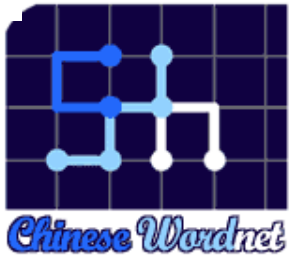



# 大數與求真

# 如何由十億字語料中學習語言知識

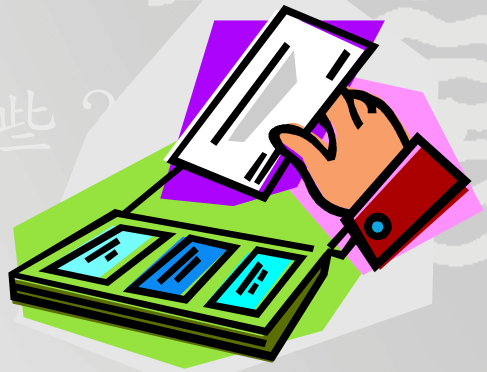
黃居仁

中央研究院語言學研究所

http://cwn.ling.sinica.edu.tw/huang/huang.htm

# 楔子：語言知識的規律化描述

## 「發票」怎麼用？



Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff | Frequency | Collocation

KWIC/Sentence | View options | Sample | Filter | Sort

Corpus: **chinese_all_trd**
Hits: **8408**
conc description

421 Go Next | Last

| | |
|---|---|
| CNA19910102.0102 | 民眾 可 蒐集 金額 新台幣 二千 元 以上 的 發票 ， 向 稅捐 單位 領取 紀念品 。 省 |
| CNA19910102.0102 | 活動 ， 除 可 繼續 提醒 民眾 保持 購物 索取 發票 的 習慣 外 ， 更 配合 年度 所得稅 的 申報 |
| CNA19910102.0102 | 宣導 將 舉辦 的 活動 包括 ： - 集 發票 兌換 紀念品 ， 凡 集滿 本 月份 發票 金額 |
| CNA19910102.0102 | - 集 發票 兌換 紀念品 ， 凡 集滿 本 月份 發票 金額 在 二千 元 以上 者 ， 可 憑 發票 於 |
| CNA19910102.0102 | 月份 發票 金額 在 二千 元 以上 者 ， 可 憑 發票 於 本 月 三十日 及 三十一日 向 各 稅捐處 |
| CNA19910102.0102 | 或 摸彩 活動 。 - 辦理 餐飲業 開立 發票 模範 商店 選拔 。 八 名 減刑 獲釋 泰人 |
| CNA19910104.0131 | 勒令 停業 一 週 。 稅捐處 已 將 漏開 發票 與 大額 欠稅戶 以 電腦 列管 ， 一 年 內 |
| CNA19910104.0131 | 欠稅戶 以 電腦 列管 ， 一 年 內 有 三 次 漏開 發票 紀錄 及 欠稅 超過 十萬 元 ， 都 將 依 規定 |
| CNA19910108.0104 | 八十年 元 月份 二 聯 式 統一發票 或 收銀機 發票 收執 聯 金額 達 二千 元 以上 者 ， 均 可 |
| CNA19910122.0254 | 將 是 一 大 損失 ， 商家 最好 誠實 開立 發票 ， 以免 因小失大 。 財政部 自 十五日 |
| CNA19910122.0254 | 凡 遭 查到 三 次 以上 未 誠實 開立 發票 的 商家 ， 將 遭 停業 處置 ， 王建（火宣 |
| CNA19910122.0254 | 一百萬 者 ， 在 誠實 開發票 前提 下 ， 使 發票 數額 提高 至 兩百萬 元 以上 ， 財政部 的 |
| CNA19910125.0259 | 進入 各 類 商店 購物 ， 如 當場 未 開立 發票 ， 即 以 「 現行犯 」 成爲 處罰 列管 對象 |
| CNA19910125.0259 | 包含 了 自己 所 繳 的 稅額 ， 如 不 索取 發票 ， 就 等於 被 不肖 商家 吃掉 ， 爲 維護 |
| CNA19910127.0102 | 打算 再 增加 四 人 。 上海 查獲 非法 發票 交易（中央社 台北 二十七日 電） 香港 |
| CNA19910127.0102 | 報導 ， 上海 工商 執法 人員 查獲 一 批 非法 發票 交易 市場 ， 當場 擒獲 多 名 票販 。 |
| CNA19910127.0102 | 用 三元 「 人民幣 」 買來 一 張 空白 非法 發票 ， 任意 填寫 金額 便 可 回去 報假賬；而 |

expand left

台省 各 縣市 舉辦 統一發票 宣導 活動（中央社 中興新村 二日 電）本 月份 購物 時 別 忘 了 索取 統一發票 。 臺灣省 各 縣市 正 舉辦 統一發票 宣導 ， 民眾 可 蒐集 金額 新台幣 二千 元 以上 的 發票 ， 向 稅捐 單位 領取 紀念品 。 省 稅務局 聯合 各 縣市 稅捐處 ， 將 在 農曆年 過 年 前 ， 共同 舉辦 多 項 統一發票 宣導 活動 ， 除 可 繼續 提醒 民眾 保持 購物 索取 發票 的 習慣 外

Bonito2 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://corpora.fi.muni.cz/chinese_all/ 移至 連結

Google Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

## Word Sketch Entry Form

**Corpus:** chinese_all_trd

**Word Form:** 發票

Sort grammatical relations: ☑

Minimum frequency: 5

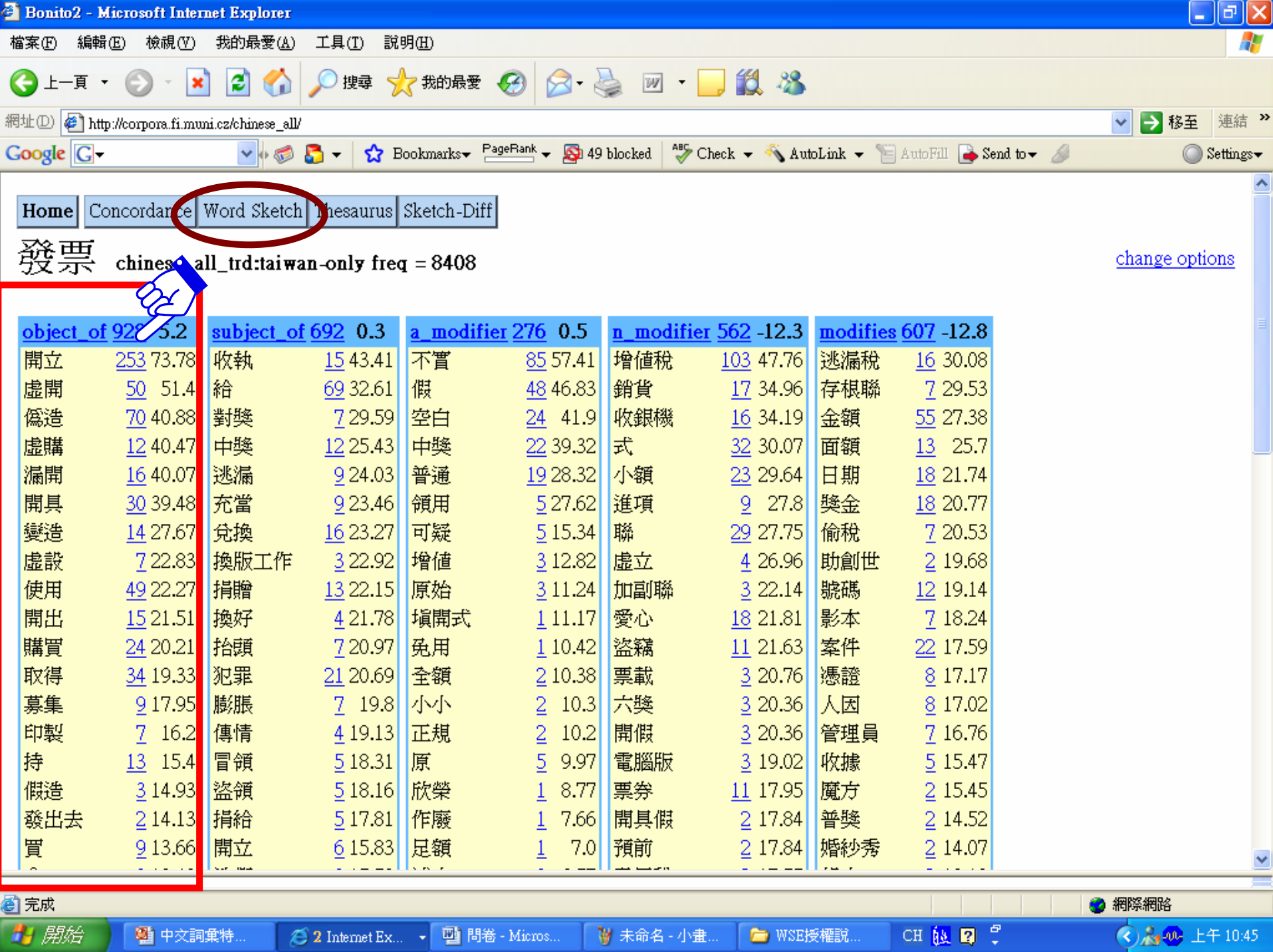
Minimum salience: 0.0

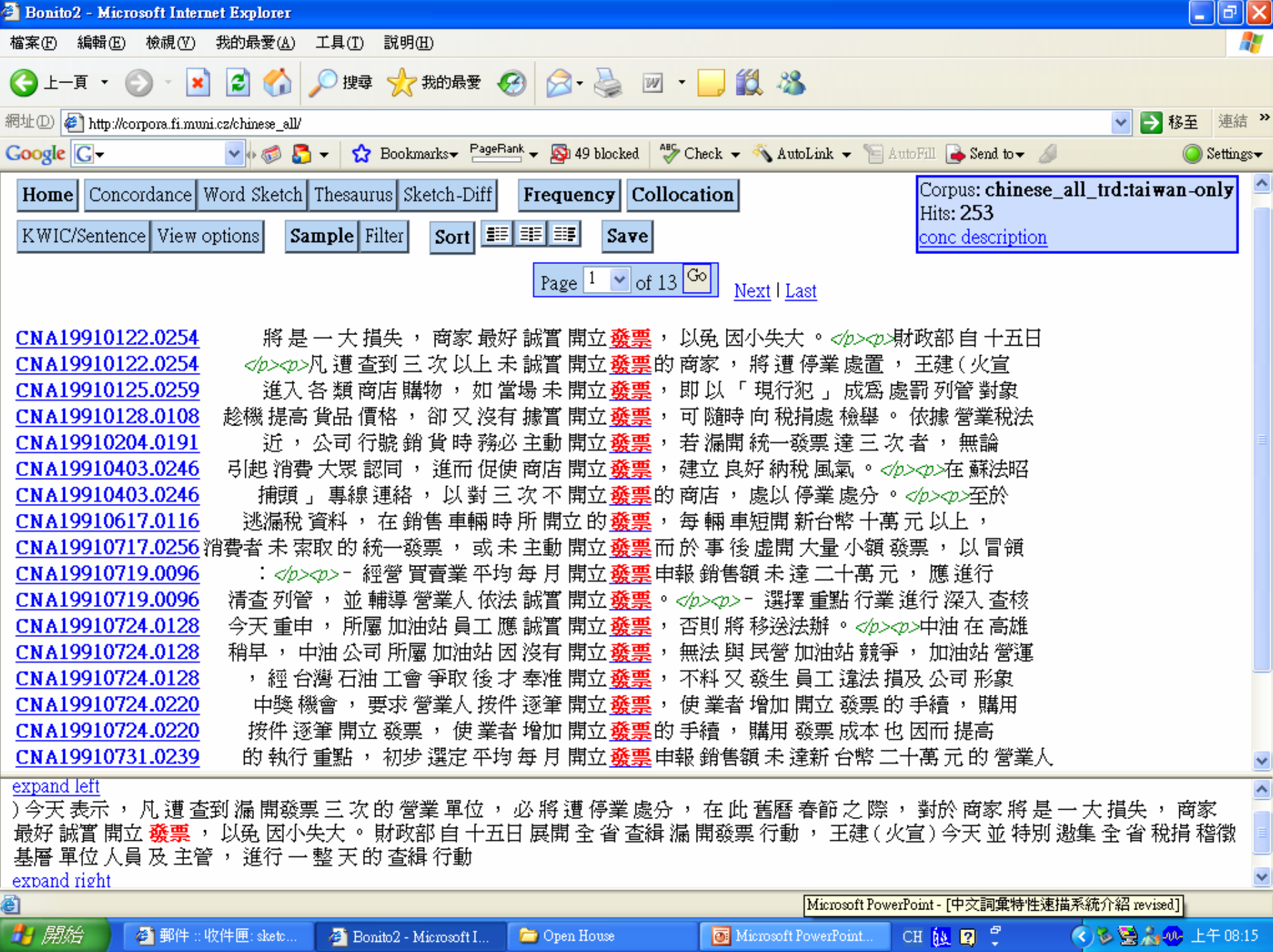
Maximum number of items in a grammatical relation: 25

Show Word Sketch

完成 網際網路

開始 中文詞彙特... 2 Internet Ex... 問卷 - Micros... 未命名 - 小畫... WSE授權說... CH 上午 10:56


Bonito2 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/
移至 連結
Google Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
Home Concordance Word Sketch Thesaurus Sketch-Diff
發票 chinese_all_trd:taiwan-only freq = 8408
change options
object_of 928 5.2
開立 253 73.78
虛開 50 51.4
僞造 70 40.88
虛購 12 40.47
漏開 16 40.07
開具 30 39.48
變造 14 27.67
虛設 7 22.83
使用 49 22.27
開出 15 21.51
購買 24 20.21
取得 34 19.33
募集 9 17.95
印製 7 16.2
持 13 15.4
假造 3 14.93
發出去 2 14.13
買 9 13.66
subject_of 692 0.3
收執 15 43.41
給 69 32.61
對獎 7 29.59
中獎 12 25.43
逃漏 9 24.03
充當 9 23.46
兌換 16 23.27
換版工作 3 22.92
捐贈 13 22.15
換好 4 21.78
抬頭 7 20.97
犯罪 21 20.69
膨脹 7 19.8
傳情 4 19.13
冒領 5 18.31
盜領 5 18.16
捐給 5 17.81
開立 6 15.83
a_modifier 276 0.5
不實 85 57.41
假 48 46.83
空白 24 41.9
中獎 22 39.32
普通 19 28.32
領用 5 27.62
可疑 5 15.34
增值 3 12.82
原始 3 11.24
填開式 1 11.17
免用 1 10.42
全額 2 10.38
小小 2 10.3
正規 2 10.2
原 5 9.97
欣榮 1 8.77
作廢 1 7.66
足額 1 7.0
n_modifier 562 -12.3
增值稅 103 47.76
銷貨 17 34.96
收銀機 16 34.19
式 32 30.07
小額 23 29.64
進項 9 27.8
聯 29 27.75
虛立 4 26.96
加副聯 3 22.14
愛心 18 21.81
盜竊 11 21.63
票載 3 20.76
六獎 3 20.36
開假 3 20.36
電腦版 3 19.02
票券 11 17.95
開具假 2 17.84
預前 2 17.84
modifies 607 -12.8
逃漏稅 16 30.08
存根聯 7 29.53
金額 55 27.38
面額 13 25.7
日期 18 21.74
獎金 18 20.77
偷稅 7 20.53
助創世 2 19.68
號碼 12 19.14
影本 7 18.24
案件 22 17.59
憑證 8 17.17
人因 8 17.02
管理員 7 16.76
收據 5 15.47
魔方 2 15.45
普獎 2 14.52
婚紗秀 2 14.07
完成
網際網路
開始 中文詞彙特... 2 Internet Ex... 問卷 - Micros... 未命名 - 小畫... WSE授權說... CH 上午 10:45


Bonito2 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/
移至 連結
Google Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
Home Concordance Word Sketch Thesaurus Sketch-Diff Frequency Collocation
KWIC/Sentence View options Sample Filter Sort Save
Corpus: chinese_all_trd:taiwan-only
Hits: 253
conc description
Page 1 of 13 Go
Next | Last
CNA19910122.0254 將是一大損失，商家最好誠實開立發票，以免因小失大。 財政部自十五日
CNA19910122.0254 凡遭查到三次以上未誠實開立發票的商家，將遭停業處置，王建(火宣
CNA19910125.0259 進入各類商店購物，如當場未開立發票，即以「現行犯」成爲處罰列管對象
CNA19910128.0108 趁機提高貨品價格，卻又沒有據實開立發票，可隨時向稅捐處檢舉。依據營業稅法
CNA19910204.0191 近，公司行號銷貨時務必主動開立發票，若漏開統一發票達三次者，無論
CNA19910403.0246 引起消費大眾認同，進而促使商店開立發票，建立良好納稅風氣。 在蘇法昭
CNA19910403.0246 捕頭」專線連絡，以對三次不開立發票的商店，處以停業處分。 至於
CNA19910617.0116 逃漏稅資料，在銷售車輛時所開立的發票，每輛車短開新台幣十萬元以上，
CNA19910717.0256 消費者未索取的統一發票，或未主動開立發票而於事後虛開大量小額發票，以冒領
CNA19910719.0096 ： - 經營買賣業平均每月開立發票申報銷售額未達二十萬元，應進行
CNA19910719.0096 清查列管，並輔導營業人依法誠實開立發票。 - 選擇重點行業進行深入查核
CNA19910724.0128 今天重申，所屬加油站員工應誠實開立發票，否則將移送法辦。 中油在高雄
CNA19910724.0128 稍早，中油公司所屬加油站因沒有開立發票，無法與民營加油站競爭，加油站營運
CNA19910724.0128 ，經台灣石油工會爭取後才奉准開立發票，不料又發生員工違法損及公司形象
CNA19910724.0220 中獎機會，要求營業人按件逐筆開立發票，使業者增加開立發票的手續，購用
CNA19910724.0220 按件逐筆開立發票，使業者增加開立發票的手續，購用發票成本也因而提高
CNA19910731.0239 的執行重點，初步選定平均每月開立發票申報銷售額未達新台幣二十萬元的營業人
expand left
)今天表示，凡遭查到漏開發票三次的營業單位，必將遭停業處分，在此舊曆春節之際，對於商家將是一大損失，商家最好誠實開立發票，以免因小失大。財政部自十五日展開全省查緝漏開發票行動，王建(火宣)今天並特別邀集全省稅捐稽徵基層單位人員及主管，進行一整天的查緝行動
expand right
Microsoft PowerPoint - [中文詞彙特性速描系統介紹 revised]
開始 郵件 :: 收件匣: sketc... Bonito2 - Microsoft I... Open House Microsoft PowerPoint... CH 上午 08:15

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

# 發票

chinese_all_trd:taiwan-only freq = 8408

change options

| object_of 928 5.2 | | subject_of 692 0.3 | | a_modifier 276 | | n_modifier 562 -12.3 | | modifies 607 -12.8 | |
|---|---|---|---|---|---|---|---|---|---|
| 開立 | 253 73.78 | 收執 | 15 43.41 | 不實 | 85 57.41 | 增值稅 | 103 47.76 | 逃漏稅 | 16 30.08 |
| 虛開 | 50 51.4 | 給 | 69 32.61 | 假 | 48 46.83 | 銷貨 | 17 34.96 | 存根聯 | 7 29.53 |
| 偽造 | 70 40.88 | 對獎 | 7 29.59 | 空白 | 24 41.9 | 收銀機 | 16 34.19 | 金額 | 55 27.38 |
| 虛購 | 12 40.47 | 中獎 | 12 25.43 | 中獎 | 22 39.32 | 式 | 32 30.07 | 面額 | 13 25.7 |
| 漏開 | 16 40.07 | 逃漏 | 9 24.03 | 普通 | 19 28.32 | 小額 | 23 29.64 | 日期 | 18 21.74 |
| 開具 | 30 39.48 | 充當 | 9 23.46 | 領用 | 5 27.62 | 進項 | 9 27.8 | 獎金 | 18 20.77 |
| 變造 | 14 27.67 | 兌換 | 16 23.27 | 可疑 | 5 15.34 | 聯 | 29 27.75 | 偷稅 | 7 20.53 |
| 虛設 | 7 22.83 | 換版工作 | 3 22.92 | 增值 | 3 12.82 | 虛立 | 4 26.96 | 助創世 | 2 19.68 |
| 使用 | 49 22.27 | 捐贈 | 13 22.15 | 原始 | 3 11.24 | 加副聯 | 3 22.14 | 號碼 | 12 19.14 |
| 開出 | 15 21.51 | 換好 | 4 21.78 | 塡開式 | 1 11.17 | 愛心 | 18 21.81 | 影本 | 7 18.24 |
| 購買 | 24 20.21 | 抬頭 | 7 20.97 | 免用 | 1 10.42 | 盜竊 | 11 21.63 | 案件 | 22 17.59 |
| 取得 | 34 19.33 | 犯罪 | 21 20.69 | 全額 | 2 10.38 | 票載 | 3 20.76 | 憑證 | 8 17.17 |
| 募集 | 9 17.95 | 膨脹 | 7 19.8 | 小小 | 2 10.3 | 六獎 | 3 20.36 | 人因 | 8 17.02 |
| 印製 | 7 16.2 | 傳情 | 4 19.13 | 正規 | 2 10.2 | 開假 | 3 20.36 | 管理員 | 7 16.76 |
| 持 | 13 15.4 | 冒領 | 5 18.31 | 原 | 5 9.97 | 電腦版 | 3 19.02 | 收據 | 5 15.47 |
| 假造 | 3 14.93 | 盜領 | 5 18.16 | 欣榮 | 1 8.77 | 票券 | 11 17.95 | 魔方 | 2 15.45 |
| 發出去 | 2 14.13 | 捐給 | 5 17.81 | 作廢 | 1 7.66 | 開具假 | 2 17.84 | 普獎 | 2 14.52 |
| 買 | 9 13.66 | 開立 | 6 15.83 | 足額 | 1 7.0 | 預前 | 2 17.84 | 婚紗秀 | 2 14.07 |

Corpus: **chinese_all_trd:taiwan-only**
Hits: 85
conc description



CNA19911103.0009 申報 案件 抽查 作業 、 營業人 開立 不實 發票 的 查緝 作業 、 進口 大宗 或 高 價值 貨物
CNA19911119.0281 顯示 ， 有 多 起 虛設 行號 開立 的 不實 發票 ， 由於 案情 重大 且 牽連 甚 廣 ， 於是
CNA19930224.0091 已 達新 台幣 二十多億 元 。 而 開立 不實 發票 的 面額 也 達到 四億五千四百三十六萬 元
CNA19930817.0082 和 廠商 協議 開發 發票 ， 取得 不實 面額 發票 ， 向 工業局 請 領 公款 ， 而 以 此 種
CNA19931129.0190 台灣 再生 公司 涉嫌 僞造 文書 ， 以 不實 發票 向 汽水 公會 領取 寶特瓶 清除 處理費 和
CNA19931130.0329 ， 外傳 該 公司 回收 廢 寶特瓶 有 以 不實 發票 領取 費用 情形 ， 是 完全 不 了解 回收
CNA19940302.0282 身分證 ， 申請 設立 公司 行號 並 開出 不實 發票 ， 金額 高達 新台幣 五十餘億 元 。 
CNA19940302.0282 台 各 地 串連 ， 形成 供應 和 販賣 不實 發票 的 銷售 網 ， 開出 發票 的 金額 高達 五十多億
CNA19950324.0326 四百多萬 元 及 十六萬 多元 面額 的 不實 發票 給 民間 業者 ， 也 被 檢方 併案 偵辦 。
CNA19950324.0326 蔡盛財 等 四 名 中油 員工 和 一 名 使用 不實 發票 的 民間 業者 范植森 等 五 人 收押 禁見
CNA19950324.0326 表示 ， 這 六 名 中油 員工 是否 交付 不實 發票 而 取得 民間 業者 的 期約 或 收受 賄賂
CNA19950327.0266 1 2 0 件 爲 涉嫌 繳驗 僞造 、 變造 或 不實 發票 虛報 完稅 價格 逃漏 進口 稅捐 案件 ， 其
CNA19951228.0287 另 被告 等 製造 假買賣 ， 以 不實 發票 、 支票 向 銀行 融資 ， 迄 八十四年 五月
CNA19960312.0210 公司 財物 經理 侯秋香 ， 相互 開立 不實 發票 向 金融 行庫 辦理 貼現 ， 經 檢方 偵辦
CNA19960510.0176 買賣 發票 逃漏 稅捐 ， 以往 廠商 開立 不實 發票 都 是 採 分散 方式 ， 人工 查核 不易 ，
CNA19960626.0318 調查 單位 查獲 十八 家 虛設 行號 開立 不實 發票 、 營業額 十五億 多元 ; 徵收 積欠 稅收

expand left
， 財政部 指出 ， 十八 個 項目 中 ， 最 重要 的 防止 營業人 漏開 統一發票 稽查 作業 、 繁華 地段 房屋 租金 偏低 案件 查核 作業 、 簽証 不實 的 營所稅 申報 案件 抽查 作業 、 營業人 開立 不實 發票 的 查緝 作業 、 進口 大宗 或 高 價值 貨物 銷售 流程 追查 作業 、 扣繳 異常 單位 檢查 作業 、 營利 事業 取得 小店戶 收據 查核 作業 等 七 項 ， 都 由 財政部 列管 ， 稽徵 機關
expand right

# 前言

## 數大就是美 ——徐志摩

- 語料與大數
  - 語言學家的任務就是要在無窮盡的語言資料中找出規律與解釋的真理來。
  - 語料庫語言學研究的挑戰之一，是當語料庫不夠大時，有些語言現象無法呈現，或呈現不出全貌。
  - 但是語料量太大時，人又無法有效消化
- 從大數中求真
  - 語料量愈大，愈較能呈現語言使用的完整面貌
  - 更重要的是能呈現詞語與詞語間的關係

# 語料量多大才算夠大？

## 關鍵詞詞彙的觀察

- 第10,000名常用詞的詞頻大約在0.001%左右
- 在五百萬詞語料庫中出現大約50次

→

- 足夠觀察到一個詞彙使用的大致狀況
- 也正好是人們可以從容處理語料的規模
- 五百萬到一千萬詞，是關鍵詞檢索語料（KWIC）觀察的最佳規模

# 古典語料庫：數量級在百萬的中文語料庫

| Corpus Name | Online Year | Data | Duration/ Content |
|---|---|---|---|
| Sinica 4.0 (Taiwan) | 1996 | 5.2 M words<br>7.9 M characters | 1990-1996<br>Fully Tagged |
| Sinica 5.0 (Taiwan) | 2006 | 10 M words | 1990-2004<br>Fully Tagged |
| Sinorama (Taiwan) | 2003 | 3.2 M English words<br>5.3 M Chinese characters | 1976 – 2000<br>(1999-2000)<br>Aligned |
| CCL (Peking) | 2003 | 85 M simplified characters | 1919 -2003<br>Partially tagged<br>(1 million) |

M= million

# 古典語料庫中如何將知識規律化

# The coloured pens method

1 arity, which will be used to take a party of under-privileged children to
2 from outside. You are invited to a party and after a couple of drinks you
3 tion, we believe politicians of all parties will listen to our views. &eq
4 ould be reaching agreement with all parties concerned, as to which event
5 lack people. I have certainly been party to one or two discussions amongs
6 . These should be discussed by both parties before entering into the relat
7 presents They had hosted a cocktail party at Kensington palace, for examp
8 akes. By midnight the end-of-course party is in full swing, but most cad
9 e should be a right for the injured party to terminate the contract. A ma
10 by the Safran Peoples ' Liberation Party. This presents the powerful nei
11 s. Ahead I could see the rest of my party plodding towards the final slope
12 cial ethic. The two main political parties - the Tories and the Liberals
13 ritish successes in Perth The small party of British players competing in
14 to help control. One member of the party went to summon the rescue team a
15 rket society fashion magazine. The party was held at his flat which was a
16 security and secrecy than any Tory Party Conference : it seems that bootl

1 political association
2 social event
3 group of people
4 person in an agreement/dispute
5 to **be party to** something...

from Kilgarriff et al. 2005

# 自動觀察預測語言規律所需要的語料量預估

- 以單一詞彙爲中心的詞彙規律
- 呈現詞與詞之間的共現關係，如
  - V+N：「開立」+「發票」
  - A+N：「不實」+ 「發票」
- 以第一萬名中頻詞爲切分點，希望能看到以這些詞爲標的語法規律。
- 在關鍵詞前後5個詞內出現（比緊鄰多10倍機率）
- 估計需要十億詞的語料，在機率上才能與其看到資料

# 當代鉅量語料庫的挑戰

鉅量：一億詞（100,000,000）以上，
到十億詞（1,000,000,000 giga）的規模

- 第一萬名中頻詞約有1,000-10,000個例句
- 足夠做自動預測
- 並提供有效樣本作分佈分析，但是
  人無法有效閱讀
  - 50個例句: 輕輕鬆鬆
  - 500個例句: 得花點時間與力氣
  - 5000個例句: 無法集中精神，頭昏眼花

更難想像由人來標記/檢查十億個詞的詞類

# 處理鉅量語料庫的解決方案

- 採取機器自動分詞與詞類標記
  - 前提：計算語言學研究成績進步，有95%以上的準確率

*Ma and Huang 2006 LREC*

➢鉅量語料庫的處理模式

- ➢必須以自動機器標記爲主
- ➢語言知識必須以自動方式取得並由機器處理
- ➢人工方式評估取樣結果
- ➢回到自動標記系統中改進

# 由鉅量語料跨越到語言規律

- 利用語法結構知識自動抽取詞與詞間的關係訊息，並進行統計
  - 跨越人類記憶力與注意力集中的限制
  - 由詞彙例句爲本，轉爲語法語意關係的分佈趨勢爲本
  - 幫電腦加上初步歸納規律的能力
  - 觀察規律與趨勢，有大量資料時；少量的例外與誤差（如5%以下）不影響大局

# 以觀察人爲比喻

- 可以觀察其身高，體重，體重，衣著。。。
  - 還是不太瞭解這個人是什麼樣的人
- 觀察他的朋友，親人，他如何吃飯，走路，交談，如何處理人際關係。。。
  - 原來他是這樣的人！
- 詞彙也是如此

You shall know a word by the company it keeps. -M. K. Halliday

# 新一代的語料庫處理工具

- 直接呈現共現句型 (collocation)
- 以統計工具計算並共現句型的顯著性
  - 觀察處理約20-40個共現句型
  - 每個句型的共現詞(collocates)，都已按顯著性排列
    - 分析由幾千具中歸納出的規律性模式，而不是由幾千句素語料開始分析

# 由例句到詞彙關係：以BNC爲例

- British National Corpus (BNC)
  - 一億（100,000,000）詞 POS-標注完成
- lemmatized
  - *assisting => assist (v)*
- Parsed （部分剖析）得到
- 七千萬（70,000,000）組關係

<object, sip, coffee>  <subject, arrive, coffee>

<and-or, tea, coffee>  <modifier, coffee, instant>

# 語言結構知識的應用：BNC爲例

如何定義詞與詞間的關係？

- Uses CQL (Corpus query language)
  - Christ and Schulze, U. Stuttgart, 1994
- defining an object:

  `v (adj|n|det|num|adv)* n`

  rewriting in CQL with BNC/CLAWS-5 tags

  [tag="VV.*"] [tag="(A[JTV]|D|O).*"]* [tag="NN.*"]

# 方案平台： WordSketch Engine

- WordSketch Engine http://sketchengine.co.uk

  Adam Kilgarriff 的研究團隊開發
- 中文詞彙特性素描系統 Chinese WordSketch (CWS) 的開發
  - **Academia Sinica, Taiwan** (Huang, Smith, Ma, Simon 黃居仁，史尚明，馬偉雲，石穆)
  - **Masaryk University, Czech** (Rychly)
  - **Technical University, Budapest** (Tugwell)
  - **Lexical Computing Ltd** (Kilgarriff)

# CWS 所採用的語料庫

十四億字的Chinese Giga-word Corpus）

含兩岸三地語料，當今最大中文語料庫，無詞類標記

- CGW Corpus (LDC 2005) 的內容分布

| CGW Corpus | K-characters | Documents |
|---|---|---|
| CNA(Taiwan) | 792,195 | 1,769,953 |
| Xinhua(PRC) | 471,110 | 992,261 |
| Zaobao(Sing) | 28,066 | 41,418 |
| TOTAL | 1,291,371 | 2,803,632 |

# CGW 語料庫基本資料

| | 來源 | 字數 | 詞數 | 文件數 |
|---|---|---|---|---|
| First Edition | CNA | 735 | 462 | 1,649 |
| | Xinhua | 382 | 252 | 817 |
| | TOTAL | 1,118 | 714 | 2,466 |
| | | | | |
| Second Edition | CNA | 792 | 497 | 1,769 |
| | Xinhua | 471 | 310 | 992 |
| | Zaobao | 28 | 18 | 41 |
| | TOTAL | 1,291 | 825 | 2,803 |

Unit: **Million**

# CWS 所用的語法知識

- 詞義語與法共現的自動抽取需要大量語言知識與理論的累積
  - 1:"V[BCJ]" "Di"? "N[abc]"? "DE"? "N[abc]"?
  - 2: "Na" [tag!= "Na"]

  - ("XXX" represents XXX is a regular expression, "XXX"? represents XXX appears zero or one time, "XXX"{a,b} represents XXX appears a~b times.)

# 整合語料標記，統計，與語法知識

- Collocations are identified with Context free rules in Word Sketch Engine
  - Collocating Pattern for Object from CSE I
  - 1:"V[BCJ]" "Di"? "N[abc]"? "DE"? "N[abc]"? 2: "Na" [tag!= "Na"]
- Challenge: Long-distance relations
- 全穀麵包，吃了很健康。

  *quan.gu mian.bao, **chi** le hen jian.kang*
- 有人嘗試要將這荷花分類，卻越分越累。

  *you ren chang.shi yao **jiang** zhe he.hua **fen.lei**, que yue fen yue lei*

# 在處理平台中引進語法知識 – 來源

- Information-based Case Grammar (ICG, Chen and Huang 1992)
- Encoded on over 40,000 verbs in Sinica Lexicon
  - ICG Basic Patterns for Stative Pseudo-transitive Verb (VI)

    EXPERIENCER<GOAL[PP[對]]<VI

    EXPERIENCER<VI<<GOAL[PP[於]]

    THEME<GOAL[PP{對、以}]<VI

    THEME<VI<<GOAL[PP[於]]

    THEME<VI<<SOURCE[PP{自、於}]

    THEME< SOURCE[PP{歸、爲}]<VI

# 在處理平台中引進語法知識 – 應用

- 村莊(object) 明天將 被 夷爲平地(VB11)

  *cunzhuang mingtian jiang bei yiweipingdi*

  – begin time1 location time1 adv? passive_prep adv_string 1:"V[BCJ].*" [tag!="DE"]

- 大量 的 遊客 破壞(VC2) 公園 景觀(object)

  *daliang de youke pohuai gongyuan jingguan*

  – 1:"VC.*" (particle|prep)? NP not_noun

  – (NP is defined as "…noun_modifier{0,2} 2:noun…".

# 在處理平台中引進語法知識－結果

- *Object Recall Comparison*

| | CSE I | CSE II |
|---|---|---|
| hong2 (red) | 0 | 0 |
| pao3 (run) | 0 | 8,704 |
| kan4 (look) | 32,350 | 64,096 |
| da3 (hit) | 26,016 | 47,182 |
| song4 (give) | 0 | 76,378 |
| shuo1 (say) | 0 | 20,350 |
| xiang1xin4 (believe) | 0 | 52,373 |
| quan4 (persuade) | 0 | 3,852 |

# 在處理平台中引進語法知識 – 結果 II

- **CSEII** 中「吃」的最顯著賓語

  **CSEII** 前**20**名，但不在**CSEI**前**20**名中的

  – *飯 fan4 rice 802 70.96 (4),*

  – *虧 kui disadvantage 329 59.24 (12)*

  – *苦頭 ku3tou2 suffering 194 58.71 (14)*

# 中文詞彙素描/CGW語料庫重要資料統計(中央社語料）

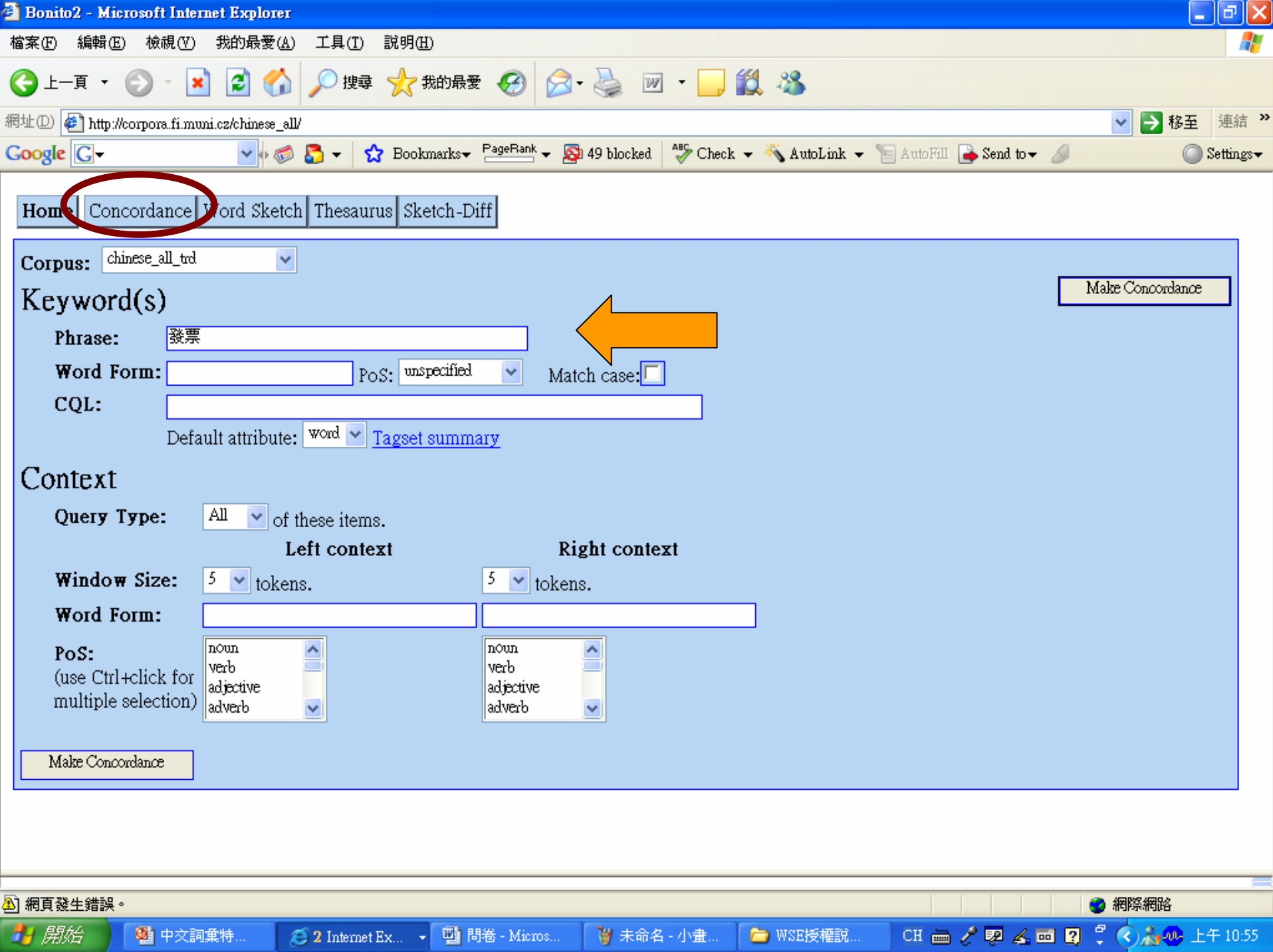
- 規模： 約5億詞 （約497,000,000）
- 斷詞後詞數：455,526,796
- 使用語法規律數： 約200 條
- 詞彙關係種類： 20 個
- 找出的詞彙關係數59,183,238

# 中文詞彙特性速描系統介紹

- Concordance
- WordSketch
- Sketch Difference
- Thesaurus
  - 如何申請使用

# Concordance

- 比傳統KWIC更強的功能


Bonito2 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/
移至 連結
Google Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
Home Concordance Word Sketch Thesaurus Sketch-Diff
Corpus: chinese_all_trd
Make Concordance
Keyword(s)
Phrase: 發票
Word Form: PoS: unspecified Match case:
CQL:
Default attribute: word Tagset summary
Context
Query Type: All of these items.
Left context
Right context
Window Size: 5 tokens. 5 tokens.
Word Form:
PoS:
(use Ctrl+click for multiple selection)
noun
verb
adjective
adverb
noun
verb
adjective
adverb
Make Concordance
網頁發生錯誤。
網際網路
開始
中文詞彙特...
2 Internet Ex...
問卷 - Micros...
未命名 - 小畫...
WSE授權說...
CH
上午 10:55

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff | Frequency | Collocation

KWIC/Sentence | View options | Sample | Filter | Sort

Corpus: **chinese_all_trd**
Hits: **8408**
conc description

421 Go Next | Last

| | |
|---|---|
| CNA19910102.0102 | 民眾 可 蒐集 金額 新台幣 二千 元 以上 的 發票 ， 向 稅捐 單位 領取 紀念品 。 省 |
| CNA19910102.0102 | 活動 ， 除 可 繼續 提醒 民眾 保持 購物 索取 發票 的 習慣 外 ， 更 配合 年度 所得稅 的 申報 |
| CNA19910102.0102 | 宣導 將 舉辦 的 活動 包括 ： - 集 發票 兌換 紀念品 ， 凡 集滿 本 月份 發票 金額 |
| CNA19910102.0102 | - 集 發票 兌換 紀念品 ， 凡 集滿 本 月份 發票 金額 在 二千 元 以上 者 ， 可 憑 發票 於 |
| CNA19910102.0102 | 月份 發票 金額 在 二千 元 以上 者 ， 可 憑 發票 於 本 月 三十日 及 三十一日 向 各 稅捐處 |
| CNA19910102.0102 | 或 摸彩 活動 。 - 辦理 餐飲業 開立 發票 模範 商店 選拔 。 八 名 減刑 獲釋 泰人 |
| CNA19910104.0131 | 勒令 停業 一 週 。 稅捐處 已 將 漏開 發票 與 大額 欠稅戶 以 電腦 列管 ， 一 年 內 |
| CNA19910104.0131 | 欠稅戶 以 電腦 列管 ， 一 年 內 有 三 次 漏開 發票 紀錄 及 欠稅 超過 十萬 元 ， 都 將 依 規定 |
| CNA19910108.0104 | 八十年 元 月份 二 聯 式 統一發票 或 收銀機 發票 收執 聯 金額 達 二千 元 以上 者 ， 均 可 |
| CNA19910122.0254 | 將 是 一 大 損失 ， 商家 最好 誠實 開立 發票 ， 以免 因小失大 。 財政部 自 十五日 |
| CNA19910122.0254 | 凡 遭 查到 三 次 以上 未 誠實 開立 發票 的 商家 ， 將 遭 停業 處置 ， 王建（火宣 |
| CNA19910122.0254 | 一百萬 者 ， 在 誠實 開發票 前提 下 ， 使 發票 數額 提高 至 兩百萬 元 以上 ， 財政部 的 |
| CNA19910125.0259 | 進入 各 類 商店 購物 ， 如 當場 未 開立 發票 ， 即 以 「 現行犯 」 成爲 處罰 列管 對象 |
| CNA19910125.0259 | 包含 了 自己 所 繳 的 稅額 ， 如 不 索取 發票 ， 就 等於 被 不肖 商家 吃掉 ， 爲 維護 |
| CNA19910127.0102 | 打算 再 增加 四 人 。 上海 查獲 非法 發票 交易（中央社 台北 二十七日 電） 香港 |
| CNA19910127.0102 | 報導 ， 上海 工商 執法 人員 查獲 一 批 非法 發票 交易 市場 ， 當場 擒獲 多 名 票販 。 |
| CNA19910127.0102 | 用 三元 「 人民幣 」 買來 一 張 空白 非法 發票 ， 任意 填寫 金額 便 可 回去 報假賬；而 |

expand left

台省 各 縣市 舉辦 統一發票 宣導 活動（中央社 中興新村 二日 電）本 月份 購物 時 別 忘 了 索取 統一發票 。 臺灣省 各 縣市 正 舉辦 統一發票 宣導 ， 民眾 可 蒐集 金額 新台幣 二千 元 以上 的 發票 ， 向 稅捐 單位 領取 紀念品 。 省 稅務局 聯合 各 縣市 稅捐處 ， 將 在 農曆年 過 年 前 ， 共同 舉辦 多 項 統一發票 宣導 活動 ， 除 可 繼續 提醒 民眾 保持 購物 索取 發票 的 習慣 外

# WordSketch

自動產生具體而微的簡易語法語意典

- 關鍵詞實際使用的完整描述
- 建立在不同的共現詞彙關係上
- 按照關係的顯著性排序

# WordSketch:以「逮捕」爲例

- It does produce the expected results with an easy to use interface.

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

逮捕 **chintag freq = 45057** change options

| object | 8502 | 5.0 | subject | 8891 | 4.5 | modifier | 8461 | 4.4 | pp 以 | 27 | 4.6 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 嫌犯 | 1538 | 72.65 | 警方 | 2769 | 67.06 | 循線 | 754 | 78.29 | 罪嫌 | 6 | 20.88 |
| 主嫌 | 197 | 50.54 | 現行犯 | 124 | 58.01 | 當場 | 1253 | 74.3 | | | |
| 行動 | 863 | 45.94 | 罪名 | 266 | 49.29 | 大肆 | 133 | 47.19 | | | |
| 現行犯 | 58 | 44.78 | 公安 | 249 | 42.62 | 一舉 | 118 | 43.6 | | | |
| 兇嫌 | 106 | 42.68 | 嫌犯 | 284 | 41.81 | 加以 | 235 | 40.04 | | | |
| 主嫌犯 | 61 | 42.47 | 當局 | 552 | 41.16 | 陸續 | 219 | 39.17 | | | |
| 特權 | 112 | 41.64 | 幹員 | 115 | 38.02 | 予以 | 239 | 38.12 | | | |
| 毒販 | 67 | 40.79 | 警網 | 50 | 34.29 | 合力 | 55 | 36.35 | | | |
| 通緝犯 | 80 | 40.52 | 間諜罪 | 22 | 33.37 | 逕行 | 63 | 36.05 | | | |

# 同義詞典：Thesaurus

- 比對素描的相似度，找出使用上最接近的詞彙來

Bonito2 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://corpora.fi.muni.cz/chinese_all/

Google Go Bookmarks PageRank 141 blocked Check AutoLink AutoFill Send to

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

## Thesaurus Entry Form

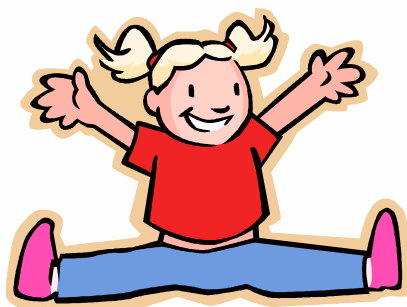
**Corpus:** chinese_all_trd

**Word Form:** 快樂

Maximum number of items: 60

Minimum similarity between cluster items: 0.15

Show Similar Words



Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

# 快樂 chinese_all_trd freq = 15438

| | | |
|---|---|---|
| 愉快 | 0.308 | 輕鬆 0.242 活潑 0.205 溫馨 0.204 溫暖 0.199 樂觀 0.199 平靜 0.193 熱鬧 0.187 忙碌 0.185 |
| 幸福 | 0.295 | 美好 0.227 痛苦 0.207 美麗 0.2 |
| 歡樂 | 0.293 | 浪漫 0.186 |
| 開心 | 0.279 | 興奮 0.249 高興 0.243 勇敢 0.242 懂得 0.235 安心 0.225 關心 0.217 熱愛 0.216 辛苦 0.215 難過 0.215 期待 0.214 失望 0.214 在一起 0.213 害怕 0.212 喜歡 0.212 覺得 0.206 激動 0.205 放心 0.204 感受到 0.201 驕傲 0.199 用心 0.194 明白 0.189 喜愛 0.188 冷靜 0.187 瞭解 0.186 講 0.185 滿意 0.185 相處 0.184 大聲 0.183 聽 0.183 |
| 有錢 | 0.221 | 聰明 0.211 富裕 0.209 舒適 0.194 乾淨 0.187 親近 0.186 賺錢 0.183 |
| 幸運 | 0.21 | 可愛 0.194 漂亮 0.188 |
| 讀書 | 0.197 | 體會 0.188 好好 0.183 |
| 歡喜 | 0.187 | |
| 健康 | 0.184 | |

Word Sketch Engine (ver:WSE-1.12-1.44)

# 快樂

快 chinese_all_trd freq = 15438

| | | |
|---|---|---|
| 愉快 | 0.308 | 輕鬆 0.242 活潑 0.205 溫馨 0.204 溫暖 0.199 樂觀 0.199 平靜 0.193 熱鬧 0.187 忙碌 0.185 |
| 幸福 | 0.295 | 美好 0.227 痛苦 0.207 美麗 0.2 |
| 歡樂 | 0.293 | 浪漫 0.186 |
| 開心 | 0.279 | 興奮 0.249 高興 0.243 勇敢 0.242 懂得 0.235 安心 0.225 關心 0.217 熱愛 0.216 辛苦 0.215 難過 0.215 激動 0.205 放心 0.204 感受到 0.201 驕傲 0.199 用心 0.194 明白 0.189 喜愛 0.188 冷靜 0.187 瞭解 0.18 |
| 有錢 | 0.221 | 聰明 0.211 富裕 0.209 舒適 0.194 乾淨 0.187 親近 0.186 賺錢 0.183 |
| 幸運 | 0.21 | 可愛 0.194 漂亮 0.188 |
| 讀書 | 0.197 | 體會 0.188 好好 0.183 |
| 歡喜 | 0.187 | |
| 健康 | 0.184 | |

# 快樂/愉快 chinese_all_trd freq = 15438/11572

change options

## Common patterns

| 快樂 | 21 | 14 | 7 | 0 | -7 | -14 | -21 | 愉快 |
|---|---|---|---|---|---|---|---|---|

| modifier | 2640 | 4052 | 5.4 | 7.1 |
|---|---|---|---|---|
| 不 | 411 | 1690 | 30.9 | 48.7 |
| 非常 | 95 | 420 | 28.2 | 46.6 |
| 很 | 354 | 466 | 39.2 | 39.6 |
| 十分 | 21 | 225 | 13.3 | 38.5 |
| 相當 | 26 | 274 | 13.2 | 38.0 |
| 更 | 189 | 51 | 30.6 | 13.1 |
| 最 | 275 | 54 | 30.4 | 9.8 |
| 又 | 117 | 50 | 27.7 | 15.6 |
| 過 | 69 | 10 | 22.7 | 4.6 |
| 能 | 106 | 37 | 22.0 | 9.4 |
| 還 | 8 | 132 | 2.2 | 21.6 |
| 著 | 71 | 33 | 18.7 | 9.4 |
| 太 | 8 | 39 | 7.4 | 18.5 |
| 並 | 15 | 98 | 5.3 | 18.3 |
| 比較 | 26 | 12 | 17.6 | 9.2 |
| 能夠 | 27 | 8 | 17.1 | 6.1 |
| 就 | 47 | 9 | 14.6 | 1.8 |
| 可以 | 36 | 9 | 14.0 | 2.9 |
| 都 | 36 | 29 | 11.5 | 7.7 |
| 也 | 40 | 42 | 8.6 | 6.6 |
| 會 | 28 | 10 | 8.2 | 0.8 |
| 更加 | 7 | 6 | 7.7 | 5.5 |
| 卻 | 9 | 7 | 7.3 | 4.5 |
| 了 | 69 | 83 | 5.7 | 4.4 |
| 是否 | 8 | 6 | 5.6 | 3.0 |

| subject | 1200 | 1656 | 2.9 | 3.4 |
|---|---|---|---|---|
| 佳節 | 242 | 241 | 78.2 | 75.5 |
| 心情 | 8 | 182 | 15.7 | 58.4 |
| 秋節 | 53 | 22 | 53.7 | 36.7 |
| 節日 | 58 | 44 | 40.3 | 34.4 |
| 生活 | 38 | 85 | 19.0 | 26.2 |
| 人 | 52 | 77 | 15.5 | 17.4 |
| 們 | 31 | 12 | 15.6 | 6.8 |
| 民眾 | 22 | 5 | 12.5 | 2.5 |

| modifies | 1541 | 631 | 1.0 | 0.4 |
|---|---|---|---|---|
| 時光 | 90 | 9 | 58.8 | 25.3 |
| 心情 | 21 | 34 | 25.3 | 36.2 |
| 氣氛 | 35 | 37 | 27.2 | 32.6 |
| 地 | 6 | 38 | 6.2 | 26.2 |
| 生活 | 67 | 10 | 23.8 | 10.1 |
| 假期 | 17 | 12 | 23.3 | 22.7 |
| 教育 | 8 | 43 | 3.9 | 20.7 |

## "快樂" only patterns

| modifier 2640 | | 5.4 |
|---|---|---|
| 天天 | 25 | 32.3 |
| 好不 | 9 | 25.1 |
| 真 | 27 | 24.6 |
| 愈 | 29 | 24.0 |
| 一點 | 13 | 20.6 |
| 極了 | 6 | 19.5 |
| 永遠 | 11 | 17.4 |
| 來得 | 8 | 16.7 |
| 一起 | 21 | 16.2 |
| 不見得 | 5 | 15.8 |
| 起來 | 15 | 15.4 |
| 真正 | 15 | 15.1 |

| subject 1200 | | 2.9 |
|---|---|---|
| 軍人節 | 20 | 42.5 |
| 老人節 | 9 | 32.9 |
| 孩子 | 35 | 28.9 |
| 小朋友 | 26 | 27.9 |
| 榮民節 | 5 | 23.4 |
| 兒童 | 41 | 23.1 |
| 青春 | 11 | 21.5 |
| 媽咪 | 5 | 21.4 |
| 孩子們 | 10 | 20.5 |
| 客輪 | 7 | 20.5 |
| 學生 | 45 | 20.4 |
| 小孩 | 11 | 19.1 |

| modifies 1541 | | 1.0 |
|---|---|---|
| 公主輪 | 37 | 64.1 |
| 捐血人 | 25 | 45.8 |
| 童年 | 34 | 43.1 |
| 炊事員 | 13 | 36.6 |
| 聯播網 | 14 | 34.9 |
| 人生 | 40 | 34.2 |
| 夏令營 | 25 | 32.8 |
| 拼財金 | 5 | 31.3 |
| 天使 | 20 | 30.9 |
| 電腦營 | 6 | 29.0 |
| 暑夏 | 5 | 28.9 |
| 捐血站 | 8 | 28.7 |

## "愉快" only patterns

| modifier 4052 | | 7.1 |
|---|---|---|
| 頗爲 | 16 | 20.4 |
| 甚爲 | 8 | 17.8 |
| 格外 | 13 | 17.7 |
| 甚 | 18 | 17.1 |
| 極爲 | 16 | 15.7 |
| 首先 | 10 | 10.8 |
| 極 | 6 | 4.8 |
| 已 | 15 | 0.5 |

| subject 1656 | | 3.4 |
|---|---|---|
| 神情 | 205 | 73.2 |
| 旅途 | 77 | 56.9 |
| 氣氛 | 60 | 34.0 |
| 感 | 21 | 32.9 |
| 精神 | 120 | 32.6 |
| 身心 | 23 | 22.5 |
| 賓主 | 6 | 19.4 |
| 生日 | 8 | 18.1 |
| 表情 | 6 | 17.6 |
| 國王 | 14 | 17.5 |
| 天皇 | 6 | 17.1 |
| 總理 | 33 | 16.5 |

| modifies 631 | | 0.4 |
|---|---|---|
| 經驗 | 95 | 39.3 |
| 事件 | 124 | 39.0 |
| 事情 | 19 | 26.6 |
| 笑容 | 9 | 25.4 |
| 回憶 | 11 | 24.6 |
| 地步 | 7 | 22.0 |
| 情事 | 10 | 20.0 |
| 神情 | 6 | 19.5 |
| 感覺 | 5 | 14.0 |
| 婚姻 | 5 | 13.9 |
| 場面 | 5 | 13.6 |
| 風波 | 5 | 13.4 |

# 提示

Chinese Wordnet

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

目標 chinese_all_trd freq = 189229

| | | |
|---|---|---|
| 任務 | 0.343 | 方面 0.284 情況 0.265 行動 0.261 過程 0.259 條件 0.258 狀況 0.238 工作 0.235 項目 0.234 需求 0.232 結果 0.229 問題 0.227 情形 0.219 |
| 目的 | 0.327 | 方案 0.318 方向 0.314 原則 0.311 計劃 0.31 措施 0.305 政策 0.303 計畫 0.29 策略 0.283 模式 0.26 立場 0.257 作法 0.252 標準 0.25 題 0.249 機制 0.249 體系 0.249 基礎 0.245 方法 0.241 制度 0.24 做法 0.239 體制 0.234 協議 0.23 辦法 0.228 經費 0.219 內容 0.217 式 0.215 |
| 指標 | 0.27 | 能力 0.243 水平 0.226 效果 0.223 效益 0.22 功能 0.22 力量 0.218 規模 0.216 |
| 理念 | 0.258 | 戰略 0.25 成果 0.249 方針 0.245 構想 0.23 進程 0.229 領域 0.223 經驗 0.218 觀念 0.218 |
| 重點 | 0.256 | 對象 0.249 範圍 0.249 |
| 要求 | 0.218 | |

# SketchDifference

詞義與使用分析對比的最佳利器

- 比對兩個詞的素描，
- 從共有的關係詞中看出分布使用趨勢
- 從特有共現詞中看出特性

# 辭典的定義

「明星」和「演員」
有甚麼不同？

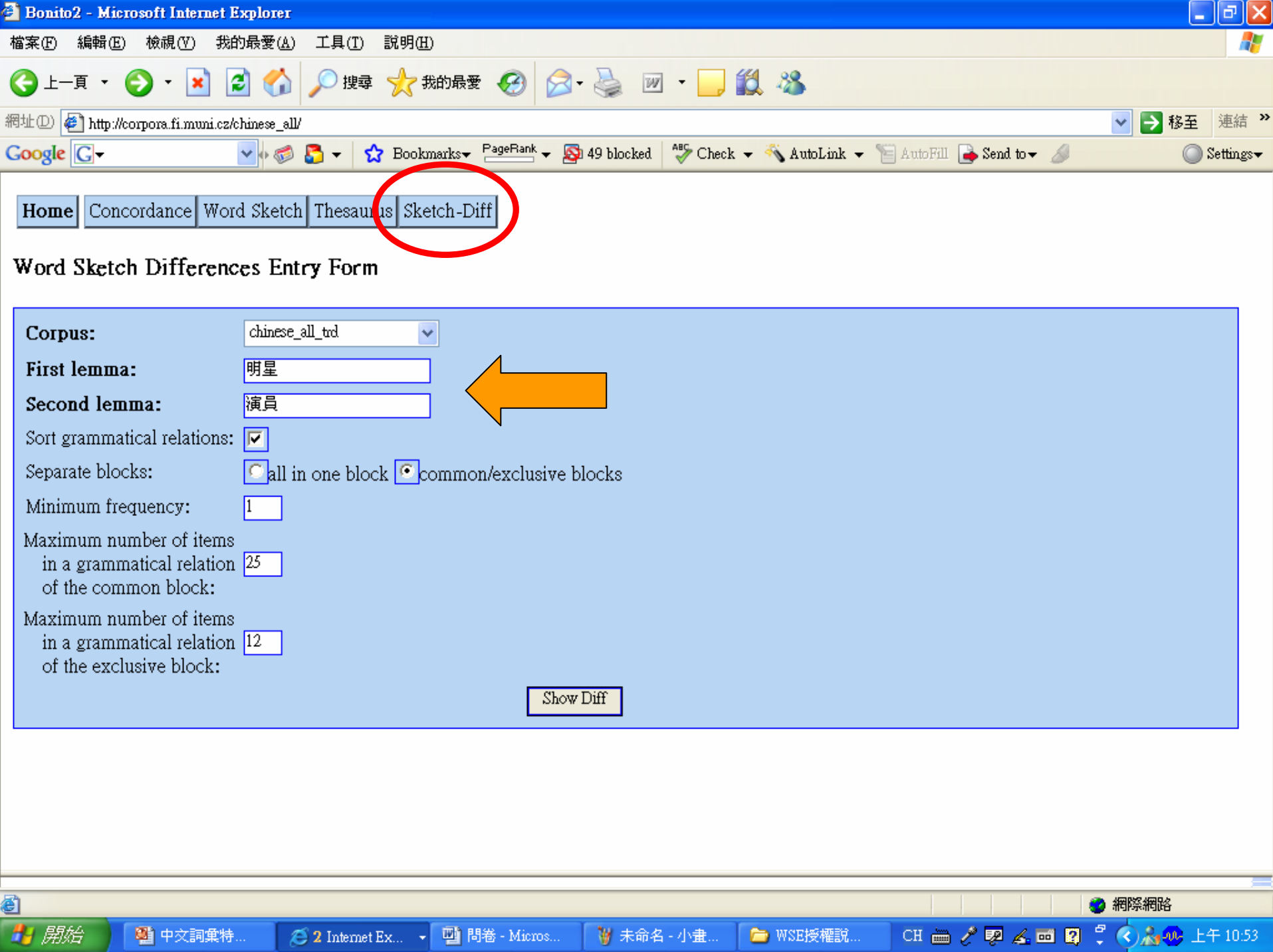
- 明星
- 比喻傑出的人物。

- 演員
- 從事演藝工作的人員。

# 語言的使用

## 「明星」和「演員」

- 什麼時候可以互換？
- 什麼時候互換會改變意義？
- 什麼時候不能互換？


Bonito2 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/
移至 連結
Google
Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
Home
Concordance
Word Sketch
Thesaurus
Sketch-Diff
Word Sketch Differences Entry Form
Corpus:
chinese_all_trd
First lemma:
明星
Second lemma:
演員
Sort grammatical relations:
Separate blocks:
all in one block
common/exclusive blocks
Minimum frequency:
1
Maximum number of items in a grammatical relation of the common block:
25
Maximum number of items in a grammatical relation of the exclusive block:
12
Show Diff
網際網路
開始
中文詞彙特...
2 Internet Ex...
問卷 - Micros...
未命名 - 小畫...
WSE授權說...
CH
上午 10:53

# 明星/演員 chinese_all_trd freq = 23923/23213

change options

## Common patterns

| 明星 | 21 | 14 | 7 | 0 | -7 | -14 | -21 | 演員 |
|---|---|---|---|---|---|---|---|---|

| a_modifier | 1729 | 4116 | 1.7 | 4.2 |
|---|---|---|---|---|
| 著名 | 84 | 738 | 39.0 | 70.4 |
| 武打 | 65 | 14 | 62.6 | 31.1 |
| 老牌 | 42 | 110 | 44.5 | 56.4 |
| 資深 | 6 | 237 | 11.2 | 53.5 |
| 知名 | 29 | 232 | 26.7 | 52.4 |
| 當紅 | 43 | 30 | 49.4 | 38.7 |
| 大 | 468 | 19 | 46.6 | 5.1 |
| 年輕 | 8 | 136 | 13.3 | 43.8 |
| 最佳 | 27 | 135 | 23.2 | 39.5 |
| 小 | 38 | 232 | 20.5 | 37.9 |
| 老 | 21 | 112 | 19.6 | 35.5 |
| 眾多 | 44 | 16 | 33.8 | 17.5 |
| 一流 | 9 | 22 | 18.5 | 25.4 |
| 名 | 11 | 198 | 4.6 | 25.3 |
| 新一代 | 12 | 13 | 22.9 | 20.6 |
| 已故 | 11 | 7 | 22.5 | 14.9 |

| measure | 688 | 1532 | 1.1 | 2.5 |
|---|---|---|---|---|
| 位 | 310 | 576 | 50.3 | 54.0 |
| 名 | 79 | 617 | 26.4 | 50.4 |
| 批 | 13 | 39 | 17.3 | 25.7 |
| 個 | 79 | 159 | 21.3 | 23.9 |
| 屆 | 13 | 7 | 12.6 | 5.7 |
| 場 | 11 | 5 | 12.5 | 4.9 |
| 次 | 9 | 17 | 8.1 | 9.5 |

| possessor | 459 | 921 | 1.0 | 2.1 |
|---|---|---|---|---|
| 知名度 | 6 | 10 | 17.6 | 21.1 |
| 國家 | 9 | 26 | 4.8 | 8.6 |

| and/or | 336 | 1517 | 0.4 | 1.7 |
|---|---|---|---|---|
| 導演 | 11 | 381 | 21.3 | 68.4 |
| 歌星 | 16 | 50 | 32.2 | 42.6 |
| 歌手 | 7 | 50 | 17.0 | 34.8 |
| 藝術家 | 6 | 36 | 15.5 | 30.3 |

| n_modifier | 13964 | 9897 | 1.6 | 1.2 |
|---|---|---|---|---|
| 大牌 | 162 | 13 | 57.7 | 21.6 |
| 偶像 | 297 | 8 | 57.1 | 11.7 |
| 喜劇 | 50 | 144 | 33.8 | 52.8 |
| 影視 | 248 | 64 | 50.8 | 32.3 |
| 好萊塢 | 190 | 70 | 50.5 | 37.0 |
| 新生代 | 25 | 133 | 24.0 | 50.3 |
| 男 | 41 | 355 | 16.1 | 46.6 |
| 演技派 | 31 | 37 | 39.8 | 44.5 |
| 歌仔戲 | 11 | 83 | 15.4 | 42.6 |
| 芭蕾舞 | 7 | 63 | 13.6 | 42.5 |
| 青年 | 26 | 564 | 5.4 | 41.7 |
| 舞蹈 | 12 | 214 | 7.9 | 41.4 |
| 電影 | 343 | 301 | 38.5 | 39.4 |
| 歌劇 | 11 | 64 | 13.4 | 35.2 |
| 男女 | 28 | 112 | 14.1 | 31.7 |
| 女 | 234 | 18 | 30.9 | 7.2 |

# 明星/演員 chinese_all_trd freq = 23923/23213

change options

## Common patterns

| 明星 | 21 | 14 | 7 | 0 | -7 | -14 | -21 | 演員 |
|---|---|---|---|---|---|---|---|---|

| a_modifier | 1729 | 4116 | 1.7 | 4.2 |
|---|---|---|---|---|
| 著名 | 84 | 738 | 39.0 | 70.4 |
| 武打 | 65 | 14 | 62.6 | 31.1 |
| 老牌 | 42 | 110 | 44.5 | 56.4 |
| 資深 | 6 | 237 | 11.2 | 53.5 |
| 知名 | 29 | 232 | 26.7 | 52.4 |
| 當紅 | 43 | 30 | 49.4 | 38.7 |
| 大 | 468 | 19 | 46.6 | 5.1 |
| 年輕 | 8 | 136 | 13.3 | 43.8 |
| 最佳 | 27 | 135 | 23.2 | 39.5 |
| 小 | 38 | 232 | 20.5 | 37.9 |
| 老 | 21 | 112 | 19.6 | 35.5 |
| 眾多 | 44 | 16 | 33.8 | 17.5 |
| 一流 | 9 | 22 | 18.5 | 25.4 |
| 名 | 11 | 198 | 4.6 | 25.3 |
| 新一代 | 12 | 13 | 22.9 | 20.6 |
| 已故 | 11 | 7 | 22.5 | 14.9 |

| measure | 688 | 1532 | 1.1 | 2.5 |
|---|---|---|---|---|
| 位 | 310 | 576 | 50.3 | 54.0 |
| 名 | 79 | 617 | 26.4 | 50.4 |
| 批 | 13 | 39 | 17.3 | 25.7 |
| 個 | 79 | 159 | 21.3 | 23.9 |
| 屆 | 13 | 7 | 12.6 | 5.7 |
| 場 | 11 | 5 | 12.5 | 4.9 |
| 次 | 9 | 17 | 8.1 | 9.5 |

| possessor | 459 | 921 | 1.0 | 2.1 |
|---|---|---|---|---|
| 知名度 | 6 | 10 | 17.6 | 21.1 |
| 國家 | 9 | 26 | 4.8 | 8.6 |

| and/or | 336 | 1517 | 0.4 | 1.7 |
|---|---|---|---|---|
| 導演 | 11 | 381 | 21.3 | 68.4 |
| 歌星 | 16 | 50 | 32.2 | 42.6 |
| 歌手 | 7 | 50 | 17.0 | 34.8 |
| 藝術家 | 6 | 36 | 15.5 | 30.3 |

| n_modifier | 13964 | 9897 | 1.6 | 1.2 |
|---|---|---|---|---|
| 大牌 | 162 | 13 | 57.7 | 21.6 |
| 偶像 | 297 | 8 | 57.1 | 11.7 |
| 喜劇 | 50 | 144 | 33.8 | 52.8 |
| 影視 | 248 | 64 | 50.8 | 32.3 |
| 好萊塢 | 190 | 70 | 50.5 | 37.0 |
| 新生代 | 25 | 133 | 24.0 | 50.3 |
| 男 | 41 | 355 | 16.1 | 46.6 |
| 演技派 | 31 | 37 | 39.8 | 44.5 |
| 歌仔戲 | 11 | 83 | 15.4 | 42.6 |
| 芭蕾舞 | 7 | 63 | 13.6 | 42.5 |
| 青年 | 26 | 564 | 5.4 | 41.7 |
| 舞蹈 | 12 | 214 | 7.9 | 41.4 |
| 電影 | 343 | 301 | 38.5 | 39.4 |
| 歌劇 | 11 | 64 | 13.4 | 35.2 |
| 男女 | 28 | 112 | 14.1 | 31.7 |
| 女 | 234 | 18 | 30.9 | 7.2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 青 | 14 | 22.6 | 戲 | 8 | 18.0 | 歌唱家 | 12 25.7 |

## "演員" only patterns

| possession 693 | | 1.6 |
|---|---|---|
| 演技 | 15 | 33.7 |
| 演出 | 18 | 21.9 |
| 手 | 21 | 21.8 |
| 功力 | 6 | 20.8 |
| 服裝 | 13 | 19.6 |
| 積極性 | 9 | 17.6 |
| 特質 | 5 | 15.9 |
| 表現 | 13 | 15.4 |
| 肢體 | 5 | 14.9 |
| 角色 | 8 | 14.4 |
| 技巧 | 5 | 13.8 |
| 動作 | 6 | 12.3 |

| subject_of 3924 | | 1.6 |
|---|---|---|
| 謝幕 | 20 | 44.9 |
| 擔綱 | 33 | 41.7 |
| 親切 | 23 | 30.1 |
| 跳起 | 11 | 29.5 |
| 演戲 | 9 | 29.4 |
| 演 | 22 | 29.1 |
| 握手 | 19 | 28.2 |
| 齊聚一堂 | 14 | 27.3 |
| 載歌載舞 | 10 | 26.3 |
| 清唱 | 8 | 24.4 |
| 拍戲 | 6 | 24.3 |
| 精湛 | 13 | 24.1 |

| n_modifier 9897 | | 1.2 |
|---|---|---|
| 京劇 | 246 | 53.7 |
| 一級 | 424 | 50.8 |
| 相聲 | 120 | 50.4 |
| 實力派 | 71 | 49.1 |
| 雜技 | 129 | 48.5 |
| 替身 | 33 | 43.7 |
| 特技 | 92 | 43.1 |
| 特型 | 21 | 42.4 |
| 越劇 | 45 | 41.0 |
| 話劇 | 59 | 38.1 |
| 舞台劇 | 45 | 37.4 |
| 歌舞伎 | 21 | 36.9 |

| modifies 8784 | | 1.0 |
|---|---|---|
| 郎雄 | 51 | 50.1 |
| 李康生 | 29 | 42.4 |
| 柯俊雄 | 29 | 42.3 |
| 金士傑 | 22 | 39.9 |
| 柯受良 | 21 | 37.7 |
| 胡軍 | 18 | 36.6 |
| 姜昆 | 19 | 36.0 |
| 鄭亞雲 | 15 | 35.9 |
| 楊貴媚 | 20 | 35.3 |
| 戴立忍 | 19 | 34.2 |
| 楊呈偉 | 14 | 33.9 |
| 濮存昕 | 13 | 33.8 |

| object_of 1535 | | 0.7 |
|---|---|---|
| 獨唱 | 50 | 54.1 |
| 客串 | 22 | 41.9 |
| 演 | 23 | 34.0 |

## "明星" only patterns

| a_modifier 1729 1.7 | | |
|---|---|---|
| 超級 | 219 | 63.7 |
| 演藝 | 78 | 51.2 |
| 美式 | 46 | 48.6 |
| 耀眼 | 26 | 46.1 |
| 閃亮 | 18 | 40.3 |
| 過氣 | 10 | 36.9 |
| 超冷 | 5 | 30.6 |
| 三棲 | 6 | 29.8 |
| 佳 | 21 | 29.3 |
| 頭號 | 19 | 27.0 |
| 溫 | 12 | 24.9 |
| 簡 | 10 | 24.6 |

| n_modifier 13964 1.6 | | |
|---|---|---|
| 籃球 | 433 | 46.3 |
| 職籃 | 179 | 45.6 |
| 足[illegible] | 675 | 44.9 |
| [illegible] | 232 | 43.0 |
| 卡[illegible] | 99 | 41.9 |
| 夢幻 | 68 | 40.8 |
| 網球 | 293 | 39.2 |
| 抗癌 | 50 | 34.9 |
| 偶像級 | 16 | 34.7 |
| 恆康 | 15 | 34.7 |
| 演藝圈 | 36 | 33.8 |
| 演藝界 | 34 | 33.8 |

| modifies 13046 1.4 | | |
|---|---|---|
| 球員 | 1095 | 59.7 |
| 對抗賽 | 346 | 59.3 |
| 聯隊 | 301 | 54.2 |
| 排名賽 | 108 | 49.0 |
| 辛浦森 | 56 | 48.4 |
| 籃球隊 | 167 | 46.9 |
| 馬拉杜納 | 30 | 43.0 |
| 白隊 | 47 | 41.2 |
| 後衛 | 141 | 40.0 |
| 前鋒 | 149 | 38.8 |
| 馬拉多納 | 39 | 38.4 |
| 王貞治 | 30 | 37.8 |

| possession 506 1.1 | | |
|---|---|---|
| 風采 | 26 | 37.4 |
| 架子 | 12 | 32.2 |
| 架勢 | 10 | 32.1 |
| 搖籃 | 12 | 30.2 |
| 名字 | 12 | 24.4 |
| 架式 | 5 | 23.8 |
| 丰采 | 5 | 21.8 |
| 光環 | 5 | 20.2 |
| 魅力 | 7 | 18.7 |
| 照片 | 9 | 18.5 |
| 故事 | 7 | 16.3 |
| 青少年 | 6 | 10.7 |

| measure 688 1.1 | | |
|---|---|---|
| 顆 | 79 | 51.3 |
| 隻 | 7 | 15.8 |
| 路 | 6 | 10.8 |
| 所 | 15 | 10.2 |
| 家 | 7 | 8.2 |

| possessor 459 1.0 | | |
|---|---|---|
| 世界級 | 11 | 25.3 |
| 名氣 | 5 | 19.9 |
| 比賽 | 21 | 17.1 |
| 聯隊 | 6 | 16.9 |
| 氣 | 5 | 14.3 |

| subject_of 2280 0.9 | | |
|---|---|---|
| 薈萃 | 53 | 49.3 |
| 雲集 | 36 | 39.5 |
| 三缺一 | 8 | 37.4 |
| 開店 | 12 | 34.6 |
| 評選 | 26 | 32.8 |

| object_of 1536 0.7 | | |
|---|---|---|
| 啓 | 44 | 48.8 |
| 崇拜 | 19 | 34.0 |
| 棲 | 11 | 33.2 |
| 考上 | 16 | 31.8 |
| 服務 | 93 | 29.0 |

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff | Frequency | Collocation

KWIC/Sentence | View options | Sample | Filter | Sort | Save

Corpus: **chinese_all_trd**
Hits: **433**
conc description

First | Previous | Page 2 of 22 Go | Next | Last

CNA19920506.0058 鍾錦隆 台北 六日 電) 韓國 前 女子 籃球 明星 朴贊淑 認爲 ， 企業界 對 球隊 的 支持 不
CNA19920506.0058 球員 ， 獲 觀眾 票選 爲最 受 歡迎 的 籃球 明星 。 曾 多 次 代表 南韓 女籃 出賽 的
CNA19920511.0029 去年 賺得 一千五百萬 美元 。 籃球 明星 喬丹 的 收入 爲 一千六百萬 美元 ， 使 他
CNA19920605.0110 愛滋病毒 而 退出 洛杉磯 湖人隊 的 前 美國 籃球 明星 「 魔術 」 強生 ， 昨晚 喜獲 麟兒 。
CNA19920712.0164 第十五 屆 威廉 瓊斯盃 國際 籃球 邀請賽 男子 明星 球員 ， 除 二 名 地主 球員 外 ， 其餘 十
CNA19920727.0158 奧運 第一 天 的 比賽 結果 ， 美國 男子 籃球 明星 代表隊 第一 場 對 安哥拉隊 的 比賽 ， 以
CNA19920728.0285 巴塞隆納 二十八日 法 新 電) 美國 籃球 明星 魔術 強森 ， 右膝 的 韌帶 受傷 之後 ，
CNA19920801.0284 讓 美國 那些 百萬 美元 身價 的 超級 籃球 明星 ， 參加 奧運 競賽 。 但是 施密特 不以爲然
CNA19920804.0252 了 。 昨天 這 個 由 美國 職業 籃球 明星 組成 的 球隊 ， 在 奧運會場 舉行 了 與賽
CNA19920805.0039 奧林匹克 籃球隊 的 喬丹 與 另外 五 位 NBA 籃球 明星 卻 與 耐吉 公司 有約 ， 使用 耐吉 公司
CNA19920926.0159 合眾 國際 電) 感染 愛滋病毒 的 美國 籃球 明星 「 魔術 」 強生 ， 昨天 退出 美國 國家
CNA19920930.0144 二十九日 合眾 國際 電) 美國 職業 籃球 明星 「 魔術 」 強生 今天 宣布 ， 他 將 在 本
CNA19920930.0161 二十九日 合眾 國際 電) 美國 職業 籃球 明星 「 魔術 」 強生 今天 宣布 ， 他 將 在 本
CNA19930111.0159 王同禹 台北 十一日 電) 今年 的 甲組 籃球 明星 對抗賽 將 於 本 月 十七日 下午 二時 ，
CNA19930621.0111 葛瑞菲絲 · 卓納 與 前 國會 議員 兼 籃球 明星 湯姆·麥米倫 主持 白宮 體育 委員會 的 業務
CNA19930718.0199 周揚力 林口 十八日 電) 瓊斯盃 籃球 邀請賽 明星 球員 出爐 ， 獲 大會 最 有 價值 球員 的
CNA19930724.0200 尤恩 是 第三 位 來華 訪問 的 美國 籃球 明星 、 也 是 第一 位 訪華 的 NBA 現役 球員
CNA19930814.0062 聯邦 調查局 官員 今天 說 ， 美國 超級 籃球 明星 邁克喬登 的 父親 ， 可能 在 綁票 事件 中
CNA19930814.0062 歲 的 詹姆斯 喬登 ， 芝加哥 公牛隊 籃球 明星 邁克喬登 的 父親 。 詹姆斯 胸部 中彈
CNA19930815.0130 職業 棒球 選手 謝長亨 、 王光熙 與 籃球 明星 鄭志龍 、 錢薇娟 ， 今天 聯手 在 台中 主持

# 吃/喝

chinese_all_trd freq = 53654/19561

change options

## Common patterns

| 吃 | 21 | 14 | 7 | 0 | -7 | -14 | -21 | 喝 |
|---|---|---|---|---|---|---|---|---|

| pp_假 | 2 | 6 | 3.4 | 25.4 |
|---|---|---|---|---|
| 酒 | 1 | 3 | 9.0 | 18.0 |

| object | 23421 | 13015 | 6.9 | 9.5 |
|---|---|---|---|---|
| 酒 | 8 | 4756 | 5.4 | 106.0 |
| 花酒 | 4 | 639 | 9.5 | 91.1 |
| 茶 | 1 | 748 | 0.6 | 74.8 |
| 藥 | 1430 | 8 | 72.0 | 7.2 |
| 春酒 | 5 | 107 | 14.6 | 62.7 |
| 牛奶 | 18 | 291 | 17.4 | 61.2 |
| 早餐 | 375 | 1 | 59.3 | 1.6 |
| 東西 | 596 | 16 | 56.1 | 12.5 |
| 消夜 | 75 | 1 | 53.5 | 4.8 |
| 喜酒 | 6 | 34 | 18.8 | 46.3 |
| 湯圓 | 84 | 1 | 45.8 | 3.3 |
| 奶 | 143 | 92 | 45.7 | 42.7 |
| 飲料 | 4 | 183 | 3.2 | 43.6 |
| 食物 | 293 | 4 | 42.6 | 3.5 |
| 湯 | 1 | 59 | 2.3 | 40.9 |
| 口水 | 2 | 51 | 5.1 | 40.1 |
| 稀飯 | 36 | 14 | 39.3 | 28.1 |
| 奶茶 | 1 | 36 | 3.3 | 38.8 |
| 農藥 | 4 | 134 | 2.9 | 37.8 |
| 習慣 | 155 | 149 | 33.2 | 36.9 |
| 巧克力 | 64 | 1 | 36.1 | 2.4 |
| 麵條 | 40 | 1 | 36.1 | 3.6 |
| 零嘴 | 19 | 1 | 33.6 | 5.4 |
| 豆漿 | 3 | 24 | 9.0 | 33.3 |
| 奶水 | 18 | 19 | 29.9 | 33.2 |

| modifier | 21712 | 6092 | 4.5 | 3.1 |
|---|---|---|---|---|
| 少 | 400 | 87 | 49.5 | 35.3 |
| 多 | 1173 | 281 | 47.5 | 36.7 |
| 一口 | 69 | 85 | 35.0 | 46.2 |
| 同 | 379 | 4 | 45.3 | 4.7 |
| 過 | 1181 | 200 | 45.3 | 30.1 |
| 不 | 3429 | 1011 | 42.0 | 36.2 |
| 常 | 227 | 97 | 38.4 | 35.2 |
| 沒 | 324 | 81 | 38.0 | 28.2 |
| 天天 | 77 | 23 | 37.0 | 27.3 |
| 連 | 177 | 21 | 36.4 | 17.9 |
| 了 | 4425 | 1125 | 36.2 | 29.3 |
| 倒 | 116 | 1 | 35.9 | 1.7 |
| 一起 | 318 | 158 | 35.3 | 35.2 |
| 不要 | 226 | 109 | 31.1 | 30.6 |
| 怎麼 | 78 | 6 | 29.2 | 9.4 |
| 不准 | 38 | 38 | 22.2 | 28.9 |
| 給他 | 56 | 10 | 28.6 | 15.4 |
| 不能 | 284 | 71 | 28.2 | 20.6 |
| 著 | 563 | 247 | 27.4 | 27.4 |
| 邊 | 146 | 26 | 27.2 | 15.8 |
| 去 | 190 | 94 | 26.8 | 27.0 |
| 誤 | 15 | 18 | 19.4 | 26.8 |
| 要 | 752 | 186 | 26.7 | 20.2 |
| 很少 | 44 | 15 | 26.4 | 20.0 |
| 不用 | 48 | 4 | 26.2 | 8.0 |

# Objects Shared by Chi1 and He1

吃/喝 chinese_all_trd_test freq = 53654/19561

Common patterns

| 吃 | 21 | 14 | 7 | 0 | -7 | -14 | -21 | 喝 |
|---|---|---|---|---|---|---|---|---|

| SentObject_of | 3859 | 1065 | 7.2 | 4.9 |
|---|---|---|---|---|
| 喜歡 | 557 | 173 | 69.7 | 57.7 |
| 試 | 371 | 22 | 68.2 | 29.2 |
| 愛 | 571 | 185 | 66.0 | 55.5 |
| 拒 | 167 | 6 | 55.5 | 15.4 |
| 嗜 | 70 | 9 | 55.5 | 27.4 |
| 顧不上 | 63 | 12 | 53.0 | 31.3 |
| 敢 | 168 | 36 | 47.1 | 31.9 |
| 捨不得 | 39 | 8 | 42.8 | 24.6 |
| 請 | 216 | 44 | 39.8 | 26.5 |
| 喜愛 | 45 | 24 | 32.7 | 30.6 |
| 怕 | 49 | 10 | 32.0 | 18.6 |
| 放心 | 29 | 6 | 30.5 | 16.6 |
| 喜 | 36 | 7 | 30.0 | 16.2 |
| 涉嫌 | 9 | 44 | 8.2 | 29.4 |
| 知道 | 53 | 24 | 25.2 | 22.6 |
| 喜好 | 10 | 10 | 20.7 | 25.1 |

| Object | 33038 | 16684 | 3.7 | 4.6 |
|---|---|---|---|---|
| 酒 | 13 | 5198 | 7.5 | 106.9 |
| 茶 | 7 | 825 | 7.0 | 75.5 |
| 藥 | 1558 | 8 | 73.0 | 7.3 |
| 牛奶 | 24 | 386 | 19.4 | 65.6 |
| 春酒 | 5 | 119 | 13.6 | 63.0 |
| 東西 | 639 | 16 | 53.3 | 11.1 |
| 食物 | 610 | 6 | 52.9 | 5.0 |
| 喜酒 | 6 | 43 | 17.7 | 48.8 |
| 奶 | 160 | 106 | 46.5 | 44.6 |
| 頓 | 167 | 5 | 43.9 | 8.0 |
| 稀飯 | 47 | 23 | 41.5 | 33.9 |
| 水 | 16 | 320 | 2.4 | 35.5 |
| 習慣 | 181 | 165 | 31.5 | 35.3 |
| 碗 | 75 | 19 | 35.2 | 21.4 |
| 奶水 | 18 | 20 | 29.8 | 34.3 |
| 母奶 | 24 | 20 | 32.3 | 32.7 |

| Modifier | 13757 | 4501 | 4.5 | 3.7 |
|---|---|---|---|---|
| 少 | 440 | 95 | 65.6 | 46.5 |
| 多 | 1289 | 304 | 61.7 | 46.6 |
| 同 | 384 | 5 | 52.2 | 7.4 |
| 不 | 1885 | 909 | 45.7 | 45.1 |
| 一起 | 317 | 159 | 44.1 | 42.0 |
| 大口 | 5 | 24 | 17.6 | 44.1 |
| 常 | 198 | 94 | 43.3 | 39.7 |
| 天天 | 76 | 23 | 42.8 | 30.9 |
| 沒 | 307 | 79 | 42.8 | 31.2 |
| 邊 | 145 | 27 | 41.7 | 25.1 |
| 連 | 179 | 20 | 41.1 | 19.4 |
| 只 | 363 | 128 | 37.0 | 31.1 |
| 不要 | 214 | 110 | 36.8 | 35.3 |
| 給他 | 54 | 10 | 34.9 | 18.9 |
| 不能 | 258 | 71 | 34.3 | 25.3 |
| 著 | 177 | 36 | 34.0 | 21.2 |

# “Chi1” only patterns

"吃" only patterns

| SentObject_of 3859 7.2 | | | Modifier 13757 4.5 | | | Subject 11519 4.3 | | | Object 33038 3.7 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 愁 | 103 | 57.9 | 倒 | 128 | 41.6 | 飯 | 718 | 78.7 | 敗仗 | 326 | 72.3 |
| 講究 | 27 | 32.5 | 津津有味 | 19 | 36.8 | 啞巴 | 30 | 42.0 | 晚飯 | 310 | 71.6 |
| 嚐試 | 13 | 26.5 | 怎麼 | 78 | 35.0 | 最愛 | 72 | 39.4 | 飯 | 802 | 71.0 |
| 忌 | 8 | 25.9 | 硬 | 41 | 32.5 | 柿子 | 27 | 33.5 | 定心丸 | 211 | 68.0 |
| 寧可 | 16 | 25.3 | 有得 | 20 | 31.8 | 糖 | 45 | 31.5 | 午飯 | 241 | 67.5 |
| 捨得 | 11 | 24.8 | 不用 | 47 | 31.0 | 金飯碗 | 14 | 31.5 | 大鍋飯 | 245 | 66.6 |
| 擔心 | 46 | 24.1 | 按時 | 35 | 29.4 | 全家 | 44 | 31.1 | 閉門羹 | 173 | 66.5 |
| 拒絕 | 45 | 22.9 | 常年 | 28 | 27.4 | 魚 | 73 | 30.0 | 年夜飯 | 270 | 65.8 |
| 寧願 | 11 | 21.2 | 該 | 47 | 25.2 | 東西 | 78 | 29.4 | 狗肉 | 190 | 61.5 |
| 討厭 | 7 | 21.0 | 年年 | 24 | 24.5 | 啞吧 | 11 | 29.4 | 肉 | 488 | 60.1 |
| 忘 | 13 | 20.8 | 一律 | 31 | 23.3 | 們同 | 10 | 29.2 | 虧 | 329 | 59.2 |
| 記得 | 12 | 20.3 | 著實 | 13 | 23.3 | 金碗 | 9 | 27.7 | 頓飯 | 84 | 59.2 |

| PP_在 165 1.5 | | | Modifies 1527 0.1 | | |
|---|---|---|---|---|---|
| 工地 | 17 | 32.3 | 東西 | 109 | 47.6 |

# “He1” only patterns

"喝" only patterns

| SentObject_of 1065 | | 4.9 |
|---|---|---|
| 涉 | 8 | 17.7 |
| 盛行 | 5 | 17.6 |
| 疑 | 5 | 15.0 |

| Object 16684 | | 4.6 |
|---|---|---|
| 花酒 | 666 | 90.8 |
| 咖啡 | 969 | 74.3 |
| 啤酒 | 421 | 58.5 |
| 下午茶 | 101 | 57.1 |
| 白開水 | 67 | 54.3 |
| 口水 | 114 | 52.5 |
| 開水 | 108 | 51.2 |
| 杯 | 237 | 50.3 |
| 飲料 | 283 | 49.6 |
| 紅酒 | 83 | 49.6 |
| 花酒案 | 33 | 49.3 |
| 綠茶 | 88 | 48.7 |

| Modifier 4501 | | 3.7 |
|---|---|---|
| 厲聲 | 13 | 35.5 |
| 猛 | 17 | 21.5 |
| 成天 | 5 | 19.7 |
| 獨自 | 7 | 16.8 |
| 有沒有 | 7 | 16.7 |
| 一口 | 5 | 16.2 |
| 盡情 | 6 | 15.9 |
| 其實 | 5 | 13.0 |
| 太 | 7 | 12.8 |
| 難免 | 5 | 12.7 |
| 總共 | 5 | 9.0 |
| 並 | 12 | 3.6 |

| Subject 3235 | | 3.0 |
|---|---|---|
| 開水 | 19 | 32.3 |
| 口水 | 16 | 30.0 |
| 胡吃海 | 5 | 28.5 |
| 馬永成 | 7 | 19.9 |
| 蕭敦仁 | 5 | 19.3 |
| 友人 | 17 | 18.9 |
| 校長 | 30 | 18.1 |
| 自來水 | 17 | 17.7 |
| 粥 | 5 | 17.6 |
| 督學 | 6 | 15.8 |
| 牛奶 | 7 | 15.8 |
| 檢察官 | 20 | 14.8 |

| Modifies 306 | | 0.1 |
|---|---|---|
| 水 | 44 | 34.5 |
| 飲料 | 15 | 28.6 |

# 素描對比也可用於兩岸詞彙對比研究

- 利用CGW包含來自大陸與台灣兩個次語料庫的特性，

# 大陸和台灣用語比較

- 出租車

台灣 109筆

大陸

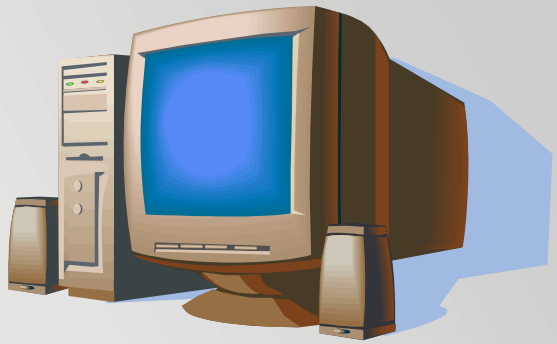
- 公安

台灣 18,641筆

大陸

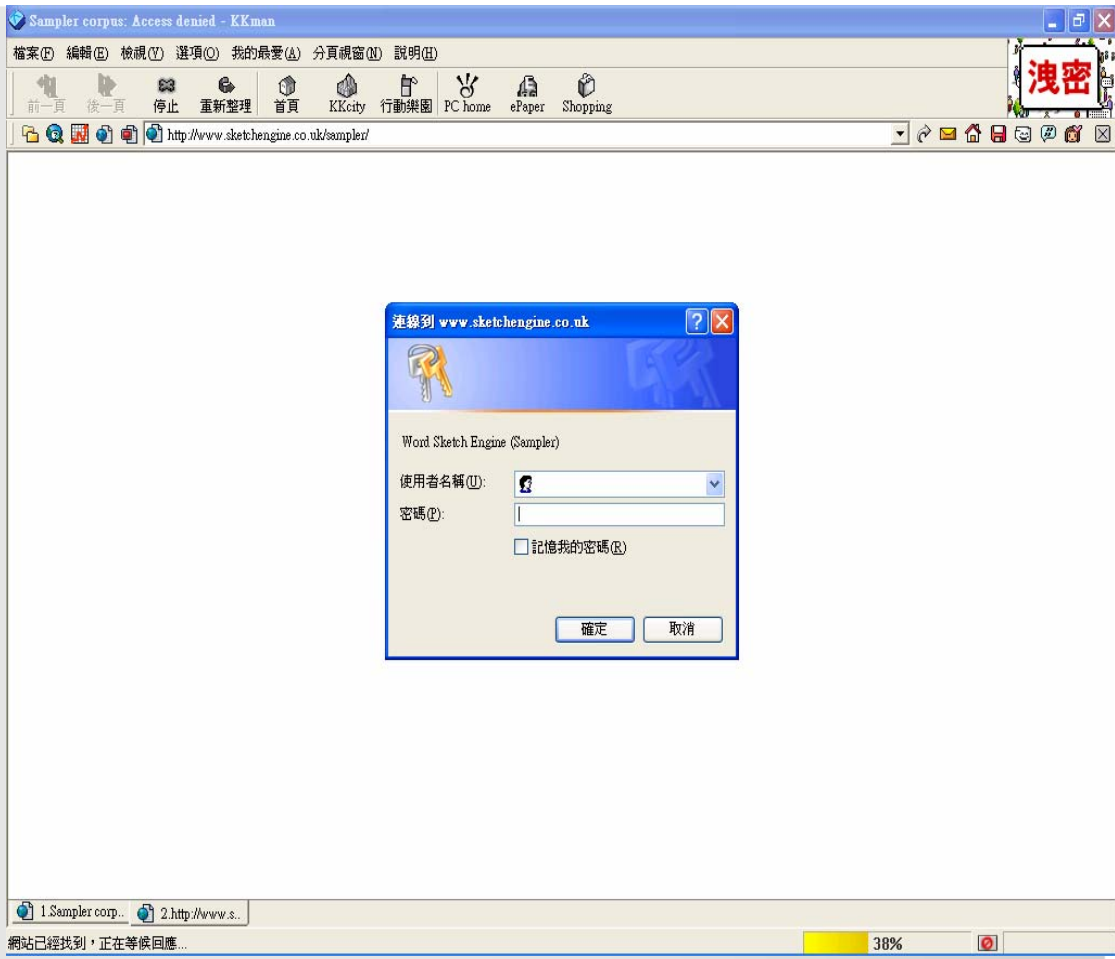
# 系統運用

- 國文學習
- 提供查詢功能
- 減少詞彙的誤用
- 輔助外國人學國文
- 大陸和台灣用語比較



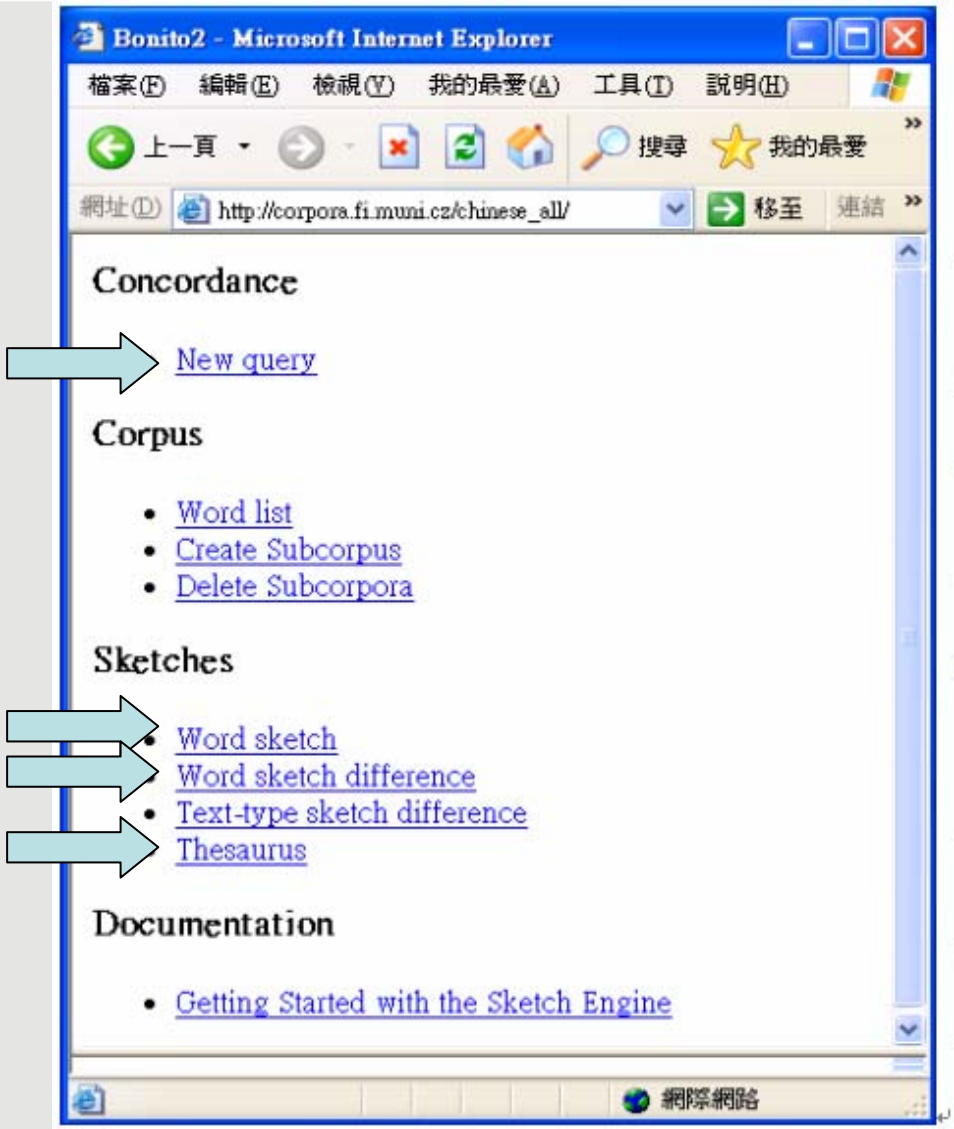
# 帳號申請

- 免費
- 使用者帳號和密碼
- 請跟我們聯絡



- 中研院版本
- http://140.109.150.73/wse/
- 帳號: guest
- 密碼: try2guest
- 捷克版本
- http://corpora.fi.muni.cz/chinese/
- 帳號: chinese
- 密碼: chinese

# 操作流程- 功能列表




檢索關鍵詞或詞組的語料

對語料庫的進階處理

檢索詞彙的進階相關訊息

其他相關訊息與使用說明

中央研究院語言學研究所
「中文詞彙特性速描系統」帳號使用切結書
【使用內容及操作方式說明】
軟體名稱：__中文詞彙特性速描系統_____
【使用對象】
學齡以上民眾得申請使用
【使用方式】

1. 免費提供學術研究或學習使用，但不包括違法及營利之使用
2. 中央研究院語言學研究所保留授權帳號之所有權
3. 論文發表或引用時，請註明資料取得來源：

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. October 14-15. Jeju, Korea.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell. 2004. The Sketch Engine. Proceedings of EURALEX, Lorient, France.

Chinese Sketch Engine

http://wordsketch.ling.sinica.edu.tw

禁止部份或全部轉移給其它第三者使用

申請單位(所屬單位)：
申請人：
計畫名稱（申請目的）：
聯絡電話：
日　　期：　　年　　月　　日

# 中央研究院語言學研究所
# 「中文詞彙特性速描系統」帳號申請表

帳號 (本欄由承辦人輸入)中文姓名英文姓名單位□個人申請　職別**E-mail Address**聯絡電話異動備註(本欄由承辦人輸入)(本欄由系統管理員輸入)※申請核准後將由系統自動產生一組帳號密碼，並以電子郵件寄送至您的信箱，帳號密碼皆無法修改且僅供單人使用，請妥善保存

申請人簽名：＿＿＿＿＿＿＿＿＿＿＿＿　承辦人：＿＿＿＿＿＿＿＿＿＿＿＿

申請日期＿＿＿＿年＿＿＿＿月＿＿＿＿日

---

異動紀錄

承辦人員：＿＿＿＿＿＿＿＿＿＿＿＿處理日期：＿＿＿＿＿＿＿＿＿＿＿＿處理結果：

# 簡報完畢

謝謝

- SINICA BOW
- http://bow.sinica.edu.tw
- Chinese  Wordnet
- http://ling.sinica.edu.tw/cwn

# 美白/解酒 chinese_all_trd freq = 542/144

change options

## Common patterns

| 美白 | 21 | 14 | 7 | 0 | -7 | -14 | -21 | 解酒 |
|---|---|---|---|---|---|---|---|---|

| modifies | 228 | 76 | 2.1 | 2.5 |
|---|---|---|---|---|
| 效果 | 30 | 5 | 33.4 | 15.6 |
| 功效 | 8 | 5 | 24.9 | 21.9 |
| 產品 | 20 | 18 | 18.8 | 22.4 |

## "美白" only patterns

| subject | 104 | 3.4 |
|---|---|---|
| 牙齒 | 23 | 46.0 |
| 義 | 13 | 32.5 |
| 雷射 | 10 | 28.9 |
| 皮膚 | 7 | 22.9 |

| modifies | 228 | 2.1 |
|---|---|---|
| 乳霜 | 12 | 48.1 |
| 保養品 | 17 | 45.2 |
| 面膜 | 6 | 31.6 |
| 化妝品 | 12 | 29.8 |
| 成份 | 11 | 26.4 |
| 霜 | 5 | 23.8 |
| 用品 | 5 | 14.7 |
| 商品 | 5 | 10.1 |