語料庫應用於教材編纂

語文教育及編譯研究中心 吳鑑城

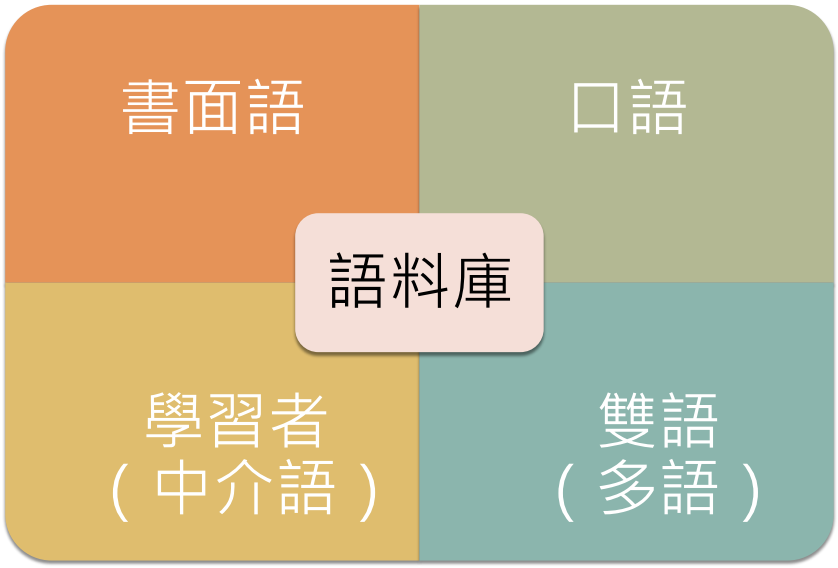
語料庫應用於教材編纂

- 語料庫概念介紹
- 語料庫技術介紹
- 語料庫工具應用於教材編纂介紹
  - 中文分詞工具
  - 華語文錯字自動偵測系統(雛型)
  - 索引典系統
  - 語義場工具

語料庫？有人用嗎？

- 藉學術語料庫提出中文學術常用詞表: 以人文社會科學為例
  - 劉貞妤，陳浩然，楊惠媚，2016
- 基於語料庫的英語專業教材詞彙研究
  - 宋曉舟，2016
- 使用專業語料庫增加籃球英語教材詞彙及搭配詞的豐富度
  - 高睿禹，2017

“語料庫是為了特定目的，根據特定原則搜集或取樣，並分門別類集合起來的一批語言材料”



畫面擷取自：https://www.youtube.com/watch?v=6fU6RIfK0vs

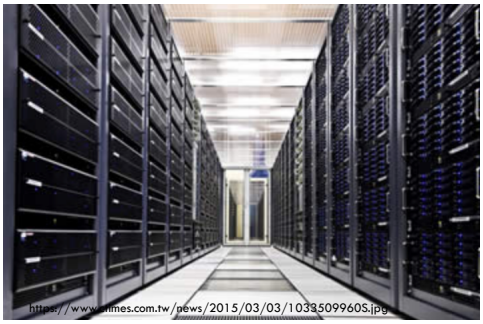
畫面節取自：https://www.youtube.com/watch?v=MPzrD-DLIgo

若我們所知，亞洲在世界上就有豐富多彩文化與歷史而且有明，也有很多人欣賞。說到文化和歷史方面，亞洲的確無人不知，尤其在手工藝的部分，自從一前手工藝的技藝-非常讓人不可界議，什麼都可做得到

圖片擷取自：https://www.youtube.com/watch?v=DYyE68IIx0U

語料庫技術！？

圖片來源: https://i1.kknews.cc/SIG=3869471/5c4800008no7646a2672.jpg



https://www.chinatimes.com.tw/news/2015/03/03/1033509960S.jpg

Google

Baidu 百度

百度一下，你就知道！

bing

機會

想要成功，要培養實力並把握機會。

我坐飛機會頭暈。

人參

人參是許多家庭選購的拜年伴手禮。

老師說明天每個人參加大隊接力時要加油。

My dog also likes eating meat.

my dog also like eat meat .

圖片來源：http://588ku.com/sucai/7850437.html

國教院分詞系統

請輸入需要分詞的文本

市面上很少有「教科書設計」的專書，因為我們總覺得那是出版社的事！然而，真的是這樣嗎？教科書設計其實與課程綱要、教師的教學、學生的學習息息相關，是課程、教學、學習三位一體間一個重要的環節，除了有教育學與學科專業等內容涵納其中，也與編輯、版式等視覺設計元素的概念有關。有鑒於此議題的重要，本院教科書發展中心邀請淡江大學課程與教學研究所陳麗華所長，於8月27日上午進行「教科書設計研究」專題演講，除了院內同仁，也邀請出版社編輯企劃相關人員參與。

開始分詞 清空

市面(Nc) 上(Ncd) 很少(D) 有(V_2) 「(PUNC) 教科書(Na) 設計(Na) 」(PUNC) 的(DE) 專書(Na) ，(PUNC) 因為(Cbb) 我們(Nh) 總(D) 覺得(VK) 那(Nep) 是(SHI) 出版社(Nc) 的(DE) 事(Na) ！(PUNC) 然而(Cbb) ，(PUNC) 真的(D) 是(SHI) 這樣(VH) 嗎(T) ？(PUNC) 教科書(Na) 設計(Na) 其實(D) 與(P) 課程(Na) 綱要(Na) 、(PUNC) 教師(Na) 的(DE) 教學(Na) 、(PUNC) 學生(Na) 的(DE) 學習(Na) 息息相關(VH) ，(PUNC) 是(SHI) 課程(Na) 、(PUNC) 教學(VA) 、(PUNC) 學習(VC) 三位一體(Na) 間(Ng) 一(Neu) 個(Nf) 重要(VH) 的(DE) 環節(Na) ，(PUNC) 除了(P) 有(V_2) 教育學(Na) 與(Caa) 學科(Na) 專業(VH) 等(Cab) 內容(Na) 涵納(VJ) 其中(Nep) ，(PUNC) 也(D) 與(P) 編輯(Na) 、(PUNC) 版式(Na) 等(Cab) 視覺(Na) 設計(Na) 元素(Na) 的(DE) 概念(Na) 有關(VJ) 。(PUNC) 有(V_2) 鑒於此(VH) 議題(Na) 的(DE) 重要(Na) ，(PUNC) 本(Nes) 院(Nc) 教科書(Na) 發展(Na) 中心(Nc) 邀請(VC) 淡江(Nb) 大學(Nc) 課程(Na) 與(Caa) 教學(Na) 研究所(Nc) 陳麗華(Nb) 所長(Na) ，(PUNC) 於(P) 8(FW) 月(Na) 27(Neu) 日(Nf) 上午(Nd) 進行(VC) 「(PUNC) 教科書(Na) 設計(Na) 研究(Na) 」(PUNC) 專題(Na) 演講(Na) ，(PUNC) 除了(P) 院(Nc) 內(Ncd) 同仁(Na) ，(PUNC) 也(D) 邀請(VC) 出版社(Nc) 編輯(VC) 企劃(Na) 相關(VH) 人員(Na) 參與(VC) 。(PUNC)

http://coct.naer.edu.tw/Segmentor/

自拍是年輕人回應生活的方式。

中文分詞

自拍 是 年輕人 回應 生活 的 方式 。

新詞發現：自拍、都更、網軍....

騎機車要戴安全帽，以確保生命安全。

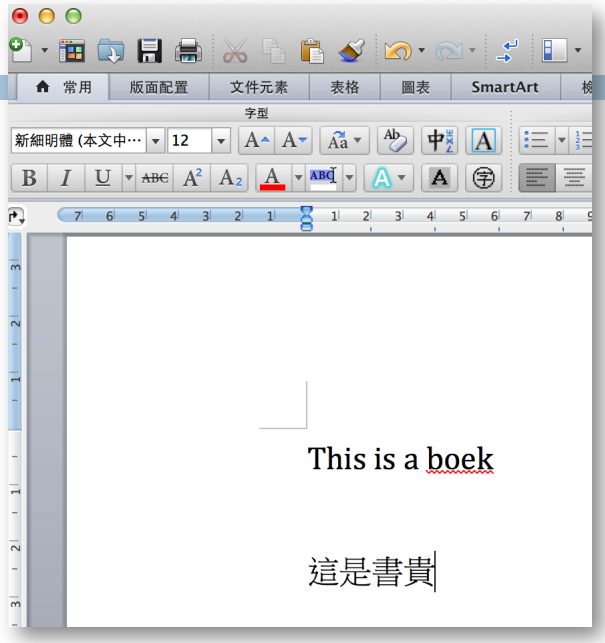
騎(VC) 機車(Na) 要(D) 戴(VC) 安全帽(Na) ，(PUNC) 以(Cbb) 確保(VE) 生命(Na) 安全(Na) 。(PUNC)

真的很機車耶，幹嘛偷看我的LINE？

真的(D) 很(Dfa) 機車(VH) 耶(T) ，(PUNC) 幹嘛(D) 偷看(VC) 我(Nh) 的(DE) LINE(FW) ？(PUNC)

新義發現！？

錯誤資料=無用資料！？



五、改錯(二題共 5 分)
1. 大仁具有高瞻遠囑的見識，做事不會猶豫不決更不會推謝責任。
2. 美美發奮練舞，和同伴切磋舞藝，不敢半途而費，盼望自己能有精綻的演出，出人頭地。

新北市板橋區江翠國中 101 學年度上學期七年級國文科第三次段考試題題目卷

教育部常用國字辨似

【芭】ㄅㄚ

〔笆〕ㄅㄚ

○ 芭蕉
╳ 笆蕉
「芭蕉」是多年生草本亞熱帶植物；「笆」是指用竹子或柳條等編成的東西。所以「芭蕉」當用「芭」而非「笆」。

【篡】ㄘㄨㄢˋ

〔簒〕ㄔㄨㄢˇ　〔竄〕ㄘㄨㄢˋ

○ 篡位
╳ 簒位　竄位
「篡」是奪取的意思，形近於「簒」；「簒」音ㄔㄨㄢˇ，是編輯的意思；「竄」是逃跑的意思。「篡位」是指古時臣下非分奪取王位，所以當用「篡」而非「簒」、「竄」。

圖片擷取自：https://www.youtube.com/watch?v=DYyE68IIx0U

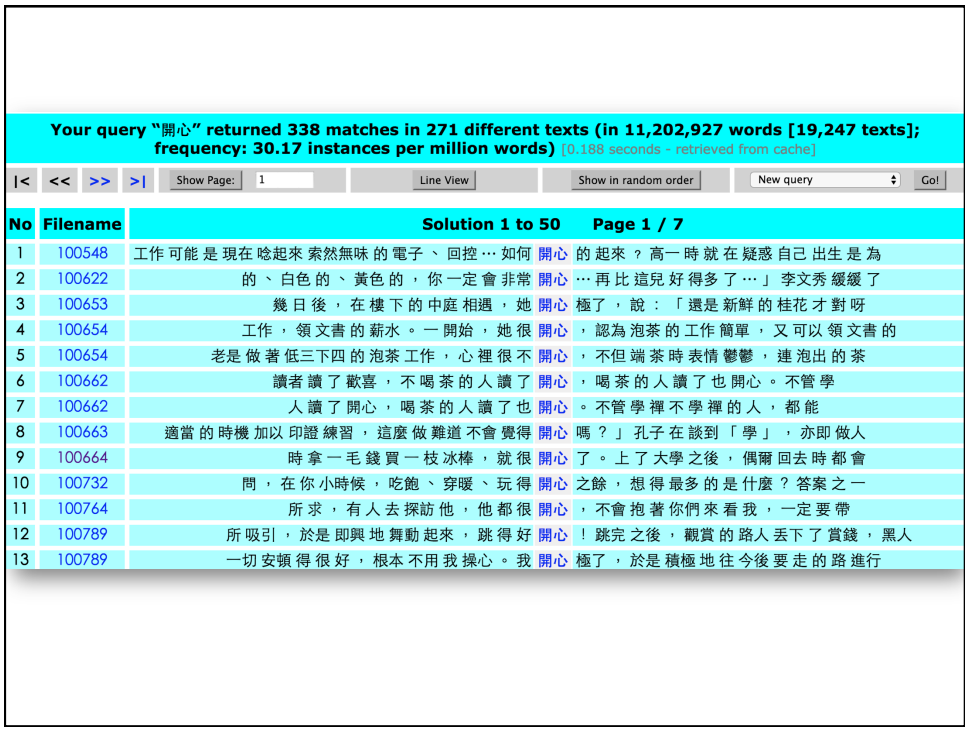
國教院索引典系統

Your query "開心" returned 3,402 matches in 2,528 different texts (in 68,809,790 words [40,960 texts]; frequency: 49.44 instances per million words) [0.165 seconds - retrieved from cache]

|< << >> >| Show Page: 1 Line View Show in random order New query Go!

No	Filename	Solution 1 to 50 Page 1 / 69
1	56f4fa66326d1b19b79ecebb	，孩子們將會有許多困難，可能經常不 開心 ，也可能因被留級而離開學校。相反的
2	56f4fa66326d1b19b79ecebe	孩童高高舉起來或倒提舉高，心想孩童會 開心 ，是一種暱愛的做法。一些對感覺
3	56f4fa66326d1b19b79ecebe	地心引力獲得大量的感覺輸入，這些輸入會產生「 開心 」，對運動和情緒的發展非常重要。例如
4	56f4fa66326d1b19b79ecebf	墊子，孩童在上面滾來滾去，常發出 開心 的笑聲，在小孩要匍匐、攀登、爬上、
5	56f4fa66326d1b19b79ecf6b	原則71：在教室中如果有人大聲拿別人 開心 ，以致干擾別人，教師可以讓這樣的孩子到
6	56f4fa66326d1b19b79ecfa0	大人也能參與他們的活動，相信孩子會更 開心 。記得曾經有一位小孩扮演醫生，跑來向
7	56f4fa66326d1b19b79ecfd2	再問他有何感覺，如何才能使小牛 開心 。要盡快採取行動，找出鬧彆扭的真正原因，
8	56f4fa66326d1b19b79ecff2	跟別人鬧得面紅耳赤。結果，不但使自己不 開心 ，而且會帶來挫敗的厄運。許多怨偶，彼此
9	56f4fa66326d1b19b79ecff6	貧窮的人未必痛苦；社會地位高的人未必 開心 ，市井小民未必不快樂。真正造成不快樂的原因
10	56f4fa66326d1b19b79ed0c6	的自我。這一長串的說明，足以為 開心 生活畫出一幅色彩光鮮的圖譜。把佛學想像成
11	56f4fa66326d1b19b79ed0c8	「心存感恩」，也都是佛家過著 開心 生活的修持法門，退一步，往往海闊天空。
12	56f4fa66326d1b19b79ed0d2	朋友看待，把最喜歡的東西與人分享， 開心 的對待宇宙的萬事萬物，知道這一點，
13	56f4fa66326d1b19b79ed0e4	不應侷限在醫療的定位，而是使人「 開心 」，這卻是宗教的本質，始於一種
14	56f4fa66326d1b19b79ed0e5	用內省的方式，把她的關心找出來：「 開心 的活著」、「健康的生存」、「

http://coct.naer.edu.tw/cqpweb/



Your query "開心" returned 338 matches in 271 different texts (in 11,202,927 words [19,247 texts]; frequency: 30.17 instances per million words) [0.188 seconds - retrieved from cache]

|< << >> >| Show Page: 1 Line View Show in random order New query Go!

No	Filename	Solution 1 to 50 Page 1 / 7
1	100548	工作 可能 是 現在 唸起來 索然無味 的 電子 、 回控 … 如何 開心 的 起來 ？ 高一 時 就 在 疑惑 自己 出生 是 為
2	100622	的 、 白色 的 、 黃色 的 ， 你 一定 會 非常 開心 … 再 比 這兒 好得多 了 … 」 李文秀 緩緩 了
3	100653	幾 日 後 ， 在 樓下 的 中庭 相遇 ， 她 開心 極了 ， 說 ： 「 還是 新鮮 的 桂花 才 對 呀
4	100654	工作 ， 領 文書 的 薪水 。 一 開始 ， 她 很 開心 ， 認為 泡茶 的 工作 簡單 ， 又 可以 領 文書 的
5	100654	老是 做 著 低三下四 的 泡茶 工作 ， 心 裡 很 不 開心 ， 不但 端 茶 時 表情 鬱鬱 ， 連 泡出 的 茶
6	100662	讀者 讀 了 歡喜 ， 不 喝 茶 的 人 讀 了 開心 ， 喝 茶 的 人 讀 了 也 開心 。 不管 學
7	100662	人 讀 了 開心 ， 喝 茶 的 人 讀 了 也 開心 。 不管 學 禪 不 學 禪 的 人 ， 都 能
8	100663	適當 的 時機 加以 印證 練習 ， 這麼 做 難道 不會 覺得 開心 嗎 ？ 」 孔子 在 談到 「 學 」 ， 亦即 做人
9	100664	時 拿 一 毛 錢 買 一 枝 冰棒 ， 就 很 開心 了 。 上 了 大學 之後 ， 偶爾 回去 時 都 會
10	100732	問 ， 在 你 小時候 ， 吃飽 、 穿暖 、 玩 得 開心 之餘 ， 想 得 最多 的 是 什麼 ？ 答案 之 一
11	100764	所 求 ， 有 人 去 探訪 他 ， 他 都 很 開心 ， 不會 抱 著 你們 來 看 我 ， 一定 要 帶
12	100789	所 吸引 ， 於是 即興 地 舞動 起來 ， 跳 得 好 開心 ！ 跳完 之後 ， 觀賞 的 路人 丟下 了 賞錢 ， 黑人
13	100789	一切 安頓 得 很 好 ， 根本 不用 我 操心 。 我 開心 極了 ， 於是 積極 地 往 今後 要 走 的 路 進行

詞彙基本搜尋

符號	意義	範例	符合之搜尋結果
?	任意的單一字	車？	車輛、車床、…
*	零～多字	*書	書、教書、保證書、…

如何查到：一板一眼、一心一意、…的例句？

如何查到：教學者、研究者、學者、…的例句？

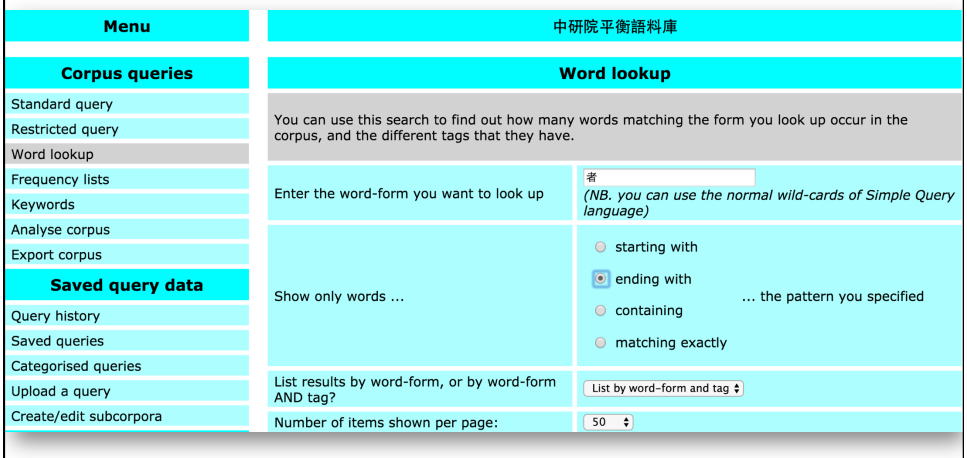
詞彙基本搜尋（詞性）

詞_詞性：搜尋符合特定詞性的詞彙

請比較下列差尋的差異性
工作
工作_VA
工作_Na

如何找到所有『臺ｏ』的專有地名？

搜尋特定詞彙列表



搜尋結果

Showing frequency breakdown of words in this query, at the query node; there are 265 different types and 30,810 tokens at this concordance position.

|< << >> >| Breakdown position: Node | Frequency breakdown of words only | Go!

No.	Search result	No. of occurrences	Percent
1	者	13695	44.45%
2	業者	3185	10.34%
3	或者	1821	5.91%
4	學者	1711	5.55%
5	記者	1539	5%
6	消費者	1423	4.62%
7	讀者	1120	3.64%
8	作者	845	2.74%
9	患者	778	2.53%

詞頻統計分析

Menu

Corpus queries

- Standard query
- Restricted query
- Word lookup
- Frequency lists
- Keywords
- Analyse corpus
- Export corpus

Saved query data

- Query history

中研院平衡語料庫

Frequency lists

You can view the frequency lists of the whole corpus and frequency lists for subcorpora you have created. Click here to create/view subcorpus frequency lists.

View frequency list for ...	Whole of 中研院平衡語料庫
View a list based on ...	Word forms

Frequency list option settings

Filter the list by *pattern* - show only words/tags ...	starting with
Filter the list by *frequency* - show only words/tags ...	with frequency between and

統計結果

Frequency list: *Word* frequencies in entire "中研院平衡語料庫", ending with "者"

|< << >> New Frequency List Go!

No.	Word	Frequency
1	者	13,695
2	業者	3,185
3	或者	1,821
4	學者	1,711
5	記者	1,539
6	消費者	1,423
7	讀者	1,120
8	作者	845

搭配詞

"「搭配詞是某語言中字彙合併的方式，以產出自然的口語和文字」
（Collocation is the way words combine in a language to produce natural-sounding speech and writing）"

Oxford Collocations Dictionary for Students of English（2002:vii）

支票

1 | Line View | Show in random order | Collocations... | Go!

Solution 1 to 50 Page 1 / 12

不會。嫌她不懂得在美國生活，連開	支票	都不會。嫌她不會開車，什麼事都要
著書包，她得了獎學金，拿著一張	支票	到銀行的櫃枱去領錢，櫃枱小姐問
領錢，櫃枱小姐問她：「這張	支票	是妳本人的嗎？」「是的，是
的書包，戴著很深的眼鏡，拿了	支票	就跑到銀行去領錢。走到櫃台，向櫃台
小姐說：「我要領錢。」然後把	支票	給小姐。小姐問：「這是你本人嗎
冊第三課？」這裡的背書指的是在	支票	後面寫上你的名字，不是背課文的背書
相信她甚於她父親。為了方便，大部分的	支票	都是以她的名字開出。婚後三年生
在工作上不敬業且不負責任，像亂開空頭	支票	。7缺乏道德、良知、冷漠無情，即使目睹
選票，不加稅，但要加福利，選舉	支票	亂開，結果債留子孫。其實道德的問題又
他賣出一批為數可觀的建材，對方以定期	支票	付款，於是陳姓商人要求對方另覓一位可靠
館子裡吃了一餐。這件事一直到	支票	遭到退票時，才發現那張支票並沒有背書
事一直到支票遭到退票時，才發現那張	支票	並沒有背書，他們只顧著高高興興的吃吃喝喝
都不是，不過是一張未填日子的	支票	，看着它固教人心跳，但能不能憑
恩恩愛愛，一齊兌了現，但又何必各拿	支票	一紙，等得那麼久耶？有些人

Choose settings for proximity-based collocations:

Include annotation:	Part-of-speech tag	Include / Exclude
Maximum window span:	+ / - 5	

Create collocation database

Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters / Go!

Extra information: **Log-likelihood** scores collocations by significance: the higher the score, the more evidence you have that the association is not due to chance. More frequent words tend to get higher log-likelihood scores, because there is more evidence for such words.

There are 1,566 different words in your collocation database for "[word="支票"%c]}". (Your query "支票" returned 566 matches in 346 different texts) [0.321 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	空頭	236	0.012	51	36	766.336
2	張	20,788	1.026	103	88	749.114
3	開	13,611	0.672	44	39	282.052
4	美元	5,384	0.266	32	24	243.648
5	兌現	169	0.008	17	17	227.111
6	給	54,016	2.666	38	33	131.66
7	銀行	4,561	0.225	18	17	122.35
8	簽	589	0.029	12	9	120.872

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	5 to the Left	Collocation window *to*:	1 to the Left
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: VC VC	Submit changed parameters / Go!

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	開	13,611	0.56	49	43	342.382
2	兌現	169	0.007	15	15	201.562
3	簽	589	0.024	14	11	150.545
4	偽造	400	0.017	7	7	70.877
5	收到	2,614	0.108	9	9	61.968
6	簽署	303	0.013	5	3	50.031
7	收下	353	0.015	5	4	48.539
8	開出	470	0.019	5	5	45.671
9	拿到	2,626	0.108	7	7	44.652
10	拿	15,450	0.635	11	11	42.046
11	寫	25,311	1.041	12	10	36.803
12	使用	17,548	0.722	5	5	10.807

國教院華英雙語索引典系統(試用版)

語料庫 光華雜誌 關鍵詞： 提供 Search

Sample Panel

1. Even Beigium had installed a sinologist in a university before this . At Cambridge , Sir Thomas Wade , who had served as an interpreter and diplomat in China and is famous for the Wade-Giles system of romanization , became professor of Chinese in 1888 after donating a collection of over 650 books to the university .
就 是 他 和 繼任 者 翟理斯 共同 創造 了 Wade-Giles 羅馬 拼音法 。 總的來說 ， 英國 在 國富 民強 的 帝國 巔峰 ， 也 只 提供 過 五 個 漢學 教授 席 而已 。
2. Unlike the multiplicity of objectivity seen in the " Taiwan · Taiwan : Facing : Faces " exhibit in the Taiwan Pavilion , the exhibition " Segmentation / Multiplication " provided Taiwanese artists with opportunities to exhibit overseas and , through the careful planning of the curators , also provided the art lovers from around the world who attended the Biennale de Venezia a chance to understand another side of contemporary art in Taiwan .
不同於 台灣 國家 館 「 面 · 目 · 全 · 非 」 客觀 多元 的 呈現 ， 另 一 個 由 策展人 以 主觀 美學 所 呈現 的 「 裂合 與 聚 生 」 展出 ， 不僅 提供 藝術家 有 更多 國際 展出 的 機會 ， 也 藉由 策展人 的 經營 ， 提供 前往 威尼斯 朝聖 的 各 國 藝術 愛好 者 ， 瞭解 台灣 當代 藝術 的 另 一 種 面目 。
3. The Commission annually provides any where from US $ 200-2000 to the majority of these schools in financial assistance . The Commission also provides textbooks free of charge to those who need them .

Translation Panel

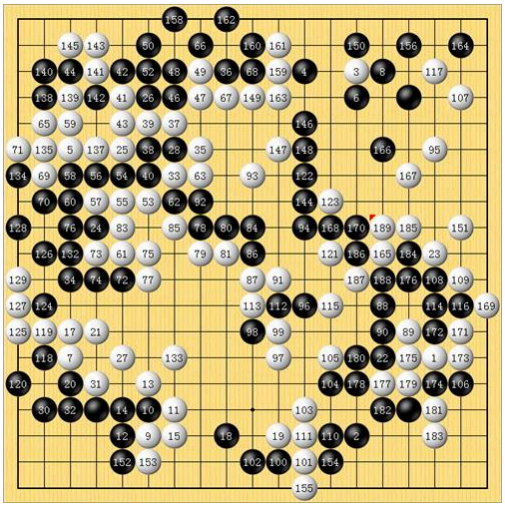
provide (531)
providing (185)
provided (184)
provides (164)
offer (128)
courtesy (107)
offering (60)
for (59)
offers (51)

Collocation Panel

提供 [Vt]
[C~]
並 ~,
[N~]
政府 ~,
[P~]
為 ~,
[Vi~]
免費 ~,

http://coct.naer.edu.tw/bc/

電腦進步到能下圍棋，
結果它還是只會算頻率跟當搜尋引擎嗎？



圖片來源：http://file.xxdao.com/2016/03/20/f3d7c7ff-d585-41b7-befa-88f1e005cf70.jpg

『依此類推』

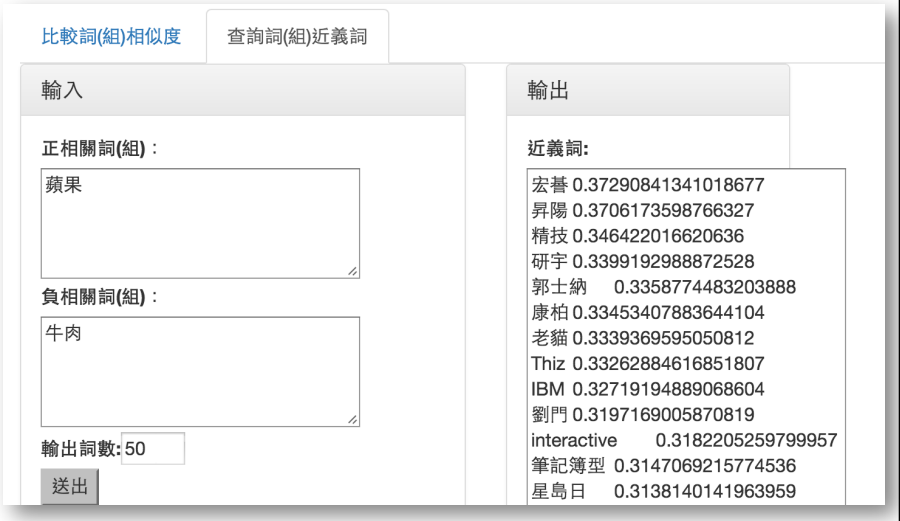
蘋果

http://a3.att.hudong.com/47/87/20300000636999135190876699345.jpg

國教院近義詞觀察系統

http://coct.naer.edu.tw/word2vec/

國教院近義詞觀察系統



國教院近義詞觀察系統

索引典系統進階語法

Case Study

同中求異，異中求同

	解決	處理
問題	27491	5186
困境	971	69
爭議	926	349
困擾	710	119
難題	675	54
危機	624	560
水患	606	20
紛爭	428	64
困難	408	86

	解決	處理
問題	27491	5186
垃圾	111	2020
廢棄物	22	1756
案	167	1361
事	405	1073
事件	105	1072
事宜	69	1054
案件	60	1018
事務	40	1015

燃眉之急	314
淹水	283
瓶頸	177
亂象	147
打者	102
不便	61
懸案	61
疑難	53
血荒	51
對手	47

後事	446
公務	193
屍體	151
遺體	105
郵件	74
油污	67
傷口	65