

語料庫應用於華語教學

張莉萍

國立台灣大學華語教學碩士學位學程

2022年3月29日於東海大學華語文教學國際碩士學位學程演講內容

本國人 (L1) 語料庫 vs. 學習者 (L2) 語料庫

可以參考這本書對現有語料庫的介紹：

陳浩然（主編），(2017)，《語料庫與華語教學》。

台北：高等教育文化事業有限公司。

應用語料庫於教學

直接應用

- Data-driven learning (DDL)

間接應用

- 教材編輯、課堂講義、辭典

選擇適用的語料庫及工具很重要

- 快速找出生詞典型例句
- 教學語法排序
- 尋找定式、定式例句
- 近義詞分析
- 偏誤分析（瞭解不同母語背景學習者的表現）

幾個實用的網站

- <http://elearning.ling.sinica.edu.tw> （全球華語文數位教與學資源中心）
 - <http://words.sinica.edu.tw/> （搜文解字）
 - <http://asbc.iis.sinica.edu.tw/> （中研院平衡語料庫4.0）
 - <http://tocfl.itc.ntnu.edu.tw/> （TOCFL學習者語料庫）
 - <https://coct.naer.edu.tw/> （國家教育研究院：華語文語料庫整合應用系統）
- 

利用語料庫查詢詞頻

<http://elearning.ling.sinica.edu.tw/CWordfreq.html>

現代漢語語料庫詞頻統計

現代漢語語料庫詞頻統計提供平衡語料庫的詞頻信息。華文教師可依據詞頻統計提供的訊息得知詞語的數量與頻率，從而決定詞語學習的先後安排，幫助教師們編寫教程。

[使用說明](#)

查閱詞頻排序

從第 到第

- 說明：排序從1 到 93,826。請輸入數字，一次最多三百。

查閱個別詞的頻率

輸入詞語

- 說明：請輸入欲查詢的詞，標記可有可無，例如，工作 或 工作(Na)。

現代漢語個別詞的頻率 Word frequency: 了

No	Rank	Word	Frequency	Percent	Cumulation
9	9	了(Di)	36062	0.739	14.018
32	32	了(T)	11499	0.236	22.484
15740	15270	了(VJ)	25	0.001	88.837
36228	33467	了(VC)	8	0.000	94.503
69256	57995	了(VC)[+vrr]	3	0.000	97.505
154486	93826	了(VK)	1	0.000	99.774
154487	93826	了(VF)	1	0.000	99.774

了₁ vs. 了₂

- <http://elearning.ling.sinica.edu.tw>

<u>Rank</u>	<u>Word</u>	<u>Frequency</u>	<u>Percent</u>
-------------	-------------	------------------	----------------

9	了(Di)	36062	0.739
---	-------	-------	-------

32	了(T)	11499	0.236
----	------	-------	-------

- 平衡語料庫（設定會話訪談語料）

<u>詞類</u>	<u>關鍵詞</u>	<u>合計</u>	<u>比例</u>
-----------	------------	-----------	-----------

T	781	781	67.15
---	-----	-----	-------

Di	382	382	32.85
----	-----	-----	-------

漢字教學

搜詞尋字

檢字

若有不知如何輸入的部首，請[download](#)附件的外字集

A. 部件搜尋	示例
所有含 的字	力 (功、加、勞、劦 ...)
字首為 的字	力 (加、劦 ...)
字尾為 的字	力 (功、勞、劦 ...)
B. 部首筆畫搜尋	示例
部首 筆畫	力部、3畫(功、加)
總筆畫	5畫(功、加、付)

☐ 簡單搜尋 ☒ 複雜搜尋

開始搜尋

清除

您現在搜尋的是有力的字 第1頁

序號	字形	注音	部首	筆劃	總筆劃	字根序
1	力	ㄌㄧˊ	力	0	2	力
2	加	ㄐㄧㄚ	力	3	5	力口
3	功	ㄍㄨㄥ	力	3	5	工力
4	另	ㄌㄧㄥˊ	口	2	5	口力
5	叻	ㄌㄨˊ	口	2	5	口力
6	幼	ㄩˋ	𠂇	2	5	𠂇力
7	劣	ㄌㄧㄝˊ	力	4	6	少力
8	肋	ㄌㄨˊ	肉	2	6	𠂇力
9	伽	ㄍㄧㄝˊ	人	5	7	𠂇力口
10	劫	ㄐㄧㄝˊ	力	5	7	土𠂇力
11	助	ㄓㄨˋ	力	5	7	且力
12	努	ㄋㄨˋ	力	5	7	女又力
13	𠂇	ㄍㄨˊ	力	5	7	𠂇口力
14	男	ㄋㄢˊ	田	2	7	田力
15	劬	ㄉㄨˊ	力	6	8	亥力
16	𠂇	ㄉㄨˊ	力	6	8	𠂇王力
17	協	ㄒㄧㄝˊ	十	6	8	十力力力

<http://words.sinica.edu.tw/> (搜文解字)

搜詞尋字

造詞搜尋

A. 一般搜尋		示例
所有含	的詞	夫 (老夫子、丈夫、夫人 ...)
詞首為	的詞	夫 (夫妻、夫子 ...)
詞尾為	的詞	性 (可靠性、重要性 ...)
B. 成語搜尋		示例
活用四字詞 有	有	一 * 而 * (一炮而紅、一飲而盡 ...)
C. 重疊詞搜尋		示例
重疊詞		AAB (走走路、洗洗澡 ...) ABB (亮閃閃、紅通通 ...) AABB (高高興興、平平安安 ...) ABAB (高興高興、研究研究 ...)
D. 字數搜尋		示例
我要 字詞		一、二、三、四或五字以上

四字格

- **Usage:**

Different four-character formats have different productivity. For example, there are more expressions in the “有A 有B” format than in the “大A大B” format. Further examples of “有A 有B” format include: 有高有低, 有快有慢, 有新有舊. (引自《當代中文課程2》第15課)

- 有聲有色
- 有模有樣
- 有說有笑
- 有板有眼
- 有條有理
- 有始有終
- 有好有壞

- 有滋有味
- 有名有姓
- ~~• 有沒有空~~
- 有為有守
- 有血有淚
- 有來有往
- 有頭有臉

COCT 書面語語料庫2019

- Your query “有+有+” returned 2,109 matches in 1,940 different texts,
- there are 67 different types and 2,109 tokens at this concordance position.
- Your query “大+大+” returned 3,016 matches in 2,622 different texts,
- there are 283 different types and 3,016 tokens at this concordance position.

Frequency breakdown

No.	Search result	No. of occurrences	Percent
1	有說有笑	443	21.01%
2	有聲有色	435	20.63%
3	有條有理	345	16.36%
4	有模有樣	255	12.09%
5	有始有終	239	11.33%

應用語料庫編輯教材

- 例1：給10個量詞（高頻）+5例。
個：朋友、老師、學生、人、妹妹
家：...
- 和「家」搭配緊密的名詞

利用語料庫找出量詞「家」後面常用名詞

<http://asbc.iis.sinica.edu.tw/>

中央研究院平衡語料庫

請輸入你所要查詢的關鍵字

關鍵詞或重疊詞		詞類	特徵
家	或 ▼	Nf (量詞) ▼	▼

送出

重填 Reset

中央研究院 [詞庫小組](#) 提供 (Ver 1.1-20130621)

MI值	freq(y)	freq(x,y)	y:詞/詞類
7.182	13	5	花店
7.069	32	11	分店
7.038	27	9	教養院
6.432	44	8	信合社
6.163	36	5	電視網
5.940	63	7	出版商
5.940	45	5	車廠
5.875	48	5	咖啡店
5.795	52	5	應召站
5.739	66	6	超級市場
5.603	63	5	固網
5.599	329	26	電視台
5.596	165	13	戲院
5.559	79	6	雜貨店

MI值	freq(y)	freq(x,y)	y:詞/詞類
4.466	9418	240	公司
-1.134	923327	94	,
4.750	2186	74	廠商
4.165	3183	60	業者
4.390	2244	53	銀行
-1.871	581824	45	的
-1.223	341014	43	。
3.612	3867	42	企業
5.450	602	41	店
3.999	2319	37	醫院
3.441	3168	29	以上
5.212	84889	28	「
2.262	9590	27	新
4.526	999	27	大型
5.599	329	26	電視台
4.921	648	26	餐廳
1.328	21446	24	大
1.856	12704	24	三
0.548	44940	23	人
4.669	673	21	航空
0.762	29894	19	以
4.463	749	19	上市

1	公司
2	店
3	餐廳
4	旅館
5	企業
6	醫院
7	銀行
8	報紙
9	工廠
10	分店
11	商店
12	大型
13	小
14	義集疏
15	書店
16	飯店
17	廠商
18	老小
19	餐館
20	出版社

結果

- 家：公司、銀行、咖啡店、醫院、廠商
- 種：水果、方法、情況、感覺、產品
- 條：魚、河、毛巾、裙子、馬路
- 件：事、案子、外套、衣服、藝術品
- 場：電影、比賽、音樂會、大雨、表演
- 隻：手、狗、鳥、腳、眼睛
- 塊：蛋糕、布、肉、玉、木頭
- 支：手機、傘、舞、球隊、廣告

(因為/由於)...加上...所以/因此/於是

加上 (+){1,9} (因此|所以|於是)

查找例句：

1. 王太太那時年紀輕，**加上**孩子小，處處需要她照顧，所以常一個人躲著哭泣。
2. 學校由於停車位不足，**加上**考量師生安全，因此決定學生不得在校園內停車。
3. 台北交通的擠和亂，主要是因為人口和車輛增加得太快了，再**加上**有許多不遵守交通規則的駕駛和行人，於是路上的情況越來越糟糕。

應用本國人語料庫找出答案

1. 本國人使用「一點都...」還是「一點也...」哪一個頻率高？學習者呢？（教學語法）
2. 和「喜酒」搭配的動詞是「吃」還是「喝」？（編寫課文遇到的問題）

No.	Search result	No. of occurrences	Percent
1	喝喜酒	80	58.82%
2	吃喜酒	56	41.18%

喝 * 喜酒91筆；吃 * 喜酒64筆

應用學習者語料庫找出答案

1. 學習者「關係子句」偏誤情況為何？
2. 能願動詞「會」，學習者的偏誤類型分布？
3. 「如果...」「如果...的話」「...的話」三種形式，哪一個學習者使用得多？不同母語背景學習者傾向一樣嗎？

中介語語料庫

- 畢業
- Your query “畢業” returned 248 matches in 168 different texts (in 704,270 words [1,963 texts]; frequency: **352.14** instances per million words)
- 本國人2019：Your query “畢業” returned 19,764 matches in 11,387 different texts (in 319,712,694 words [168,090 texts]; frequency: **61.82** instances per million words)

中介語對比分析方法 (contrastive interlanguage analysis, CIA)

- CIA是語料庫語言學與第二語言習得研究結合後的新領域。
- 主要是針對兩類語料進行對比，一是本國人語料庫和學習者語料庫的比較；一是不同母語背景學習者語料庫之間的比較（Granger, 1998: 12）。
- 應用前者對比分析可以找出學習者少用 (underuse) 或過度使用 (overuse) 的語言特徵；應用後者對比分析，則可能找出針對不同母語背景學習者的語言特徵。

<http://tocfl.itc.ntnu.edu.tw/>

TOCFL 學習者語料庫檢索系統

說明:「TOCFL 學習者語料庫」所收集的語料是母語非華語的外籍人士參加華語文能力測驗(TOCFL)所寫的作文。寫作時間為2006到2012年。共收集4567篇,約150萬字。其中2837篇(989045字)經過偏誤標記處理。如欲進一步參考語料庫建置內容請參考[「TOCFL作文語料庫的建置與應用」](#)。

中文斷詞及詞性標註採中央研究院CKIP自動斷詞系統。不論自動或人工偏誤標記謬誤難免。開放此系統,僅供研究目的。使用本系統所完成之研究報告或論文,請加註語料出處或致謝。致謝詞範例:本研究感謝國立台灣師範大學TOCFL學習者語料庫團隊提供之語料及檢索系統。

帳號:

密碼:

登入

*如無帳號而欲使用此系統請聯絡:[張老師](#)

2021/4/22

瀏覽人次: 90839

TOCFL 學習者語料檢索系統

登出

偏誤檢索

字串檢索

關鍵詞檢索

考試等級: ☒ A2 ☐ B1 ☐ B2 ☐ C1 | 考生母語: 全部 | 體裁: 全部 | 考生分數: ☒ 3 ☐ 4 ☐ 5 | 左右字數: 15

星號(*)表示為必填項目(註.主類或次類至少必須選擇一個) 標記說明

*主類: 全部 * | 次類: 全部 查詢

左邊 個字內包含 , 不包含 且 或 右邊 個字內包含 , 不包含

過濾

偏誤：四大主類

- 遺漏偏誤（Missing，[M]）

我真的[Madv]緊張要跟林美美見面

→我真的很緊張要跟林美美見面

- 冗贅偏誤（Redundant，[R]）

我知道妳又很[Radv]聰明又用功

→我知道妳又聰明又用功。

- 選用偏誤（Misselection，[S]）

我還不[Sadv]決定要在哪一家餐廳吃

→我還沒決定要在哪一家餐廳吃。

- 語序偏誤（Misordering，[W]）

今天早上我起床六點半[Wtime]。

→今天早上我六點半起床。

36 次類

- 「詞」層次（共16類）：動作動詞(v)、能願動詞(aux)、狀態動詞(vs)、名詞(n)、代名詞(pron)、連接詞(conj)、介詞(p)、數詞(num)、指示詞/定詞(det)、量詞(cl)、句末助詞(sp)、時貌助詞(asp)、副詞(adv)、de 助詞(de)、疑問詞(que)、複數詞綴(plural)；
- 「語」的層次（11 類），如主語(sub)、賓語(obj)、名詞短語(np)、動詞短語(vp)、介詞短語(pp)、修飾語(mod)、時間詞語(time)、地方詞語(loc)、[數/定+量]結構(dm)、離合結構(vo)、及物性(tran)；
- 句型層次（7 類），如複雜名詞子句(rel)、特殊句型（把ba、被bei、讓rang、是shi、有you）、其它句式(pattern)。
- 混合類（2 類），形式錯誤(form)，表達方式錯誤(sentence)。

複雜名詞子句錯誤（rel）

我覺得餐廳在你家旁邊[Srel]好像[Madv]好吃。

→我覺得在你家旁邊的餐廳好像很好吃。

希望我們的計畫要見面[Srel]沒有甚麼問題

→希望我們要見面的計畫沒有甚麼問題

我覺得野餐是最好的辦法認識新朋友[Srel]。

→我覺得野餐是認識新朋友的最好辦法。

英、日、韓學習者「如果…的話」 使用比例統計

	如果...的話	如果...	...的話	小計
英	15.51%	71.62%	12.87%	100%
日	45.45%	18.18%	36.36%	100%
韓	29.60%	25.37%	45.03%	100%
本國人				

英、日、韓學習者「如果…的話」 使用比例統計

	如果...的話	如果...	...的話	小計
英	15.51%	71.62%	12.87%	100%
日	45.45%	18.18%	36.36%	100%
韓	29.60%	25.37%	45.03%	100%
本國人	6.51%	80.72%	12.77%	100%

直接應用於教學 Data-driven learning 示例

▣ 呈現語料的方法

- 建議有一定的量（例如，5句、10句為一個單位）
- 不同語法性質，可能適合不同呈現方式
 - ✓ 由短到長、依搭配詞語、依不同結構、依語境（正向/負向）

▣ 教學活動（搭配語料呈現所設計的教學活動）

▣ 學習單設計

Data-driven learning 示例

一面...一面... （當代中文課程五語法點）

- 步驟一：呈現語料庫例句（由短到長）
- 一面看，一面笑
- 一面加麵粉，一面加水
- 一面打工當家庭教師，一面上學
- 一面伸了個懶腰，一面在桌前坐下
- 一面抽著煙，一面舒舒服服地躺在床上
- 一面看他的武俠小說，一面又讀《明報月刊》
- 一面用語言哄她，一面卻想要用毒針將她殺死
- 一面維護自己的自尊，一面使勁撕破對方的面子
- 林燕妮一面在美容院做頭髮，一面看倪匡的《我看金庸小說》

步驟二：設計練習

- 示範一：隨機拿掉句子中「一面A」、「一面B」一部分，並把拿掉的部份做成選項，讓學生自行判斷挖空處填入何者最合適。例如：
- 選項：(1)一面用眼睛在看 (2) 一面還是忍不住流下淚來
(3) 一面吃飯 (4) 一面親切地提醒顧客在哪裡簽名
- 一面用耳朵在聽，_____。
- 程英一面說甚麼人有聚有散，_____。
- 丈夫在餐桌上_____，一面看報。
- _____，一面不停地讚美顧客。

- 示範二：情境創作（不但...反而...）
- 展示語料庫中不含上下文的例句，請學生根據例句，創造一個適合此語料的情境。例如：
 - _____，不但沒有替她高興，反而滿口不屑之詞。
 - _____，不但沒有想像中的快樂，反而十分茫然。
 - _____，不但沒有減輕她的恐懼，反而讓她更害怕了。

無論...都... 依不同結構展示

(當代中文課程4語法點)

- 「無論」+A或(是)B：
- 無論過去或是現在都是害羞的。
- 無論是生理或心理上的變化都是漸進的。
- 無論是父母或子女都得學習怎樣來適應這些變遷才行。
- 無論從青少年或成人的立場來看性問題，都是有討論的意義的。
- 無論是南美、非洲或中東地區，都有人在研究這種Nervos的「生活苦痛」。

無論...都...

- 「無論」+A還是B：
- 無論他是男性還是女性都一樣。
- 無論是可愛的新生兒還是男人女人，都該是受到疼愛的寶貝。
- 在青樸遇到的人，無論是僧侶還是普通人，都對我們敞開了靈魂深處。
- 「北方澤鵠」與「灰澤鵠」，無論外型還是習性，都沒有很大的差別。
- 無論是價值觀、行為模式還是教養方法，都會使孩子在成長過程中，希望或選擇排斥成為和我們相類似的人。

無論...都...

- 「無論」+ A (B) 不A (B) :
- 無論你愛不愛，都得擺個笑臉。
- 無論你喜不喜歡，都必須接受。
- 無論他結不結婚，他都懂得男女間角色轉變的趨勢。
- 無論同不同意，都必須相信，人是時間串連的作品。
- 大丈夫一言既出，駟馬難追，無論願不願意，都得兌現。

無論...都...

- 「無論」+疑問詞：
- 無論發生任何事，我都願意幫助妳。
- 我認為，無論哪種人，都該接受他們的存在
- 無論碰到誰遇到什麼事，都要分個勝敗輸贏的意思。
- 無論你採用何種處理態度，對他都有舉足輕重的影響。
- 原始的族羣中，無論發生什麼病，都會先試試這樣的方法。
- 無論最後的結局如何，他都在人生的最後五十七天中，用一種比較清澄的心靈過日子。

- 「無論」+反義複合詞：

- 無論好壞，我們都曾伸出雙手，在中間做過一點事。
- 現在年輕人無論男女，生活圈子都比過去人要大得多。
- 無論東西方，男性對性的享受和婚姻，都比女性有更多的自由和特權。
- 快樂本身就是生命的實現，無論貧富、尊卑、男女、老幼，都是平等的。
- 無論中外，研究少年犯罪問題的調查報告中，家庭原因都是佔了最大的比率。