

發展主題語料庫以輔助華語教學 —以 2019 新型冠狀病毒語料庫為例*

白明弘

國家教育研究院語文教育及編譯研究中心

陳浩然¹

國立臺灣師範大學英語學系

林鶯

國立臺灣師範大學英語學系

摘要

2019 新型冠狀病毒 (COVID-19) 對人類產生巨大的影響，歐洲及美國皆有團隊建立英文 COVID-19 語料庫，但華語圈目前尚未有類似語料庫。因此，本文希望能補足此一缺口，建立「中文 COVID-19 主題語料庫」供研究人員、老師及學生使用。本研究的研究問題有二：(1) 中文 COVID-19 語料庫和 No Sketch Engine 平臺能否提供有用的資訊？(2) 此語料工具有何優缺點？本研究以 WebBootCat 技術建構 COVID-19 主題語料庫，也產出各種教學素材：(1) 詞頻、(2) 關鍵詞、(3) 常見 N 連詞、(4) 搭配詞。此研究發現 WebBootCat 技術可有效生成主題語料庫，此庫有以下優點：(1) 即時性、(2) 廣泛涵蓋率、(3) 真實語言、(4) 豐富語境。然而，此語料庫是爬取網路資料所建成，不免納入不相關的雜訊，而平臺上仍有許多重要工具有待開發。

* 謝辭：本研究獲教育部核定之國立臺灣師範大學「高等教育深耕計畫」經費補助。感謝本期刊兩位匿名審查者提供寶貴建議，使本文更臻完善。文中如有疏漏，悉由作者負責。

¹ 本文通訊作者。

關鍵詞：N 連詞分析 主題華語 字詞頻率 搭配詞 網路作為語料庫
關鍵字詞分析

1. 前言

嚴重特殊傳染性肺炎(英語:Coronavirus disease 2019,縮寫:COVID-19),另稱武漢肺炎、新型冠狀病毒肺炎(新冠肺炎),是指在 2019 年至 2020 年間由嚴重急性呼吸系統症候群冠狀病毒 2 (SARS-CoV-2) 所引發的全球大流行疫情。疫情最初在 2019 年 12 月於中國湖北省武漢市被發現,隨後在 2020 年初迅速擴散至全球多國,逐漸變成一場全球性大瘟疫,被多個國際組織及傳媒形容為自第二次世界大戰以來全球面臨的最嚴峻危機。截至 2020 年 5 月 31 日,全球已有 220 多個國家和地區累計報告逾 600 萬名確診個案,逾 37 萬名患者死亡。由於 COVID-19 潛伏期長達 14 天,有個案則長達 24 天。感染者即使沒有感染跡象或僅有輕微感染跡象也可能將病毒傳染給他人,且無法有效篩查,因此比中東呼吸症候群 (MERS) 或嚴重急性呼吸道症候群 (SARS) 疫情更難控制,僅花四分之一的時間就造成 SARS 事件十倍의 確診數字,對全球航空、旅遊、娛樂、體育、石油市場、金融市場等方面造成巨大影響,也引發全球媒體的廣泛重視,產生巨量的相關報導。這些報導目前已吸引幾批專家學者建立了不同的 COVID-19 英文語料庫 (Davis 2020; Wicke and Bolognesi 2020; COVID-19 2020),作為研究與教學的基礎。在網路上也能看見一些英文老師針對 COVID-19 的特有單字跟片語製作詳細的影片介紹,如果能進一步善用語料庫資源,這些單字及片語都能夠在語料庫中查詢到實際的例子以作為教學素材。

相較英文的豐富資源,目前華語圈中,雖然中文相關新聞報導也很多,然而無論臺灣或其他華語地區都沒有類似的中文 COVID-19 主題語料庫。一般新聞報導常見主題包括戰爭、政府、政治、教育、衛生、環境、經濟、商業、時尚和娛樂以及體育賽事。由於新聞華語的重要性及普及性,國內幾個主要的語言中心都開設了相關的課程,如成功大學開設的「新聞聽力訓練」、政治大學提供新聞選讀課程、東海大學開設「新聞選讀」以及輔仁大學提供「讀報學華語(一)(二)(三)」及「新聞專題選讀」。此次 COVID-19 對全球多方面造成了巨大影響,由 COVID-19 的語料所建構的語料庫可以視為一個中型的華語新聞語料庫,因為該語料庫能含括的主題相當廣,包括政府、政治、教育、衛生、環境、疾病、經濟、商業、交通及旅遊等多個重要的主

題。中文 COVID-19 主題語料庫可以作為一個華語新聞教學的重要學習資源。

Leech (1997) 將語料庫在語言教學的應用劃分為直接應用和間接應用。直接應用指學生和教師在語言課堂上使用語料庫輔助教學，從而影響教與學的方式，即「數據驅動學習」(Data-Driven Learning, DDL)。語料庫亦可直接應用在語言教學中。教師可先向學生說明語料庫及語料分析工具等的使用方式，引導學生使用語料庫進行資料分析。學生一旦掌握了語料庫的檢索方式，就可以透過語料庫進行相關詞彙句法的比較、進行詞語搭配、詞彙語義和話語分析等，從而培養自主學習的能力。間接應用指將語料庫運用於工具書的出版、教材開發和語言測試等，幫助擬定教學內容。語料透過中心詞檢索 (concordancing)、關鍵字分析 (keyword analysis)、詞塊分析 (cluster analysis) 等方法，可從詞彙、詞頻、搭配詞、句型、語意及語用等方面進行分析，這些分析結果可直接或間接地應用在語言課程大綱設計、測驗、教材編寫等方面。

對華語教學研究而言，若有中文 COVID-19 主題語料庫，在編寫新聞類教材，以及提供相關學習輔助素材，如關鍵詞表或工具上，必定會有相當大的助益。此外，此語料庫也可以讓學生或者老師直接查詢使用。方便學生找尋單詞例句或單字、片語的各種不同用法。因此本研究針對此一缺口，嘗試建立 COVID-19 主題語料庫，提供各地學者專家或學校師生使用。本文除提出語料庫建置的方式外，也會介紹如何利用搜尋平臺來應用語料，並提出實際分析 COVID-19 中文語料並應用於教材或工具書編寫的方式。希望本研究建置語料庫的技術及如何融入開放平臺的方法能夠拋磚引玉，為國內更多相關教學研究提供參考。²

2. 文獻探討

2.1 COVID-19 主題語料庫現況

目前國際上的 COVID-19 主題語料庫仍不多見，較知名的當屬 English-Corpora.org 語料庫所發佈的冠狀病毒語料庫 (<https://www.english-corpora.org/corona/>)、英國 Word Sketch 團隊建立之 COVID-19 主題語料庫，以及數位學者利用推特 (Twitter) 建置之語料庫。以下分別介紹：

² 本文所建置之 COVID-19 主題語料庫搜尋平臺的網址為：http://corpus.eng.ntnu.edu.tw/bonito/run.cgi/first_form

2.1.1 English-Corpora.org 發佈之冠狀病毒語料庫

冠狀病毒語料庫 (Coronavirus Corpus) 是 English-Corpora.org 於 2020 年 5 月發佈的語料庫，其目標是希望成為記錄 2020 年後冠狀病毒 (COVID-19) 在社會、文化和經濟各面向影響的集大成者。該語料庫顯示的是 20 個不同英語國家的線上報紙及雜誌中，人們實際書寫的內容。目前的規模約為 3.31 億字 (2020 年 5 月 31 日)，並且每天以 3-4 百萬字的速度持續累積。按照這個速度，到 2020 年 8 月，它的規模可能會達到 5-6 億字。通過此語料庫，使用者可以觀察到自 2020 年 1 月以來以 10 天為單位所累進的相關詞彙和詞組頻率，如：social distancing (社交距離)、flatten the curve (拉平曲線)、Zoom、Wuhan (武漢)、hoard (囤積)、toilet paper (衛生紙)、pandemic (大流行)、reopen (解封)、defy (抗爭) 等；也可以從搭配詞得知當前最熱門的話題，例如：~virus (~病毒)、ban~ (禁止~)、stockpile~ (囤積~)、disinfect~ (消毒~)、或 remotely~ (遠距~) 等等。

此外，從不同時段比較這些詞彙頻率的消長，可以看到人們對事物的看法隨著時間的推移而發生變化，譬如 social~ (社交~) 或 economic~ (經濟~) 這樣的詞組在一、二月比四、五月更常見，或者 ban (禁止) 或 obey (遵循) 的詞在四、五月比一、二月更常見；從不同區域來看，則可觀察語料庫中 20 個國家 (美國，英國，澳大利亞，印度等) 對冠狀病毒的不同看法。而且，與大多數線上語料庫一樣，冠狀病毒語料庫能查看上下文中依詞類 (PoS) 排序的關鍵字詞或短語置中 (KWIC) 關鍵詞語索引 (concordance)，還可以根據文本、國家／地區、日期、出版來源等後設資料，快速輕鬆地為特定主題創建個人化的「虛擬語料庫」(Davis 2020)。

2.1.2 Word Sketch 建立之 COVID-19 主題語料庫

Word Sketch (詞彙特性速描系統) 最早由英國國家語料庫 (British National Corpus) 所提出 (Kilgariff and Tugwell 2001)，是從語料庫中抽取詞彙特定語法關係上的搭配詞和頻率的搜尋引擎，目的在作為詞彙學家編輯辭典、語言教師編輯教材、學生自學語言時的工具。Word Sketch 於 2020 年 5 月發佈 COVID-19 主題語料庫，語料來源為 COVID-19 開放研究數據集 (CORD-19)³ 中的部分英文文本，規模約為 2 億 2 千萬詞，所提供的功能包括

³ COVID-19 開放研究數據集 (CORD-19)。2020。版本 2020.05.02。該語料庫是屬於「開放使用」的範疇，http://ske.li/covid_19。

列出每一個檢索詞的詞彙特性素描及近義詞集、關鍵詞分析、字表分析、N 連詞分析、例句等等。這個語料庫建立的目標是希望能藉由自然語言處理技術，使研究者能加速理解這種傳染病，進而有效對抗及預防它。隨著新的研究論文陸續提出，目前該語料庫也在定期更新中 (COVID-19 2020)。

2.1.3 利用 Twitter 建置之語料庫

Wicke 與 Bolognesi (2020) 基於 Twitter 建立了一個規模為 20 萬條推文的語料庫，該語料庫蒐集時間為 2020 年 3 月和 4 月，主旨是 COVID-19 話題。其建立的原因是作者觀察到在 COVID-19 病毒傳播期間，戰爭譬喻常被用來作為流行病和抗疫作為的描述框架，不僅在公共話語和媒體上，在非大眾傳播專家的推文中也會使用，為了分析戰爭框架在流行病上的使用，他們利用推文可以按話題分類的特性，應用語言分析指出戰爭框架最常用來指涉病毒治療。若以怪物 (MONSTER)、風暴 (STORM) 和海嘯 (TSUNAMI) 等三個隱喻框架，以及一個不用譬喻的字面框架 (FAMILY) 作為對照組，來測量、比較語料庫中戰爭框架與另四個框架的受歡迎程度，結果顯示，雖然字面框架覆蓋了更多的語料，但在隱喻框架中，戰爭框架是最常用的，因此也可以說是最受歡迎的框架。作者認為雖然戰爭譬喻並不能完整地闡述疫情過程中所涉及的各種面向，但是它在病毒大流行期間仍有助於提供社交媒體的使用者表達他們的感受、意見和想法。例如其中一則推文說「這幾周，醫生和護士們在戰壕裡忙碌著，與一個新的隱形敵人 COVID-19 病毒戰鬥。城市被封鎖，平民被圍困在自己的家中，以防止病毒的傳播。」即為其例。

有鑑於以上三個英文 COVID-19 主題語料庫剛剛問世，對其應用的研究很少。到目前為止，還沒有類似的中文語料庫。所以，在本研究中，我們希望開發一個新的 2019 新型冠狀病毒主題語料庫 (COVID-19 主題語料庫)，並通過 No Sketch Engine 平臺 (Rychlý 2007) 提供語料庫分析的功能。在本研究中，我們想進一步研究網路語料庫產生的各種輸出以及相關工具是否對華語二語教學 (Chinese as a Second Language, CSL) 的學生和教師有用。以下為本研究之研究問題：

1. COVID-19 主題語料庫和 No Sketch Engine 平臺能否為華語教師和學生提供有用的資訊？
2. 這個網路平臺就語料內容及分析上的優點和缺點為何？有何可改進之處？

3. 研究方法

本文研究方法主要分為兩個部分：一是藉由爬取網路資料，自動建置 COVID-19 主題語料庫；二是將建好的主題語料庫整合入智慧搭配詞搜尋引擎平臺，利用平臺提取常用詞、關鍵詞、N 連詞及詞彙搭配等語言訊息，以檢視其相關度及可用度。以下分別說明：

3.1 建置 COVID-19 主題語料庫

過去，建立語料庫是一件十分困難的工作，往往曠日持久，以教學領域為例，通常是由教師一篇一篇的收集文章，而不同老師彼此間也難以分享資料。目前雖有眾多網路資料，但要一窺某一主題的全貌也有相當的難度。隨著網路科技的日新月異，近年來語料庫建置的方式已有了技術上的突破。早在 1999 年，Resnik (1999) 就以網路作為語料來源，發表了以多語版本的網頁建置平行語料庫的方法。Fujii 與 Ishikawa (2000) 則是利用網路語料揭示了自網頁中蒐集專業科技術語定義的方法。又如 Jones 與 Ghani (2000) 以塔加拉族語料庫的建構為例，展示了運用網路來建構語料庫的方法。到了 2003 年，Kilgariff 與 Grefenstette (2003) 直接提出「將網路作為語料庫」(Web as Corpus) 的概念。他們指出，因為網路資料量大，加上取得容易的優勢，使得善用搜尋引擎蒐集網路上數量龐大的語料，成為建置語料庫更便捷、有效率的方式。與此同時，愈來愈多的語言科學家及技術專家也傾向將網路作為語言資料的來源。學者們在 2005 年組成了「將網路作為語料庫」的工作坊，至 2020 年，已舉辦了 12 次的研討會，發表了不少應用網路語料庫 (Web Corpus) 進行研究的範例。

反觀國內卻較少相關的研究，現存之華語語料庫也多半採用傳統作法，直接從教科書、書信、備忘錄、新聞、報紙、傳單等蒐集語料。為增進蒐集語料的速度與效率，也就是能在短時間內蒐集大量各式各樣語料，並大幅縮短語料庫建置的時間與人力成本，本研究嘗試透過將網路作為語料庫的概念蒐集語料，以 WebBootCat 為工具，建造一個網路 COVID-19 主題語料庫。並進一步將此 COVID-19 主題語料庫分享於開放式語料搜尋引擎 (No Sketch Engine)。

WebBootCat 意指 Bootstrapping Corpora and Terms from the Web，是以 BootCat 軟體為基礎所開發的線上工具，為 Baroni 與 Bernardini 於 2004 年所提出的語料蒐集方式。BootCat 軟體主要透過種子詞 (seed words) 在網路資

源中進行搜尋，然後將符合種子詞的網頁內容集結起來形成語料庫。透過 BootCat 軟體蒐集語料最大的特色，在於整個語料蒐集的過程皆可透過電腦自動處理，且語料庫內容可根據自訂的種子詞而有所調整，更有助於專業領域的語料蒐集。Baroni、Kilgarriff、Pomikálek 與 Rychlý（2006）為方便使用者免於軟體下載、安裝等步驟，進而開發了 WebBootCat，也就是線上版的 BootCat 軟體，可供使用者直接在 WebBootCat 網頁上執行語料蒐集工作。

透過 WebBootCat 軟體蒐集語料，首先須選定種子詞。本研究首先分析維基百科對 COVID-19 的描述內容，經團隊討論後找出較核心的 50 個關鍵詞，並將這 50 個關鍵詞輸入 WebBootCat 的語料爬取系統。選擇繁體中文為搜尋語言，並設定需要的資料量和搜尋的檔案類型，WebBootCat 將這 50 個種子詞隨機結合成為一百組的檢索詞組合，如：「病毒」、「醫院」與「武漢」為一組，結合起來在網路中進行搜尋。搜尋結果會自動列出包含這組詞的網頁，這些網頁中的所有文字即為本研究中 COVID-19 主題語料庫之語料來源。此外，蒐集主題語料庫一方面必須考慮盡可能涵蓋多面向的次主題，例如，和新冠肺炎相關的醫藥疫苗、隔離生活、防疫衛生、病毒知識等報導；另一方面，又要避免主題過於發散以致蒐集到和主題無關之語料。例如只以「疫情」單一詞搜尋，會蒐集到各種傳染病相關主題的報導。所以結合關鍵詞的數目必須適當選擇。本研究根據一般研究如 Sketch Engine 的商業版的建議 (Baroni et al. 2006; Lui and Cook 2013)，選擇結合三個種子關鍵詞成為檢索詞進行搜尋。另外，在檢索詞的數量上，50 個種子詞可以窮舉結合成 12 萬 5 千個檢索詞，如此龐大的搜尋次數在實踐上十分困難，在語料蒐集上意義也不大。而以隨機結合產生檢索詞的方法，其背後原理類似隨機抽樣的概念。每一組檢索詞都可能涵蓋一到兩個次主題，透過上百個隨機產生的檢索詞，應該足以涵蓋和新冠肺炎相關的重要次主題。

接著，我們利用第一波搜尋出來的語料，從中自動抽取出更多關鍵詞後，再度將這批新的種子詞輸入到語料爬取系統。如此來回重複做上數次後，就可以在網路上收集到大量語料。總計本研究所建之 COVID-19 語料量約為 43,142,007 詞。這些語料藉中央研究院中文斷詞系統自動斷詞⁴處理完成後，統整為本研究之 COVID-19 主題語料庫。上述方式不僅步驟簡單，而且可以

⁴ 本研究採用中央研究院新開發之 python 版斷詞系統進行批次斷詞處理，該系統為免費授權軟體，可從 github 下載 (<https://github.com/ckiplab/ckiptagger>)。

將大量的語料即時提供給前臺的各地師生使用，並進一步利用網絡平臺 No Sketch Engine 呈現語言分析結果，方便老師、學生及研究人員使用。

3.2 No Sketch Engine 平臺

我們發展的網路平臺是 No Sketch Engine 的中文版本 (Rychlý 2007)。Sketch Engine 是由 Lexical Computing Ltd. 所開發的一套語料庫管理與文本分析系統。其目的是讓辭典語言學家、語料庫語言學家、翻譯者或語言學習者能夠在大量的文本中，使用複雜的語言查詢規則來進行檢索。No Sketch Engine 則是 Sketch Engine 的簡化版本，Lexical Computing Ltd. 將 Sketch Engine 的部分功能刪除後釋出為 No Sketch Engine。儘管已經刪除了包含 word sketch、thesaurus 等在內的重要功能，No Sketch Engine 仍保留語料庫系統最重要的語料庫管理、關鍵詞查詢、KWIC 及搭配詞分析等功能（參見圖 1）。在中文版 No Sketch Engine 中我們運用了以下功能以便應用在華語教學上：

第一，詞頻分析：詞頻分析是從語料中統計出每一個詞語的出現次數。在語料庫的研究中，一般都認為較高頻的詞應是學生優先學習的目標，而作為教學的重點，此系統的詞頻計算能提供最基礎且有用的資訊。

第二，關鍵詞分析：由於詞表中參雜一般高頻核心詞，降低主題語料庫凸顯關鍵詞的效能，因此，關鍵詞分析的功能主要是排除核心詞。透過主題關鍵性 (keyword-keyness) 統計法來找出主題關鍵詞彙，如圖 2 所示。主題關鍵性統計法是通過詞語的關鍵性分析 (keyness) 來找出某一主題文本的詞語特徵 (William 1976) 的方法，早期多應用於分析文類或作者風格。但主題關鍵性分析的統計概念，即找出不同子語料庫中相對高頻或低頻的關鍵性詞語 (Baker 2006:139)，也有助於分析不同語料的用詞特徵，因此也被用作提取關鍵詞的方法。

Rayson 與 Garside (2000) 指出目前利用主題關鍵性方法所做的研究主要有兩種類型：第一類是把兩個語料庫相比，其中較小的語料庫是被觀察語料庫，較大的語料庫則作為參照標準。比較的目的是透過顯著不同的高頻詞來擷取被觀察語料庫的特徵；第二類是對比兩個差不多大小的語料庫，目的是想要找出可以區辨兩個語料庫的詞彙特徵。它們的基本原理都是利用統計的方法，藉由比對兩個語料庫所產出的詞表，檢視哪些詞是屬於不尋常高頻或不尋常低頻的主題詞 (Scott 1997:236)，進而分析文體類型或作者風格。原

則上，這些研究方法都基於一個參照語料庫及一個被觀察語料庫，而參照語料庫通常要大於被觀察語料庫 (Scott and Tribble 2006:58)。由於它不是分析單一語料庫的字詞頻，因此能凸顯專門用語在特定話題中扮演的角色，有助於教學研究上的應用分析。

NoSketch Engine COVID19

Query 疫情 122,629 > GDEX 122,629 (2,211.14 per million) ⓘ

Page 1 of 6,132 Go Next | Last

00001 Reston 伊波拉 病毒 死亡 的 疫情 。在 非洲，果蝠（尤其是

00001 2007 與 2012 年 烏干達 發生 疫情 。2013 年 12 月 西非 發生 疫情

00001 剛果 及 剛果 民主 共和國 數 度 發生 疫情 ； 2007 與 2012 年 烏干達

00001 之 致死率 平均 約 50%，依 過去 疫情 經驗 約 在 25% - 90%。二

00001 綜上 所述，高度 懷疑 此次 疫情 與 野生 動物 交易 有關。對 核酸

00001，2004 年 蘇丹 南部 省份 爆發 疫情，同 年 在 俄羅斯 及 美國 亦 曾

00001 發生。2014 年 造成 西非 大規模 疫情 的 病毒 株 即 為 Zaire 病毒。；

00001 等於 新增 確診 病例 數。目前 看，疫情 防控工作 取得 階段性 成效。全 國

00001 元。13. 補助 受 疫情 影響 中小 企業 創新 研發 等 所需 經費

00001 全力 守護 國人 健康。為 有效 防治 疫情，政府 已 採行 各項 因應 措施

00001 的 野生 動物。武漢 新型 冠狀 病毒 疫情 早期 確診 的 病例，大多 來自 武漢

00001，追查 感染 源 及 找出 接觸 者。疫情 調查 及 感染 源 調查 工作 事項 請 見

00001。武漢 肺炎 (Covid-19) 疫情 仍在 持續，中國 官方 稱 新增

00001 元。13. 補助 受 疫情 影響 中小 企業 創新 研發，提升 既有

00001 等地，陸續 有 大小 不等 的 疫情 爆發。其中 以 1995 年 在 薩伊

00001 曾 在 非洲 地區 造成 數 起 大規模 疫情 發生。2014 年 造成 西非 大規模

00001 法 訂定 「因應 嚴重 特殊 傳染 性 肺炎 疫情 整備 應變 計畫」。另 依據 國際

00001) 下午 召開 新聞 發佈 會，介紹 為 疫情 防控、復工 復產 和 實體 經濟 發展

00001 剛果 民主 共和國 再度 發生 疫情，世界 衛生 組織 並 於 2019 年

00001 Reston 伊波拉 病毒 感染 疫情 美國 於 1989 年 與 1990 年，

Page 1 of 6,132 Go Next | Last

圖 1：No Sketch Engine 平臺的搜尋介面

NoSketch Engine COVID19 defaults					
<div>Home</div> <div>Search</div> <div>Word list</div> <div>Ngram list</div> <div>Corpus info</div> <div>My jobs</div> <div>User guide</div>					
<div>Save</div> <div>Change options</div>					
Menu position					
<div>Word list</div> <div>Corpus: COVID19</div> <div>Reference corpus: TWWaC</div> <div>Switch focus and reference (sub)corpus</div> <div>Page 1 Go Next ></div>					
<div>COVID19</div> <div>TWWaC</div>					
word	average reduced frequency	average reduced frequency/mill	average reduced frequency	average reduced frequency/mill	Score
疫情	38,351.00	691.5	2,638.90	1.8	245.4
確診	17,172.90	309.6	710.00	0.5	208.4
肺炎	28,322.90	510.7	2,610.80	1.8	182.6
武漢	17,139.20	309.0	1,512.40	1.0	151.7
冠狀	8,410.70	151.7	245.20	0.2	130.6
新冠	4,957.20	89.4	2.20	0.0	90.2
病例	17,483.50	315.2	4,417.70	3.1	78.1
為了	4,196.00	75.7	0	0.0	76.7
防疫	12,436.60	224.2	3,664.20	2.5	63.8
病毒	48,299.70	870.9	21,838.10	15.1	54.2
冠狀病毒	2,696.80	48.6	0	0.0	49.6
新冠肺炎	2,659.00	47.9	0	0.0	48.9
感染	48,498.90	874.5	25,020.40	17.3	47.9
口罩	12,818.60	231.1	6,134.90	4.2	44.3
檢疫	4,607.80	83.1	1,377.20	1.0	43.1
呼吸道	8,447.10	152.3	3,706.60	2.6	43.1
COVID-19	2,204.00	39.7	0	0.0	40.7
世衛	2,483.60	44.8	247.30	0.2	39.1
防控	2,126.80	38.3	18.20	0.0	38.9

圖 2：No Sketch Engine 平臺提供的 COVID-19 主題語料庫關鍵詞擷取畫面

第三，高頻 N 連詞：N 連詞 (N-gram)，是指「由三個或以上的字詞所組合且重複出現的詞組排列」(Biber, Johansson, Leech, Conrad and Finegan 1999)。它在語言學習中日漸受到重視。有些研究者從認知科學的觀點指出，N 連詞的習得可以幫助學習者提升語言處理的效率 (Wray and Perkins 2000; Simpson-Vlach and Ellis 2010)。Biber 與 Barbieri (2007) 則表示 N 連詞會依據語言交流目的不同而有差異。同時，許多研究者 (e.g. Nattinger and DeCarrico 1992; Lewis 1997; Wray and Perkins 2000; Willis 2003; Simpson-Vlach and Ellis 2010) 都認為熟悉 N 連詞有助於學習者學習語言，非母語者適時地使用 N 連詞不僅可以使措辭更加接近母語者的語用，也能讓語言更加自然。

為滿足學習者查詢 N 連詞之需求及時機，原本 No Sketch Engine 平臺並未提供擷取 N 連詞的功能，本研究在 No Sketch Engine 平臺的基礎上開發了 N 連詞的提取功能。透過統計連續 N 個詞在語料庫中重複出現的次數，然後，再依出現頻率由高而低，排列 N 連詞的結果（參見圖 3）。

Corpus: COVID19		
N-gram range: from 3-gram to 6-gram.		
Total number of n-grams: 164,851		
Total number of pages: 5,496		
[Download n-gram list]		
Page	1	/5,496 Next >
SN.	Word	Count
1	新型 冠狀 病毒	5,953
2	世界 衛生 組織	3,254
3	是 一 個	2,078
4	疫情 指揮 中心	1,906
5	是 一 種	1,798
6	流行 疫情 指揮	1,636
7	流行 疫情 指揮 中心	1,626
8	中央 流行 疫情	1,617
9	中央 流行 疫情 指揮	1,596
10	中央 流行 疫情 指揮 中心	1,586
11	武漢 肺炎 疫情	1,415
12	最 大 的	1,415
13	華南 海鮮 市場	1,365
14	特殊 傳染性 肺炎	1,193
15	嚴重 特殊 傳染性	1,192
16	嚴重 特殊 傳染性 肺炎	1,187
17	無 症狀 感染	1,077
18	冠狀 病毒 肺炎	1,073
19	的 情況 下	1,067
20	境 外 移入	1,055
21	新型 冠狀 病毒 肺炎	1,035
22	症狀 感染 者	1,027
23	的 一 個	1,012
24	了 一 個	1,012
25	無 症狀 感染 者	1,009
26	冠狀 病毒 感染	985

圖 3：No Sketch Engine 平臺提供的 COVID-19 N 連詞擷取畫面

第四，搭配詞：根據 Lewis (2000:245)：「搭配詞為經常以可預測的模式共同出現的字⁵」。另外，根據牛津的線上搭配詞辭典 (Online Collocation Dictionary, <https://www.freecollocation.com/>)，搭配詞為「常見的單詞組合，

⁵ 原文：Collocations might be described as the words that are placed or found together in a predictable pattern.

並且是自然語言中不可或缺的基本組成單元⁶。」例如哪些介詞與特定的動詞搭配使用，或者哪些動詞與名詞搭配使用。它們基於語言習慣而一起出現，且較少和其他近義詞共現，於是成為彼此的搭配詞，例如：「開 支票」，「新年 快樂」和「划 龍舟」，我們鮮少說成「新年 高興」、「簽 支票」或「駕 龍船」等等。搭配詞在二語學習領域被廣泛認為是詞彙能力的關鍵指標，掌握搭配詞不僅可以促進學習者的語言表達和整體理解，而且還可以增進學習者的二語流利程度，從而達成像本地人使用母語一樣程度的學習目標。相較於國內著名的中研院平衡語料庫，將專門主題的語料庫整合到 No Sketch Engine 平臺後，可提取出更適用的搭配詞，因為這些搭配詞是由大量的相關報導中提取出來，搭配詞更豐富，類型也更加多元（參見圖 4）。

Collocation candidates						
Page <input type="text" value="1"/>		<input type="button" value="Go"/>	Next >			
	<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>T-score</u>	<u>MI</u>	<u>logDice</u>	
P N 病毒	25,355	169,581	156.520	5.875	11.384	
P N 者	16,274	124,532	125.084	5.681	10.970	
P N 症狀	9,515	73,840	95.617	5.661	10.501	
P N 肺炎	9,782	93,655	96.493	5.358	10.414	
P N 管制	6,317	15,953	78.968	7.280	10.363	
P N 新型	6,422	33,678	79.067	6.226	10.232	
P N 冠狀	6,223	32,705	77.830	6.223	10.195	
P N 無	6,773	50,820	80.726	5.709	10.174	
P N 呼吸道	5,703	27,899	74.577	6.327	10.109	
P N 感染	8,801	141,205	89.981	4.613	9.996	
P N 被	6,769	82,106	79.733	5.017	9.956	
P N 人	9,096	165,762	90.947	4.429	9.923	
P N 內	6,826	89,743	79.854	4.900	9.919	
P N 可能	6,396	78,910	77.462	4.992	9.895	
P N 院	4,440	16,674	65.996	6.708	9.847	
P N 有	13,425	336,909	108.462	3.968	9.845	
P N 風險	4,514	26,145	66.195	6.083	9.787	
P N 群聚	3,960	10,408	62.507	7.223	9.741	

圖 4：No Sketch Engine 平臺提供的搭配詞擷取畫面

⁶ 原文：Collocations/collocation.....- common word combinations are the essential building blocks of natural-sounding English.

下一節將依序介紹我們如何利用 COVID-19 主題語料庫及網路語料庫平臺分析 COVID-19 的相關語料。主要從四個面向來討論：一是有關高頻詞的提取，二是有關關鍵詞的提取，三是有關常見 N 連詞的抽取，四是有關搭配詞的搜尋。

4. 分析結果與討論

如前所述，以下將根據本研究提出的兩個問題，呈現分析結果。

4.1 研究問題一：COVID-19 主題語料庫和 No Sketch Engine 平臺能否為華語教師和學生提供有用的資訊？

4.1.1 高頻詞表建置

應用我們所建置的大型 COVID-19 主題語料庫來產生詞表十分簡便，教師或學生只要選取 No Sketch Engine 平臺裡詞表的 make word list 鍵，系統就會產生最高頻的單詞表，並依其在語料庫中出現的詞頻由高而低排序。表 1 列出的是 COVID-19 主題語料庫前一百個最高頻詞。然而，這樣產生的華語詞表固然有教學的重要性，但是其計算方式會包含大量的核心詞彙，如：助詞「的」、繫詞「是」、存現動詞「在」等。核心詞不管在哪一類型的語料庫中位置都高居榜首。這些詞通常與主題不相關，列在最高頻而跟主題較相關的詞反而不多。如表 1 所顯示出來與 COVID-19 相關的詞大約只有十多個，例如：病毒、感染、疫情、肺炎、治療等（以粗體標示）。此方法反而比較不容易看到 COVID-19 主題語料庫獨特的詞彙。因此，在下一節中會介紹另一種擷取關鍵詞的方式，它能產生跟 COVID-19 主題相關性更高的詞表。

表 1：COVID-19 詞表的前 100 詞（前 300 高頻字表參見附錄）

1	的	21	會	41	醫院	61	說	81	年
2	是	22	等	42	治療	62	各	82	還
3	在	23	就	43	大	63	天	83	我們
4	有	24	後	44	應	64	確診	84	如
5	一	25	者	45	被	65	次	85	來
6	及	26	疫情	46	要	66	其	86	本
7	之	27	上	47	使用	67	可以	87	前
8	了	28	並	48	我	68	至	88	台灣

表 1：COVID-19 詞表的前 100 詞（前 300 高頻字表參見附錄）（續）

9	和	29	中國	49	可能	69	每	89	其他
10	或	30	對	50	種	70	病例	90	由
11	不	31	於	51	人員	71	出現	91	從
12	與	32	可	52	他	72	沒有	92	表示
13	個	33	都	53	例	73	美國	93	研究
14	也	34	以	54	已	74	衛生	94	工作
15	病毒	35	時	55	症狀	75	醫療	95	相關
16	人	36	但	56	最	76	疾病	96	中心
17	為	37	而	57	患者	77	到	97	所
18	這	38	將	58	能	78	進行	98	兩
19	感染	39	肺炎	59	病人	79	多	99	很
20	中	40	內	60	武漢	80	口罩	100	名

4.1.2 關鍵詞表建置

本研究透過 No Sketch Engine 平臺所提供的主題關鍵性統計法來找出 COVID-19 的主題關鍵詞彙。主題關鍵性統計法是通過詞語的關鍵性分析來找出某一主題文本的詞語特徵 (William 1976) 的方法。它的基本原理是藉由比對兩個語料庫的詞彙，以統計法找出不尋常高頻或不尋常低頻的主題詞 (Scott 1997:236)。原則上，這些研究方法都基於一個參照語料庫及一個被觀察語料庫。依據 Scott 與 Tribble (2006:58) 的建議，參照語料庫要大於被觀察語料庫。依此，本研究以包含約 1 兆 1 億 3 千萬詞的台灣網路語料庫 (Taiwan Web Corpus) (吳鑑城、陳浩然、張俊盛 2017) 作為參照語料庫，包含 4 千 3 百萬詞的 COVID-19 主題語料庫作為被觀察語料庫；然後利用 No Sketch Engine 平臺的關鍵性分析計算比較兩者並篩除核心詞後，再將不尋常高頻的詞彙提取出來作為關鍵詞，成為主題語料庫中 300 個與主題最相關的關鍵詞彙。

表 2：COVID-19 中的前 100 詞關鍵詞表

1	確診	21	防疫	41	呼吸道	61	醫療	81	報道
2	冠狀	22	口罩	42	呼吸器	62	症狀	82	感染症
3	新冠	23	檢疫	43	第 2	63	病例數	83	移入
4	武漢	24	李文亮	44	病原體	64	衛福部	84	大紀元
5	疫情	25	抗疫	45	漂白水	65	隔離	85	進行
6	肺炎	26	疾控	46	傳染病	66	川普	86	專家組
7	防控	27	の	47	毎日	67	陽性	87	X 光
8	世衛	28	李斌	48	檢體	68	負壓	88	疫區
9	採檢	29	福利部	49	接觸史	69	に	89	插管
10	新冠肺炎	30	衛健委	50	陰性	70	治療	90	を
11	冠狀病毒	31	爲	51	MERS	71	第	91	病原
12	為了	32	收治	52	病毒	72	呼吸機	92	王陽
13	COVID-19	33	傳染性	53	湖北省	73	COVID-	93	る
14	疾管署	34	冠肺炎	54	病毒性	74	病室	94	鍾南山
15	病例	35	飛沫	55	不	75	潛伏期	95	拭子
16	圖博館	36	第 1	56	湖北	76	感染	96	發病
17	旅遊史	37	流感	57	Id":	77	航母	97	宿主
18	武漢市	38	譚德塞	58	WHO	78	SARS	98	洗手
19	蘇灼	39	陳時中	59	jpg"	79	核酸	99	高衍
20	管制署	40	而言	60	伊波拉	80	名詞	100	C 型

表 2 列出前 100 個關鍵詞，我們發現其中夾雜少數雜訊（例如日文假名和非詞的符號），這是因為爬取網頁時，日文漢字的關鍵詞和中文沒有差別所致，例如：

「武漢市は集中隔離觀察施設の隔離者とニーズがある在宅觀察中の濃厚接觸者に約 52 萬 1,000 人分の漢方煎じ薬と 24 萬 8,000 人分の漢方製剤を配り、隔離者の多くが漢方を服用している。」

（武漢市向集中隔離觀察所的檢疫人員和居家觀察期間有需求的集中接觸者發放了約 52 萬 1,000 份中藥煎劑和 24 萬 8,000 份中藥製劑，不少隔離者正在服用中藥。）

上文中，「武漢、隔離、接觸」是爬取網頁時重要的關鍵詞，而其日文和中文寫法完全相同，致使少數日文文獻會被收入。在關鍵詞分析時，由於這些日文假名在統計分布上和主題關鍵詞相同，又因為在參照的中文語料庫中沒有相應的核心詞能篩除它們，所以即使這些詞只出現少數幾次，也會被視為是此語料庫獨有的詞，因此極容易出現在關鍵詞表中。要改善此問題，可以利用調整系統內建之頻率特徵 (average-reduced-frequency, ARF) 來提升查詢成效，得到表 3 的結果。

表 3：COVID-19 中的前 100 詞關鍵詞表（調整頻率特徵後）

1	疫情	21	傳染性	41	咳嗽	61	新型	81	宿主
2	確診	22	傳染病	42	疫苗	62	衛健委	82	移入
3	肺炎	23	而言	43	病原體	63	疑似	83	感染源
4	武漢	24	旅遊史	44	第 1	64	病原	84	報道
5	冠狀	25	流感	45	福利部	65	病人	85	入院
6	新冠	26	疾管署	46	抗體	66	病毒性	86	負壓
7	病例	27	隔離	47	潛伏期	67	體溫	87	X 光
8	為了	28	陰性	48	接觸史	68	呼吸器	88	N95
9	防疫	29	每日	49	檢體	69	群聚	89	醫護
10	病毒	30	武漢市	50	湖北省	70	COVID-	90	接種
11	冠狀病毒	31	陽性	51	湖北	71	陳時中	91	漂白水
12	新冠肺炎	32	飛沫	52	核酸	72	衛福部	92	染病
13	感染	33	抗疫	53	疾控	73	疫區	93	CDC
14	口罩	34	收治	54	急性	74	SARS	94	傳染力
15	檢疫	35	第	55	肺部	75	症狀	95	傳染給
16	呼吸道	36	管制署	56	病例數	76	流行病學	96	救治
17	COVID-19	37	發病	57	第 3	77	入境	97	感染症
18	世衛	38	冠肺炎	58	洗手	78	截至	98	爆發
19	防控	39	傳染	59	病房	79	加護	99	插管
20	採檢	40	WHO	60	患者	80	MERS	100	譚德塞

4.1.3 各種長度之 N 連詞分析

N 連詞指的是「由三個或以上的詞所組合且重複出現的詞組排列」(Biber et al. 1999)，換句話說，N 連詞即為有固定組合模式的多字詞。N 連詞可從字面看出語意，也可能是不完整的結構 (Biber 2006; Lin 2011)。許多研究都指出母語者在語言表達中頻繁地使用 N 連詞，習慣且非常自然地將這種詞彙組合融入於語言表達中，然而在教學上，由於無法完整解釋 N 連詞，很少教師會教導由多字詞組構成的 N 連詞，因而成為非母語者在語言學習上的難點之一。

目前，N 連詞已成為研究英文多字詞組領域中的研究重點之一 (Howarth 1998)。相關成果例如：研究證實熟悉 N 連詞不僅能輔助學習者閱讀上的理解、提供寫作上的指引，以及提升語言處理的效率，更可以協助二語學習者的語言表達更貼近母語者 (e.g. Nattinger and DeCarrico 1992; Lewis 1997; Wray and Perkins 2000; Willis 2003; Simpson-Vlach and Ellis 2010)。諸如此類的英文 N 連詞相關研究 (Simpson-Vlach and Ellis 2010; Salazar and Joy 2011) 不勝枚舉，然而針對中文 N 連詞的探究迄今少見，特別是新興主題 N 連詞方面的研究。現在，透過語料庫工具的輔助，研究者得以大量蒐集、整合不同的語料，探究 N 連詞在不同主題中的語言角色及功能。這將使 N 連詞的研究從語言學的範疇擴展至語言教學的領域，華語教育者得以開始探究 N 連詞習得在語言學習上的影響。

以 COVID-19 主題語料庫為例，本研究在 No Sketch Engine 平臺的基礎上開發了 N 連詞的提取功能。透過統計連續 N 個詞在語料庫中重複出現的次數，再依出現頻率由高而低，統計並產出 3 字詞／4 字詞／5 字詞／6 字詞之 N 連詞列表（參見表 4-5）。N 連詞列表提供各 N 連詞出現的頻率，從數量上顯示出 COVID-19 之 N 連詞以 3 字詞與 4 字詞 N 連詞較多。為了提升詞表的精確性，本研究透過人工篩選，將因斷詞錯誤而產生的錯誤 N 連詞，如：「的另一」、「上所述」、「了一種」等 N 連詞刪除，整理出依頻率及文本分佈率篩選後的高頻 3 字與 4 字 N 連詞，提供師生研究或自學時使用。研究人員、教師及學習者也可自行登入 No Sketch Engine 平臺自行檢索其所需長度的 N 連詞。

表 4：COVID-19 主題語料庫中 3-5 字 N 連詞表

3 字詞 N 連詞	4 字詞 N 連詞	5 字詞 N 連詞
新型 冠狀 病毒	流行 疫情 指揮 中心	中央 流行 疫情 指揮 中心
世界 衛生 組織	中央 流行 疫情 指揮	新型 冠狀 病毒 感染 的
疫情 指揮 中心	嚴重 特殊 傳染性 肺炎	冠狀 病毒 感染 的 肺炎
是 一 種	新型 冠狀 病毒 肺炎	中國 疾病 預防 控制 中心
流行 疫情 指揮	無 症狀 感染 者	人類 免疫 缺乏 病毒 感染
中央 流行 疫情	新型 冠狀 病毒 感染	流行 疫情 指揮 中心 今
武漢 肺炎 疫情	病毒 感染 的 肺炎	急性 病毒性 C 型 肝炎
最 大 的	冠狀 病毒 感染 的	新型 冠狀 病毒 肺炎 疫情
華南 海鮮 市場	2019 新型 冠狀 病毒	病毒 感染 的 肺炎 疫情
特殊 傳染性 肺炎	新型 冠狀 病毒 的	疫情 指揮 中心 指揮官 陳時中
嚴重 特殊 傳染性	衛生 福利部 疾病 管制署	國際 公共 衛生 緊急 事件
無 症狀 感染	疾病 預防 控制 中心	流行 疫情 指揮 中心 指揮官
冠狀 病毒 肺炎	的 新型 冠狀 病毒	關注 公共 衛生 緊急 事件
的 情況 下	嚴重 急性 呼吸道 症候群	在 全球 報紙 版面 上
境 外 移入	公共 衛生 緊急 事件	國際 關注 公共 衛生 緊急
症狀 感染 者	2019 冠狀 病毒 疾病	的 突發 公共 衛生 事件
的 一 個	突發 公共 衛生 事件	國際 關注 的 突發 公共
冠狀 病毒 感染	人類 免疫 缺乏 病毒	關注 的 突發 公共 衛生
	例 境 外 移入	監測 與 邊境 管制 措施
	居家 檢疫 14 天	疫情 監測 與 邊境 管制

表 5：COVID-19 主題語料庫中 6 字 N 連詞表

6 字詞 N 連詞
新型 冠狀 病毒 感染 的 肺炎
中央 流行 疫情 指揮 中心 今
中央 流行 疫情 指揮 中心 指揮官
冠狀 病毒 感染 的 肺炎 疫情
關注 的 突發 公共 衛生 事件
國際 關注 的 突發 公共 衛生

表 5：COVID-19 主題語料庫中 6 字 N 連詞表（續）

6 字詞 N 連詞
疫情 監測 與 邊境 管制 措施
流行 疫情 指揮 中心 指揮官 陳時中
加強 疫情 監測 與 邊境 管制
持續 加強 疫情 監測 與 邊境
依 指示 配戴 口罩 儘速 就醫
主動 告知 醫師 旅遊史 及 接觸史
並 依 指示 配戴 口罩 儘速
疾管署 持續 加強 疫情 監測 與
同時 主動 告知 醫師 旅遊史 及
美國 疾病 控制 與 預防 中心
病毒 感染 的 肺炎 疫情 防控
撥打 免 付費 防疫 專線 1922
特殊 傳染性 肺炎 防治 及 紓困
嚴重 特殊 傳染性 肺炎 防治 及

Biber 等人（1999）曾建議 N 連詞之篩選標準應以詞頻為主，所採用之標準為每百萬字中至少出現 10 次以上才收錄；也有學者（Biber and Barbieri 2007; Chen and Baker 2010）提出若語料量較少，研究者應依語料庫特性及研究需求自行訂定篩選標準。本研究之中文語料僅 4,300 萬餘詞，包含網頁中的表格，且本平臺所觀察的 N 連詞長度最多可達個數 9，故自訂頻率為至少出現 100 次以上，作為 N 連詞的篩選標準。

本研究為進一步驗證產出的關鍵詞表及 N 連詞表的效益，我們徵求了一些華語老師的意見，下面我們說明他們對於關鍵詞表及 N 連詞表的一些看法。關鍵詞表的部分，他們指出其優勢在於詞表中大部分的詞和新冠肺炎主題高度相關，涵蓋的面向也較廣，如表 6 所示：

表 6：涵蓋多個主題面向的關鍵詞

病毒	
病毒知識	新冠肺炎、病毒、冠狀病毒、病原體、病原、宿主、冠肺炎、基因組、病毒株、病毒性、突變、分離出、受體、核酸、抗原、RNA
傳染	飛沫、染病、感染、潛伏期、傳染力、傳染給、傳染、感染症、感染性、接觸
相關疾病	流感、SARS、MERS、伊波拉、腦炎、麻疹、腦膜炎、禽流感、HIV、肝炎、結核病
流行病學	感染源、傳染性、致死率、死亡率、發病率、感染率、病例數、病例
防疫措施	
篩檢	檢體、陰性、陽性、體溫、抗體、檢疫、疑似、採檢、特異性、篩查、試劑、病毒量、篩檢、送驗、拭子
限制	隔離、封城、復工、停課
疫調	旅遊史、返國、邊境、境外、本地、口岸、隱瞞、入境、移入、接觸史、高危險群
防疫物資	防護衣、隔離衣、N95、口罩、漂白水、佩戴、醫用、防護服
防疫衛生知識	防疫、抗疫、洗手、消毒劑、消毒
疫苗	接種、疫苗、免疫、免疫力
醫療	
症狀	血氧、休克、併發症、痰、敗血症、嘔吐、高燒、發熱、乾咳、康復、腹瀉、衰竭、鼻水、打噴嚏、分泌物、病徵、病程、痰液、浸潤、症狀、咳嗽、發燒、併發、死亡
設備	病室、ICU、急診、呼吸機、X 光、負壓、呼吸器、病房、加護、感染科
身體器官	鼻腔、心肺、口鼻、黏膜、喉嚨、支氣管、氣管、肺部、呼吸道、肺泡
治療	治療、就醫、藥物、就診、重症、急促、診斷、探病、病況、出院、住院、收治、救治、插管、入院
趨勢	大規模、擴散、蔓延、暴發、肆虐、傳播
行為	恐慌、微信、臉書、群聚、直播
相關機構	衛福部、疾管署、CDC、世衛、WHO、衛生部、公衛、醫管局、醫療院所、國務院
相關人名	陳時中、譚德塞、鍾南山、李文亮、川普、莊人祥、張上淳、習近平

上述十多種不同次主題的用語，為編輯華語中高級程度的主題式教材提供了多元豐富的素材。此外，透過 No Sketch Engine 的介面，在出習題或考題時教師也可以直接應用每個關鍵詞的例句、使用情境、搭配用法等。

不過，老師們也指出關鍵詞表也有部分的缺點，首先是無法包含比較長的概念，例如：急性呼吸道症候群、新型冠狀病毒、世界衛生組織、流行疫情指揮中心、負壓隔離病房等，這些概念都是由短語所組成。其次，華語老師亦建議，系統若能將關鍵詞依照次主題自動分群，對準備教材將有更大的幫助。

除此之外，為了多字詞的分析，我們也開發了 N 連詞查詢系統。例如：負壓隔離病房、急性呼吸道症候群、新型冠狀病毒、無症狀感染者等等重要主題，都可以透過 N 連詞系統自動抽取。其次，這種短語結構在教學上也常凸顯出其特殊性，例如：「感染」這個動詞，在 N 連詞中的功能多是當中心語或定語用，如：新型冠狀病毒「感染」、無症狀「感染」、「感染」者等等，熟悉這些連結關係能培養華語「定中結構」語感，不和「動賓結構」混淆而導致錯解文意。另外，有一些短語雖然和主題關聯性較低，但卻是引介或連接主題的重要成分，例如：「另一個」、「而不是」、「並不是」、「有一個」、「也就是」、「最常見的」、「最大的」等，這些短語也可適當的融入到教材中，以輔助主題教學。在 N 連詞表的缺點上，主要的問題是 N 連詞所涵蓋的主題較偏狹，不像關鍵詞表的涵蓋面那麼廣，內容偏於公共衛生政策、組織名、疾病名稱、相關地名等面向的短語。所以它應該和關鍵詞表一併使用、相輔相成。總之，以人工的方式去彙整這些詞組並不容易，而電腦程式能夠有系統地找出這些實用的表達方式，對老師及學生都有相當大的助益。

4.1.4 搭配詞表建置

根據 Lewis (2000:245) 所言：「搭配詞為經常以可預測的模式共同出現的字⁷。」最早提出搭配這一概念的是英國語言學家 Firth (1957)，他將 Collocation 稱為一種「結伴關係」，其他學者的定義也與其相似，如 Halliday、McIntosh 與 Strevens (1964) 都說明搭配是詞彙的共現關係。陸國強 (1983:144) 則在《現代英語詞彙學》一書中指出「詞的搭配關係主要指詞與詞之間的橫組合關係，即什麼詞經常與什麼詞搭配使用」，他還進一步提

⁷ 原文：Collocations might be described as the words that are placed or found together in a predictable pattern.

出，英語學習到一定階段時，提升語言能力的方法之一就是環繞中心內容進行聯想，釐清詞與詞之間的搭配關係，建立「聯想場」的概念。Sinclair 與 Renouf(1988)在談到詞彙教學大綱時亦指出，詞彙學習的焦點應該指向：(1)最常用的詞形；(2)這些詞形的主要使用模式；(3)這些詞形所構成的典型組合。換句話說，學習者不但要學會詞彙的發音、拼寫和詞典裡的意思，還要學習它最常用的詞形和搭配。歷來許多研究者皆肯定搭配詞的重要性，Woolard(2000)亦在文章中提出「要學習更多的詞彙並不只是要學習新詞彙，而是要學習舊詞彙的新的搭配用法。」Lewis(2000)也點出他觀察到許多學生常造出合乎文法但母語者覺得奇怪的句子，原因是學生並不熟悉詞彙的搭配用法。Nation(2001)同樣指出要使用流利適切的語言，搭配詞的知識是不可或缺的。由此可知，學習者在學習二語詞彙的過程中，若不能將詞彙和它的搭配詞合併起來學習，便常會受母語思維的負遷移影響，以母語方式來組合漢語、進行詞彙搭配，這樣將不利於培養學習者漢語詞彙聯想能力，難以習得正確的華語表達方式。

然而，Bahns 與 Eldaw(1993)、Farghal 與 Obiedat(1995)、Gitsaki(1996)等都曾透過翻譯測驗及克漏字測驗來檢視學習者的搭配詞使用能力，結果指出英語學習者無論國別皆有使用搭配詞的困難。詞彙搭配不僅是英語學習者的難點，亦是華語學習者難以克服的障礙，許多華語教學研究者都指出詞彙搭配不當是華語學習者學習華語時極為常見的問題(王建勤 1997；胡明揚 2006；馬玉汴 2006；彭增安 2007；全香蘭 2008；高燕 2008；董政、鄭艷群 2008；劉亞菲、鄭艷群 2008；蕭頻、張妍 2008)。

就主題式語言教學而言，搭配詞無疑更應是詞彙教學的重點，因為對華語學習者來說，諸如「疫情、病毒、抗體」這些關鍵詞都是陌生的，應該如何選用動詞來表達事件，或用什麼定語來修飾它們，甚至是與之搭配的特定量詞等等都是難點，且不容易在非主題語料庫或辭典中找到正確的搭配詞資訊。因此，建立主題語料庫在學習語言搭配上顯得特別重要，如果學習者希望能從語意的理解提升到語用的高度，由了解詞彙功能而提高語言綜合運用能力，就必須借助主題語料庫中的豐富搭配知識來達成。

在 No Sketch Engine 平臺上使用者只要輸入查詢詞(可限制詞性)，並在產生 KWIC 畫面後選取搭配統計，就可依相關度(log dice)產出搭配詞表。例如：本研究選用 5 個與 COVID-19 相關的名詞作為關鍵詞，透過與中央研

究院平衡語料庫搭配詞檢索結果做對比⁸，來凸顯主題語料庫於搭配詞擷取上的優勢。(參見表 7) 在層層因素的干擾之下，為讓結果更精確，我們將所有可列出來的搭配詞經過人工的挑選後製作出此表。

表 7：中研院平衡語料庫與 COVID-19 主題語料庫之搭配詞表比較

語料庫 中心詞	中研院平衡語料庫	COVID-19 主題語料庫
疫情	腸病毒~, 傳出~, ~中心	肺炎~, ~指揮, 流行~, ~中心, 中央~, 武漢~, ~防控~, ~爆發~, 中國~, 新冠~, 全球~, ~持續, ~擴散, ~嚴重, ~蔓延, 因應~, 新冠肺炎~, 病毒~, 控制~, ~發生, ~工作, ~措施, ~影響~, 台灣~, 國際~, 中共~, SARS~, 宣布~, ~發展, ~地區, 通報~, ~擴大, ~期間, ~調查~, 隱瞞~, 大陸~, ~延燒, 公布~, 國內~, 美國~, ~傳播~, ~相關~, 世界~, ~感染, ~嚴峻, ~專家, ~監測~, ~資訊, 次~, 級~, 波~
病毒	~入侵, ~血清, ~抗原, ~抗體, ~受體, ~侵入, ~突變, ~基因, ~宿主, ~帶原者, ~蛋白, ~傳染給, ~感染~, ~複製, HIV~, 分離出~, 天花~, 抑制~, 肝炎~, 乳突狀~, 流行性~, 帶有~, 散佈~, 登革~, 傳染~, 傳染病~, 愛滋病~, 感冒~, 腸~, 漢他~, 濾過性~	~研究~, ~基因, ~傳播~, ~檢測, SARS~, 抗~, 武漢~, 肺炎~, 冠狀~, 流感~, 發現~, 感染~, 新冠~, 新型~, ~引起, 呼吸道~, 肝炎, 伊波拉, 確診, 核酸, 疫苗, 抗體, 細菌, 傳染, 治療

⁸ 「中央研究院現代漢語平衡語料庫」(4.0 版) 簡稱「中研院平衡語料庫」, 是第一個經過完整詞類標記的漢語平衡語料庫。為求語料的平衡性, 語料的蒐集以六大主題為主軸, 其收錄內容比例為: 哲學占 10%、科學占 10%、社會占 35%、藝術占 5%、生活占 20%、文學占 20%, 可說是現代漢語無窮多語句中一個代表性的樣本。對於華語教學研究、漢語語言研究者而言, 中研院平衡語料庫可說是目前最具代表性的綜合語料庫, 當代學者無論在詞表研擬、詞彙分析或教材編纂等方面, 皆以此為首要的參考依據。有鑑於中研院平衡語料庫之代表性和重要性, 本研究以此作為對比分析的語料依據, 以此驗證 COVID-19 語料庫具有有別於一般語料庫的特性。

表 7：中研院平衡語料庫與 COVID-19 主題語料庫之搭配詞表比較（續）

語料庫 中心詞	中研院平衡語料庫	COVID-19 主題語料庫
肺炎	支氣管炎~, 非典型性~, 感染~	~疫情, ~個案, ~原因, ~病例, ~病毒, ~症狀, ~患者, COVID-19~, 不明~, 出現~, 武漢~, 非典型~, 冠狀~, 冠狀病毒~, 特殊~, 病毒性~, 傳染性~, 感染~, 新~, 新冠~, 新型~, 確診~, 爆發~
口罩	手套~, 戴~	N95~, 一般~, 手套~, 外科~, 生產~, 防疫~, 防護~, 佩戴~, 使用~, 兒童~, 配戴~, 買~, 需要~, 戴~, 戴上~, 購買~, 醫用~, 醫療~
抗體	單胞~, 抗原~, C 型~, 血清~, 母體~, 帶原者~, 肝炎~, 腫瘤~, 表面~, 核心~, 病毒~, 細胞~, 產生~, ~作用, ~存在~	~檢測~, 血清~, 產生~, ~特異性, 陽性~, 株~, 免疫~, ~反應, 肝炎~, ~中和, 表面~, 檢驗~, 結合~, 病毒~, ~細胞, 核心~, 新冠~, ~存在, 單株~, 代表~, 製造~, 人體~, 血漿~, ~作用, 具有~, ~測試~, 保護性~, ~濃度, 篩檢~, 出現~, ~檢查~, ~試驗~, 球蛋白~, ~針對~, 確認~, 藥物~, 免疫力~, 自身~, 研發~, 驗出~, 發現~

總體而言，兩個語料庫檢索出的名詞搭配詞在筆數上有明顯的差異，中研院平衡語料庫的搭配詞較少。5 個關鍵詞中，有 4 個在本研究所建置的 COVID-19 主題語料庫呈現出較為豐富多樣的搭配詞檢索結果，可提供教師及教材編輯者豐富大量的 COVID-19 搭配範例。如：本網路 COVID-19 主題語料庫檢索出的「肺炎」搭配詞有 20 筆以上；而中研院平衡語料庫搭配詞檢索結果僅有 3 筆。再從內容上來看，以「疫情」一詞為例，在中研院平衡語料庫中，搭配詞僅有「腸病毒、傳出、中心」3 詞。然而在 COVID-19 語境中，「疫情」經常搭配使用的詞彙很多，動詞有：

~防控~, ~爆發~, ~持續, ~擴散, ~蔓延, 因應~, ~控制~, ~發生, 通報~, ~擴大, 隱瞞~, ~傳播~, ~調查~, ~發展, ~監測~, ~嚴峻, ~影響~, ~延燒, 公布~, 宣布~, ~嚴重

常修飾「疫情」的定語則包括「疾病+疫情」和「地區+疫情」，並以後者居多。如：

肺炎~, SARS~, 新冠肺炎~, 新冠~, 世界~, 全球~, 武漢~, 台灣~, 國際~, 中共~, 美國~, 國內~, 大陸~, 中國~, 中央~, ~相關~, 流行~

常被「疫情」修飾的中心語包括時間、地點、人以及名物化的動詞。如：

~中心, ~資訊, ~專家, ~期間, ~控制, ~監測

常搭配的量詞包括：

次~, 級~, 波~

這些搭配皆出現在 COVID-19 檢索結果中，教師在講授「疫情」一詞時可以參照補充，使「疫情」的詞彙教學更豐富多元。

從語料內容的層面來看，中研院現代漢語平衡語料庫有其優點，其語料來自哲學、文學、生活、社會、科學、藝術六大領域，面面俱到，豐富多樣。然也因為取材廣泛，平衡語料庫裡某一特定領域的語言使用在種類及數量上則沒有特殊主題語料庫多。反觀本研究採用的語料內容來源為網路 COVID-19 文本，因此所得搭配結果較具有針對性。對於編寫各種 COVID-19 華語教材而言，COVID-19 主題語料庫能更準確地反映出網路上 COVID-19 實際使用相關詞彙和用語的情形，更加切合真實之網路 COVID-19 情境。COVID-19 語料的搭配詞搜尋結果較一般語料庫更為相關，華語教師及教材編纂者能參考搭配詞的共現頻率資料，從相關的檢索結果中選擇共現頻率高的搭配詞作為優先教學、編寫教材的參考。

4.2 研究問題二：這個網路平臺就語料內容及分析上的優點和缺點為何？有何可改進之處？

首先，「網路作為語料庫」理念所蒐集的主題語料庫在華教應用上有以下幾點優勢：（1）即時性：教科書的編纂參與者有限，且曠日耗時，新興主題無法即時納入教材中讓學習者現學現用（陳燕秋 2006；黃琬華 2014）。主題語料庫則相反，能夠在短時間內自動持續收集即時性的新聞事件（吳鑑城等 2017），新聞事件的知識性與時效性為語言教學的重要取材來源（謝佳玲、李家豪 2011）。（2）涵蓋率：主題語料庫從網路上取得大量的語言使用樣本，

特性在於它直接呈現巨量資料，不像教科書是以挑選、編寫符合語法說明的例句為主。高頻的詞彙或搭配能提示優先學習的順序，改變教學大綱，與教科書較固定的教學順序不同（吳鑑城等 2017；陳浩然、潘依婷 2017）。(3) 真實語言：主題語料庫為真實語言之取樣，透過計算與分析得以描繪出真實語言的使用情形，在詞彙的選用上更貼近當代用法，搭配選擇也更多。網路作為語料庫的語料更為豐富並與日常使用的詞彙更加接近（陳浩然、潘依婷 2017）。(4) 豐富情境：教科書的優勢在於能清楚陳述華語語法，並以編寫過的例句呈現。然而在實際應用上，學習者會遇到多元的使用情境，未必符合教科書上的說明。主題語料庫收集範圍很廣，包含政治、經濟、科技、生活等議題文章，透過即時而多元的時事題材，足以反映時代現況，從中延伸多元的學習內容，可以呈現豐富語言訊息（謝舒凱 2017；謝佳玲、吳欣儒 2018），能更進一步與他人交流對此事件、議題的看法（黃琬華 2014），藉以彌補教科書不足之處，兩者相輔相成。

其次，No Sketch Engine 網路平臺可匯入不同的語料庫，且建置大型語料庫的速度也比電腦語料庫軟體如 *AntConc* (Anthony 2004) 和 *WordSmith* (Scott 2008) 快並較少有當機的情況，此優勢則可使研究人員、教師、學生及教材編纂者蒐集更豐富的資源和有較佳的使用經驗。此外，結合數量龐大的網路語料庫，No Sketch Engine 平臺的分析能力可以挖掘語料庫豐富的語言現象，包括抽取常用的關鍵詞彙、多字詞及搭配詞等等。再配合主題式或專業的華語語料可以針對特定主題，進行有深度且有效的分析和處理。

除了供教師及教材編纂者取材外，No Sketch Engine 平臺還可以直接應用在課堂上進行 DDL 式的語言教學。根據 Boulton 與 Cobb (2017) 最近的後設分析研究，DDL 對外語學習者的語言學習成效有正面的影響。根據他們的研究，其效果值 (Cohen's d effect size)，無論是在語言學習的效果 ($d = 1.5$) 或效率 ($d = 0.95$) 方面皆有高度效果。他們的分析亦顯示，隨著使用 DDL 方式學習的時間增加，平均效果則跟著提高。因此，他們對於學習者自主使用 DDL 的方式學習語言抱持樂觀的態度。藉由主題語料庫豐富的資料找出相關字詞的真實使用情境，這對利用 DDL 學習道地的語言將有莫大的幫助。

至於此網路語料庫平臺的弱點則為 (1) 從網路上爬取資料可能會納入一些不相關的雜訊，例如，本研究就爬取到一些日文漢字語料，而過多的雜訊可能造成語言分析上的困擾。(2) 在平臺工具方面，本研究已開發 N 連詞的檢索工具，但尚有一些其他的語料分析工具是我們急切需要的，例如，前文

已提到之關鍵詞自動分群，可提供情境式教材編撰。又如在檢索功能上仍有改進之處，例如依檢索結果擇定詞性後再檢索，將可以進一步排除不相關的資訊，提升搭配詞檢索的效能。這方面還需仰賴資訊工程相關的團隊成員幫忙才能在技術上有所突破。

5. 結語

整體而言，目前可用的主題式華語語料庫仍是寥寥可數。有鑑於此，本研究以「網路作為語料庫」的方法，應用 COVID-19 常用詞彙作為種子詞，利用 WebBootCat 語料蒐集軟體有效率地蒐集大量網路華語語料，完成網路 COVID-19 主題語料庫的建置工作。並進一步介紹利用 No Sketch Engine 語料庫搜尋引擎檢索該語料庫，方便華語教師或教材編纂人員將檢索出的相關 COVID-19 的語言素材運用於華語教學及教材編寫上。相較於一般綜合性語料庫之檢索功能，此 No Sketch Engine 語料庫搜尋引擎提供不同檢索功能，且速度快，結果也較豐富多樣，更能協助教師及教材編輯者找出更多不同詞性的搭配詞，提供大量搭配詞的範例，以彌補搭配詞學習上常見的語感侷限，對華語教師及教材編輯者而言是極佳的輔助工具。

此外，COVID-19 主題語料庫檢索系統也可以讓華語學生直接進行學習檢索。讓學生運用語料庫觀察真實語料，從大量語料中觀察到某一語言現象，然後討論和分享在語料中的發現，形成與語料庫的互動。接著再歸納該語言現象的規則，並在教師的指導下，通過觀察更多語料，逐步修正歸納出的規則，即 DDL 的學習模式。除此之外，語言學家可以進一步再對 COVID-19 主題語料庫進行相關研究。例如，社會語言學家可能對整個語料庫的社會論點感興趣，即可以針對此文本中的不同社會主題進行分析。

本研究仍有許多侷限及受限之處需加以改進，如：在標註語料庫文本的詞性時，受詞性標註軟體功能上的限制，會有詞性標註的錯誤，故所得搭配結果會出現搭配錯誤的情形，需再以人工方式加以篩選。另一方面，雖然取材為不同之網路文本，但目前語料仍不夠大。期許在未來的研究中，可增加更多語料（如突破一億詞），進而建置更全面且更多元之 COVID-19 主題語料庫，以提升研究成果之廣度。

而本研究期許此 COVID-19 主題語料庫的建置及 No Sketch Engine 開源語料庫檢索系統能提供國內外主題式、專業領域科目的華語語料庫建置及研究工作作為參考。期望可促成更多元、更完備之網路主題語料庫之建置。另

一方面，No Sketch Engine 提供使用者便捷的檢索功能和介面，可應用於辭典編纂、教學現場、教材設計和搭配詞研究等方面，鼓勵華語教師及學生多多應用。期望相關研究和應用得以藉 No Sketch Engine 搜尋引擎作為基礎，拓展華語教學的深度與廣度。

引用文獻

- Anthony, Laurence. 2004. AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of An Interactive Workshop on Language e-Learning*, 7-13. Tokyo, Japan.
- Bahns, Jens, and Moira Eldaw. 1993. Should we teach EFL students collocations? *System* 21.1: 101-114.
- Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.
- Baroni, Marco, and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the Web. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 1313-1316. Lisbon, Portugal.
- Baroni, Marco, Adam Kilgarriff, Jan Pomikálek, and Pavel Rychlý. 2006. WebBootCaT: Instant domain-specific corpora to support human translators. *Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation*, 247-252. Oslo, Norway.
- Biber, Douglas. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: Benjamins.
- Biber, Douglas, and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26.3: 263-286.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.
- Boulton, Alex, and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67.2: 348-393.
- Chen, Yu-hua, and Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14.2: 30-49.
- COVID-19 Open Research Dataset (CORD-19). 2020. Version 2020.05.02. Accessed online, May 2, 2020. <https://pages.semanticscholar.org/coronav>

- irus-research. doi:10.5281/zenodo.3715505
- Davis, Mark. 2020. The Coronavirus Corpus. Accessed online, May 2, 2020. <https://www.english-corpora.org/corona/>
- Farghal, Mohammed, and Hussein Obiedat. 1995. Collocations: A neglected variable in EFL. *IRAL-International Review of Applied Linguistics in Language Teaching* 33.4: 315-332.
- Firth, John. 1957. *Papers in Linguistics, 1934-1951*. Oxford: Oxford University Press.
- Fujii, Atsushi, and Tetsuya Ishikawa. 2000. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 488-495. Hong Kong: Association for Computational Linguistics.
- Gitsaki, Christina. 1996. *The Development of ESL Collocational Knowledge*. Brisbane: The University of Queensland Ph. D. dissertation.
- Halliday, Michael A. K., Angus McIntosh, and Peter Stevens. 1964. *The Linguistic Sciences and Language Teaching*. London: Longman.
- Howarth, Peter. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19.1: 24-44.
- Jones, Rosie, and Rayid Ghani. 2000. Automatically building a corpus for a minority language from the web. *Annual Meeting-Association for Computational Linguistics*, 29-36.
- Kilgariff, Adam, and David Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography. *Proceedings of the Workshop on Collocation: Computational Extraction, Analysis and Exploitation*. Toulouse, France.
- Kilgariff, Adam, and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29.3: 333-347.
- Leech, Geoffrey. 1997. Teaching and language corpora: A convergence. *Teaching and Language Corpora*, eds. by Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 1-23. London: Routledge.
- Lewis, Michael. 1997. *Implementing the Lexical Approach: Putting Theory into*

- Practice*. United Kingdom: Heinle.
- Lewis, Michael. 2000. *Teaching Collocation: Further Development in Lexical Approach*. England: The Language Teaching Publication, LTP.
- Lin, Yu-hsiu. 2011. *A Corpus-based Analysis of the Use of Lexical Bundles in English Academic Writing*. Taipei: National Taiwan Normal University MA thesis.
- Lui, Marco, and Paul Cook. 2013. Classifying English documents by national dialect. *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, 5-15.
- Nation, I. S. Paul. 2001. *Learning Vocabulary in Another Language*. Ernst Klett Sprachen. Cambridge: Cambridge Press.
- Nattinger, James R., and Jeanette S. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.
- Rayson, Paul, and Roger Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora 9*: 1-6. Hong Kong, China.
- Resnik, Philip. 1999. Mining the web for bilingual text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 527-534. Maryland, U.S.A.
- Rychlý, Pavel. 2007. Manatee/Bonito- a modular corpus manager. *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65-70. Brno, Czech Republic.
- Salazar, Lorenzo, and Danica Joy. 2011. *Lexical Bundles in Scientific English: A Corpus-based Study of Native and Non-native Writing*. Barcelona: Universitat de Barcelona Ph. D. dissertation.
- Scott, Mike. 1997. PC analysis of key words- and key key words. *System* 25.2: 233-245.
- Scott, Mike. 2008. WordSmith Tools version 5. Liverpool: Lexical Analysis Software. <http://www.lexically.net/wordsmith/version5/index.html>
- Scott, Mike, and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Simpson-Vlach, Rita, and Nick C. Ellis. 2010. An academic formulas list: New

- methods in phraseology research. *Applied Linguistics* 31.4: 487-512.
- Sinclair, John, and Antoinette Renouf. 1988. A lexical syllabus for language learning. *Vocabulary and Language Teaching*, 140-160.
- Wicke, Philipp, and Marianna M. Bolognesi. 2020. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *arXiv Preprint arXiv: 2004.06986*.
- William, Raymond. 1976. *Keywords: A Vocabulary of Culture and Society*. New York: Oxford University Press.
- Willis, Dave. 2003. *Rules, Patterns and Words: Grammar and Lexis in English Language Teaching*. Cambridge: CUP.
- Woolard, George. 2000. Collocation- encouraging learner independence. *Teaching Collocation: Further Development in the Lexical Approach*, ed. by Michael Lewis. Oxford: Oxford University Press.
- Wray, Alison, and Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20.1: 1-28.
- 王建勤。1997。《漢語作為第二語言的習得研究》。北京：北京語言大學出版社。[Wang, Jian-qin. 1997. *Studies in Second Language Acquisition of Chinese*. Beijing: Beijing Language and Culture University Press.]
- 全香蘭。2008。〈韓語漢字詞對學生習得漢語詞語的影響〉，《基於中介語語料庫的漢語詞彙專題研究》，132-139。[Quan, Xiang-lan. 2008. The impact of Korean Hanja on the acquisition of Chinese words. *Interlanguage Corpus-based Chinese Vocabulary Thematic Studies*, 132-139.]
- 吳鑑城、陳浩然、張俊盛。2017。〈網路語料庫介紹與應用〉，《語料庫與華語教學》，陳浩然（主編），49-70。臺北：高等教育文化事業有限公司。[Wu, Jian-cheng, Howard Hao-jan Chen, and Jason S. Chang. 2017. *Wang lu yu liao ku jie shao yu ying yong. Corpus and Teaching Chinese as a Second Language*, ed. by Howard Hao-jan Chen, 49-70. Taipei: Higher Education Publishing.]
- 胡明揚。2006。〈詞彙教學理論〉，《對外漢語詞彙及詞彙教學研究》，203-225。[Hu, Ming-yang. 2006. Vocabulary teaching theory. *Research in Chinese Vocabulary and Vocabulary Teaching*, 203-225.]
- 馬玉汭。2006。〈詞彙教學方法〉，《對外漢語詞彙及詞彙教學研究》，263-291。

- [Ma, Yu-bian. 2006. Vocabulary teaching method. *Research in Chinese Vocabulary and Vocabulary Teaching*, 263-291.]
- 高燕。2008。《對外漢語詞彙教學》。上海：華東師範大學出版社。[Gao, Yan. 2008. *Teaching Chinese as a Foreign Language*. Shanghai: East China Normal University Press.]
- 陳燕秋。2006。〈新聞選讀團體語言教學法實例〉，《台灣華語文教學》，第 1 期，36-39。[Chen, Yan-qiu. 2006. Community language learning applied to newspaper readings. *Teach Chinese as a Second Language* 1: 36-39.]
- 陳浩然、潘依婷。2017。〈語料庫與華語教學〉，《語料庫與華語教學》，陳浩然（主編），6-39。臺北：高等教育文化事業有限公司。[Chen, Howard Hao-jan, and I-ting Pan. 2017. Corpus and teaching Chinese as a second language. *Corpus and Teaching Chinese as a Second Language*, ed. by Howard Hao-jan Chen, 6-39. Taipei: Higher Education Publishing.]
- 黃珣華。2014。〈從新聞語體特色談華語新聞教學〉，《華語學刊》，第 16 期，80-86。[Huang, Chu-hua. 2014. On teaching news Chinese in the perspective of news language characteristics. *A Journal of the Association of Teaching Chinese as a Second Language* 16: 80-86.]
- 陸國強。1983。《現代英語詞彙學》。上海：上海外語教育出版社。[Lu, Guo-qiang. 1983. *Modern English Lexicology*. Shanghai: Shanghai Foreign Language Education Press.]
- 彭增安。2007。《跨文化的語言傳通—漢語二語習得與教學》。上海：學林出版社。[Peng, Zeng-an. 2007. *Cross-cultural Language Communication-Chinese Second Language Acquisition and Teaching*. Shanghai: Academia Press.]
- 董政、鄭艷群。2008。〈歐美學生漢語量詞的使用情況〉，《基於中介語語料庫的漢語詞彙專題研究》，89-98。[Dong, Zhen, and Yan-qun Zheng. 2008. The Chinese quantifiers usage of students in Europe and America. *Interlanguage Corpus-based Chinese Vocabulary Thematic Studies*, 89-98.]
- 劉亞菲、鄭艷群。2008。〈韓國學生漢語量詞的使用情況〉，《基於中介語語料庫的漢語詞彙專題研究》，79-88。[Liu, Ya-fei, and Yan-qun Zheng. 2008. The Chinese quantifiers usage of Korean students. *Interlanguage Corpus-based Chinese Vocabulary Thematic Studies*, 79-88.]

- 蕭頻、張妍。2008。〈印尼學生漢語單音節動詞語義偏誤〉，《基於中介語語料庫的漢語詞彙專題研究》，62-72。[Xiao, ping, and Yan Zhang. 2008. The monosyllabic verb semantic errors of Indonesian students. *Interlanguage Corpus-based Chinese Vocabulary Thematic Studies*, 62-72.]
- 謝佳玲、李家豪。2011。〈臺灣電視新聞標題研究與教學啟示〉，《華語文教學研究》，第 8 卷第 3 期，79-114。[Hsieh, Chia-ling, and Jia-hao Li. 2011. A study on Taiwanese television news headlines and their pedagogical implications. *Journal of Chinese Language Teaching* 8.3: 79-114.]
- 謝佳玲、吳欣儒。2018。〈以華語電視新聞為材料的語篇研究及聽力教學應用〉，《臺灣華語教學研究》，第 16 期，91-124。[Hsieh, Chia-ling, and Xin-ru Wu. 2018. A discourse study of Chinese television news and its application to listening pedagogy. *Taiwan Journal of Chinese as a Second Language* 16: 91-124.]
- 謝舒凱。2017。〈中文語料與詞彙知識地圖〉，《語料庫與華語教學》，陳浩然（主編），66-77。臺北：高等教育文化事業有限公司。[Hsieh, Shu-kai. 2017. *Zhong wen yu liao yu ci hui zhi shi di tu. Corpus and Teaching Chinese as a Second Language*, ed. by Howard Hao-jan Chen, 66-77. Taipei: Higher Education Publishing.]

[審查：2020.6.15 修改：2020.8.6 接受：2020.9.11]

白明弘

Ming-Hong BAI

10644 臺北市大安區和平東路一段 179 號

國家教育研究院語文教育及編譯研究中心

Research Center for Translation, Compilation and Language Education

National Academy for Educational Research

No.179, Sec. 1, Heping E. Rd., Taipei City 10644, Taiwan

mhbai@mail.naer.edu.tw

華語文教學研究

陳浩然

Howard Hao-Jan CHEN

10610 臺北市大安區和平東路一段 162 號 國立臺灣師範大學英語學系

Department of English

National Taiwan Normal University

No.162, Sec. 1, Heping E. Rd., Taipei City 10610, Taiwan

hjchen@ntnu.edu.tw

林鶯

Ying LIN

10610 臺北市大安區和平東路一段 162 號 國立臺灣師範大學英語學系

Department of English

National Taiwan Normal University

No.162, Sec. 1, Heping E. Rd., Taipei City 10610, Taiwan

stella242630@gmail.com

附錄 1：高頻字表含華測八千詞及國教院詞表之分級（前 300）

（註：為了節省版面空間，本文將華測八千詞的等級以數字表示如下：1→準備一級，2→準備二級，3→入門級，4→基礎級，5→進階級，6→高階級，7→流利級。）

詞	華測 等級	國教 等級	詞	華測 等級	國教 等級	詞	華測 等級	國教 等級
的	1	1	疫情			人員	6	5
是	1	1	上	2	1	他	1	1
在	2	1	並	6	5	例	7	3
有	1	1	中國	1	1	已	5	3
一	1	1	對	4	1	症狀	6	5
及	5	5	於	7	5	最	4	1
之	6	5	可	5	3	患者	7	5
了	1		都	1	1	能	1	1
和	2	1	以	5	5	病人	4	2
或	4	3	時	5	3	武漢		
不	1	1	但	5	2	說	1	1
與	5	4	而	5	5	各	4	2
個	1	1	將	6	5	天	2	1
也	1	1	肺炎	7		確診		
病毒	6	5	內	5	3	次	2	2
人	1	1	醫院	2	2	其	6	6
為	5	4	治療	7	5	可以	1	1
這	1	1	大	1	1	至	6	5
感染	6	5	應	5	1	每	2	1
中	5	1	被	4	2	病例		6
會	2	1	要	2	1	出現	4	4
等	1	1	使用	5	4	沒有	4	1
就	4	1	我	1	1	美國	1	1
後	2	1	可能	2	1	衛生	6	5
者	6		種	4	2	醫療	7	5

詞	華測 等級	國教 等級
疾病	6	5
到	1	1
進行	5	5
多	2	1
口罩		4
年	1	1
還	4	1
我們	1	1
如	5	
來	1	1
本	2	1
前	2	1
台灣	1	1
其他	4	2
由	6	4
從	2	1
表示	5	4
研究	5	3
工作	2	1
相關	5	4
中心	4	3
所	5	2
兩	1	1
很	1	1
名	5	2
下	2	1
目前	5	4
無	5	5
國家	3	1

詞	華測 等級	國教 等級
該	5	4
新	2	1
健康	3	2
隔離		6
更	4	2
名詞	5	4
此	5	5
個案		6
防疫		
全	5	2
則	6	5
讓	4	2
發生	4	3
時間	4	1
措施	6	6
公司	2	2
組織	6	5
因		
高	1	1
發現	4	2
政府	6	4
再	1	1
三	1	1
死亡	6	4
你	1	1
地	4	2
嚴重	5	3
接觸	5	4
未	6	5

詞	華測 等級	國教 等級
包括	6	3
因為	2	1
歲	1	1
才	4	2
外	2	1
細胞	6	6
醫師	6	1
用	2	1
系統	5	5
流感		
提供	5	4
著	2	6
自己	4	1
需要	4	2
呼吸	6	4
較	5	2
世界	4	2
問題	2	1
造成	5	4
且	6	3
如果	4	2
已經	4	2
請	1	1
藥物	6	
若	7	
只	1	1
得	1	1
全球	5	4
新型		

發展主題語料庫以輔助華語教學－以 2019 新型冠狀病毒語料庫為例

詞	華測 等級	國教 等級
家	1	1
主要	5	4
又	4	1
過	1	1
冠狀		
通報	7	6
疫苗		6
項	5	5
向	4	2
好	1	1
開始	2	2
第一		2
檢測		6
病患		5
影響	4	3
檢查	4	4
國	4	1
因此	5	4
元	2	1
管理	4	4
以及	6	5
臨床	7	6
增加	5	4
月	2	1
做	1	1
情況	5	4
建議	5	3
起	4	3
這些	4	

詞	華測 等級	國教 等級
發燒	4	3
需	6	2
他們	1	1
民眾	6	4
國際	6	4
仍	5	4
是否	6	5
傳染病		
服務	4	3
其中	5	3
去	1	1
地區	5	4
經	5	
以上	5	3
單位	6	5
方式	5	4
結果	5	3
香港		
呼吸道		
機構	6	5
約	4	3
持續	7	4
傳播	6	5
那	1	1
環境	4	3
日	2	1
所有	4	2
依	7	
風險	6	5

詞	華測 等級	國教 等級
二	1	1
認為	4	2
免疫	7	6
流行	4	3
規定	6	4
報告	4	3
避免	5	4
處理	5	4
曾	5	4
非	5	5
接受	5	4
安全	4	3
條	4	2
無法	5	5
資料	5	3
預防	7	5
期間	5	4
同時	5	3
一般	4	3
把	4	2
功能	5	4
病房	6	4
重要	3	2
照護		6
位	4	1
技術	6	4
大學	1	1
對於	6	5
控制	6	5

華語文教學研究

詞	華測 等級	國教 等級
導致	7	6
引起	5	4
看	1	1
小	1	1
發展	6	4
產生	5	4
根據	5	4
不同	4	3
當	4	2
指出	6	
檢驗	6	5
同	5	
部分	5	2
中央	6	5
自	7	5
現在	1	1
市場	2	2

詞	華測 等級	國教 等級
卻	5	4
由於	5	4
必須	4	3
所以	2	1
有關	6	4
中共		
沒	1	1
方法	4	2
活動	4	2
檢疫		
減少	5	4
新增		
原因	5	3
新冠		
使	6	5
她	1	1
專家	6	4

詞	華測 等級	國教 等級
急性		
報導	6	5
受	6	4
居家		6
診斷	7	5
大陸	5	3
均	7	6
亦	7	6
SARS		
小時	3	1
日本	1	1
反應	5	4
超過	5	4
比	2	1
醫生	2	1
醫護		6
防護		6

附錄 2：關鍵詞表（前 300）

（註：為了節省版面空間，本文將華測八千詞的等級以數字表示如下：1→準備一級，2→準備二級，3→入門級，4→基礎級，5→進階級，6→高階級，7→流利級。）

詞	華測 等級	國教 等級	詞	華測 等級	國教 等級	詞	華測 等級	國教 等級
疫情			疾管署			湖北		
確診			隔離		6	核酸		
肺炎	7		陰性		7	疾控		
武漢			每日			急性		
冠狀			武漢市			肺部		
新冠			陽性		6	病例數		
病例		6	飛沫			第 3		
為了	4		抗疫			洗手		
防疫			收治			病房	6	4
病毒	6	5	第 2			患者	7	5
冠狀病毒			管制署			新型		
新冠肺炎			發病		6	衛健委		
感染	6	5	冠肺炎			疑似		7
口罩		4	傳染	6	5	病原		
檢疫			WHO			病人	4	2
呼吸道			咳嗽	4	3	病毒性		
COVID-19			疫苗		6	體溫	6	5
世衛			病原體			呼吸器		
防控			第 1			群聚		7
採檢			福利部			COVID-		
傳染性			抗體			陳時中		
傳染病			潛伏期			衛福部		
而言	7		接觸史			疫區		
旅遊史			檢體			SARS		
流感			湖北省			症狀	6	5

詞	華測 等級	國教 等級
流行病學		
入境	6	4
截至		6
加護		
MERS		
宿主		
移入		
感染源		
報道		
入院		
負壓		
X 光		
N95		
醫護		6
接種		
漂白水		
染病		
CDC		
傳染力		
傳染給		
救治		
感染症		
爆發	7	5
插管		
譚德塞		
第 4		
感染性		
共有		
死亡率		

詞	華測 等級	國教 等級
出院	6	4
致死率		
通報	7	6
衛生部		
川普		
有些	6	
臉書		
抗原		
發燒	4	3
病情	7	4
流行病		
境外		
症		
專家組		
體液		
用於		
住院	6	4
個案		6
傳人		7
大規模		
消毒	7	6
疫調		
擴散	7	6
第 5		
血清		
病毒株		
整個		
當局	7	5
措施	6	6

詞	華測 等級	國教 等級
伊波拉		
衰竭		
痰液		
死亡	6	4
抗藥性		
染疫		
監測		
RNA		
流行性		
例	7	3
呼吸	6	4
肺泡		
病徵		
物資	7	6
病程		
打噴嚏		4
鍾南山		
分泌物		
蔓延	7	6
疾病	6	5
公衛		
特異性		
腹瀉		7
病患		5
鼻水		3
免疫	7	6
急促		6
隔離衣		
李文亮		

發展主題語料庫以輔助華語教學－以 2019 新型冠狀病毒語料庫為例

詞	華測 等級	國教 等級
圖博館		
防護衣		
衛生	6	5
麻疹		
基因組		
中共		
佩戴		
臨床	7	6
重症		
指揮	7	5
停課		
抗擊		
醫用		
就是	5	
撥打		
病室		
篩查		
併發	7	
傳染源		
接觸	5	4
第 7		
抗生素		7
早前		
醫院	2	2
呼吸機		
康復	7	6
腦炎		
第 6		
感染科		

詞	華測 等級	國教 等級
消毒劑		
受體		
乾咳		
患病		5
公主號		
拭子		
窘迫		7
微信		
恐慌		6
醫務		
併發症		7
莊人祥		
衛生		
隱瞞		6
鏈球菌		
張上淳		
國務院		
中方		
治療	7	5
感染率		
處置		6
HIV		
就醫		6
爲		
意大利		
不		
急診	7	4
ICU		
痰	7	

詞	華測 等級	國教 等級
病況		
習近平		
高燒		
送驗		
習近		
遏制		
鼻腔		
PCR		
科研		
敗血症		
肆虐		7
跟著		
指揮官		
應對		6
不應該		
氣管	7	
嚴峻		7
醫療院所		
直播		
病史		
結核病		
病毒量		
血氧		
新華社		
探病	7	
手液		
休克		
高危險群		
禽流感		

詞	華測 等級	國教 等級
口鼻		
心肺		
疫病		
返國		6
TOCC		
淋巴細胞		
黏膜		
分離出		
防護服		
復工		
陪病		
支氣管		
第一線		
藥物	6	
浸潤		
嘔吐		6
血漿		

詞	華測 等級	國教 等級
傳播	6	5
細菌	6	5
慢性		
免疫力		
嚴重	5	3
喉嚨	6	4
試劑		
發布會		
診斷為		
邊境		6
SARS-CoV-2		
抑制劑		
細菌性		
突變		7
本地		
肝炎		
發熱		

詞	華測 等級	國教 等級
暴發		
金銀潭		
口岸		
就診		
醫管局		
症狀		
同住		
發病率		
腦膜炎		
篩檢		
封城		
着		
不到	6	
炎症		
診斷	7	5
癥狀		
暴露	7	5

附錄 3：N 連詞表（3 字詞 N 連詞到 6 字詞 N 連詞）

3 字詞 N 連詞（前 100 個）

1	新型 冠狀 病毒	35	疾病 預防 控制	69	境 外 輸入
2	世界 衛生 組織	36	最 好 的	70	注 意 的 是
3	疫情 指揮 中心	37	使用 我們 的	71	個人 防護 裝備
4	是 一 種	38	個 月 內	72	最 嚴 重 的
5	流行 疫情 指揮	39	更 好 的	73	國 家 和 地 區
6	中央 流行 疫情	40	在 內 的	74	衛 生 緊 急 事 件
7	武漢 肺炎 疫情	41	2019 新 型 冠 狀	75	最 常 見 的
8	最 大 的	42	全 球 大 流 行	76	就 是 一
9	華 南 海 鮮 市 場	43	的 一 種	77	冠 狀 病 毒 疾 病
10	特 殊 傳 染 性 肺 炎	44	武 漢 肺 炎 的	78	受 試 者
11	嚴 重 特 殊 傳 染 性	45	這 是 一	79	嚴 重 急 性 呼 吸 道
12	無 症 狀 感 染	46	衛 生 福 利 部 疾 病	80	隔 離 14 天
13	冠 狀 病 毒 肺 炎	47	每 個 人	81	重 要 的 是
14	的 情 況 下	48	福 利 部 疾 病 管 制 署	82	的 就 是
15	境 外 移 入	49	這 就 是	83	世 界 各 國
16	症 狀 感 染 者	50	新 增 確 診 病 例	84	的 新 型 冠 狀
17	的 一 個	51	很 大 的	85	公 共 衛 生 事 件
18	冠 狀 病 毒 感 染	52	傳 染 病 防 治 法	86	急 性 呼 吸 道 症 候 群
19	有 一 個	53	這 也 是	87	衛 生 健 康 委 員 會
20	密 切 接 觸 者	54	中 國 疾 控 中 心	88	的 各 種
21	是 不 是	55	疫 情 防 控 工 作	89	感 染 管 制 措 施
22	也 就 是	56	一 段 時 間	90	突 發 公 共 衛 生
23	並 不 是	57	負 壓 隔 離 病 房	91	世 界 各 地
24	一 個 人	58	中 央 主 管 機 關	92	小 細 胞 肺 癌
25	病 毒 感 染 的	59	不 明 原 因 肺 炎	93	一 次 性
26	新 冠 肺 炎 疫 情	60	在 這 個	94	這 一 點
27	一 個 月	61	全 民 健 康 保 險	95	於 今 日 確 診
28	冠 狀 病 毒 的	62	預 防 控 制 中 心	96	更 大 的

29	院 內 感 染	63	公 共 衛 生 緊 急	97	值 得 注 意 的
30	自 主 健 康 管 理	64	的 醫 護 人 員	98	24 小 時 內
31	最 重 要 的	65	重 症 患 者	99	這 個 問 題
32	而 不 是	66	2019 冠 狀 病 毒	100	的 最 新
33	另 一 個	67	中 華 人 民 共 和 國		
34	感 染 的 肺 炎	68	還 有 一		

4 字詞 N 連詞（前 100 個）

1	流 行 疫 情 指 揮 中 心	51	武 漢 新 型 冠 狀 病 毒
2	中 央 流 行 疫 情 指 揮	52	國 際 關 注 公 共 衛 生
3	嚴 重 特 殊 傳 染 性 肺 炎	53	最 重 要 的 是
4	新 型 冠 狀 病 毒 肺 炎	54	主 動 告 知 醫 師 旅 遊 史
5	無 症 狀 感 染 者	55	發 燒 或 呼 吸 道 症 狀
6	新 型 冠 狀 病 毒 感 染	56	動 手 做 相 關 實 驗
7	病 毒 感 染 的 肺 炎	57	關 注 的 突 發 公 共
8	冠 狀 病 毒 感 染 的	58	武 漢 肺 炎 確 診 病 例
9	2019 新 型 冠 狀 病 毒	59	全 國 各 地
10	新 型 冠 狀 病 毒 的	60	更 重 要 的 是
11	衛 生 福 利 部 疾 病 管 制 署	61	醫 療 照 護 相 關 感 染
12	疾 病 預 防 控 制 中 心	62	配 戴 口 罩 儘 速 就 醫
13	的 新 型 冠 狀 病 毒	63	與 邊 境 管 制 措 施
14	嚴 重 急 性 呼 吸 道 症 候 群	64	監 測 與 邊 境 管 制
15	公 共 衛 生 緊 急 事 件	65	疫 情 監 測 與 邊 境
16	2019 冠 狀 病 毒 疾 病	66	為 境 外 移 入
17	突 發 公 共 衛 生 事 件	67	肺 炎 疫 情 防 控 工 作
18	人 類 免 疫 缺 乏 病 毒	68	加 強 疫 情 監 測 與
19	例 境 外 移 入	69	值 得 一 提 的
20	居 家 檢 疫 14 天	70	依 指 示 配 戴 口 罩
21	個 國 家 和 地 區	71	醫 師 旅 遊 史 及 接 觸 史
22	感 染 新 型 冠 狀 病 毒	72	福 利 部 疾 病 管 制 署 編 號
23	新 型 冠 狀 病 毒 疫 情	73	症 狀 感 染 者 的

發展主題語料庫以輔助華語教學－以 2019 新型冠狀病毒語料庫為例

24	醫療 聯盟 資料 來源	74	境 外 移入 病例
25	中國 疾病 預防 控制	75	告知 醫師 旅遊史 及
26	提供 更 好 的	76	華南 海鮮 批發 市場
27	免疫 缺乏 病毒 感染	77	並 依 指示 配戴
28	非 小 細胞 肺癌	78	流感 併發 重 症
29	國家 衛生 健康 委員會	79	新型 冠狀 病毒 核酸
30	疫情 指揮 中心 今	80	指示 配戴 口罩 儘速
31	世界 衛生 組織 的	81	如 有 疑似 症狀
32	指揮 中心 指揮官 陳時中	82	呼吸 症候群 冠狀 病毒
33	疫情 指揮 中心 指揮官	83	中東 呼吸 症候群 冠狀
34	病毒性 C 型 肝炎	84	可 防 可 控
35	急性 病毒性 C 型	85	疾管署 持續 加強 疫情
36	感染 的 肺炎 疫情	86	中國 國家 衛生 健康
37	冠狀 病毒 肺炎 疫情	87	醫事 人員 繼續 教育
38	境 外 移入 個案	88	持續 住院 隔離 中
39	的 密切 接觸 者	89	因應 嚴重 特殊 傳染性
40	國際 公共 衛生 緊急	90	特殊 傳染性 肺炎 防治
41	疫情 最 嚴重 的	91	武漢 肺炎 疫情 持續
42	較 去年 同 期	92	長期 照護 醫事 人員
43	我們 的 候選 藥物	93	照護 醫事 人員 繼續
44	武漢 華南 海鮮 市場	94	新 冠肺炎 確診 病例
45	醫療 照護 工作 人員	95	急性 病毒性 A 型
46	華南 海鮮 市場 的	96	及時 診斷 及 通報
47	關注 公共 衛生 緊急	97	同時 主動 告知 醫師
48	對 新型 冠狀 病毒	98	由 衛生 單位 安排
49	由 中央 主管 機關	99	新冠 肺炎 疫情 防控
50	等 所 需 經費	100	收治 負壓 隔離 病房

5 字詞 N 連詞（前 100 個）

1	中央 流行 疫情 指揮 中心	51	特殊 傳染性 肺炎 防治 及
2	新型 冠狀 病毒 感染 的	52	傳染性 肺炎 防治 及 紓困

3	冠狀病毒感染的肺炎	53	在這種情況下
4	中國疾病預防控制中心	54	由衛生單位安排就醫
5	人類免疫缺乏病毒感染	55	例境外移入及
6	流行疫情指揮中心今	56	衛生單位安排就醫採檢
7	急性病毒性C型肝炎	57	嚴重特殊傳染性肺炎中央
8	新型冠狀病毒肺炎疫情	58	肺炎中央流行疫情指揮
9	病毒感染的肺炎疫情	59	病毒感染的肺炎病例
10	疫情指揮中心指揮官陳時中	60	特殊傳染性肺炎中央流行
11	國際公共衛生緊急事件	61	傳染性肺炎中央流行疫情
12	流行疫情指揮中心指揮官	62	新型冠狀病毒核酸檢測
13	關注公共衛生緊急事件	63	自主健康管理14天
14	在全球報紙版面上	64	中國醫藥大學附設醫院
15	國際關注公共衛生緊急	65	嚴重特殊傳染性肺炎病例
16	的突發公共衛生事件	66	呼吸症候群冠狀病毒感染症
17	國際關注的突發公共	67	通報機場及港口檢疫
18	關注的突發公共衛生	68	新型冠狀病毒肺炎的
19	監測與邊境管制措施	69	主動通報機場及港口
20	疫情監測與邊境管制	70	應主動通報機場及
21	加強疫情監測與邊境	71	人與人之間的
22	告知醫師旅遊史及接觸史	72	肺炎疫情工作領導小組
23	衛生福利部疾病管制署編號	73	專家諮詢小組召集人張上淳
24	並依指示配戴口罩	74	國家衛健委高級別專家組組長
25	持續加強疫情監測與	75	流行疫情指揮中心宣布
26	值得一提的是	76	確診個案中6人
27	指示配戴口罩儘速就醫	77	的肺炎疫情防控工作
28	中東呼吸症候群冠狀病毒	78	一種新型冠狀病毒
29	無症狀感染者的	79	在人與人之間
30	依指示配戴口罩儘速	80	可撥打免付費防疫
31	主動告知醫師旅遊史及	81	4.1 傳染病檢體採檢手冊
32	疾管署持續加強疫情監測	82	中心專家諮詢小組召集人
33	因應嚴重特殊傳染性肺炎	83	流行疫情指揮中心將

發展主題語料庫以輔助華語教學－以 2019 新型冠狀病毒語料庫為例

34	長期 照護 醫事 人員 繼續	84	指揮 中心 專家 諮詢 小組
35	嚴重 特殊 傳染性 肺炎 防治	85	需 居家 檢疫 14 天
36	以利 及時 診斷 及 通報	86	開放 空間 不用 戴 口罩
37	同時 主動 告知 醫師 旅遊史	87	給付 項目 及 支付 標準
38	臺灣 大學 醫學院 附設 醫院	88	有 下列 情形 之 一
39	中國 國家 衛生 健康 委員會	89	進出 醫院 者 要 戴
40	疾病 控制 與 預防 中心	90	請 務必 告知 醫師 旅遊史
41	其餘 持續 住院 隔離 中	91	確診 感染 新型 冠狀 病毒
42	美國 疾病 控制 與 預防	92	接觸 者 持續 追蹤 中
43	感染 的 肺炎 疫情 防控	93	嚴重 特殊 傳染性 肺炎 通報
44	的 無 症狀 感染 者	94	醫院 者 要 戴 口罩
45	新型 冠狀 病毒 感染 肺炎	95	疫情 最 嚴重 的 國家
46	嚴重 特殊 傳染性 肺炎 疫情	96	目前 收治 負壓 隔離 病房
47	有 發燒 或 呼吸道 症狀	97	定點 醫療 機構 隔離 治療
48	免 付費 防疫 專線 1922	98	病毒 感染 的 肺炎 診療
49	撥打 免 付費 防疫 專線	99	病毒 感染 的 肺炎 確診
50	肺炎 防治 及 紓困 振興	100	感染 的 肺炎 診療 方案

6 字詞 N 連詞（前 100 個）

1	新型 冠狀 病毒 感染 的 肺炎	51	蔓延 28 個 國家 和 地區
2	中央 流行 疫情 指揮 中心 今	52	新型 冠狀 病毒 感染 的 肺炎
3	中央 流行 疫情 指揮 中心 指揮官	53	並 請 務必 告知 醫師 旅遊史
4	冠狀 病毒 感染 的 肺炎 疫情	54	返國 後 14 天 內 如
5	關注 的 突發 公共 衛生 事件	55	其他 接觸 者 持續 追蹤 中
6	國際 關注 的 突發 公共 衛生	56	中央 流行 疫情 指揮 中心 今天
7	疫情 監測 與 邊境 管制 措施	57	4.1 傳染病 檢體 採檢 手冊 頁碼
8	流行 疫情 指揮 中心 指揮官 陳時中	58	機場 及 港口 檢疫 人員 並
9	加強 疫情 監測 與 邊境 管制	59	國家 衛健委 高級別 專家組 組長 鍾南山
10	持續 加強 疫情 監測 與 邊境	60	及 港口 檢疫 人員 並 配合
11	依 指示 配戴 口罩 儘速 就醫	61	防治 及 紓困 振興 特別 條例

12	主動 告知 醫師 旅遊史 及 接觸史	62	疫情 指揮 中心 專家 諮詢 小組
13	並 依 指示 配戴 口罩 儘速	63	在 人 與 人 之間 傳播
14	疾管署 持續 加強 疫情 監測 與	64	及 其他 應 遵行 事項 之
15	同時 主動 告知 醫師 旅遊史 及	65	亂 丟 口罩 毋通 亂 丟
16	美國 疾病 控制 與 預防 中心	66	中央 流行 疫情 指揮 中心 表示
17	病毒 感染 的 肺炎 疫情 防控	67	港口 檢疫 人員 並 配合 防疫
18	撥打 免 付費 防疫 專線 1922	68	武漢 爆發 的 新型 冠狀 病毒
19	特殊 傳染性 肺炎 防治 及 紓困	69	檢疫 人員 並 配合 防疫 措施
20	嚴重 特殊 傳染性 肺炎 防治 及	70	新型 冠狀 病毒 肺炎 診療 方案
21	傳染性 肺炎 防治 及 紓困 振興	71	國務院 聯防 聯控 機制 召開 新聞
22	特殊 傳染性 肺炎 中央 流行 疫情	72	醫學 觀察 的 密切 接觸 者
23	嚴重 特殊 傳染性 肺炎 中央 流行	73	新型 冠狀 病毒 肺炎 疫情 防控
24	傳染性 肺炎 中央 流行 疫情 指揮	74	後 14 天 內 如 出現
25	肺炎 防治 及 紓困 振興 特別	75	天 內 如 出現 疑似 症狀
26	中東 呼吸 症候群 冠狀 病毒 感染症	76	免疫 缺乏 病毒 感染 急性 病毒性
27	冠狀 病毒 感染 的 肺炎 病例	77	人類 免疫 缺乏 病毒 感染 急性
28	通報 機場 及 港口 檢疫 人員	78	14 天 內 如 出現 疑似
29	肺炎 中央 流行 疫情 指揮 中心	79	有關 國際 公共 衛生 的 緊急
30	主動 通報 機場 及 港口 檢疫	80	新冠 肺炎 疫情 工作 領導 小組
31	由 衛生 單位 安排 就醫 採檢	81	在 一 份 聲明 中 說
32	應 主動 通報 機場 及 港口	82	國際 公共 衛生 的 緊急 事件
33	確診 個案 中 6 人 死亡	83	海鮮 市場 確診 了 7 例
34	感染 的 肺炎 疫情 防控 工作	84	機構 因應 嚴重 特殊 傳染性 肺炎
35	中央 流行 疫情 指揮 中心 宣布	85	新型 冠狀 病毒 核酸 檢測 試劑盒
36	可 撥打 免 付費 防疫 專線	86	居家 檢疫 或 自主 健康 管理
37	中央 流行 疫情 指揮 中心 將	87	國光 生物 科技 股份 有限公司 國內

發展主題語料庫以輔助華語教學－以 2019 新型冠狀病毒語料庫為例

38	中心 專家 諮詢 小組 召集人 張上淳	88	做好 不明 原因 肺炎 救治 工作
39	指揮 中心 專家 諮詢 小組 召集人	89	保險 醫療 服務 給付 項目 及
40	進出 醫院 者 要 戴 口罩	90	依 中央 流行 疫情 指揮 中心
41	病毒 感染 的 肺炎 診療 方案	91	衛生 防護 中心 傳染病 處 主任
42	冠狀 病毒 感染 的 肺炎 確診	92	空氣 不 流通 的 公共 場所
43	冠狀 病毒 感染 的 肺炎 診療	93	發現 明確 的 人 傳人 證據
44	病毒 感染 的 肺炎 確診 病例	94	各 款 情形 之 一 者
45	報告 新型 冠狀 病毒 感染 的	95	不明 原因 肺炎 救治 工作 的
46	台灣 中央 流行 疫情 指揮 中心	96	防治 及 紓困 振興 特別 預算案
47	累計 報告 新型 冠狀 病毒 感染	97	華南 水果 海鮮 市場 確診 了
48	中央 流行 疫情 指揮 中心 公布	98	社團 法人 台灣 感染 管制 學會
49	新型 冠狀 病毒 感染 肺炎 疫情	99	政府 嚴重 特殊 傳染性 肺炎 防治
50	依 指示 戴 口罩 儘速 就醫	100	中央 政府 嚴重 特殊 傳染性 肺炎

Developing a Topic-Specific Web Corpus, COVID-19, for Chinese Language Teaching and Learning

Ming-Hong BAI

**Research Center for Translation, Compilation and Language Education
National Academy for Educational Research**

Howard Hao-Jan CHEN

**Department of English
National Taiwan Normal University**

Ying LIN

**Department of English
National Taiwan Normal University**

Abstract

The coronavirus disease 2019 (COVID-19) has had a serious impact on people around the globe. However, teams in Europe and the United States worked hard in developing English COVID-19 corpora. Yet, there is no such corpus available in Chinese. Therefore, this paper aimed to fill this gap by building a Chinese COVID-19 corpus for researchers, teachers, and students. The two research questions are as follows: (1) Can the Chinese COVID-19 corpus and No Sketch Engine platform provide useful information for Chinese teachers and students? (2) What are the advantages and disadvantages of this web platform in terms of the contents and analyses of the corpus? A Chinese COVID-19 corpus was built with WebBootCat. This study also generated raw data for assisting Chinese teaching and learning by analyzing the corpus: (1) top-frequency vocabulary items, (2) keywords, (3) n-grams, (4) collocations. This study found that using WebBootCat could efficiently generate a topic-specific corpus, which has the following advantages: (1) immediacy, (2) wide coverage, (3) authentic language, and (4) rich language contexts. However, as data were crawled from

the web, irrelevant noises might be detected. Moreover, more tools need to be developed in No Sketch Engine.

Keywords: Chinese for specific topics, collocation, keyword analysis, N-gram analysis, web as corpus, word frequency