LANGUAGE AND LINGUISTICS 9.2:221-238, 2008
2008-0-009-002-000259-1

# 語料庫爲本的兩岸對應詞彙發掘

洪嘉馡　　黃居仁
國立台灣大學　　中央研究院

近年來，語言學界對於漢語詞彙的研究，不論在語音、語義或語用上的分析，發現兩岸對使用漢語時的詞彙差異越來越顯著。這些差異無疑造成了知識與信息交流的障礙。而兩岸卻又的確是使用漢字體系的書寫系統，只有字形上有可預測的規律性對應。本文在共同文字系統的基礎上，以兩岸詞彙對比的特性，來探討一些與詞彙語義相關的基本問題。

本文由語料庫爲出發點，探索兩岸對於漢語詞彙在使用上的差異現象，例如：相關共現詞彙 (collocation) 的差異、較容易與台灣詞彙共同出現或與大陸詞彙共同出現的差異、特定語境下的特殊用法的差異、語言使用習慣的差異等等。並由這些分析中建立從語料庫中抽取兩岸對應詞彙的研究方法。

關鍵詞：GigaWord Corpus，Chinese Word Sketch，兩岸詞彙對應，共現詞彙

## 1. 前言

兩岸詞彙差異的問題，在一般兩岸人民的交流時，早已呈現出許多無法溝通、理解困難，或者是張冠李戴等狀況。探討兩岸詞彙的差異性，不僅讓大量使用詞彙的記者們，感受到兩者的差異（如：華夏經緯網 2004，南京語言文字網 2004，廈門日報 2004），亦成爲漢語詞彙學與詞彙語義學上研究的重要課題（如：竺家寧 1995，王鐵昆、李行健 1996，姚榮松 1997，Hong & Huang 2006 等等）。以往，不論語言學學者或文字工作者注意到這個問題時，僅能就所觀察到特定詞彙的局部對應，來提出分析與解釋，缺乏全面系統性的研究。

本文所提出的研究方法將有兩個不同以往對於兩岸詞彙分析的重要創新：第一、是能由大量語料中自動發掘新的詞彙差異；第二、是同時可以取出差異詞彙在語法與語義上的對比。

本文將以 GigaWord Corpus 爲主要語料來源，而以 Chinese Word Sketch（CWS，中文詞彙速描）爲工具 (Huang et al. 2005)，並參考 Hong & Huang (2006)

對於兩岸詞彙對比的研究分析，來探討兩岸詞彙的差異。主要研究的問題爲：(1) 分析兩岸詞彙特有的共現詞彙與句型；(2) 比對後特有共現詞彙中，抽取出新的兩岸對比詞彙；(3) 比對後特有的詞彙中，也可抽取出只與大陸與台灣詞彙搭配、共同出現或有特殊用法的詞彙；(4) 討論說明兩岸詞彙差異的對比與分布情形。

## 2. 研究動機與目的

自兩岸交流日趨頻繁之後，知識與信息交流的障礙，莫過於兩岸詞彙使用的差異。等同的詞形，卻代表不同的詞義；或相同的語義，卻有兩種不同的表達詞彙。這種問題，已經讓許多文字工作者費盡心思，試圖來解決這樣的窘境；而語言學者對於這種現象，也試圖從語音、語義、語用等方面著手，希望從各種與語言相關的角度，來探究兩岸詞彙的差異。事實上，對於兩岸詞彙差異的研究，到目前爲止，都僅限於局部的觀察與描述，而缺乏全面系統性分析，更遑論理論性的架構與解釋了。

在研究議題上，光是觀察到兩岸選擇以不同的詞形來代表相同的語義，如下述兩例，是不夠的。在詞彙語義學研究上，我們必須進一步追究，這些對比的動機，對比與語言的詞彙與詞義演變的動力是否相關，對比有無系統性的解釋等。

(1) 台灣的「煞」、大陸的「非典」
（「Sars (SEVERE ACUTE RESPIRATORY SYNDROME)」「嚴重急性呼吸道綜合症」的翻譯）
(2) 台灣的「計程車」、大陸的「出租車」

如要追究動機與解釋等理論架構問題，當然不能只靠少數觀察到的例子，而必須建立在語料的全面深入分析上。本文試圖建立以語料庫爲本，客觀發掘兩岸對應詞彙的研究方法。

## 3. 文獻探討

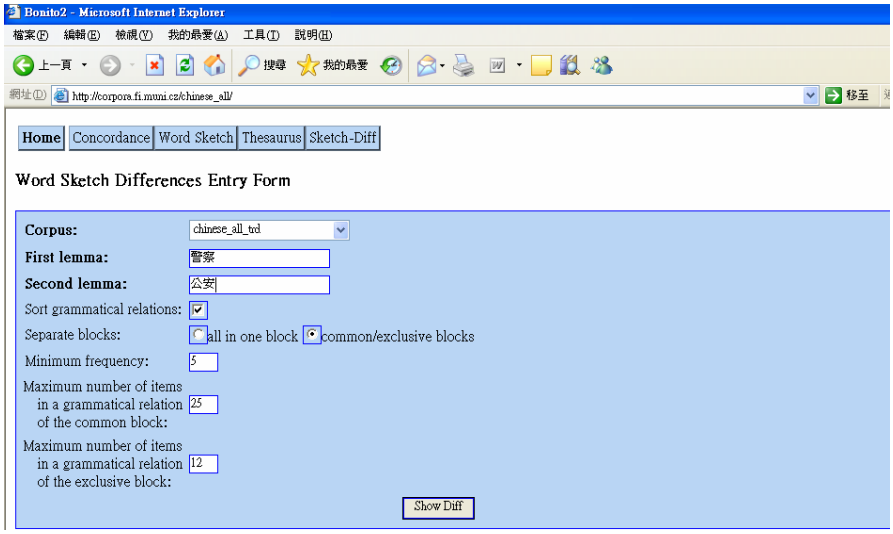
對於兩岸詞彙對比的探討，過去的研究，多半著重在語言特徵的區別。如列舉語音方面、詞彙方面的對比（南京語言文字網 2004）；或以語音、詞彙、語法

及表達方式等方面來分析語言差異的現象（許斐絢 1999，姚榮松 1997，王鐵昆、李行健 1996，竺家寧 1995 等）；比較新的研究方法，則是以 WordNet 爲基礎，取兩岸語料庫資料作比較，進而分析兩岸詞彙的對比（如：Hong & Huang 2006）。

## 4. 研究方法

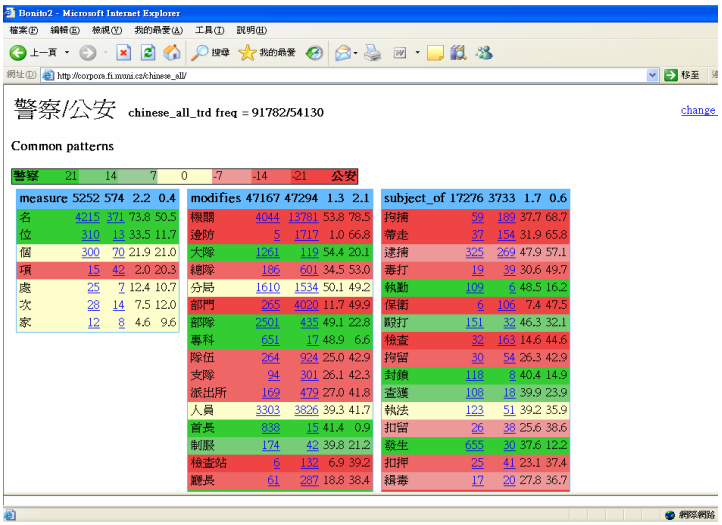
本文的研究出發點，是以約 11 億字的 Chinese GigaWord Corpus 爲主要語料來源，以中文詞彙速描 (Chinese Word Sketch) 爲搜尋語料工具。Chinese GigaWord Corpus 包含了分別來自兩岸的大量語料，包括近 4 億字來自大陸新華社的資料，及 7 億餘字來自台灣中央社的資料。因此，可以提供兩岸詞彙差異的大量詞彙證據。

中文詞彙速描則提供了以語料庫爲本，詞彙差異比較的工具。可以看出兩岸對於同一概念而使用不同詞彙的實際狀況與分布，也可以看出同一語義詞彙在兩岸的實際語料中，所呈現的相同點與差異性。我們主要利用中文詞彙速描中詞彙速描差異 (word sketch difference) 的功能。在此功能下，可以同時輸入兩岸個別使用的詞彙，以探究兩詞彙的使用狀況與分布，如：相關共現詞彙、語法功能、獨特搭配詞彙等等，進而了解兩岸詞彙的實際現象，以進行本研究的分析。詞彙速描對比的實際介面畫面如〈圖 1〉：

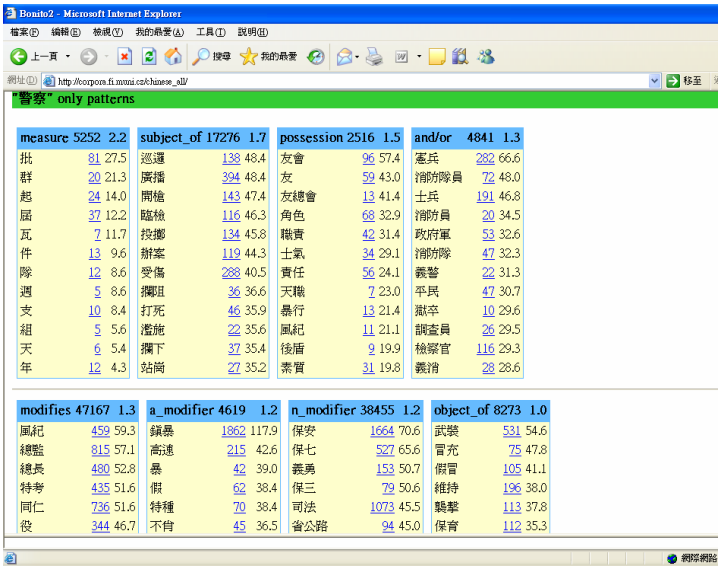



〈圖 1〉中文詞彙速描 Engine 下的詞彙速描對比

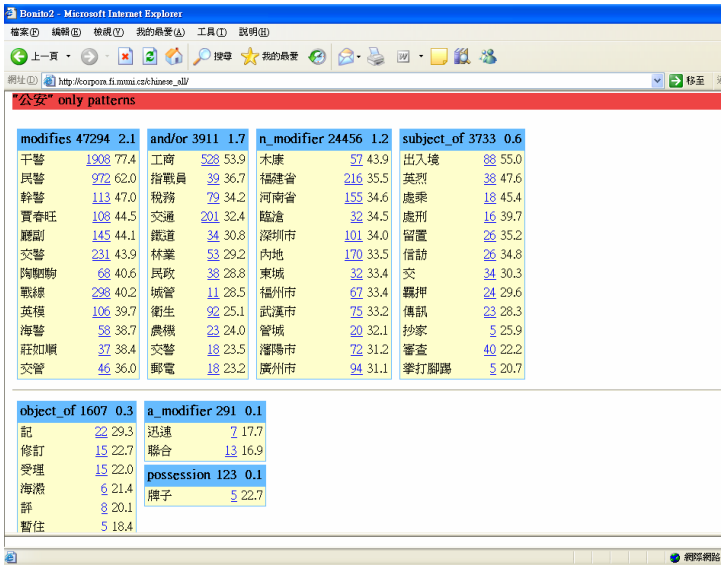
本文中，我們提出的方法，是由事先設定的一組已知對比詞彙出發，再利用上述介紹的「詞彙速描差異功能」進行兩詞彙的對比。藉由〈圖 1〉所呈現的對比功能，設定相關共現詞彙的最小詞頻爲 5 的條件下，進行兩岸詞彙的比較，可取得的詞彙對比資料，呈現顯示如〈圖 2〉：




〈圖 2〉「詞彙速描差異」呈現「警察/公安」兩詞彙的比較




〈圖 3〉「詞彙速描差異」呈現「警察」的使用分布




〈圖 4〉「詞彙速描差異」呈現「公安」的使用分布

藉由「詞彙速描差異」所提供的訊息，即可以從顯示的數據與相關共現詞彙裡，發掘出兩岸詞彙在實際語言中，使用上的種種不同現象。

# 5. 語料分析

## 5.1 語法功能的句式搭配

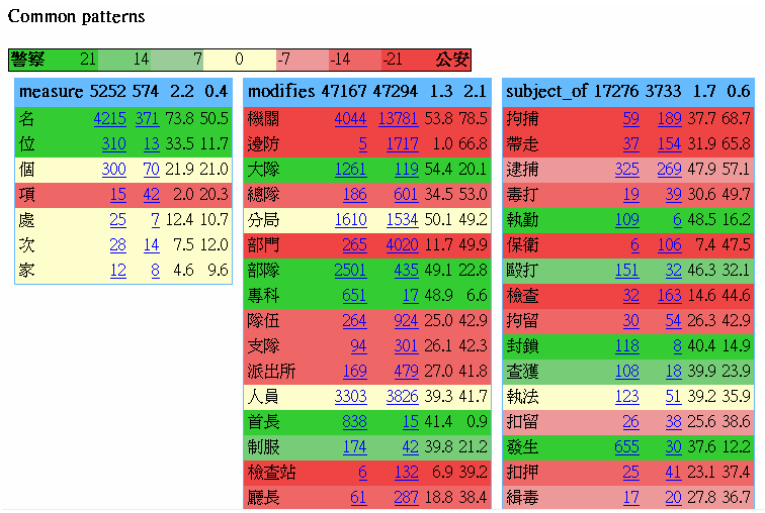
本文將以 GigaWord Corpus 的資料爲語料，並依照中文詞彙速描下的詞彙速描差異所呈現的訊息進行兩岸詞彙比對。假設是，已知兩岸對比詞彙的共現相關詞彙 (collocation) 中，如果產生對比差異，也能是兩岸對比的詞彙。依此分析原則，本文選定概念相同的名詞詞彙組，如：「警察：公安」；以及概念相同的動詞詞彙組，如：「做：搞」，進行分析、探討。

本文分析的研究方法具有不同以往的重要創新，原因在於：第一、是由一個具大量語料，且同時擁有大陸與台灣的語料，藉此自動發掘新的詞彙差異；第二、在發掘新的詞彙差異之際，又同時可以取出差異詞彙在語法與語義上的對比。

根據中文詞彙速描的語料，可以看出，兩岸詞彙在名詞詞彙組的對應中，詞彙的出現頻率、構詞共現的共同排列，大致上可呈現的語法功能 (grammatical

function) 的句式搭配有 measure、and/or、modifier、subject、object、possession 等，兩岸詞彙的構詞共現頻率 (frequency) 及其顯著率 (salience)；同樣地，在動詞詞彙組的對應中，除了有詞彙的詞頻，另外也有句式搭配的語法功能：subject、object、modifier 等，以及構詞共現頻率及其顯著率的分布數據。

中文詞彙速描查詢兩岸詞彙的差異，已知「警察/公安」爲兩岸對於同一概念所使用不同的兩詞彙，以此爲名詞詞彙組的對應，「警察」，出現的頻率爲 91,782 次；「公安」，出現的頻率爲 54,130 次，在兩對比詞彙的通用模組 (common patterns) 中，兩詞彙搭配各種不同的語法功能，其分布情形如下：




〈圖 5〉「警察/公安」以語法功能爲主的相關共現詞彙

從〈圖 5〉的分析結果得知，「警察/公安」這組兩岸對比詞彙，在通用模組中，出現的語法功能有：

(3) 「警察」較常搭配的：measure、modifier、subject_of、and/or、possession、a_modifier、object_of、possessor。
(4) 「公安」較常搭配的：measure、modifier、subject_of、and/or、n_modifier、object_of、possessor。
(5) 兩詞彙皆常使用的：measure、modifier、subject_of、and/or、a_modifier、object_of、possessor。

又在詞彙速描差異的比對結果中，呈現的另一個訊息是只與個別詞彙搭配、共同出現的模組 (only patterns)，亦傳達出搭配各種不同語法功能的分布，如下：

"警察" only patterns

| measure 5252 2.2 | | subject_of 17276 1.7 | | possession 2516 1.5 | | and/or 4841 1.3 | |
|---|---|---|---|---|---|---|---|
| 批 | 81 27.5 | 巡邏 | 138 48.4 | 友會 | 96 57.4 | 憲兵 | 282 66.6 |
| 群 | 20 21.3 | 廣播 | 394 48.4 | 友 | 59 43.0 | 消防隊員 | 72 48.0 |
| 起 | 24 14.0 | 開槍 | 143 47.4 | 友總會 | 13 41.4 | 士兵 | 191 46.8 |
| 屆 | 37 12.2 | 臨檢 | 116 46.3 | 角色 | 68 32.9 | 消防員 | 20 34.5 |
| 瓦 | 7 11.7 | 投擲 | 134 45.8 | 職責 | 42 31.4 | 政府軍 | 53 32.6 |
| 件 | 13 9.6 | 辦案 | 119 44.3 | 士氣 | 34 29.1 | 消防隊 | 47 32.3 |
| 隊 | 12 8.6 | 受傷 | 288 40.5 | 責任 | 56 24.1 | 義警 | 22 31.3 |
| 週 | 5 8.6 | 攔阻 | 36 36.6 | 天職 | 7 23.0 | 平民 | 47 30.7 |
| 支 | 10 8.4 | 打死 | 46 35.9 | 暴行 | 13 21.4 | 獄卒 | 10 29.6 |
| 組 | 5 5.6 | 濫施 | 22 35.6 | 風紀 | 11 21.1 | 調查員 | 26 29.5 |
| 天 | 6 5.4 | 攔下 | 37 35.4 | 後盾 | 9 19.9 | 檢察官 | 116 29.3 |
| 年 | 12 4.3 | 站崗 | 27 35.2 | 素質 | 31 19.8 | 義消 | 28 28.6 |

〈圖 6〉在語料中，只與「警察」搭配出現的語法功能之分布

(6) 在語料中，只與「警察」搭配出現的模組中，呈現的語法功能爲：measure、subject_of、possession、and/or、modifier、a_modifier、n_modifier、object_of、possessor。

"公安" only patterns

| modifies 47294 2.1 | | and/or 3911 1.7 | | n_modifier 24456 1.2 | | subject_of 3733 0.6 | |
|---|---|---|---|---|---|---|---|
| 干警 | 1908 77.4 | 工商 | 528 53.9 | 木康 | 57 43.9 | 出入境 | 88 55.0 |
| 民警 | 972 62.0 | 指戰員 | 39 36.7 | 福建省 | 216 35.5 | 英烈 | 38 47.6 |
| 幹警 | 113 47.0 | 稅務 | 79 34.2 | 河南省 | 155 34.6 | 處乘 | 18 45.4 |
| 賈春旺 | 108 44.5 | 交通 | 201 32.4 | 臨滄 | 32 34.5 | 處刑 | 16 39.7 |
| 廳副 | 145 44.1 | 鐵道 | 34 30.8 | 深圳市 | 101 34.0 | 留置 | 26 35.2 |
| 交警 | 231 43.9 | 林業 | 53 29.2 | 內地 | 170 33.5 | 信訪 | 26 34.8 |
| 陶駟駒 | 68 40.6 | 民政 | 38 28.8 | 東城 | 32 33.4 | 交 | 34 30.3 |
| 戰線 | 298 40.2 | 城管 | 11 28.5 | 福州市 | 67 33.4 | 羈押 | 24 29.6 |
| 英模 | 106 39.7 | 衛生 | 92 25.1 | 武漢市 | 75 33.2 | 傳訊 | 23 28.3 |
| 海警 | 58 38.7 | 農機 | 23 24.0 | 管城 | 20 32.1 | 抄家 | 5 25.9 |
| 莊如順 | 37 38.4 | 交警 | 18 23.5 | 濯陽市 | 72 31.2 | 審查 | 40 22.2 |
| 交管 | 46 36.0 | 郵電 | 18 23.2 | 廣州市 | 94 31.1 | 拳打腳踢 | 5 20.7 |

〈圖 7〉在語料中，只與「公安」搭配出現的語法功能之分布

(7) 在語料中，只與「公安」搭配出現的模組中，呈現的語法功能爲：modifier、and/or、n_modifier、subject_of、object_of、a_modifier、possession。

同樣的分析方法，我們取一組已知兩岸詞彙比對的動詞詞組——「做/搞」進行比較，試圖探討兩詞彙間的一些關係。我們一樣藉由「詞彙速描差異」的對比功能進行兩岸詞彙的對比，得知「做」出現的頻率爲 237,908 次；「搞」出現的頻率爲 38,543 次，其相關語法功能訊息如下所示：

Common patterns

| 做 | 21 | 14 | 7 | 0 | -7 | -14 | -21 | 搞 |
|---|---|---|---|---|---|---|---|---|

| object | 71929 | 13307 | 4.6 | 6.4 |
|---|---|---|---|---|
| 事 | 9073 | 38 | 77.9 | 11.3 |
| 生意 | 561 | 18 | 46.5 | 14.2 |
| 廣告 | 760 | 5 | 37.7 | 1.6 |
| 實驗 | 445 | 20 | 33.9 | 10.7 |
| 工作 | 3353 | 92 | 31.1 | 4.9 |
| 調研 | 12 | 48 | 8.4 | 29.9 |
| 試點 | 15 | 84 | 2.6 | 26.0 |
| 試驗 | 241 | 79 | 25.2 | 23.8 |
| 宣傳 | 306 | 53 | 23.6 | 16.0 |
| 目的 | 257 | 5 | 22.7 | 1.5 |
| 環保 | 347 | 22 | 22.2 | 7.1 |
| 勞務 | 6 | 39 | 1.5 | 21.7 |
| 動作 | 147 | 9 | 21.3 | 6.1 |
| 政治 | 122 | 269 | 0.7 | 20.8 |
| 原子彈 | 7 | 16 | 7.3 | 20.2 |
| 時候 | 110 | 8 | 19.6 | 6.2 |

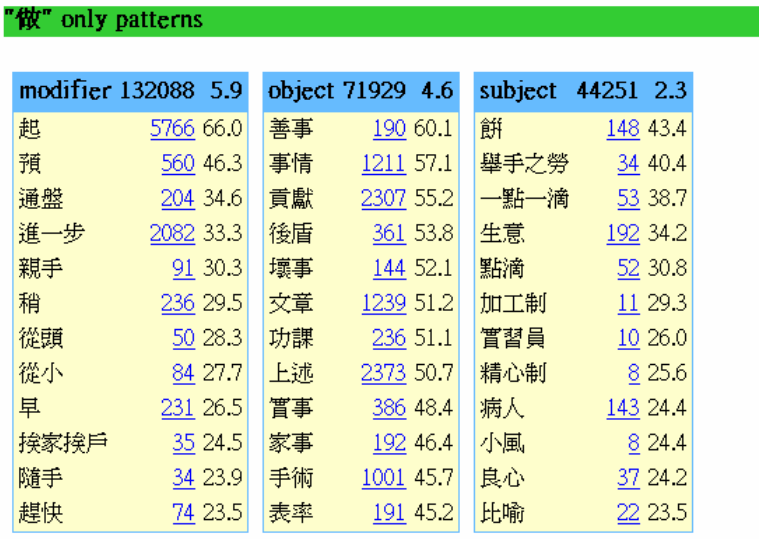
| modifier | 132088 | 13367 | 5.9 | 4.5 |
|---|---|---|---|---|
| 這麼 | 3787 | 29 | 71.1 | 16.7 |
| 怎麼 | 1459 | 100 | 60.4 | 35.7 |
| 去 | 4019 | 241 | 57.3 | 33.7 |
| 所 | 14068 | 32 | 56.9 | 1.7 |
| 該 | 1736 | 5 | 54.8 | 4.0 |
| 不 | 6081 | 4235 | 29.4 | 51.4 |
| 多 | 3730 | 167 | 47.6 | 23.6 |
| 些 | 597 | 11 | 45.5 | 11.6 |
| 過 | 3428 | 205 | 43.4 | 24.3 |
| 再 | 4963 | 329 | 43.2 | 25.9 |
| 不能 | 815 | 596 | 26.2 | 43.2 |
| 了 | 24005 | 1111 | 41.8 | 21.2 |
| 要 | 7008 | 869 | 41.2 | 33.5 |
| 不准 | 21 | 97 | 8.0 | 37.2 |
| 不要 | 777 | 253 | 32.6 | 36.5 |
| 應該 | 1248 | 24 | 35.7 | 8.4 |

| subject | 44251 | 6286 | 2.3 | 2.5 |
|---|---|---|---|---|
| 精力 | 11 | 80 | 7.2 | 37.7 |
| 事 | 583 | 11 | 33.7 | 5.9 |
| 方面 | 1585 | 30 | 32.3 | 5.1 |
| 工作 | 2343 | 272 | 31.2 | 20.9 |
| 事情 | 175 | 34 | 30.8 | 22.8 |
| 農民 | 80 | 171 | 5.7 | 26.8 |
| 全民 | 35 | 69 | 7.4 | 24.9 |
| 才能 | 360 | 10 | 24.5 | 4.3 |
| 人 | 860 | 286 | 18.6 | 22.5 |
| 企業 | 233 | 287 | 4.0 | 21.8 |
| 勢力 | 8 | 41 | 1.2 | 21.4 |
| 資金 | 60 | 114 | 3.1 | 21.3 |
| 心思 | 5 | 12 | 7.2 | 20.8 |
| 錢 | 130 | 34 | 20.6 | 18.2 |
| 農戶 | 6 | 24 | 2.7 | 20.0 |
| 家家戶戶 | 6 | 9 | 9.8 | 19.4 |

〈圖 8〉「做/搞」以語法功能爲主的相關共現詞彙

從〈圖 8〉的分析結果顯示，得知「做/搞」的搭配語法功能爲：

(8) 「做」較常搭配的：object、modifier、subject。
(9) 「搞」較常搭配的：object、modifier、subject。
(10) 兩詞彙皆常使用的：object、modifier、subject。

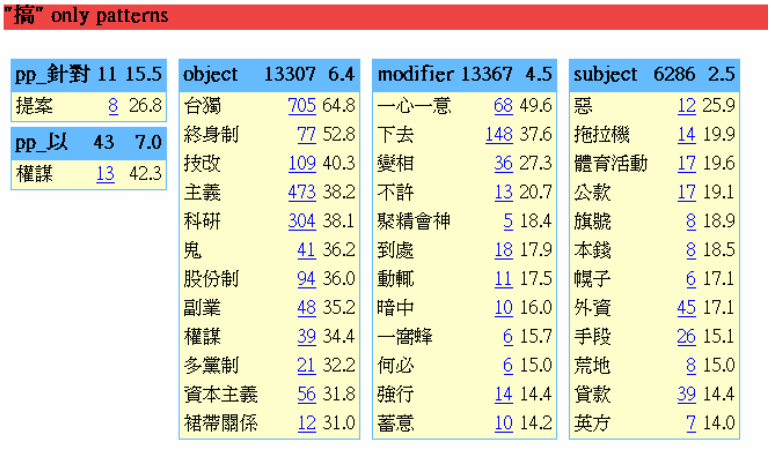


"做" only patterns

| modifier 132088 5.9 | | object 71929 4.6 | | subject 44251 2.3 | |
|---|---|---|---|---|---|
| 起 | 5766 66.0 | 善事 | 190 60.1 | 餅 | 148 43.4 |
| 預 | 560 46.3 | 事情 | 1211 57.1 | 舉手之勞 | 34 40.4 |
| 通盤 | 204 34.6 | 貢獻 | 2307 55.2 | 一點一滴 | 53 38.7 |
| 進一步 | 2082 33.3 | 後盾 | 361 53.8 | 生意 | 192 34.2 |
| 親手 | 91 30.3 | 壞事 | 144 52.1 | 點滴 | 52 30.8 |
| 稍 | 236 29.5 | 文章 | 1239 51.2 | 加工制 | 11 29.3 |
| 從頭 | 50 28.3 | 功課 | 236 51.1 | 實習員 | 10 26.0 |
| 從小 | 84 27.7 | 上述 | 2373 50.7 | 精心制 | 8 25.6 |
| 早 | 231 26.5 | 實事 | 386 48.4 | 病人 | 143 24.4 |
| 挨家挨戶 | 35 24.5 | 家事 | 192 46.4 | 小風 | 8 24.4 |
| 隨手 | 34 23.9 | 手術 | 1001 45.7 | 良心 | 37 24.2 |
| 趕快 | 74 23.5 | 表率 | 191 45.2 | 比喻 | 22 23.5 |

〈圖 9〉在語料中，只與「做」搭配的語法功能之分布

(11) 在語料中，只與「做」搭配、共同出現的模組中，呈現的語法功能爲：modifier、object、subject。



"搞" only patterns

| pp_針對 11 15.5 | | object 13307 6.4 | | modifier 13367 4.5 | | subject 6286 2.5 | |
|---|---|---|---|---|---|---|---|
| 提案 | 8 26.8 | 台獨 | 705 64.8 | 一心一意 | 68 49.6 | 惡 | 12 25.9 |
| **pp_以 43 7.0** | | 終身制 | 77 52.8 | 下去 | 148 37.6 | 拖拉機 | 14 19.9 |
| 權謀 | 13 42.3 | 技改 | 109 40.3 | 變相 | 36 27.3 | 體育活動 | 17 19.6 |
| | | 主義 | 473 38.2 | 不許 | 13 20.7 | 公款 | 17 19.1 |
| | | 科研 | 304 38.1 | 聚精會神 | 5 18.4 | 旗號 | 8 18.9 |
| | | 鬼 | 41 36.2 | 到處 | 18 17.9 | 本錢 | 8 18.5 |
| | | 股份制 | 94 36.0 | 動輒 | 11 17.5 | 幌子 | 6 17.1 |
| | | 副業 | 48 35.2 | 暗中 | 10 16.0 | 外資 | 45 17.1 |
| | | 權謀 | 39 34.4 | 一窩蜂 | 6 15.7 | 手段 | 26 15.1 |
| | | 多黨制 | 21 32.2 | 何必 | 6 15.0 | 荒地 | 8 15.0 |
| | | 資本主義 | 56 31.8 | 強行 | 14 14.4 | 貸款 | 39 14.4 |
| | | 裙帶關係 | 12 31.0 | 蓄意 | 10 14.2 | 英方 | 7 14.0 |

〈圖 10〉在語料中，只與「搞」搭配的語法功能之分布

(12) 在語料中，只與「搞」搭配、共同出現的模組中，呈現的語法功能爲：pp_針對、pp_以、object、modifier、subject。

從〈圖 5〉至〈圖 10〉，不論是名詞詞彙組的對比，還是動詞詞彙組的對比，可以很清楚發現，兩岸對比詞彙在實際語料的使用上，搭配各種不同語法功能的狀況。

而從分析的結果顯示，在通用模組中，雖然有相同搭配的語法功能，但仍舊有比例偏重的現象，例如：名詞組——「警察/公安」的對比，搭配「measure」的語法功能，以 2：1 顯示「警察」多過於「公安」；搭配「n_modifier」的語法功能中，則以 16：9 的比例說明「公安」佔的比例大過於「警察」。動詞組——「做/搞」的對比裡，搭配「object」的語法功能，以 15：6 呈現「做」的比例多於「搞」；而「subject」的語法功能搭配裡，則以 15：7 說明「搞」所佔的比例大於「做」。

這樣的數據說明，特定的詞彙在特定的語義下，可能在實際語言使用上，會偏重搭配某一類語法功能的詞彙，然而，「警察」與「公安」這兩個屬於概念相同的詞彙，卻在實際搭配語法功能的使用情形，有著不盡相同的微妙現象。理論上，如果是相同的語義，不同的詞彙，其搭語法功能的詞彙與使用情形，應該都是相同的，但是，我們在兩個來自兩岸對比的詞彙——「警察」和「公安」，發現了彼此差異的端倪。這也說明了，兩岸對於相同概念的表達，除了會以不同詞彙來呈現之外，其搭語法功能不同，更傳達了語言在使用習慣上的差異性。

另外，值得一提的是，在語料中，儘管這些語料呈現，可以清楚了解大陸和台灣在使用詞彙的共同性與個別性，但是，在這個語料裡，仍然有一些不符合語法功能的用法，例如〈圖 8〉至〈圖 10〉，語法功能是「subject」的詞彙分別有：精力、方面、一點一滴、手段等，就是違反語法規則的證明。這些結果，確實存在於這個語料庫中，是我們在使用這個語料進行研究時，需要注意、避免使用錯誤的地方。

## 5.2 相關共現詞彙的搭配

藉由中文詞彙速描 Engine 中「詞彙速描差異」的功能，我們除了釐清詞彙的搭配語法功能，更可以清楚看到在不同語法功能中所搭配的相關共現詞彙。分析結果，分爲兩大部分：其一是通用模組 (common patterns) 的共現詞彙、其二是在語料庫中，只與個別詞彙搭配、共同出現的模組 (“X” only patterns)。在通用模組裡，會有比較的兩詞彙皆常搭配的相關共現詞彙、有個別詞彙較常搭配的共現詞彙；而在只與個別詞彙搭配、共同出現的模組裡，則分別將個別詞彙，依著不同的搭配語法功能，以呈現相關共現詞彙。以名詞詞彙組「警察」和「公安」爲

例，其各種共現詞彙及語法功能的搭配使用，透過「詞彙速描比對」的功能進行比對，其分布情形及實際語料，如下所示：

〈表 1〉「警察/公安」皆常使用的相關共現詞彙

| | 「警察」較常搭配 | 兩詞彙皆常搭配 | 「公安」較常搭配 |
|---|---|---|---|
| measure | 名、位 | 個、次、處、家 | 項 |
| modifier | 部隊、大隊、學校… | 人員、分局 | 機關、邊防、支隊… |
| subject_of | 執勤、取締、封鎖… | 執法、嚴密、盤問 | 保衛、拘押、拘捕… |
| and/or | 軍隊、軍人、部隊… | 司法、地方、駐軍… | 武警、官兵、邊防… |
| possession | 形象 | | |
| n_modifier | 轄區、便衣、模範… | | 中共、級、上級… |
| a_modifier | 優秀、消防 | 基層、新、結合… | 老、原 |
| object_of | 成立、協助、樹立… | 遭到、會同、加強、受到… | |
| possessor | 制服 | 地、工作、城市 | |

(13) 內政部長率同有關人員，就軍火管制與<警察>執勤(subject_of)業務到會報告，並備質詢。

(14) 在去年的「人民警察愛人民」系列活動中，這家派出所把愛民行動具體落實到<公安>保衛(subject_of)工作中去。

(15) a. 警政署已派員秉公調查，現場如確有<警察>執法(subject_of)不當，警政署絕不姑息。

b. 湖北省公安廳最近派出十餘支警務督察隊，分赴全省各地對<公安>執法(subject_of)進行現場督察。

〈表 2〉在語料中，只與「警察」搭配的相關共現詞彙

| | “警察” only patterns |
|---|---|
| measure | 批、屆、隊、群、週… |
| subject_of | 廣播、受傷、開槍、臨檢、巡邏… |
| possession | 友會、角色、職責、素質、風紀… |
| and/or | 憲兵、義消、義警、檢察官、調查員… |
| modifier | 人事權、同仁、總監、總長、特考… |
| a_modifier | 鎮暴、績優、專業、特種、高速… |
| n_modifier | 保七、秘密、刑事、司法、警政署… |
| object_of | 保育、設置、假扮、攻擊、殺害… |
| possessor | 勤務、治安、防彈衣、衝鋒鎗… |

(16) 郝柏村下午五時在台北賓館接見來訪的約旦王國阿不都拉親王後，驅車返回行政院，抵達院區時，他看到<警察>及憲兵(and/or)在炎熱的氣溫中認真執勤。

〈表 3〉在語料中，只與「公安」搭配的相關共現詞彙

| | “公安” only patterns |
|---|---|
| modifier | 干警、幹警、廳副、戰線、民警… |
| and/or | 指戰員、城管、郵電… |
| n_modifier | 內地、管城、深圳、福建省、河南省… |
| subject_of | 英烈、留置、信訪、處乘、抄家… |
| object_of | 修訂、獻身、砸、評、海淀… |
| a_modifier | 迅速、聯合 |
| possession | 牌子 |

(17) 他還會見了到海南投資的港、澳、台人士和海外僑胞，看望了駐瓊部隊指戰員(and/or)、<公安>干警和武警官兵。

至於，在動詞詞彙組「做」和「搞」，一樣透過「詞彙速描差異」的功能進行兩詞彙的比對，所分析出來的實際使用分布及實際使用語料，如下所示：

〈表 4〉「做/搞」皆常使用的相關共現詞彙

| | 「做」較常搭配 | 兩詞彙皆常使用 | 「搞」較常搭配 |
|---|---|---|---|
| object | 事、工作、廣告… | 研究、試驗、運動、遊戲 | 政治、試點、調研… |
| modifier | 所、這麼、多、去… | 不要、越 | 不、不能、不准… |
| subject | 工作、方面、事… | 人、錢、地 | 企業、農民、精力… |

(18) 有的一般性消毒洗衣粉，<做>廣告(object)時卻吹噓有殺菌作用，誤導消費者該產品屬於消毒藥劑。

(19) 田風東接任廠長後，提出要依靠幹部職工發揚艱苦創業精神，眼睛向外<搞>市場調研(object)，來振興企業。

(20) a. 成功大學希望藉兩岸大學實質的學術合作，除設備較好的一方提供另一方<做>科學研究(object)，也希望大陸優秀的教授來台教學及與國內年輕的教授合作。

b. 勞倫斯勸他留下來<搞>研究(object)，並說每年至少拿八九萬美元工資，妻子和 3 個女兒也來「集體勸說」。

〈表 5〉在語料中，只與「做」搭配的相關共現詞彙

| | “做” only patterns |
|---|---|
| modifier | 起、進一步、預、通盤… |
| object | 上述、貢獻、文章、事情… |
| subject | 生意、餅、病人、一點一滴、舉手之勞… |

(21) 許文志強調，如果都市計畫細部計畫<做>通盤(modifier)檢討時，將容積率寫到說明書中，才會開始實施，否則不會實施。

〈表 6〉在語料中，只與「搞」搭配的相關共現詞彙

| | “搞” only patterns |
|---|---|
| pp_針對 | 提案 |
| pp_以 | 權謀 |
| object | 台獨、主義、科研、技改… |
| modifier | 下去、一心一意、變相、到處… |
| subject | 外資、貸款、手段、公款… |

(22) 各級政府和部門都不得以檢查驗收和評比等形式，變相(modifier)<搞>這類達標升級活動。

在上述對於兩岸對比詞彙的分析與實際例子的呈現，不難發現台灣使用的「警察」、「做」與大陸使用的「公安」、「搞」，雖然兩兩互爲概念相同的詞彙，但是經過比對後，其差異性即可清楚顯示出來。以〈表 4〉至〈表 6〉「做」與「搞」的比對爲例，同樣是搭配 object 語法功能的共現詞彙，兩岸都會有「做研究」或「搞研究」的用法，不過，如果是從事特定研究時，台灣與大陸的使用就會出現差異，其中，可探討出兩個現象：

(23) a. 大陸的用法：「搞」＋研究 → 說明一般性的研究
b. 大陸的用法：「搞」＋調研 → 強調特定目標的研究
(24) a. 台灣的用法：「做」＋科學研究 → 強調的重點是研究本身
b. 大陸的用法：「搞」＋科研 → 強調的重點在科學類

從這樣的實際語料現象，可以說明兩岸對於使用相同概念、不同詞形的詞彙時，所顯現出來的差異性以及個別的特殊性，也可以看出兩岸使用語言的文化差異性以及區域差異性。

## 5.3 比對後特有的詞彙

經過上述對於詞彙速描差異的對比，讓我們很清楚看到，兩岸使用語言獨特性，在「警察/公安」這組名詞組當中發現，與「警察」較常共現的詞彙，如：憲兵、軍人、執勤、臨檢……等等；與「公安」較常共現的詞彙，如：武警、指戰員、保衛、留置……等等。在「做/搞」這組名詞組當中發現，與「做」較常共現的詞彙，如：工作、生意、事情……等等；與「搞」較常共現的詞彙，如：試點、調研、台獨、外資……等等。

藉由事先確定的已知對比詞彙，來探討兩岸使用同一概念不同詞形的詞彙的同時，除了可以探究其所搭配的語法功能與相關共現詞彙之外，更可以從相關共現詞彙中，再發掘出兩岸在使用上的特有詞彙，以致於比對出更多兩岸詞彙的差異性與特殊性。

這個議題，是值得語言學界的重視，是值得對於探究兩岸語言的對比分析，也是值得對於探究兩岸詞彙的對比發掘。

## 5.4 兩岸詞彙差異的對比與分布

藉由中文詞彙速描 Engine 對於上述分析名詞詞彙組「警察/公安」的詞彙速描差異，可以得到各項簡單的歸納：

一、兩個詞彙特有的共現詞彙與句型，正顯示了兩岸語言的對比。
二、比對後特有共現詞彙中，可以抽取出新的兩岸對比詞彙，如：「憲兵/武警」。
三、比對後特有的詞彙中，也可抽取出大陸與台灣較常使用或有特殊用法的

詞彙，如：大陸的「邊防、民警、指戰員」，及台灣的「轄區、保七、警政署」。

四、詞彙速描差異也可看出兩岸語言使用習慣的差異，如：台灣的「警察」，最常用的動詞是「執勤」，是用於描述他的工作狀態；而大陸的「公安」，最常用的動詞是「保衛（人民）」，則是揭櫫其目標。

同樣的，在動詞詞彙組「做/搞」的詞彙速描差異，同樣地可以得到一些清楚的說明：

一、比對後特有共現詞彙中，可以抽取出新的兩岸對比詞彙，如：「實驗/試點」。

二、比對後特有的詞彙中，也可抽取出大陸與台灣獨用或有特殊用法的詞彙，如：大陸的「試點、調研、勞務、拖拉機」，及台灣的「民眾、選民、環保、決議」。

三、詞彙速描差異也可看出兩岸語言使用習慣的差異，如：台灣的「做」，最常用的名詞是「事」，是用於說明一個事件的進行；而大陸的「搞」，最常用的名詞是「政治」，則是點明了其國家性與文化性的終極目標。

## 6. 結論

經由以上分析、統計的數據呈現，我們可以很清楚得知，兩岸詞彙在使用上的差異，以及在詞彙搭配使用的異同。從對比詞彙的特有共現詞彙中，我們更可以進一步發掘出新的兩岸對比詞彙。

總而言之，本文一方面區分並釐清兩岸詞彙的個別語義架構，一方面再就兩者的語義概念了解其搭配語法功能中的共現詞彙，進而增加我們對於漢語詞彙語義系統性演變脈絡的理解。

# 引用文獻

Hong, Jia-Fei, and Chu-Ren Huang. 2006. WordNet based comparison of language variation: a study based on CCD and CWN. Paper presented at Global WordNet (GWC-06). Jeju Island, Korea.

Huang, Chu-Ren, Adam Kilgarriff, Yicing Wu, Chih-Min Chiu, Simon Smith, Pavel Rychlý, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese sketch engine and the extraction of collocations. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 48-55. Jeju Island, Korea.

Kilgarriff, Adam. 1996a. Which words are particularly characteristic of a text? A survey of statistical approaches. *Proceedings of AISB'96 Workshop on Language Engineering for Document Analysis and Recognition*, 33-40. Brighton: Nottingham Trent University.

Kilgarriff, Adam. 1996b. Why chi-square doesn't work, and an improved LOB-Brown comparison. *Proceedings of ALLC-ACH'96*, 169-172. Bergen: University of Bergen.

Kilgarriff, Adam, Chu-Ren Huang, Michael Rundell, Pavel Rychly, Simon Smith, David Tugwell, and Elaine Uí Dhonnchadha. 2005. Word sketches for Irish and Chinese. Paper presented at Corpus Linguistics 2005. Birmingham, UK.

Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, and David Tugwell. 2005. Chinese Word Sketches. Paper presented at ASIALEX 2005: Words in Asian Cultural Context. Singapore.

王鐵昆, 李行健. 1996.〈兩岸詞彙比較研究管見〉,《華文世界》81:24-31。

竺家寧. 1995.〈論兩岸詞彙的比較與詞典的編纂〉,收錄於《第一屆兩岸漢語語彙文字學術研討會論文專集》,23-30。台北：中華語文研習所。

姚榮松. 1997.〈論兩岸詞彙差異中的反向拉力〉,第五屆世界華語文教學研討會論文。台北：劍潭青年活動中心。

徐丹暉. 1995.〈兩岸詞語差異之比較〉,收錄於《第一屆兩岸漢語語彙文字學術研討會論文專集》,31-35。台北：中華語文研習所。

許斐絢. 1999.《台灣當代國語新詞探微》,國立台灣師範大學碩士論文。

許學仁. 1995.〈海峽兩岸新詞語中若干詞義衍生和規範的考察〉,收錄於《第一屆兩岸漢語語彙文字學術研討會論文專集》,228-238。台北：中華語文研習所。

曾榮汾. 1995.〈兩岸語言詞彙整理之我見〉,收錄於《第一屆兩岸漢語語彙文字學術研討會論文專集》,1-11。台北：中華語文研習所。

黎運漢. 1995.〈略論兩岸漢語語彙差異之文化因素〉，收錄於《第一屆兩岸漢語語彙文字學術研討會論文專集》，23-25。台北：中華語文研習所。
戴凱峰. 1996.《從語言學的觀點探討台灣與北京國語間之差異》，政治作戰學校碩士論文。

## 網路資源

GigaWord Corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09.
華夏經緯網. 2004.〈趣談海峽兩岸詞彙差異〉，http://www.huaxia.com/wh/zsc/00162895.html。
南京語言文字網. 2004.〈兩岸普通話大同中有小異〉，http://njyw.njenet.net.cn/news/shownews.asp?newsid=367。
廈門日報. 2004.〈趣談兩岸詞彙差異〉，http://www.csnn.com.cn/csnn0401/ca213433.htm。

[Received 8 January 2007; revised 28 September 2007; accepted 1 November 2007]

洪嘉馡
國立台灣大學語言學研究所
106 台北市羅斯福路四段 1 號
jiafei@gate.sinica.edu.tw

黃居仁
中央研究院語言學研究所
115 台北市研究院路二段 130 號
churen@gate.sinica.edu.tw

# A Corpus-Based Approach to the Discovery of Cross-Strait Lexical Contrasts

Jia-Fei Hong
*National Taiwan University*

Chu-Ren Huang
*Academia Sinica*

Studies of cross-strait lexical contrasts in the use of Mandarin Chinese reveal that a divergence has become increasingly evident. This divergence is apparent in phonological, semantic, and pragmatic analyses and has become an obstacle to knowledge-sharing and information exchange. Given the wide range of divergences, it seems that Chinese character forms offer the most reliable regular mapping between cross-strait usage contrasts. We propose a new approach to discovery of cross-strait contrasts in this paper anchored on the regular correspondences of characters.

Our approach is corpus-based and collocation-driven. We use known contrast pairs as seeds in a corpus containing data from both the PRC and Taiwan. Collocation patterns in terms of both lexical lists and grammatical functions of these contrast pairs are studied to semi-automatically discover additional contrast pairs. This approach obtains both NLP applicability and linguistic felicity since it yields both the contrast pairs as well as their usage contexts.

Key words: GigaWord Corpus, Chinese Word Sketch, cross-strait lexical contrasts, collocation