中研院臺灣當代華語語料庫

摘要

中研院臺灣當代華語語料庫收錄多達7千多萬則臺灣正體華語詞彙(73007511)。取得授權的語料來源為2015至2020年間出版的報章雜誌(總共86267296個字，50486010個詞及238361篇文章)及兒童讀物(總共37162935個字，22297533個詞及112204篇文章)。內容涵蓋了生活、社會、金融、科學、文學、文化、教育以及哲學等多種類型的文章。本語料庫所有文本的內容都依照詞彙斷開並標示詞類標記及實體辨識標記。進一步彙整每個詞彙的詞類標記，以及計算其詞彙頻率(出現次數、百萬詞頻、百萬詞頻取對數)，語境變異(出現次數、出現次數取對數、百萬詞頻、百萬詞頻取對數、千詞頻)，以及語意變異等三項指標。

技術優勢

1. 2015至2020年間當代臺灣華語文本大數據分析
2. 提供詞彙頻率、語境變異及語義變異等三項指標
3. 依據不同年齡層閱讀文本的類型，可分別建立適合小孩、成人及老年人等閱讀經驗的語料庫

	A	B	C	D	E	F	G	H	I	J	K
1	word 詞彙	WF 詞頻	WF_ppm 百萬詞頻	log(WF_ppm) 百萬詞頻取對數	CD 語境變異	log(CD) 語境變異 (取對數)	CD_ppm 語境變異 (百萬詞頻)	log(CD_ppm) 語境變異 (百萬詞頻取對數)	CD_ppk 語境變異 (千詞頻)	SD 語境變異	POS 詞類標記
2	向	55781	764.04	2.88	43711	4.64	124376.85	5.09	124.38	1.53	[('P', 55256), ('D', 79), ('VJ', 441), ('Na', 5)]
3	跟	55694	762.85	2.88	38821	4.59	109893.58	5.04	109.89	1.28	[('Caa', 5034), ('P', 48825), ('VC', 1835)]
4	國家	55550	760.88	2.88	33938	4.53	96568.40	4.98	96.57	1.47	[('Na', 55550)]
5	像	54891	751.85	2.88	41544	4.62	118210.79	5.07	118.21	1.27	[('VG', 25824), ('P', 28739), ('Na', 328)]
6	發現	54821	750.90	2.88	40648	4.61	115661.28	5.06	115.66	1.43	[('VE', 53889), ('Na', 928), ('Nv', 4)]
7	成為	54330	744.17	2.87	43946	4.64	125045.53	5.10	125.05	1.52	[('VG', 54330)]
8	未	53356	730.83	2.86	39256	4.59	111700.43	5.05	111.70	1.56	[('D', 53353), ('Nd', 3)]
9	指出	53020	726.23	2.86	41279	4.62	117456.75	5.07	117.46	1.58	[('VE', 53020)]
10	月	52575	720.13	2.86	37958	4.58	108007.06	5.03	108.01	1.53	[('Na', 52206), ('Nd', 61), ('Nf', 308)]
11	曾	52545	719.72	2.86	41501	4.62	118088.44	5.07	118.09	1.47	[('D', 52304), ('Nb', 241)]
12	發展	52187	714.82	2.85	33716	4.53	95936.72	4.98	95.94	1.51	[('VC', 48184), ('Na', 2400), ('Nv', 1603)]
13	許多	52119	713.89	2.85	42569	4.63	121127.36	5.08	121.13	1.45	[('Neqa', 50411), ('Dfb', 1708)]
14	事	52016	712.47	2.85	39482	4.60	112343.50	5.05	112.34	1.31	[('Na', 51970), ('VC', 46)]
15	世界	51375	703.69	2.85	34874	4.54	99231.73	5.00	99.23	1.43	[('Nc', 51375)]
16	因此	50909	697.31	2.84	42958	4.63	122234.24	5.09	122.23	1.49	[('Cbb', 50909)]
17	比	50544	692.31	2.84	35022	4.54	99652.86	5.00	99.65	1.51	[('P', 38716), ('Caa', 2984), ('VC', 8392), ('Na', 439), ('Nc', 9), ('Nv', 2), ('Dfa', 2)]
18	學習	50110	686.37	2.84	26298	4.42	74829.27	4.87	74.83	1.34	[('VC', 48070), ('Nv', 2040)]
19	四	50078	685.93	2.84	35323	4.55	100509.33	5.00	100.51	1.58	[('Neu', 49114), ('Cbb', 569), ('Nd', 296), ('Neqa', 52), ('Na', 35), ('Nc', 12)]
20	未來	49330	675.68	2.83	37118	4.57	105616.89	5.02	105.62	1.55	[('Nd', 49330)]
21	總統	49260	674.73	2.83	22428	4.35	63817.44	4.80	63.82	1.32	[('Na', 49260)]
22	公司	49025	671.51	2.83	25044	4.40	71261.10	4.85	71.26	1.42	[('Nc', 49025)]
23	該	48801	668.44	2.83	36654	4.56	104296.61	5.02	104.30	1.54	[('Nes', 29759), ('D', 19008), ('VL', 34)]
24	應	48795	668.36	2.83	34566	4.54	98355.34	4.99	98.36	1.54	[('D', 47712), ('VC', 360), ('P', 587), ('VJ', 97), ('VE', 39)]
25	大家	48779	668.14	2.82	36220	4.56	103060.14	5.01	103.09	1.37	[('Nh', 48712), ('Na', 67)]
26	位	48647	666.33	2.82	34387	4.54	97846.01	4.99	97.85	1.41	[('Nf', 48111), ('VCL', 241), ('Na', 295)]
27	教育	48418	663.19	2.82	24605	4.39	70011.95	4.85	70.01	1.41	[('Na', 47378), ('VC', 1040)]
28	現在	47679	653.07	2.81	38774	4.59	110328.93	5.04	110.33	1.38	[('Nd', 47679)]
29	本	47378	648.95	2.81	33750	4.53	96033.46	4.98	96.03	1.61	[('Nf', 12506), ('Nes', 31810), ('D', 2608), ('Na', 389), ('VJ', 65)]
30	提供	47356	648.65	2.81	36106	4.56	102737.31	5.01	102.74	1.57	[('VD', 47321), ('Nv', 34), ('VC', 1)]
31	知道	46941	642.96	2.81	36053	4.56	102586.50	5.01	102.59	1.22	[('VK', 46940), ('Nv', 1)]
32	起	46759	640.47	2.81	37159	4.57	105733.55	5.02	105.73	1.62	[('Nf', 3676), ('Ng', 30388), ('Di', 8664), ('VC', 2851), ('VA', 247), ('VJ', 867), ('VE', 59), ('VK', 7)]
33	去年	46737	640.17	2.81	32850	4.52	93472.57	4.97	93.47	1.57	[('Nd', 46737)]
34	社會	46611	638.44	2.81	29492	4.47	83917.60	4.92	83.92	1.42	[('Na', 46611)]
35	媽媽	46430	635.96	2.80	20231	4.31	57566.01	4.76	57.57	1.14	[('Na', 46430)]

圖1. 中研院臺灣當代華語語料庫:詞彙頻率、語境變異、語義變異、詞類標示 \

本院覽號

24T-1120207

公告日期

2023-01-03

智財權狀態

know-how

應用範圍

- 1. 提供中文相關學術研究之實驗材料篩選語料
- 2. 提供發展學童、醫院臨床或長照機關病人及長者中文語言能力測驗題庫的參考語料
- 3. 提供學校教師編制給學童或是華語文第二外語學習者的中文教材參考語料
- 4. 其他中文字詞選取之可能應用

創作人

李佳穎