

# *A Very Large Synchronous Corpus as Monitoring Corpus:*

Benjamin K. Tsou 鄒嘉彥

Research Centre on Linguistics and Language Information Sciences

The Hong Kong Institute of Education

btsou@ied.edu.hk

# Functions of **Large Synchronous Corpora**

## 1. rich resources for

- natural language processing
  - segmentation, parsing, MT, etc.
- human and automatic content analysis

## 2. As *monitoring corpora*, also for

- A. - Synchronous linguistic variations (lex, syn, pragm)
  - Cross-communal comparison of language developments
  - Historical developments in language and spread in time
- B. - Tracking and analyzing salient cultural items of speech communities

# 3 kinds of corpora

1. LIVAC

2. Sentiment Analysis

3. Bilingual/Trilingual  
(English, Chinese, Japanese) Patents



# 漢語共時語料庫

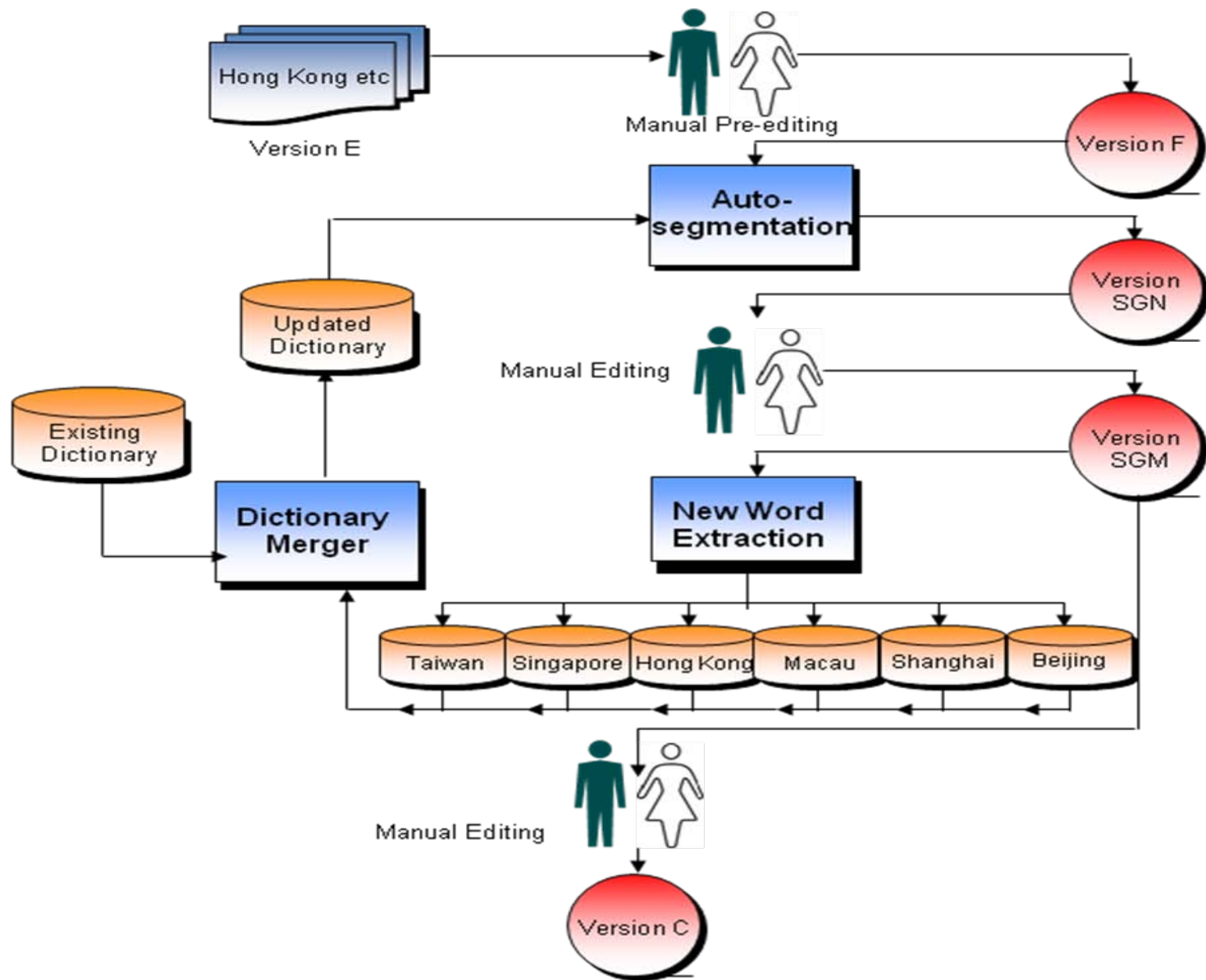
Linguistic Variations Across Chinese Communities

<http://www.livac.org>

# Design of LIVAC

Geographical coverage:	Hong Kong, Taipei, Macau, Singapore, Shanghai, Beijing, Canton, Shenzhen, (Japan)
Data source:	Representative local (Chinese) media texts
Time span:	Since 1995 ( $\geq 16$ years)
Coverage:	Main news, editorials, international news, local news, entertainment, sports, finance, etc.
Corpus size:	Up to 2010 : 400M character, 1.5M word types

# LIVAC segmentation system workflow



# 1. Segmentation Problems in Chinese

## 歧義與分詞的重要性

上海市長江大橋

上海 市長 江大橋

*The mayor of Shanghai Zhang Daqiao*

上海市 長江大橋

*Shanghai Yangtze River Bridge*

## 2. POS Tagging



LVAC ID:

Password:

[Log In](#)

The LIVAC (Linguistic Variations in Chinese Speech Communities) synchronous corpus, pioneered by the [Research Centre on Linguistics and Language Information Sciences](#) at [The Hong Kong Institute of Education](#), contains texts from representative Chinese newspapers and electronic media of Hong Kong, Taiwan, Beijing, Shanghai, Macau and Singapore. The collection of materials from the diverse communities is synchronized, and so offers an innovative "Window" approach for a whole variety of comparative studies and useful IT applications.

Analyzed by various linguistic units (e.g. characters, words, sentences), the LIVAC corpus serves many purposes. In particular, it provides an important database and means for in-depth investigation of lexical development, including the evolution of new concepts and their expressions, in contemporary Chinese.

All corpus texts have undergone automatic segmentation, and the results have been manually verified. A lexical database is derived from the segmented texts. Apart from ordinary words, those expressing new concepts or undergoing sense shifts, as well as regionalistic words from the six communities, are singled out. The database is thus a rich resource for research into linguistics, sociolinguistics, and Chinese language and society. Up to date, quantitative data on the Chinese language are also particularly useful for applications in the field of Information Technology, including the development of search engines and machine translation systems in language engineering.

Fresh textual materials for the corpus have been collected every four days since July 1995, with a 10-year time span planned for the collection to capture salient pre- and post-millennium evolving cultural and social fabrics of the diverse Chinese speech communities. Up to January 2010, the unique and growing corpus contains over 150 million Chinese characters and over 720,000 word types, and is still expanding.

The Centre has launched a bi-weekly Celebrity Roster listing the top 25 celebrities in Beijing, Shanghai, Hong Kong and Taiwan according to their media exposure in Chinese newspapers, and similar indices for place names and common words. Comments and feedback are welcome.

Try LIVAC now







[English](#) | [繁體](#) | [簡體](#)

## Word Browser

Input a word or a word fragment. Change parameters below as needed. Only traditional Chinese characters are accepted.

Enter query

Search

Match Mode

☒ Partial Match ☐ Exact Match ☐ As Prefix Only ☐ As Suffix Only ☐ As Infix Only

Word Length

From  To  Characters

☒ No Tags ☒ Personal Names ☒ Place Names ☒ Organizational Names

Word Class

[\[select all\]](#) [\[clear all\]](#)

☒ Reduplication ☒ Alias ☒ Haplology ☒ Number

☒ Other Nouns ☒ Works Names

[\[Reset query parameters\]](#)

[\[Clear results and Start over\]](#)

Copyright © 2010 Research Centre on Linguistics and Language Information Sciences, Hong Kong Institute of Education

2898 results have been retrieved with the above query parameters.

Please select words for geographical distributions by checking the boxes beside them. Specify the target date range, and click "Show Geographical Distribution" button

**Note:** your current access level entitles you to access corpus data spanning from 1995-07-01 to 2009-06-20.

Shortcuts for selection:

[select all] [clear all] [toggle all]

Show Geographical Distribution

From: 1995-07-01



To: 2009-06-20



- |                                  |                      |
|----------------------------------|----------------------|
| (1) <input type="checkbox"/> 登車  | Other Classes (其他詞類) |
| (2) <input type="checkbox"/> 重車  | Other Classes (其他詞類) |
| (3) <input type="checkbox"/> 竊車  | Other Classes (其他詞類) |
| (4) <input type="checkbox"/> 飆車  | Other Classes (其他詞類) |
| (5) <input type="checkbox"/> 攔車  | Other Classes (其他詞類) |
| (6) <input type="checkbox"/> 快車  | Other Classes (其他詞類) |
| (7) <input type="checkbox"/> 針車  | Other Classes (其他詞類) |
| (8) <input type="checkbox"/> 尬車  | Other Classes (其他詞類) |
| (9) <input type="checkbox"/> 肋車  | Other Classes (其他詞類) |
| (10) <input type="checkbox"/> 救車 | Other Classes (其他詞類) |
| (11) <input type="checkbox"/> 進車 | Other Classes (其他詞類) |

(1)	<input type="checkbox"/> 登車	Other Classes (其他詞類)
(2)	<input type="checkbox"/> 重車	Other Classes (其他詞類)
(3)	<input type="checkbox"/> 竊車	Other Classes (其他詞類)
(4)	<input type="checkbox"/> 飆車	Other Classes (其他詞類)
(5)	<input type="checkbox"/> 攔車	Other Classes (其他詞類)
(6)	<input type="checkbox"/> 快車	Other Classes (其他詞類)
(7)	<input type="checkbox"/> 針車	Other Classes (其他詞類)
(8)	<input type="checkbox"/> 尬車	Other Classes (其他詞類)
(9)	<input type="checkbox"/> 肋車	Other Classes (其他詞類)
(10)	<input type="checkbox"/> 救車	Other Classes (其他詞類)

	Beijing	Shanghai	HK	Taiwan	Macau	Singapore	Shenzhen	Guangzhou
(1) 客車 (其他詞類)	1139 (25.05%)	1758 (38.66%)	86 (1.89%)	203 (4.46%)	479 (10.53%)	72 (1.58%)	238 (5.23%)	572 (12.58%)
(2) 轎車 (其他詞類)	1840 (22.24%)	3229 (39.03%)	152 (1.84%)	455 (5.50%)	226 (2.73%)	484 (5.85%)	763 (9.22%)	1124 (13.59%)
(3) 奶車 (其他詞類)	3 (60.00%)	2 (40.00%)	0	0	0	0	0	0
(4) 炮車 (其他詞類)	6 (23.08%)	10 (38.46%)	2 (7.69%)	3 (11.54%)	0	2 (7.69%)	0	3 (11.54%)
(5) 糞車 (其他詞類)	7 (41.18%)	10 (58.82%)	0	0	0	0	0	0

Balanced corpus

(平衡語料庫)

**Avoid repetition of content**

內容避免重覆

---

**Synchronous/Homothematic corpus**

(共時/同題語料庫)

**Purposeful repetition of content**

內容刻意重覆

# LIVAC 漢語共時語料庫

- **Gigantic, Homothematic and Synchronous Chinese corpus**
- 大規模, 同題, 共時 漢語語料庫
- It can **monitor**
  - Pan-Chinese language development  
(漢語語言發展), and
  - Underlying Pan-Chinese changing cultural trends  
(不同漢語地區文化差異)

# Synchronous corpus

## 共時語料庫

Variables:      Time      (diachronic)  
                    Space      (synchronic)  
                    Domain

“WINDOW 視窗” approach

# WINDOW Approach

- Observations through WINDOW reflect broader and deeper developments
- Consistency in WINDOW approach
- Comparability in WINDOW size



# I. Monitoring Chinese lexical development

- The Chinese ‘吧’ BAR word family

# II. Comparing Chinese and Japanese

- 車 (VEHICLE) word family

# ‘吧’ BAR: Monitoring Bar culture in Chinese

- 吧 (*ba*):  
phonetic replica of *bar* (吧檯)
- An establishment for consuming liquor in a congenial setting
- 酒吧 (*jiu-ba*): liquor-bar [disyllabic word]
- 吧女 (bar-girl), 吧娘 (bar-woman), 吧檯 (bar-table)



# Examples of bar-words in LIVAC

<b>HK</b>	基吧 ‘gay-bar’	<b>BJ</b>	森林氧吧 ‘forest-oxygen-bar’ 公話話吧 ‘public-phone-bar’
<b>MC</b>	卡拉OK吧 ‘karaoke-bar’ 雪糕吧 ‘ice-cream-bar’	<b>TW</b>	影吧 ‘film-bar’ 玻璃吧 ‘glass-bar’ 琴吧 ‘piano-bar’

# Examples of bar-words from LIVAC

SH	啤酒吧 ‘beer-bar’ 搖頭吧 ‘head-shaking-bar’ 自助吧 ‘self-service-bar’ 電腦網吧 ‘computer-internet-bar’ 咖（啡）吧 ‘coffee-bar’ 網絡咖啡吧 ‘cyber-coffee-bar’	軟陶吧 ‘soft-clay-bar’ 黃陶吧 ‘yellow-clay-bar’ 舊吧 ‘vintage-bar’ 木吧 ‘wooden-bar’ 紙吧 ‘paper-bar’ 涼吧 ‘cool-bar’ 綠吧 ‘green-bar’
SG	飲料吧 ‘beverage-bar’ 奶吧 ‘milk-bar’ 茶吧 ‘tea-bar’ 沐浴吧 ‘shower-bar’	

# Distributions of “BAR” words in LIVAC

(%)	HK	MC	SH	BJ	TW	SG	TOTAL
Types	10.4	14.3	42.9	14.3	16.9	15.6	100
Tokens	30.1	20.8	23.6	12.7	4.4	8.4	100

- *Shanghai*
  - greatest variety in lexical derivatives of bar
  - greatest variety of relative cultural developments related to BAR
- *Hong Kong*
  - the highest concentration of smaller range of bar related events
  - greater penetration of bar culture

# Meaning extension of BAR in Chinese

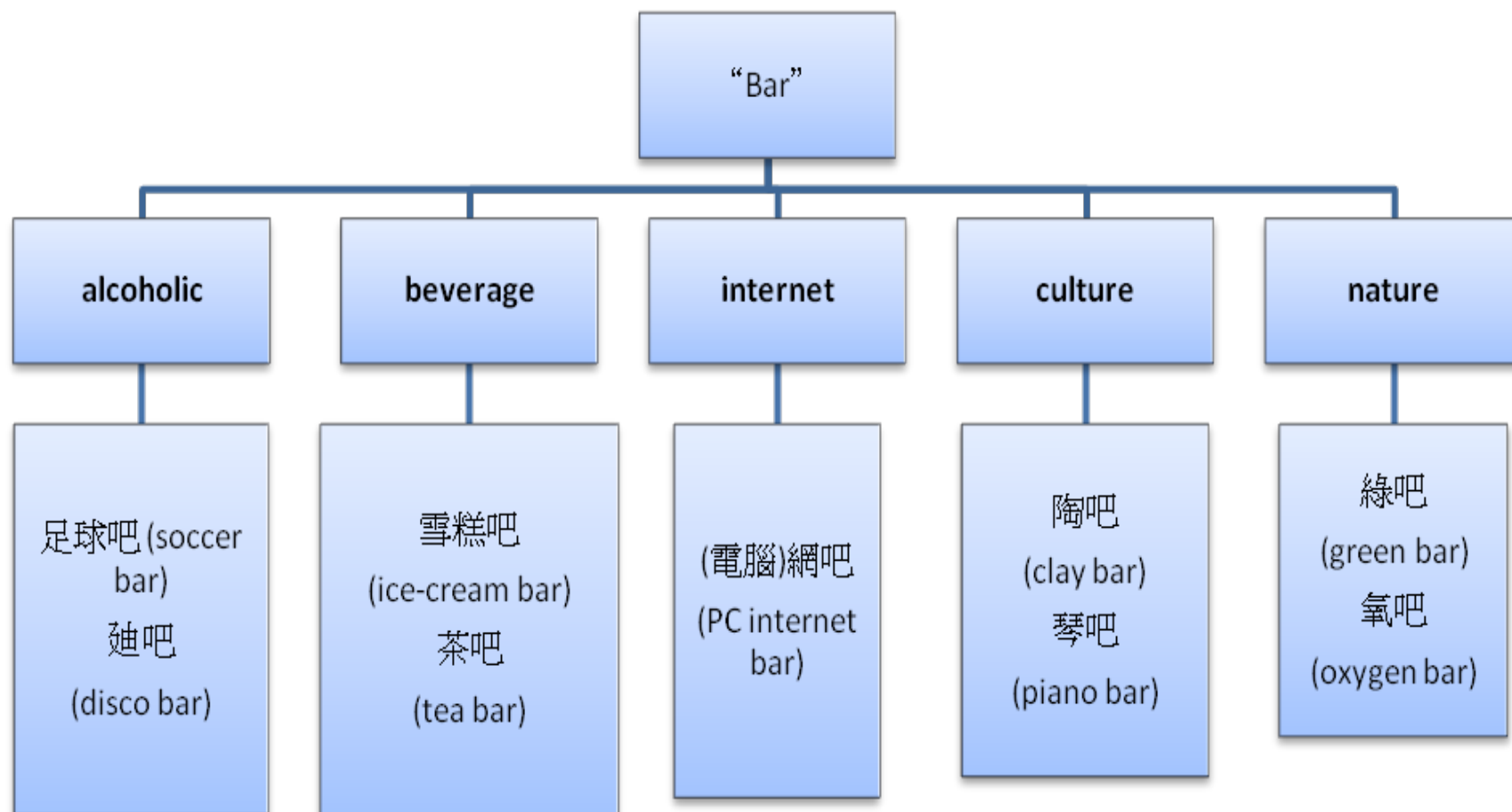
- Beyond the traditional association with *liquor*
- 網吧 (internal-bar / Internet Cafe)
- 咖啡吧 (coffee-bar)

“liquor-drinking places”



“Leisure places which have specific functions or facilities”

# Five categories of 'Bar'



# Distribution of bar-word *tokens*

<b>Tokens (%)</b>	<b>HK</b>	<b>MC</b>	<b>SH</b>	<b>BJ</b>	<b>TW</b>	<b>SG</b>
<b>Alcoholic</b>	<i>56.3</i>	<i>61.1</i>	43.2	22.6	<i>64.6</i>	<i>61.7</i>
<b>Beverage</b>	0.5	2.8	1.6	0	1.2	4.4
<b>Internet</b>	43.2	35.8	<i>47.4</i>	<i>70.3</i>	26.8	32.7
<b>Culture</b>	0	0	<i>5.3</i>	<i>0.8</i>	<i>4.9</i>	0
<b>Nature</b>	0	0.3	<i>1.6</i>	<i>2.9</i>	0	0
<b><i>TOTAL</i></b>	<b><i>100</i></b>	<b><i>100</i></b>	<b><i>100</i></b>	<b><i>100</i></b>	<b><i>100</i></b>	<b><i>100</i></b>



# 吧 (BAR):

## Observations from HK, MC, TW, SG

- Mainly *alcoholic* and *internet establishments*



# Distribution of bar-word *types*

Type (%)	HK	MC	SH	BJ	TW	SG	Total
<i>Alcoholic</i>	17.2	20.7	27.6	10.3	10.3	13.8	<i>100</i>
<i>Beverage</i>	7.1	14.3	42.9	0	7.1	28.6	<i>100</i>
<i>Internet</i>	13.3	13.3	20	20	13.3	20	<i>100</i>
<i>Culture</i>	0	0	66.7	6.7	26.7	0	<i>100</i>
<i>Nature</i>	0	12.5	50	37.5	0	0	<i>100</i>

# 吧 (BAR):

## Observations from SH, BJ

- Shanghai and Beijing:
  - mostly related to *culture* and *nature*
  - more innovative in extending semantic range
- Shanghai: more acculturated towards foreign culture than Beijing

## II. Comparing Chinese and Japanese: Vehicular ~~車~~ word

- Chinese (LIVAC) vs. Japanese (Asahi Shinbun)
- 2 one-month WINDOWS: 1996 vs. 2006 (10 years apart)
- *Mutual Intelligibility*

# VEHICLE 車-related words

## in 1996 and 2006 WINDOWs: Chinese

No. of items WINDOW content	Pan-Chinese	HK	MC	TW	SH	SG
1996	190	76	63	55	92	55
Items retained from 1996 in 2006	108	45	42	30	38	35
New in 2006	112	32	42	39	42	38
TOTAL in 2006	220	77	84	69	80	73

# Observations

- 44% from 1996 WINDOW retained in 2006 WINDOW
- 51% new items in 2006 WINDOW not found in 1996 WINDOW
- Shanghai WINDOW: most content change

# VEHICLE-related words appear in the Pan-Chinese context

	1996	2006
1	車	車
2	車輛	車輛
3	汽車	汽車
4	貨車	貨車
5	火車	火車
6	卡車	卡車
7	列車	列車
8	裝甲車/坦克	坦克/裝甲車
9	消防車	消防車
10	救護車	救護車
12	轎車	轎車
13	軍車	中巴
14	客車	巴士
15	警車	賽車
16		車子

} Generic items

} 巴 comes from “bus” and was used in Cantonese speech community only

VEHICLE-related words  
in 1996 and 2006 WINDOW:

Japanese (車/カー)

Items	Japanese
No. in the 1996 WINDOW	112
No. retained from 1996 WINDOW	37
No. of others found in 2006 WINDOW	43
Total no found in the 2006 WINDOW	80



# Observations

- Items retained from 1996-WINDOW: CORE vehicle items
  - 自転車 (bicycle)
  - ダンプカー (dump car)
  - 列車 (train)
  - 電車 (electric train)
  - 救急車 (ambulance)
  - 消防車 (fire engine)
  - バス (bus)
  - バイク (bike)
  - タクシー (taxi)

# Observations

- New items in 2006-WINDOW can reflect an increasing concern with different social issues in Japan:
  - 環境対応車 (car powered by electricity)
  - 禁煙車 (non-smoking car)
  - ダンプカー (green car)
  - 自爆車 (suicidal explosion car)
  - 装甲車両 (armored car)

# Comparison of mode of adaptation of VEHICLE words

- **Phonetic adaptation:**

- ダンプカー (dump car), バス (bus), バイク (bike), タクシー (taxi)

- **Semantic adaptation:**

- 消防車 (fire-engine), 救護車 (ambulance)

- **Hybrid:**

- キャンペーン車 (campaign car), クレーン車 (crane), 坦克車 (tank),

# Cultural Compatibility: Hypothesis

- Correlation between linguistic adaptation method and Cultural Compatibility of new concepts/artifacts

1. Accessibility

親近情況

2. Agreeability

投合情況

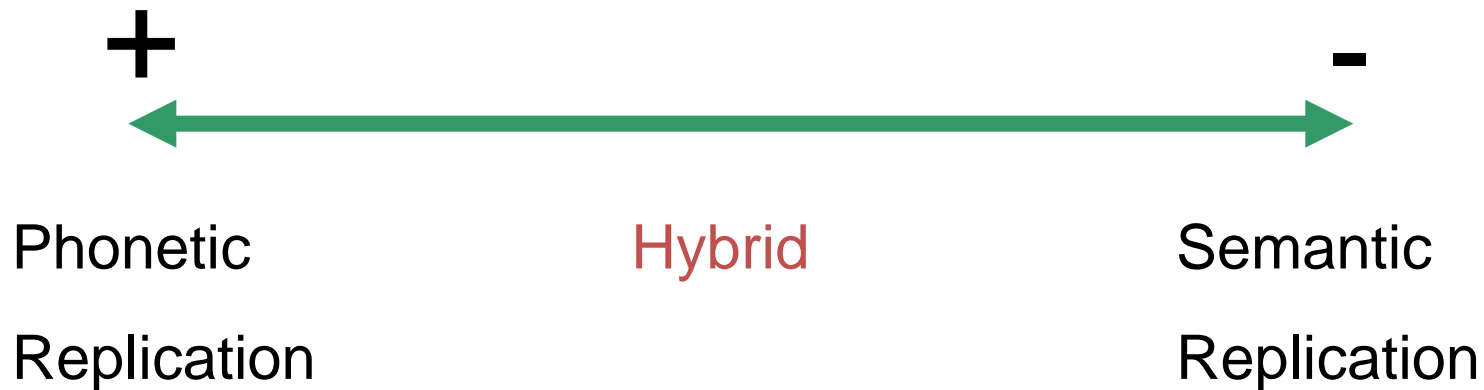
3. Familiarity

熟悉情況

(Tsou 2000)

# Cultural Compatibility

- a) Accessibility
- b) Agreeability
- c) Familiarity



# Distribution of 3 adaptation methods: VEHICLE

%	Phonetic		Hybrid		Semantic	
	1996	2006	1996	2006	1996	2006
Japan	38	15.2	2.8	21.5	59.2	63.3
Hong Kong	18.8	7	6.2	19.2	75	73.7
Macau	4.6	6.5	6.2	9.7	89.2	83.9
Taiwan	6.5	4.3	6.5	10.6	87	85.1
Shanghai	3.8	7.7	6.3	5.8	90	86.5
Singapore	7.4	7.5	5.6	9.4	87	83

# Observations

- Japan and Hong Kong have significant **reduction** in using purely **phonetic** adaptation
- **Hybrid** forms **most increase**

# Comparative distribution of VEHICLE–related words using KANJI

	<b>1996</b>	<b>2006</b>
Kanji	74 (66%)	52 (65%)
Kana	19 (17%)	18 (22.5%)
Kanji + Kana	19 (17%)	10 (12.5%)

1. Kanji, reflecting semantic adaptation, remains stable
2. Kana, representing phonetic adaptation, has a slight increase



# Mutual Intelligibility

- A measure of 2-way understanding of words between 2 languages
- E.g. Japanese Kanji vs. Chinese characters
- “a Chinese speaker or a Japanese speaker can usually understand words represented by common Kanji/Chinese characters ”

# Distribution of Mutual Intelligibility of VEHICLE-related terms

<b>%</b>	<b>China → Japan</b>	<b>Japan → China</b>	<b>Hong Kong → Shanghai</b>	<b>Shanghai→ Hong Kong</b>
<b>1996</b>	78	51.3	80.7	91.7
<b>2006</b>	98	25	98.1	82.5

E.g. “China → Japan” means “Chinese understanding of Japanese kanji words”

# Observations


- Language understanding can be **non-mutual** between 2 languages

- Significant decrease of Japanese intelligibility of Chinese items in 2006 (i.e. 51.3% → 25%): WHY?
  - Many Chinese VEHICLE-words are phonetically adapted or in hybrid forms: 吉普車 (jeep-car), 坦克 (tank), 摩托 (motor), 的士 (taxi), etc.
  - Meanings are not transparent for phonetic adaptation
    - 坦克 ≠ 坦 + 克      的士 ≠ 的 + 士
  - Japanese kanji 漢字 retain basic semantic link to (original) Chinese

## Oxford Chinese Dictionary

- The largest and most authoritative English-Mandarin Chinese/Mandarin Chinese-English bilingual dictionary available, produced in association with Foreign Language Teaching & Research Press (FLTRP) in China
- Comprehensive, in-depth coverage for English and Chinese speakers, with accurate and up-to-date translations, plus extensive grammar and usage help
- Additional resources for travel, work, or life abroad with sample letters and guides to the Internet and email, plus hundreds of cultural notes

- 最大、最权威的英汉-汉英双语词典，与外语教学与研究出版社合作编写
- 收词广度与深度并重，翻译准确、地道，辅以大量的语法及用法信息，有效帮助说英语和汉语的人士学习和使用英汉两种语言
- 词典附录中提供的信函范例、互联网和电子邮件指南，加上数以百计的文化知识，有助读者在国外旅行、工作或生活时解决语言沟通中的问题

 Buy this book and get 12 months' access\*  
to *Oxford Language Dictionaries Online*  
at [www.oxforddictionaries.com/access](http://www.oxforddictionaries.com/access)

670,000 words, phrases, and translations

 Powered by  
Oxford Corpus

Powered by the Linguistic Variations in Chinese Speech  
Communities Corpus and the Oxford English Corpus

Chinese language only (based on the Pocket Oxford Chinese Dictionary). Full terms and conditions are on the website.

OXFORD  
UNIVERSITY PRESS

[www.oup.com](http://www.oup.com)



# Oxford Chinese Dictionary



OXFORD

# Oxford Chinese Dictionary

英汉·汉英



in association with  
FLTRP Beijing

The world's most  
trusted dictionaries

wherever you are



 LiVac.org

Powered by  
Oxford Corpus

Powered by the Linguistic Variations in Chinese Speech  
Communities Corpus and the Oxford English Corpus



# 全球华语 新词语词典

QUANQIU HUAYU XINCIYU CIDIAN

邹嘉彦 游汝杰 编著

## 全球华语新词语词典

商务印书馆  
The Commercial Press



商务印书馆  
The Commercial Press

- ◎ 以巨型电脑语料库 (LIVAC) 为编写基础
- ◎ 收录2000年以后产生或流行的华语各地新词1600多条
- ◎ 所收条目精选自LIVAC语料库所见近2万条新词
- ◎ 以语料库的量化资料为基础, 说明新词在各地使用情况的差异
- ◎ 对绝大多数新词有背景知识介绍并附外来词原文
- ◎ 例句取自语料库所见各地华语的当代报刊

# Dazao “打造” “fabricate” (verb)

## 演變 *Selectional Restriction*

- I) 「打造」: fabricate concrete objects  
(動補結構, 以實物為對象賓語)
  - (a) “金縷玉衣” (jade suit)
  - (b) “家具” (furniture)
  - (c) “船隻” (ships)“
  
- II) 半虛化 (metaphoricalization):
  - a) 打造汽車業的「航空母艦」  
fabricate an “aircraft carrier” in  
automotive manufacturing
  - b) 打造與網絡有關的全新航空母艦  
fabricate a brand new “aircraft carrier” of  
the internet







# Full relaxation of S-R:

“打造”全虛化的延伸

a)新台灣	<i>New Taiwan</i>
b)新願景	<i>New vista</i>
c)京劇	<i>Peking opera</i>
d)未來	<i>Future</i>
e)新專輯	<i>New special edition</i>
f)新希望	<i>New hope</i>
g)新品牌	<i>New brand</i>
h)生活空間	<i>Leisure time</i>

# “打造” 語法進化過程



# Some other uses of the monitoring corpus

A. Content analysis and monitoring

B. Sentiment Analysis

- Report on political personalities
- ..... Products
- .....

C. Cultural icon

Identity and ethnicity

D. Linguistic changes

# Vote USA 2004



## Sources

- The Economist magazine
- Time Magazine

The Home Stretch



輸贏差一線 克里處下風

民調專家：克里不出錯可當選

巴黎

法國力挺凱瑞

八十%支持度宛如在選法國總統 主要是擔心

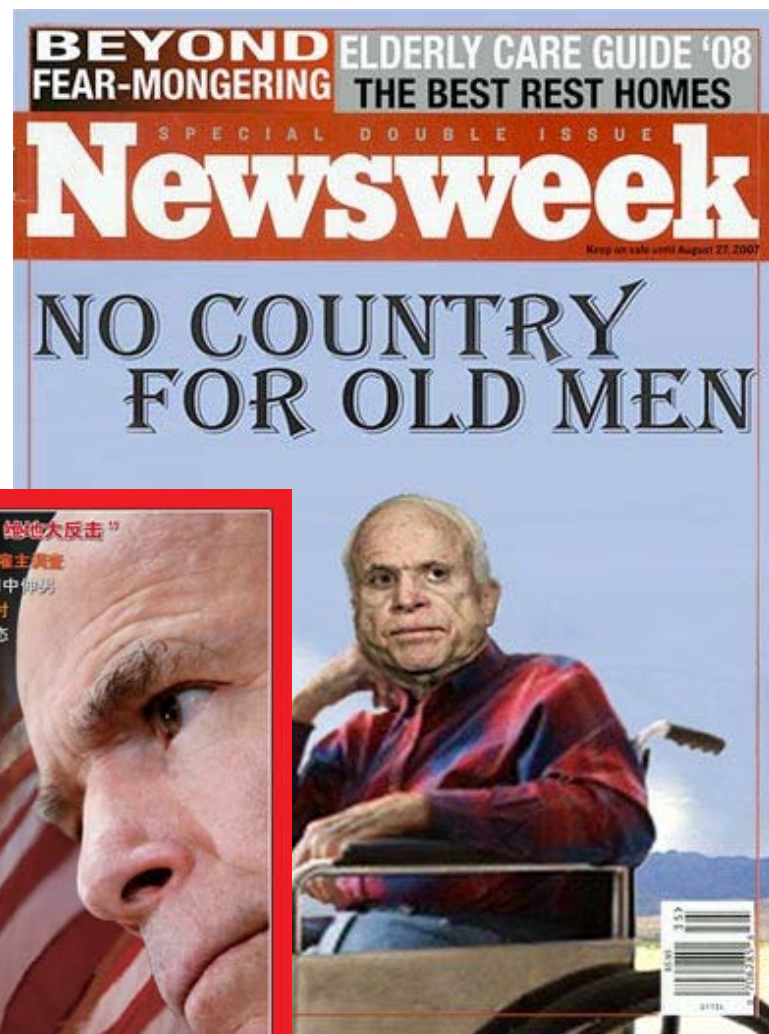
期待又怕傷害

世界危險



拉丹亮相谷起布殊支持度





# Polarity of celebrities

## 新聞人物褒貶指數

- Celebrities from 600 news from **LIVAC** 2004 (i.e. highest exposure in newspaper)
- 選自2004年**LIVAC共時語料庫**600篇報導中的名人 (即曝光率最高的新聞人物)





# 新聞人物褒貶指數

## Polarity of celebrities

人物	北京 BJ	香港 HK	台北 TW
鄧小平 Deng Xiaoping	10	10	10
克里 John Kerry	10	0.6	1.4
小泉純一郎 Junichiro Koizumi	-8.4	-10	-4.6
劉翔 Lu Xiang	9.6	10	10
陳水扁 Chen Shuiben	-10	-6.2	-4.6
董建華 Tung CH	10	-2.6	-7.0

# *National Identity & Ethnicity*

# Titles: 稱謂

1. (中國)國家主席 江澤民/ 胡錦濤  
(China) President Jiang Zemin/ Hu Jiantao
2. 李鵬總理/ 朱鎔基總理/ 溫家寶總理  
Prime Minister Li Peng/ Zhu Rongji/ Wen jiabao
3. 李總理/ 朱總理/ 溫總理  
Prime Minister Li/ Zhu/ Wen
4. 李總/ 朱總/ 溫總

# Media use of titles on National figures: HK (1995-2002)

## 香港媒體稱呼國家領導人

### - 總理 Premier

李鵬、朱鎔基、溫家寶

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

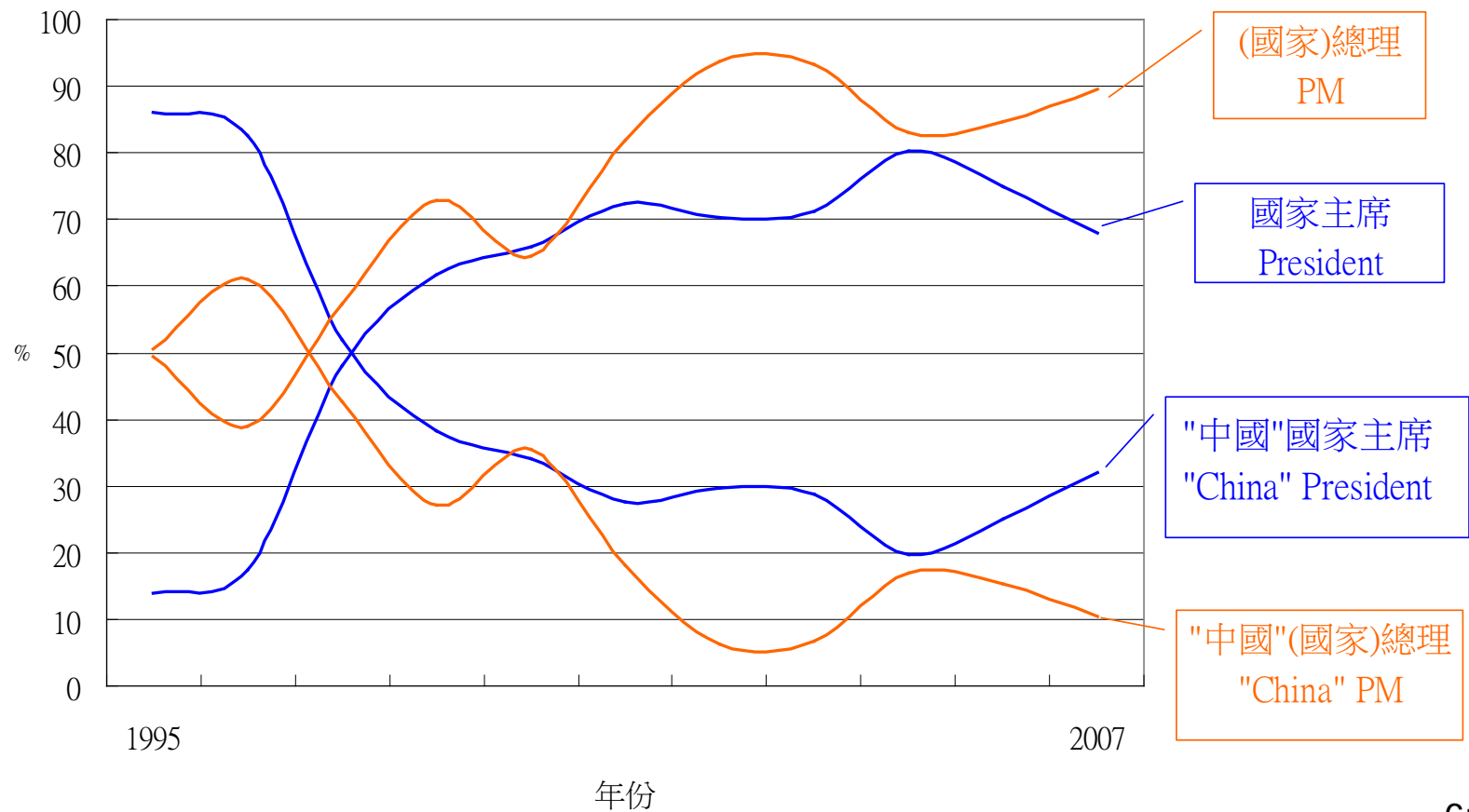
### - 國家主席 President

江澤民

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

# National Identity of Hong Kong people : Reference to National Leaders

## 香港人的民族情感: 媒體稱呼國家領導人



# Titles: 稱謂

英國首相 梅杰/ 布莱尔/ 布朗/ 卡梅倫

**Prime Minister** Major/ Blair/ Brown/ Cameron

[恆常]

(always)

# Media use of titles on National figures: **Macau (1995-2002)**

## 澳門媒體稱呼國家領導人

### - 總理 **Premier**

李鵬、朱鎔基、溫家寶

年份	冠有中國 + “Chinese” (%)	不冠中國 - “Chinese” (%)
95	46.97	53.03
96	46.81	53.19
97	29.41	70.59
98	25.71	74.29
99	33.33	66.67
00	36.36	63.64
01	44.44	55.56
02	37.50	62.50

### - 國家主席 **President**

江澤民

年份	冠有中國 + “Chinese” (%)	不冠中國 - “Chinese” (%)
95	65.31	34.69
96	60.61	39.39
97	78.41	21.59
98	54.43	45.57
99	44.05	55.95
00	40.00	60.00
01	56.52	43.48
02	59.38	40.63

# The first Chinese astronaut: Yang Liwei





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

京港台滬四地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

	HONG KONG	TAIWAN	BEIJING	SHANGHAI
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

# Chinese- English Bilingual Patents:

1.160k patent pairs

2.503k sentence pairs for abstracts,  
150k of very good quality

3.5.9M for claims, 1.9M of very good quality

4.38M for descriptions, 11M of very good quality

Japanese- English patents: 100K+

# Through Direct or Indirect Vendors



访问诺基亚其它站点

诺基亚中国



# 為網頁提供檢索功能



WWW.TOM.COM



火速填問卷  
獎您免費遊亞洲!

TOM.COM  
大學生交流室

17 Aug 2001 Fri  
29°C 76%

新聞財經 | 財 | 樂 | 電影 | 音樂 | 遊戲 | 科技 | 電訊 | PDA | 潮流 | 生活 | 健美 | 星相 | 中國機  
遇 | 教育 | 房產

社群服務 | 漫談室 | 討論區 | 留言 | 投票 | AV 播放 | 心意卡 | 下載 | 書店 | 消費購物

我的 TOM >> 更多

記事簿 行程表  
六合彩 股票報價  
星座運程 天氣報導

自家電郵

自家人名  
密碼  
登入 | 忘記密碼  
登出 | 修改資料  
新人登記

搜尋   ☒ 全球中文網頁 ☐ 全球中文網站

HOT Palm 人參 蜀山 容祖兒 王昭君 複製羊 教科書 超級電腦



華裔科學家在美國的真實處境  
連載小說《超光速運行》講述了一個和「李文和案」十分類似的故事，讀者們可能以為這是一部以該案為背景的小說，因為無論是故事情節或最  
華建控股停牌以公布中期業績 [10:02]

即時新聞

新聞財經

娛樂地帶

熱門推介

「奕妹」熱浪  
未來手電的界面設計  
(財俊) 瓦格納  
大學生交流室

TOM 著數區 >> 更多

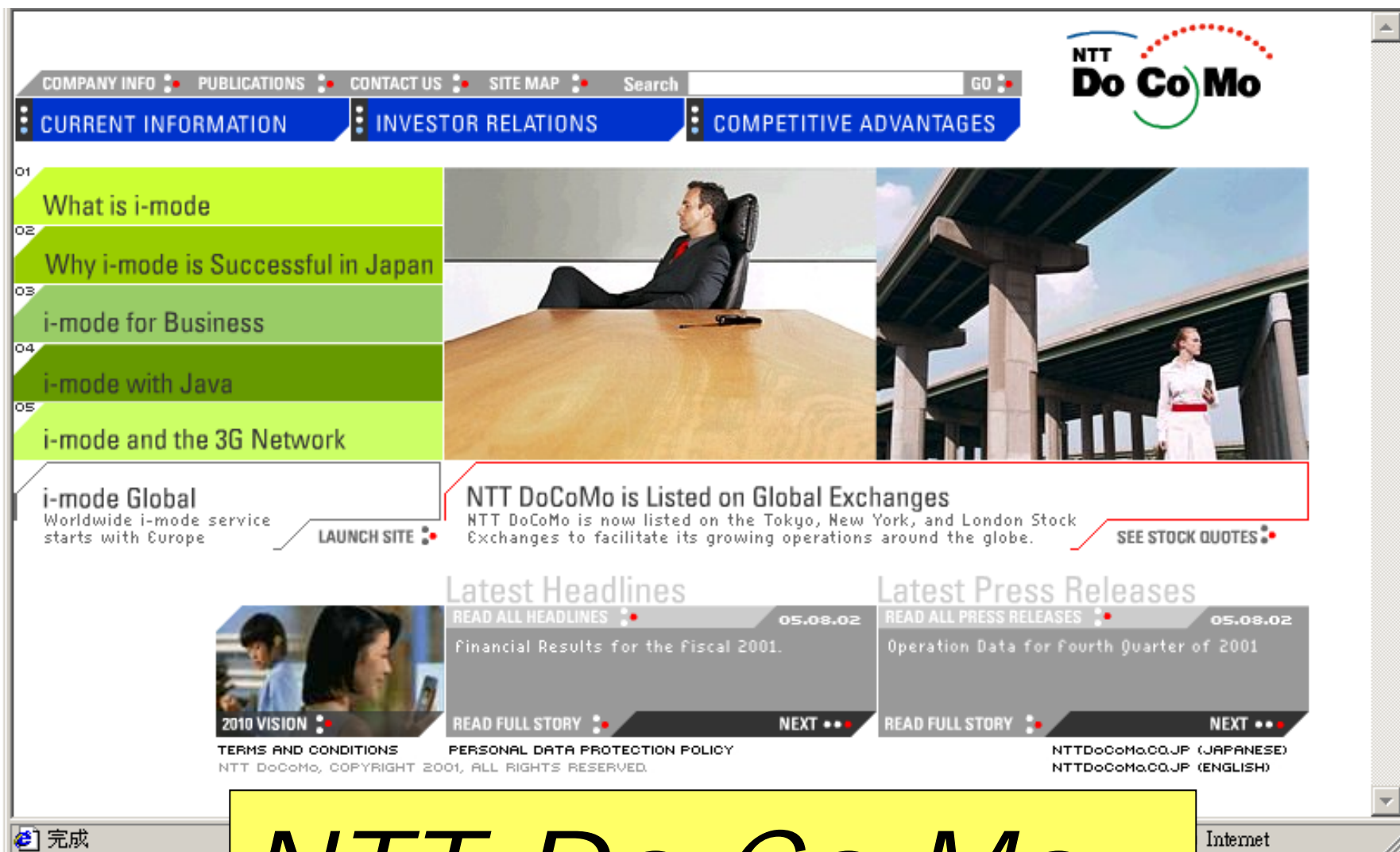
送《多啦 A 夢——大雄的太陽王傳說》精品

香港節目通 >> 更多

節目推介  
請選擇類別



# TOM.COM



*NTT Do Co Mo*

# Conclusion

- LIVAC, as a Homothematic and synchronous corpus, is very useful as a *monitoring corpus*
  - To track language development across Chinese communities;
  - To explore social and cultural changes *quantitatively* and *qualitatively*;
  - To explore changes in relative cultural orientations and mutual influences among the Chinese communities,
  - and between Chinese and Japanese societies

# References

- Cheng, C.M., Kwong O. Y., Tsou Benjamin. (2004). “Pan-Chinese Variation on Verbal Synonymy: A Study of Common Reportage Verbs in News Texts.” In *Recent Advancement in Chinese Lexical Semantics, Proceeding of 5th Chinese Lexical Semantics Workshop (CLSW-5)*, edited by Ju Dong Hong, Lua Kim Teng and Wang Hui, pp.213-219
- Chow, Ka Po, Chin, Andy, and Tsoi W. F. (2005). “Maximal Match Chinese Segmentation Augmented by Post-Processing Using Resources Generated from a Very Large Dictionary.” *Proceedings of The Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, pp.176-179.
- Kwong, O.Y., Tsou, Benjamin. (2005). “Data Homogeneity and Semantic Role Tagging in Chinese.” In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pp.1-9.
- Kwong, O.Y., Tsou, Benjamin. (2005). “Semantic Role Tagging for Chinese at the Lexical Level.” *Lecture Notes on Artificial Intelligence (LNAI)*, Vol. 3651, pp.804-814. Berlin: Springer-Verlag.

# References

- Kwong, O.Y., Tsou, Benjamin. (2005). “A Synchronous Corpus-Based Study of Verb-Noun Fluidity in Chinese.” *Journal of Chinese Language and Computing*, 13(3):227-278.
- Kwong, O.Y., Tsou, Benjamin. (2006). “Feasibility of Enriching a Chinese Synonym Dictionary with a Synchronous Chinese Corpus.” *Lecture Notes in Computer Science*, Vol. 4139, pp.322–332. Springer-Verlag.
- Kwong, O.Y. and Tsou, Benjamin. (2007). “Extending a Thesaurus in the Pan-Chinese Context.” In *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic.
- Kwong, O.Y., Tsou, Benjamin. (2008). “Extending a Thesaurus with Words from Pan-Chinese Sources.” In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, Manchester.
- Lu, Bin, Tsou, Benjamin, Zhu, Jingbo, Jiang Tao, Kwong, O.Y. (2009). “The Construction of a Chinese-English Patent Parallel Corpus.” Paper presented at the *MT Summit XII, 3rd Workshop on Patent Translation*, Ottawa, 2009.8.26-30.



# References

- Sun, Maosung, and Tsou, Benjamin. (1998). "Chinese Word Segmentation Without Using Lexicon and Hand-crafted Training Data." *Proceedings of the Conference of COLING-ACL*.
- Sun, Maosung, Xu Dongliang, Tsou, Benjamin, and Lu Huaming. (2006). "An Integrated Approach to Chinese Word Segmentation and Part-of-Speech Tagging." *Lecture Notes in Computer Science (LNCS)*, pp.299-309. Berlin/Heidelberg: Springer.
- Tsou, Benjamin K. (2000). "A WINDOW on Re-lexification in Chinese." In *In Memory of Professor Li Fang-Kuei: Essays on Linguistic Change and the Chinese Dialects*, edited by Ting Pang-hsin and Anne Yue, pp.53-72. Taipei/Seattle: Institute of Linguistics, Academia Sinica / University of Washington.
- Tsou, Benjamin. (2001). "Language Contact and Lexical Innovation." In *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*, edited by M. Lackner, I. Amelung and J. Kurtz, pp.35-56. Leiden: Brill.

# References

- Tsou, Benjamin, Chin, Andy, Chan, Kam Pong. (in press). *An English-Chinese Glossary of Latin Terms in Law* (法律常用拉丁詞匯手冊). Beijing: Commercial Press.
- Tsou, Benjamin, Kwong, O.Y., Wong W. L., and Lai, Tom. (2005). "Sentiment and Content Analysis of Chinese News Coverage." *International Journal of Computer Processing of Oriental Languages*, 18(2):171-183.
- Tsou, Benjamin, and Kwong, O.Y. (2006). "Toward a Pan-Chinese Thesaurus." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.
- Tsou, Benjamin, Lai Tom, and Chow, K. P. (2004). "Comparing Entropies Within the Chinese Language." *Proceedings of the 1st International Joint Conference on Natural Language Processing IJCNLP (LNAI)*, pp.475-481, Sanya, 2004. Also reprinted in *Lecture Notes in Artificial Intelligence*, edited by K. Y. Su, J. Tsujii, J. H. Lee, O. Y. Kwong, Vol. 3248, (2005), pp.466-475.

# References

- Tsou, Benjamin, Yuen Raymond, Kwong O.Y., Lai, Tom, and Wong, W. L. (2005).  
“Polarity Classification of Celebrity Coverage in the Chinese Press.”  
*Proceedings of 2005 International Conference on Intelligence Analysis*,  
Virginia
- Yuen, Raymond, Chan, Terence, Lai, Tom, Kwong, O.Y., and Tsou Benjamin.  
(2004). “Morpheme-based Derivation of Bipolar Semantic Orientation of  
Chinese Words.” *Proceedings of 20th International Conference on Chinese  
Linguistics (COLING)*, pp.1008-1014.

# References

- 孫茂松, 鄒嘉彥. (2001). 漢語自動分詞研究評述. 《當代語言學》, 1(3):23-32.
- 鄒嘉彥, 馮良珍. (2000). 漢語(五地)與日語新概念詞語對比研究 – 從新聞視窗看詞彙衍生與重整, 《語言研究》, 第3期, pp.51-70.
- 鄒嘉彥, 黎邦洋. (2003). 漢語共時語料庫與資訊開發 (Chinese synchronous corpus and data mining). 《中文資訊處理若干重要問題》 (Critical Issues in Chinese Information Processing). pp.147-165. 北京: 科學出版社.
- 鄒嘉彥, 游汝杰. (2003). 當代漢語新詞的多元化趨向和地區競爭. 《語言教學與研究》, 第2期.
- 鄒嘉彥, 游汝杰. (2007). 《21世紀華語新詞語詞典》 (*A 21st Century Dictionary of Chinese New Words*). 上海: 復旦大學出版社.
- 鄒嘉彥, 游汝杰. (2010). 《全球華語新詞語詞典》. 北京: 商務印書館.