讓「主題關鍵性分析工具」幫你找出語言特徵

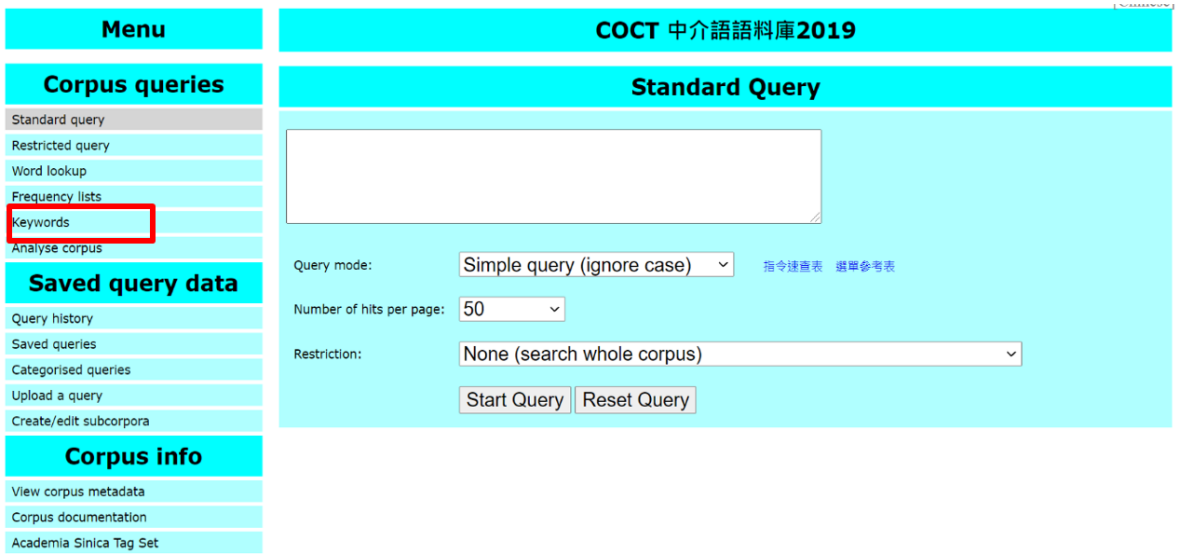
【語文教育及編譯研究中心助理研究員　吳欣儒】

「特徵」，意思是一事物異於其他事物的特點，我們可以從一個人的外表來歸納他的外顯特徵，例如長得很高、頭髮很長、眼睛很大，這是一種透過比較而得的相對概念，例如林書豪很高，但若將他跟姚明相互比較，林書豪就不算高了。在高個如雲的NBA，林書豪不能以「高」作為他的特徵。應用這個概念，你知道一個文本（泛指任何的文字或口語的語言材料，長如一本書，短如一個段落，都可以稱為文本）也能歸納出它的語言特徵嗎？什麼是語言特徵？要怎麼做呢？

我們有時會在聽一個人說話的時候，發現他的口頭禪是「然後」、「隨便啦」、「奇怪欸你」，短的文本或許我們還可以自己歸納用詞特徵，但如果長如一本書、一部電影，甚至是上千萬、上億字數的文本時，我們就必須依靠電腦來處理了。現在的語料庫技術，有一種叫主題關鍵性（keyword keyness）的分析工具，可幫助我們歸納文本的語言特徵。這個工具的原理簡單來說，就是需要兩組語料庫，一組作為參照用（NBA），一組為觀察用（林書豪），觀察用的是我們要歸納特徵的對象。電腦以統計去計算這兩組語料庫每個詞語的出現頻次，並且以參照用的為基準，去看觀察用的語料庫中，哪些詞語的使用率是不尋常的高或不尋常的低（比預期的出現率高或低），用這個方式來歸納、推論屬於某個主題文本的詞語特徵。

現在這種技術的應用很廣泛，例如有些社會人文學科以此探究某一類文本的可能主題或語義脈絡中的詞彙使用，用這種方式去詮釋語言在建構社會意義所扮演的角色，但更多的是應用在語言教學上，透過主題關鍵性分析工具，我們可以得知某一類群體的語言表達特點，再將之應用在教學上。現在，我們以國家教育研究院 COCT 中介語語料庫的主題關鍵性分析工具為例，來看如何使用。

圖 1、國教院 COCT 中介語語料庫（https://coct.naer.edu.tw/cqpweb/learners_2019/）



「中介語」語料庫指的是學習者語料庫，華語中介語語料庫就是將學習華語的外國人所產出的華語作文集結而成的語料庫。目前臺灣較有規模的中介語語料庫為國教院 COCT 中介語語料庫：https://coct.naer.edu.tw/cqpweb/learners_2019/（見上圖 1）。在上述圖 1 的頁面點選左邊 Corpus queries 選項中的 keywords，可看到如下圖 2 的頁面。圖 2 上面的兩個紅色方框，代表的是前面提到的參照用（NBA）與觀察用（林書豪）的語料庫。左邊紅框是觀察用語料庫，也就是要歸納特徵的語料庫，右邊紅框是參照用語料庫。操作的時候，先下拉左邊紅框的選單，選擇要觀察的對象，例如選擇 1a_Korean，韓國人的語料，而右邊紅框選擇 1b_non_Korean，意思是把韓國人跟非韓國人的語料相比，最後點下最下面的 Calculate keywords，即可得知韓國人的用詞特徵，如圖 3。

圖 2、以 keywords 功能分析韓國人的詞語特徵

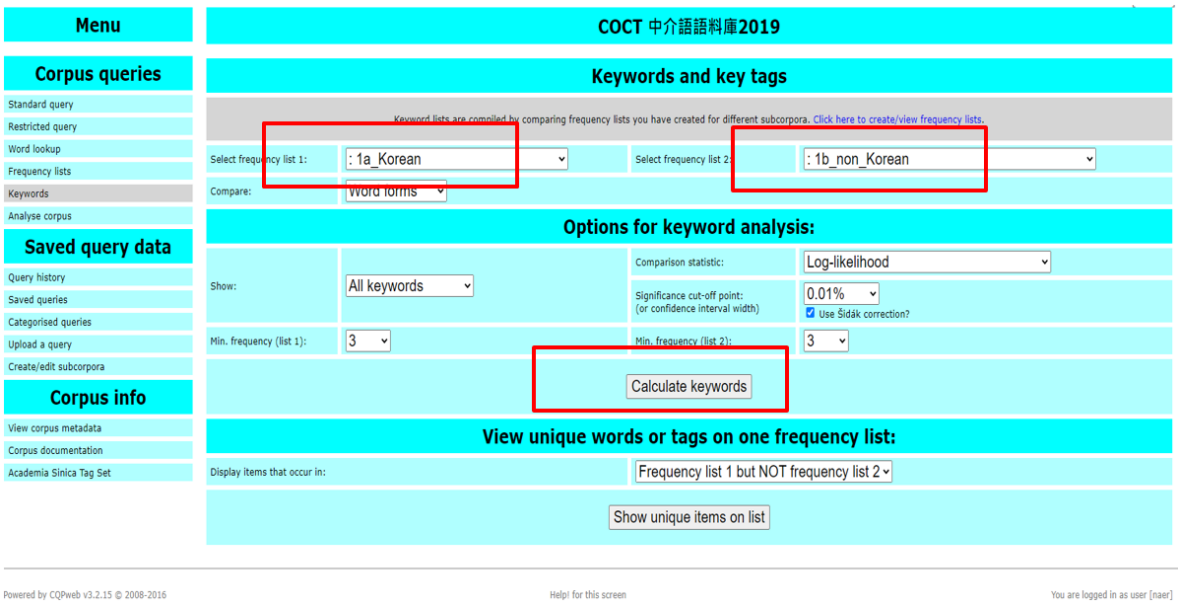


圖 3、語料庫呈現韓國人的用詞特徵

Keyword list for subcorpus "1a_Korean" compared to subcorpus "1b_non_Korean";
using log-likelihood statistic, significance cut-off 0.01% (adjusted LL threshold = 31.95);
items must have minimum frequency 3 in list #1 and 3 in list #2.

|< << >> Download whole list Go!

No.	Word	In subcorpus "1a_Korean":		In subcorpus "1b_non_Korean":		+/-	Log likelihood
		Frequency (absolute)	Frequency (per mill)	Frequency (absolute)	Frequency (per mill)		
1	韓國	572	7,313.83	377	602.25	+	1330.82
2	韓國人	87	1,112.42	62	99.04	+	194.72
3	所以	604	7,723.00	2,979	4,758.88	+	106.2
4	.	217	2,774.65	3,327	5,314.80	-	104.95
5	韓國	29	370.81	4	6.39	+	104.04
6	整型	34	434.74	14	22.36	+	94.8
7	日本	42	537.03	1,144	1,827.51	-	90.97
8	韓	31	396.38	13	20.77	+	85.91
9	台灣	35	447.52	21	33.55	+	84.7
10	'	43	549.82	42	67.09	+	81.08
11	-	4,337	58,011.97	31,376	50,441.86	+	79.11
12	%	37	473.10	29	46.33	+	78.95
13	男	53	677.68	75	119.81	+	76.99
14	電影	24	306.87	8	12.78	+	71.39
15	影片	33	421.95	26	41.53	+	70.22
16	首	20	255.73	4	6.39	+	67.23
17	韓國	467	5,971.26	2,445	3,905.83	+	64.4
18	韓國	37	473.10	41	65.50	+	64.37
19	以後	202	2,582.86	834	1,332.29	+	62.09
20	最近	72	920.62	173	276.36	+	60.49
21	電影	95	1,214.71	277	442.50	+	60.11
22	最近	34	434.74	44	70.29	+	52.97
23	旅行	115	1,470.44	401	640.59	+	52.44
24	的	4,897	62,615.08	35,199	56,229.51	+	51.42
25	日本之	18	230.16	7	11.18	+	51.12
26	很重	31	396.38	41	65.50	+	47.5
27	然後	180	2,301.55	787	1,257.21	+	47.1
28	韓事	33	421.95	48	76.68	+	46.86
29	減肥	14	179.01	3	4.79	+	46.4
30	時候	137	1,751.74	547	873.82	+	45.91
31	韓星	30	383.59	41	65.50	+	44.81
32	韓韓	15	191.80	5	7.99	+	44.62
33	:	34	434.74	715	1,142.19	-	41.11
34	日本人	4	51.15	271	432.92	-	39.61
35	字太	14	179.01	7	11.18	+	36.45
36	已	17	217.37	455	726.85	-	35.48

如圖 3 所示，藍底表示韓國人與其他國家的人相比，用得特別多的詞語，灰底表示用得特別少的詞語。韓國人用得特別多的詞語整理如表 1 所示，可發現韓國人談論的話題可能多圍繞與政治時事（包含國家）有關，如「韓國」、「首爾」、「北韓」、「總統」、「戰爭」、「南韓」，或與娛樂休閒有關，如「整型」、「影片」、「電影」、「旅行」、「減肥」、「主角」等。我們用前述提及的方式操作日本、越南的華語學習者的用詞特徵，發現日本學習者常用「櫻花」、「溫泉」、「漢字」、「料理」等，泰語學習者常用「佛教」、「大象」、「節日」、「創業」等，越語學習者則是「廟」、「龍」、「神明」、「菩薩」等的主題詞語。

表 1、韓國人使用的不尋常高頻詞語

編號	詞語	編號	詞語	編號	詞語
1	韓國	11	爾	21	投資
2	韓國人	12	時候	22	而且
3	所以	13	婚禮	23	戰爭
4	首爾	14	以後	24	減肥
5	整型	15	最近	25	的話
6	韓	16	電影	26	機車
7	北韓	17	總統	27	南韓
8	狗	18	旅行	28	子女
9	濟州島	19	的	29	主角
10	影片	20	百分之	30	部

藉由這個範例，我們可以知道主題關鍵性這個語料庫工具的使用方式及其結果所代表的意義。透過這樣的功能，主題關鍵性分析能增進華語教師、教材編者、測驗研發人員等對學習者語言的掌握，也可作為教師授課或編輯國別化教材的參考，未來也可能為編輯華語學習者辭典提供選詞參照。

近年來，隨著電腦運算能力與人工智慧的進步，電腦輔助的文本分析工具也日臻成熟，例如電腦可幫我們進行詞頻分析、詞彙搭配關係分析、主題關鍵性分析等，未來你想要挖掘更多文本特徵，不妨來試試國教院語料庫的各種好用工具喔！

資料來源

吳欣儒（2020）。**華語文中介語詞語、語法及語篇特徵研究。**國家教育研究院個別型計畫案成果報告（NAER-2019-029-C-1-1-B5-04）。國家教育研究院。連結網址：https://rh.naer.edu.tw/handle/59kxv