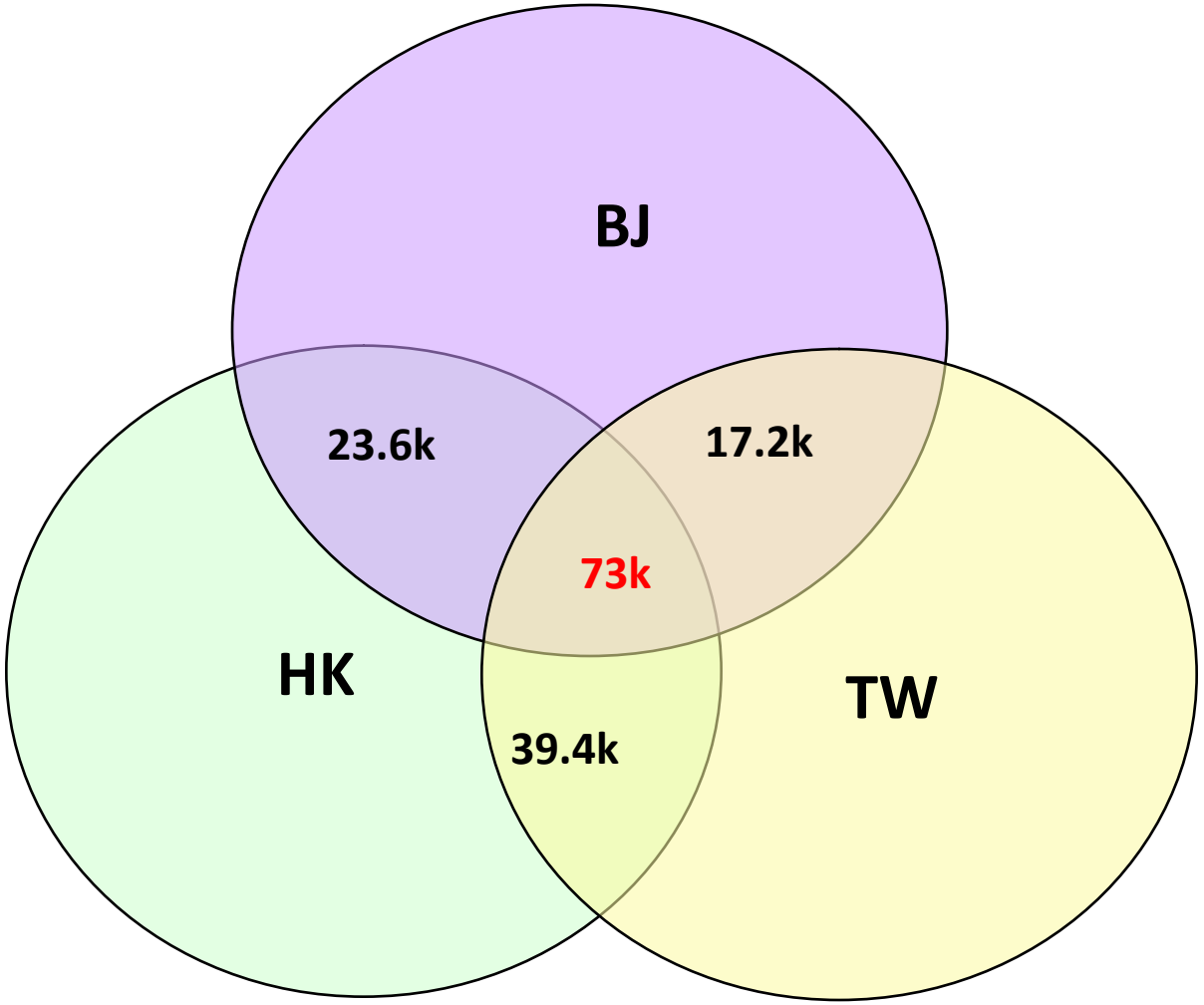
- 的 在 是 和 了 有 他 不 年 与 个 为 对
  将 这 说 人 也 但 中国 多 以 及 会 要
  中 日 表示 后 到 国 都 等 时 已 政府
  上 第 我 而 发展 就 美国 大 并 于 各
  名 被 最

京，港，台三地共现单音节词

北京，香港，台湾
的 在 是 有 他 不 及 会 和 但 年 个 与 人 了 将 说 对 多 后 时 为 而 已 这 名 以 中 于 也 上 被 大 要 并 至 到 她 向 元 该 都 更 前 由 内 可 我 最 本 次 就 第 新 日 能 等 又 每 来 其 下 再 很 或 才 各 之 国 地 过 从 全 高 所 把 位 应 据 此 还 好 比 副 达 项 家 给 着 得 用 种 出 天 做 让 起 几 使 党

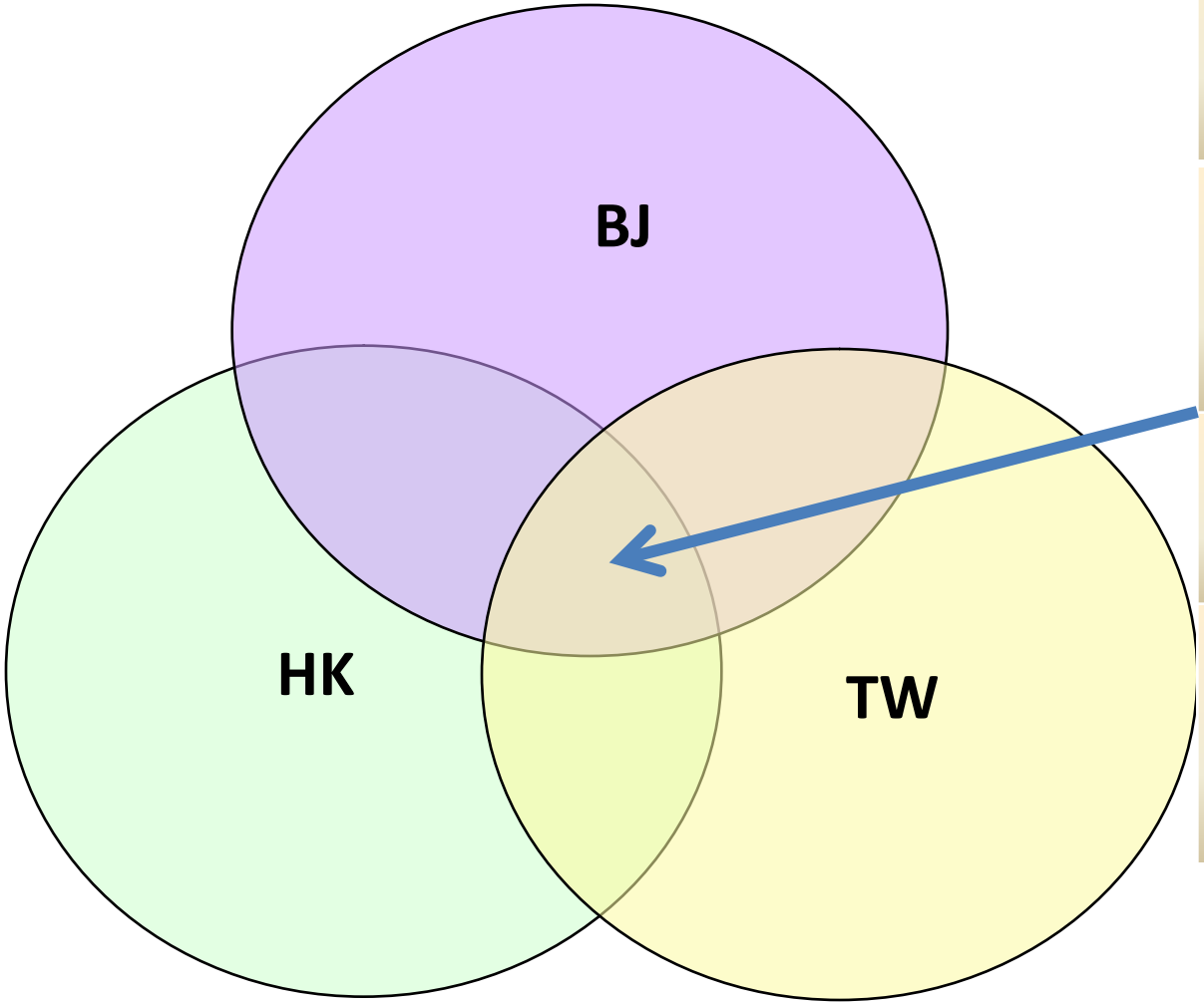
香港和台湾
则 仍 曾 较 成 岁 因 却 月 约 未 无 均 只 外 作 间 分 若 数 另 如 正 事 即 点 自 场 案 美 先 连

京，港，台三地词汇比较
(1995 - 2009)



京，港，台三地词汇比较

(1995 - 2009)



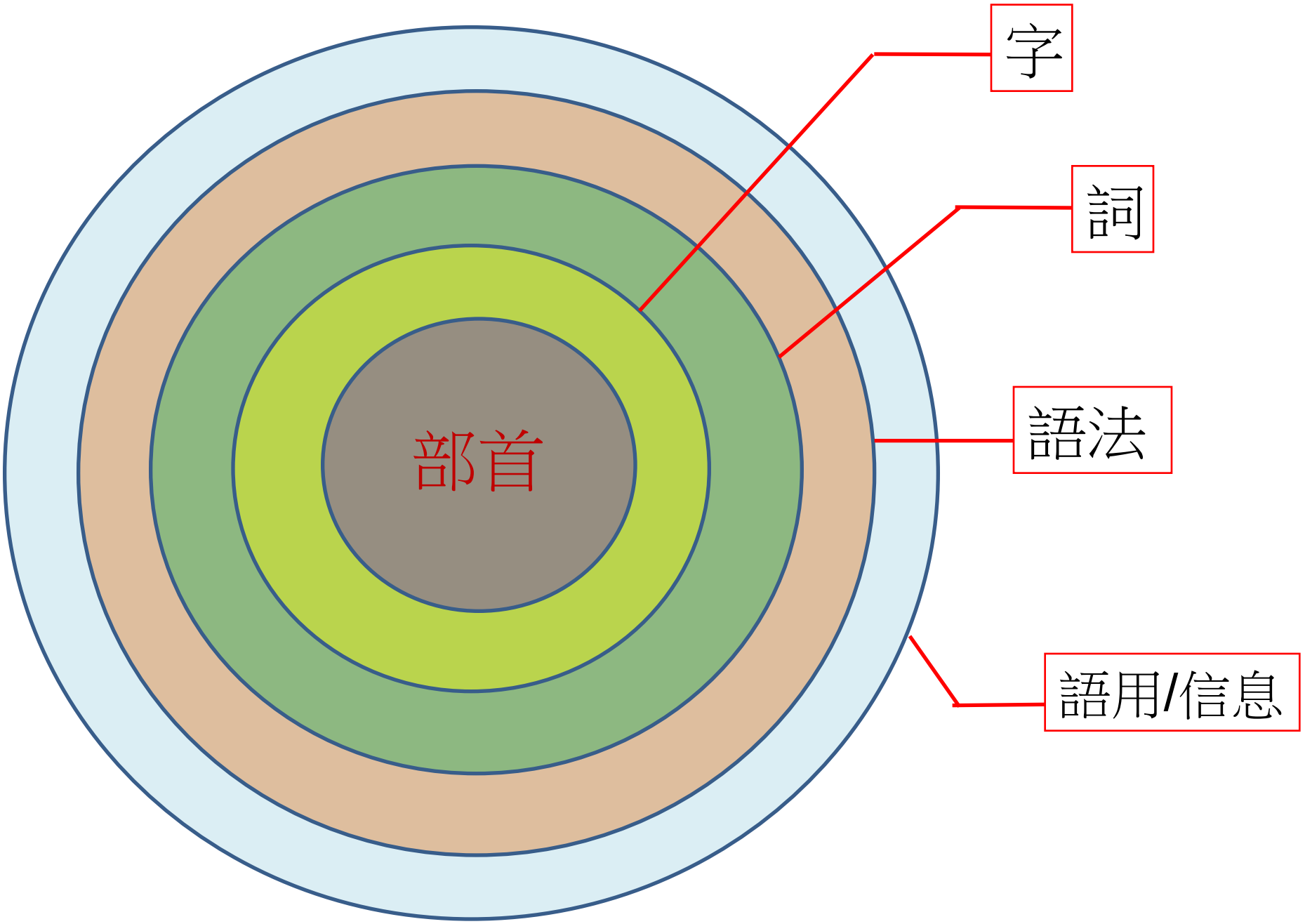
≧5%	58k	校产, 潜势, 脱口, 发卡行, 稳健型, 白搭
≧10%	43.3k	牵累, 结售, 研定, 区域主义, 累垮, 复建
≧20%	17.9k	周转率, 转报, 谋福, 打制, 拼组
≧30%	2.2k	增压, 老妪, 高架桥, 索赔, 投资热, 奥运, 谴责

语法

语言信息

新闻人物褒贬指数

人物	北京	香港	台北
邓小平 Deng Xiaoping	10	10	10
克里 John Kerry	10	0.6	1.4
小泉纯一郎 Junichiro Koizumi	-8.4	-10	-4.6
刘翔 Lu Xiang	9.6	10	10
陈水扁 Chen Shuiben	-10	-6.2	-4.6
董建华 Tung CH	10	-2.6	-7.0


字
詞
語法
語用/信息
部首

語言	熵
法語	3.98 (字母)
意大利語	4.00 (字母)
西班牙語	4.01 (字母)
英語	4.03 (字母)
俄語	4.35 (字母)
漢語	**9.71 (字)**

部首

汉语亲属关系词汇比较

I.	婆 po	maternal mother	爺 ye	paternal grandfather
	姥 lao	maternal mother		
II.	媽 ma	mother	父 fu	father
	娘 niang	mother	爸 ma	father
	姨 yi	maternal aunt	伯 bo	older uncle
			叔 shu	younger uncle
III.	姐 jie	older sister	兄 xiong	older brother
	姊 jie	older sister	弟 di	younger brother
	妹 mei	younger sister		

妣 “a deceased mother”

娣 “sister”

妹 “sister”

甥 “nephew/niece” (from sister)

姪 “nephew” (from brother)

嫁 jia marry (- to take husband)

娶 qu marry (- to take wife)

婿 xu son-in-law

媳 xi daugher-in-law

媒 mei marriage - broker

始 "origin"

姓 "surname"

北京、香港、台北汉语词「熵」

	北京五年		台北五年		香港五年		京港台五年		六地一年		六地四年		六地五年	
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	F2	G1	G2
H_0	**11.45**	**11.11**	**11.69**	**11.36**	**11.96**	**11.64**	11.96	11.60	12.11	11.72	11.94	11.57	12.03	11.64
N	119054	82553	112149	79113	123916	89713	246379	159348	132745	123572	313414	191239	394751	232016
Hmax	16.86	16.335	16.777	16.27	16.92	16.455	17.91	17.28	17.52	16.92	18.26	17.55	18.59	17.83
Hq	0.679	0.680	0.697	0.698	0.707	0.707	0.668	0.671	0.691	0.693	0.654	0.659	0.647	0.653
語料(字)	10M	---	10M	---	10M	---	31M	---	10M	---	43M	---	64M	---

(表中 $H_{max} = \log_2 N$; $H_q = H_0 / H_{max}$.)

楊利偉 (Yang Liwei)

Media coverage of the first Chinese astronaut 楊利偉 (Yang Liwei)

京港台區四地2003年新聞名人榜

香港、台灣、北京、上海四地見報率最高的名人

	香港	台灣	北京	上海
1	小布殊	小布希	胡錦濤	姚明
2	碧咸	陳水扁	溫家寶	薩達姆
3	薩達姆	哈珊	布什	布什
4	董建華	胡錦濤	江澤民	陳良宇
5	劉德華	劉泰英	吳邦國	韓正
6	謝霆鋒	李登輝	薩達姆	胡錦濤
7	張國榮	游錫	李肇星	哈恩
8	張柏芝	溫家寶	吳儀	貝克漢姆
9	梅艷芳	江澤民	阿巴斯	奧尼爾
10	胡錦濤	馬英九	曾慶紅	科比
11	王見秋	張國榮	賈慶林	楊利偉
12	溫家寶	布萊爾	李長春	巴金
13	梁錦松	謝深山	楊利偉	小威廉姆斯
14	王菲	宋楚瑜	希拉克	阿拉法特
15	葉劉淑儀	連戰	姚明	小泉純一郎
16	鄭秀文	呂秀蓮	李元龍	吳金貴
17	唐英年	劉德華	唐家璇	吳承瑛
18	陳水扁	宋安雄	毛澤東	烏代
19	陳冠希	阿諾	鮑威爾	成耀東
20	江澤民	李遠哲	阿拉法特	雷鋒
21	梁朝偉	郝龍斌	布萊爾	陳貞虎
22	楊永強	吳國棟	李瑞環	馬良行
23	李克勤	林全	郁建興	江澤民
24	貝理雅	游盈隆	黃菊	阿加西
25	楊利偉	證嚴	陳衛國	阿巴斯

香港媒体称呼国家领导人 HK (1995-2002)

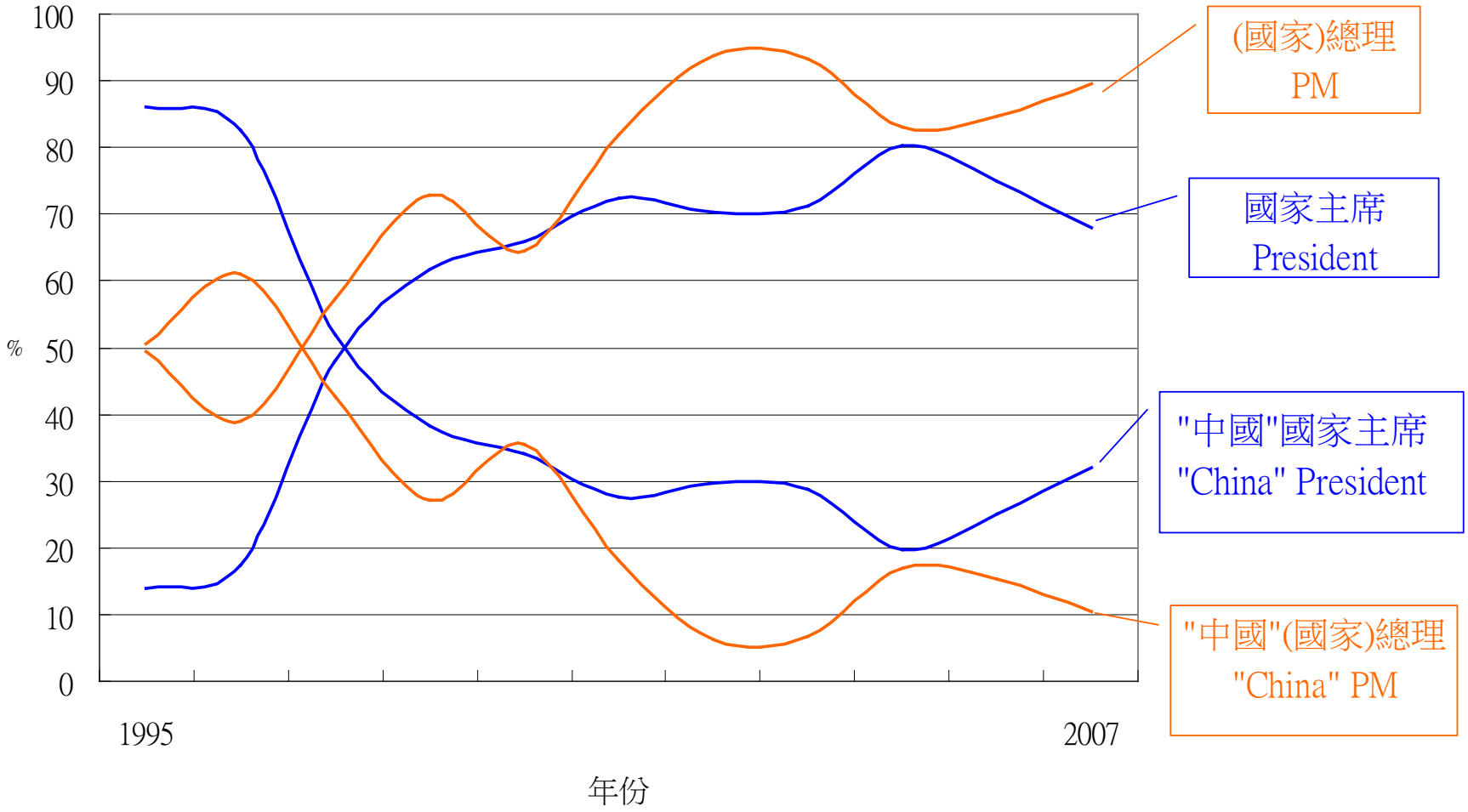
- 总理 Premier
李鹏、朱镕基、温家宝

年份	冠有 “中國” + “Chinese” （%）	不冠 “中國” - “Chinese” （%）
95	58.33	41.67
96	67.80	32.20
97	48.08	51.92
98	26.13	73.87
99	31.15	68.85
00	17.46	82.54
01	4.05	95.95
02	8.33	91.67

- 国家主席 President
江泽民

年份	冠有 “中國” + “Chinese” （%）	不冠 “中國” – “Chinese （%）
95	93.75	6.25
96	85.90	14.10
97	57.14	42.86
98	40.28	59.72
99	41.98	58.02
00	22.62	77.38
01	22.37	77.63
02	19.05	80.95

香港人的民族情感: 媒体称呼国家领导人



Linguistic Variation Across Chinese Communities

LIVAC 共時語料庫

http://www.livac.org

中国自然语言处理研究对象

- 汉藏 ~藏
- 汉蒙 ~蒙
- 汉维 ~维
- 汉((老)壮文) ~壮
- ⋮

- 汉英
- 汉日
- ⋮


120°W
歐洲
北美
俄羅斯
哈薩克
蒙古
烏茲別克
吉爾吉斯
土庫曼
塔吉克
北韓
日本
阿富汗
中國
中東
不丹
巴基斯坦
尼泊爾
南韓
台灣
印度
緬甸
香港
泰國
孟加拉
菲律賓
越南
柬埔寨
10°N
10°N
斯里蘭卡
汶萊
馬來西亞
馬爾地夫
新加坡
印度尼西亞
大洋洲
東帝汶
120°W
N

结语

- “字”与“词”本体以外的深层信息也重要

- 中国的自然语言处理：
  大有空间向内、外扩展并前程无限

参考书目

邹嘉彦. (2005). 21世纪初的中文处理 (黄居仁导读 吕学强翻译). 《计算语言学前瞻》, pp209-258. 北京: 商务印书馆.

邹嘉彦. (2004). 汉语的熵与信息开发：从共时语料库说起, 《“依旧悠然见南山” -- 香港城市大学20周年 文史论文集》, 郑培凯编, pp.97-122. 香港: 香港城市大学出版社.

邹嘉彦、莫宇航. (待刊). “汉语书面语的历史与现状：海峡两岸汉语书面语近年演变 - 以语料库为出发点”.《汉语书面语的历史与现状》, 商务印书馆出版.

邹嘉彦, 黎邦洋. (2003). “汉语共时语料库与信息开发”.《中文信息处理若干重要问题》 徐波 孙茂松 靳光瑾 主编, pp.147-165. 科学出版社.

邹嘉彦、邝蔼儿、路斌、蔡永富. (2011). “汉语共时语料库与追踪语料库: 语料库语言学的新方向”.《中文信息学报 - 庆祝中文信息处理学会成立三十周年会议论文集》 2011年第6期, 38-45页, 北京: 中国中文信息学会.

Tsou, Benjamin. (1981). “A Sociolinguistic Analysis of the Logographic Writing System of Chinese”, , 9: 1-19.

謝謝