

臺灣華語文語料庫在華語文教育的應用*

吳鑑城 白明弘 林慶隆¹

國家教育研究院語文教育及編譯研究中心

摘要

國家教育研究院臺灣華語文語料庫(Corpus of Contemporary Taiwanese Mandarin, COCT)語料包括書面語、口語、華英雙語及華語中介語。本文目的主要為應用華語文語料庫研發華語文漢字、詞語及語法點分級及研發語料庫整合應用系統。

本文應用華語文語料庫語料的詞頻、覆蓋率、分布均勻度、類詞綴、語義場關聯詞、構詞率及組字力的統計分析結果，輔以學者專家和資深華語文教師諮詢，完成華語文漢字、詞語及語法點分級標準。此外，整合應用華語文分級標準成果及語料庫科技研發建置了「語料庫索引典系統」、「語義場關聯詞查詢系統」、「作文錯別字自動批改系統」及「例句編輯輔助系統」等系統。

最後，本文並對未來華語文語料庫在通用詞頻表的建置、基礎詞彙表的建構、及華語文搭配詞結構分析等之研究，提出建議。

關鍵詞：華語文教學 華語文分級標準 臺灣華語文語料庫

* 本文部分成果曾於 2019 年中央研究院語言學研究所主辦之《第二屆中研語言學論壇》報告及討論，題目為〈臺灣華語文語料庫建構與應用〉，感謝與會者的寶貴建議。而且，感謝本期刊三位匿名審查委員提供寶貴建議，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。

1. 前言

臺灣對外華語文教學需要有共同的華語文學習者能力指標及漢字、詞語及語法點分級標準，以作為教學、教材及評量發展的參照依據。而且，華語文能力指標及分級標準的建置若能植基於民眾生活中實際會使用的語言文字，則更能符合華語文學習者學習華語的需求，因此，能提供準語言實際使用情境的華語文語料庫便成為必要的工具（柯華葳等 2013）。

語料庫為一個語言的知識庫，不僅可以客觀顯示語言真實自然的使用情境和使用頻率等資訊，也可作為語言應用研究資源。語言學習者可透過真實語料，省視在語文學習上的語誤及盲點，而改善語感的侷限。研究人員更可藉由語料，探究語言在不同情境和語域的變化。一個大型的語料庫結合科技可分析語言的規律和用法，為語言科技和電腦輔助語言教學的核心。

基於提升華語文教學的效能，教育部於其華語文教育八年計畫（2013-2020）中將「建置應用語料庫及標準體系」列為建構永續發展基礎計畫，由國家教育研究院研究人員自 2013 年開始執行。6 年來，所建置的臺灣華語文語料庫(Corpus of Contemporary Taiwanese Mandarin, COCT)，語料數量包括書面語 4 億 1,000 萬字，口語 2,530 萬字，華英雙語 1,160 萬字，華語中介語 142 萬字。已經開放線上查詢語料內容，整合入口網址為：<https://coct.naer.edu.tw/>；而且，應用在建置華語文漢字、詞語及語法點分級標準、基礎詞彙及研發建置了各類華語文教學整合應用系統。本文目的主要為探討華語文語料庫在研發華語文漢字、詞語及語法點分級標準及研發語料庫整合應用系統的應用（柯華葳等 2013, 2014, 2015, 2016；許添明等 2017；郭工賓等 2018）。

2. 文獻探討

2.1 語料庫應用系統的研發

語料庫的應用必須運用語料庫分析工具提供檢索及分析。Key word in context (KWIC) 的概念最早由 Luhn (1966) 提出，其概念是將含關鍵詞的例句從語料庫中抽出，以關鍵詞置中對齊排列呈現例句，以便於觀察關鍵詞的語境。至今，KWIC 仍是語料庫檢索工具最常用的呈現方式。

雖然 KWIC 呈現例句有利於語境的觀察，但是當例句樣本超過一定的數量時，詞典專家也難以逐一閱讀，於是進一步語言樣本分析技術應運而生。

例如透過詞彙共現的分析，可以統計出常用的搭配詞。例如：Church 與 Hanks (1990) 提出交互資訊(mutual information)從語料庫中評估搭配關係的強度；Smadja (1993) 提出以詞彙距離的平均變異數(means variance)來計算搭配關係的強度。Pecina (2005) 收集了數十種搭配關係強度的計算公式，比較分析這些公式的優缺點。

以上的搭配詞研究都著重在詞彙之間的共現強度，但搭配詞在句法中擔任不同的角色，以共現強度來計算搭配詞可能排擠到詞頻不高，但在某些角色中極為重要的搭配詞。因此 Kilgariff 與 Tugwell (2001) 提出了詞彙特性速描系統(Word Sketch Engine)，該系統進一步考量詞彙之間的角色關係，直觀呈現關鍵詞和搭配詞的語法結構。例如：主詞、受詞、形容詞等語法關係描述。詞彙特性速描系統目前已被廣泛應用在包括牛津英語詞典及麥克米倫英語詞典等許多詞典的編輯。

語料庫分析除了透過例句提供詞典編輯知識之外，同時也可以用來抽取近義詞資訊。在英文的近義詞抽取研究上，最主要的理論基礎來自於近義詞具備近似語境的假設。根據語境分布相似度的計算，此方法可提供一組語境分布相似的詞彙做為近義詞的候選清單(Salton 1989; Ruge 1992; Alshawi and Carter 1994; Grishman and Sterling 1994)。為提高近義詞精確率，除了語境相似度之外，有些研究者利用文法關係加以限制(Lin 1998; Curran and Moens 2002; van der Plas and Bouma 2004)。然而，許多研究者發現語境相似的詞不全然是近義詞，也包括反義詞、上位詞及下位詞等，而且無法透過語境分布來區別近義關係與非近義關係(van der Plas and Tiedemann 2006)。所以，另有一些研究者以雙語語料做為抽取近義詞的來源，其背後的假設為：近義詞在其他語言中會以相似的詞彙翻譯。所以只要比較詞語翻譯成其他語言所使用的詞彙分布，即可得到一組近義詞的候選清單(Curran and Moens 2002; van der Plas and Tiedemann 2006; Wang, Cao and Zhou 2015)。

Deerwester、Dumais、Furnas、Landauer 與 Harshman (1990) 提出潛在語義分析(Latent Semantic Analysis)的方式計算近義詞，之後有許多研究者針對潛在語義分析提出了改進的方法。隨著類神經網路理論的成熟，類神經網路理論也被應用到自然語言處理的研究領域。Turian、Ratinov 與 Bengio (2010) 利用類神經網路的技術提出了詞彙語義向量(word representation)的計算方法，這個方法可以從大量語料庫中，自動估算出每一個詞彙的語義向量。這些語義向量包含了該詞彙的一些語義及語法特徵，可以用來計算詞彙

之間的一些關係。例如：使用餘弦公式計算詞彙的語義相似度，或運用於推估詞彙之間的語義及語法關係，包括詞彙的語法變化、反義詞及上下位等。隨後許多研究者也提出了不同的詞彙向量表示計算法 (Mikolov, Chen, Corrado and Dean 2013; Pennington, Socher and Manning 2014)；這些語義向量的研究可以提供非常豐富的語義訊息。

在反義詞的抽取研究上，由於語境分布的差異無法區別近義詞與反義詞，所以許多研究者以基於模版(pattern-based)的方式抽取反義詞(Wang, Thomas, Sheth and Chan 2010; Al-Yahya, Aldhubayi and Al-Malak 2014)，使用模版的好處是，近義詞和反義詞的模版不會完全一樣，但缺點是，模版的建立必需依賴人工的觀察與建立，很難照顧到不同類型的反義詞。

2.2 語料庫在華語文教學的應用

語料庫語料的詞語頻率代表每一個詞語在語料庫中的常用性，是詞語常用性指標。儘管一些研究者對語料庫頻率做為常用詞的依據存在疑慮，但是依據語料庫詞頻選詞的原則仍是常用的做法（張莉萍、陳鳳儀 2006）。因為較高頻的詞，語料庫覆蓋率也勢必較高，故選擇高頻詞關係到教學的效率。儘管如此，一些研究發現，僅依據詞頻統計，有部分高頻詞未必是常用詞。有些高頻詞只集中重複出現在特定領域的文本，頻率雖高卻不是常用詞。為解決此一問題，許多學者提出不同的常用詞計算公式 (Juilland and Chang-Rodríguez 1964; Carroll 1970; Rosengren 1971; Huang, Zhang and Yu 2005; Bai, Wu, Chien, Huang and Lin 2016)。這些研究的方法主要是透過均勻度的計算，抑制高頻但只分布在少數文本的詞語，使用均勻度的值趨近於常用度。

語料庫累計覆蓋率代表該等級的學習者必須熟悉多少日常用語，是字詞表分級的重要依據。鄭昭明（1997）對常用字和常用詞，依字詞在語料庫累計覆蓋率的 50%、70%、95%、99%、100%，將其分為 5 個等級。張郁雯（2003）依據詞語在觀察的語料庫中的累計覆蓋率（涵蓋整個語料庫詞語總數的比例），設立 75%、85%、90%、95% 的分界點，將詞語分為 5 個等級。

除了字詞表的分級，一些現代詞典的編撰也得力於語料庫的發展。Verlinde 與 Selva（2009）比較語料庫為本的編撰法和專家直覺式(intuition-based)編撰法，發現語料庫為本的詞典能較詳盡的列出重要日常詞彙，而且能得知多少詞彙量足以覆蓋多數文本，讓學習者能從基礎詞彙學習，

較快掌握生活用語。Summers (1996) 提出五項詞頻統計對現代詞典編撰產生的重要影響包括：(1) 詞頻與詞義的排序有關；(2) 詞頻與詞條的呈現方式有關；(3) 詞頻需針對不同詞類進行統計，因為不同詞類的同一詞彙可能被分在不同等級的學習詞表中，例如：口語中不視動詞 *bid* 為基礎詞彙，書面語中卻認為 *bid* 是基礎詞；(4) 詞組也應統計頻率，例如常見詞組 *the other day* 因為在語料中出現頻率非常高，故應收錄為詞及(5) 慣用語也應統計詞頻，作為參照之用，如：口語語料庫可以輔助收集同義不同形的「慣用語」，如 *don't care* 的同義詞包括：*not give a damn*、*couldn't care less*、*be past caring*、*so what* 等。

3. 研究方法

本文在華語文語料庫建置、語料庫應用於華語文分級標準建置、及語料庫應用系統研發所應用的方法，說明如下。

3.1 華語文語料庫建置研究方法

在華語文語料庫(Corpus of Contemporary Taiwanese Mandarin, COCT)建置，執行步驟包括文獻分析、文件分析、專家諮詢、語料收集、處理及建置等(柯華葳等 2013, 2014, 2015, 2016；許添明等 2017；郭工賓等 2018)。此外，由於中文分詞與詞性標注為中文語料處理最基礎亦是最重要的步驟之一，本計畫自 105 年起逐年修訂優化現有的中文分詞(含詞性標注)系統程式，以取得最佳化效果。專家諮詢採定期諮詢及不定期諮詢。

語料庫後設資料設定主要參考中央研究院漢語平衡語料庫、美國當代英語語料庫(COCA)及英國國家語料庫(BNC)。書面語後設資料包括作者、出版者、出版年、語式、語文、著(譯)作、主題、媒體、著作權等 9 類資訊。語料文本主題參考美國當代語料庫、英國國家語料庫，再參考中文圖書分類，整合為總類、哲學及宗教類、科學類、應用科學類、社會科學類、史地類、語言文學類、藝術類、商業及金融類、休閒類等 10 大類。口語語料內容涵蓋法政軍事、財經、時事、科學、生活時尚、文教藝術等電視節目的字幕及其對應音檔。華英雙語語料內容包括文學、科學、財經、藝術、思想、文化、全球、休閒等類型文章。

書面語語料、口語語料及華英雙語語料以臺灣使用的華語文為限，書寫文字需以正體字呈現，書面語語料和口語語料為 2008 年以後語料，華英雙語

為 1998 年以後語料。語料的蒐集大部分依據政府採購法招標取得利用授權，少部分為國立大學和政府機構授權，及國家教育研究院語料。華語中介語語料來源為 2014 年以後各大學華語文教學中心的非華語母語者學生所授權之作文及國家華語測驗推動工作委員會授權之限時考試語料。

3.2 語料庫應用於華語文分級標準建置研究方法

華語文分級標準建置應用「學者專家諮詢」、「語料庫統計」及「資深華語文教師諮詢」三步驟循環進行，過程中除了研究人員研究外，而且總共召開近百場諮詢會議討論，才完成研發（柯華葳等 2013, 2014, 2015, 2016；許添明等 2017；郭工賓等 2018）。

華語文漢字分級表研發，經文獻分析後再諮詢學者專家，確定字表定位為「學習者必須達到識讀程度並掌握字的書寫」，決定使用的語料庫與統計指標，字表調校的各项指標，包括：參照能力指標、構字率、構詞率、語義聯想、常用詞群類聚、認寫難易度、專名移除、功能詞級別調整及異體字整併等項目，並決定收錄的各級字數。「語料庫統計」由研究人員依據指標，並運用華語文語料庫中的教材語料庫及七份其他字表進行統計分析，統計指標依序為：各級華語教材漢字累計覆蓋率（字頻）、全級華語教材漢字累計覆蓋率（字頻）、七份字表涵蓋率、各級華語教材漢字篇章分布率及全級華語教材漢字篇章分布率。「資深華語文教師諮詢」，依據上述指標逐字討論、決定是否收錄及收錄的級別，最後與計畫「詞語分級表」各級內容拆詞為字的結果進行比較分析及再調整。

華語文詞語分級表研發，經文獻分析後再諮詢學者專家，確定詞表定位為「以溝通使用為主」，決定收詞與整併等原則、使用的語料庫與統計指標、詞表收錄詞語分級表和類詞綴表、和收錄的各級詞數。「語料庫統計」由研究人員依據指標，應用華語文語料庫的書面語語料，統計詞頻與覆蓋率，挑選出達語料庫涵蓋率 90% 的 17,329 詞，再根據原則完成詞頻加計、詞條刪併後，得出 14,379 詞的統計詞表，再加入口語語料詞頻及高頻教材生詞；最後，加入「詞頻及分布均勻度」、「詞性分布」、「注音及釋義」、「對應漢字等級」及「多音詞」等參考資訊。「資深華語文教師諮詢」依據收詞與整併原則及參考資訊，逐詞討論，再與計畫「漢字字表」及「語法點表」比較分析調整。最後再與「99~104 年常用語詞」及「華測會華語八千詞表」分析比較及調整。

華語文教學語法點分級研發，經文獻分析後再諮詢學者專家，確定語法點以「教學語法」為收錄原則，因此僅收錄「基本」和「常用」的語法點。語法點來源包括《當代中文課程》、《新版實用視聽華語》、《新實漢語課本》和《漢語教程》及語料庫。範圍涵蓋詞語、短語、固定結構及話語標記，結構或語義較艱澀用法則不收錄，如：四字格、慣用語、諺語；一般動詞、一般副詞或量詞、及表達語氣之嘆詞等。語法點分級指標包括國教院「華語文能力指標」、「語料庫統計資料」、「《語法等級大綱》和教材冊課」及「教材交集」。在「語料庫統計」，研究人員根據上述指標，應用華語文語料庫書面語及口語語料及「中央研究院平衡語料庫」，應用 CQP (Corpus Query Processor) 查詢語法，計算結構和非結構類語法點頻率。再應用隨機抽樣統計理論，分析語法點的使用情形、共現關係及互見訊息 (MI) 等資訊。在「資深華語文教師諮詢」，針對「語法點選錄項目」、「語法點分類類別」、「語法點分級級數」及「語法點例句編纂」逐一分析討論。最後本語法點表再和計畫「漢字表」及「詞語表」比較分析及調整。

3.3 語料庫應用系統研發方法

語料庫應用系統研發發主要採用文獻分析、文件分析、專家諮詢及程式撰寫及測試 (柯華葳等 2013, 2014, 2015, 2016; 許添明等 2017; 郭工賓等 2018)。由於標準體系包括漢字、詞語及語法點等，已經逐漸產生具體的成果。這些具體成果除了以書面或電子檔的型式提供給華語文教學者之外，如果結合自然語言分析技術，能夠產生更多的加值應用。在研發上可以分成兩個階段：

第一階段提供標準體系線上查詢系統：標準體系線上查詢系統是將目前標準體系的成果提供線上查詢的功能，包括：漢字分級查詢系統、詞語分級與情境詞查詢系統、語法點查詢系統。查詢的方法可以分為關鍵字詞查詢、部首查詢、拼音查詢、詞類查詢及情境查詢等。

第二階段漢字、詞語、語法點級別自動標注系統：將標準體系包括漢字、詞語、語法點的分級自動標注在資料上。依標注的標的，又可分為 (1) 華語篇章級別自動標注系統：使用者將篇章或例句輸入，系統自動標注漢字、詞語及語法點級別。(2) 搭配詞級別檢索系統：從語料庫中統計出搭配詞時，同時將搭配詞標注級別，並提供級別過濾與分群排序的功能。(3) 語義場關聯詞級別檢索系統：從語料庫中統計出語義場關聯詞時，同時將關聯詞標注

級別，並提供級別過濾與分群排序的功能。

4. 結果與討論

4.1 語料庫應用於華語文分級標準建置

華語文漢字、詞語、語法點分級標準建置過程，我們充分發揮語料庫可作為實際語言情境代表且大量的語料可彌補學者專家難免有所遺漏的特性，實際將華語文語料庫的資訊應用在各個層面，包括：（1）字詞表分級指標、（2）字詞常用度、（3）類詞綴整理、（4）情境類聚、（5）異體字及同義異形詞的類聚與合併、（6）構詞率與組字力的運算（柯華葳等 2013, 2014, 2015, 2016；許添明等 2017；郭工賓等 2018）。而且，運用學者專家及資深華語文教師諮詢，完成 1-7 級 3,100 個漢字，14,267 個詞語和 73 個詞綴的分級，及 1-5 級 520 個語法點的分級（吳鑑城、白明弘、林慶隆 2019）。

4.1.1 字詞表分級指標

語料庫覆蓋率是字詞表分級最重要的參考指標之一，語料庫覆蓋率代表字詞表在生活中涵蓋多少日常用語。例如覆蓋率 90% 的詞表，代表完整學習該詞表後，能辨識 90% 的日常用語。覆蓋率訊息可以和能力指標配合，成為語言標準體系的制定標準。在我們的計畫中，實際應用了語料庫覆蓋率作為詞表詞彙量的訂定參考。計算的語料庫以華語文語料庫取 90% 的覆蓋率作為詞表總詞彙量的標準。

4.1.2 字詞常用度

語料庫詞頻有時不能真正反應詞語常用度，有些高頻詞只在特定領域文本中集中重複出現，造成許多詞的頻率雖高卻不是常用詞的現象。例如「皇帝」一詞，在小說及故事書中頻繁的出現，但在其它領域的文本中卻幾乎不曾出現。為解決語料庫詞頻的問題，針對字詞常用度我們改採分布均勻度公式，同時考慮篇章、主題、時間、語式等分類的分布均勻度。透過均勻度的計算，可以有效抑制高頻但非常用詞的現象。

4.1.3 類詞綴整理

類詞綴的概念最早由呂叔湘（1979）所提出，是指語義上未完全虛化，介於詞根和詞綴間過渡階段的成分。由於類詞綴衍生性高，常與自由語素結合為合成詞，且同一詞綴的功能及用法可推衍，故在教學上可以達到事半功

倍的效能。然而，過去的華語文教材及詞表鮮少單獨編列詞綴，且以往研究偏重於詞綴的生成、定義、功能及類型（尹海良 2007；沈光浩 2014；湯廷池 2014）。本詞表依據語料庫統計，可以得知每一個類詞綴在語料庫中所衍生詞的頻率總和。再依詞表的分級流程，可得出各級詞彙中應習得詞綴的比例為多少，以作為學習難易度的參考。而且，本詞表收錄能產性較高，帶有固定意義的詞首、詞尾，並訂定各個類詞綴的級別及收錄相關詞語。

4.1.4 情境類聚

為因應教材編輯、測驗出題等實際教學需求，需要依據情境呈現各級詞表的詞彙內容，詞彙作情境分類主要目的為便於學習相近或同一語義場的詞，詞表依情境排序可將相關詞彙聚集。詞語應用語料庫統計分析方法產出的「語義場關聯詞詞彙」輔助詞彙分類，如「高興」一詞的語義場關聯詞有「開心、難過、感動...」等，可歸為「個人資訊」項下的「偏好及情感」。

4.1.5 異體字及同義異形詞的類聚與合併

應用語義場關聯詞統計分析，包含異體字的詞如「裏面／裡面」、「裡頭／裏頭」、「公布／公佈」、「注定／註定」、「長歎／長嘆」、「品嚐／品嚐」、「宣佈／宣布」、「台灣／臺灣」等，會自動類聚在一起。異體字選取依據教育部異體字字典及重編國語辭典修訂本。而同義異形的詞如「機車／摩托車」、「腳踏車／自行車」、「比方說／比如說／譬如說」、「一塊／一塊兒」、「星期天／星期日／週日」、「剛剛／剛」、「哥哥／哥」等，也會自動類聚在一起。應用語義場關聯詞自動類聚的功能，可以讓編輯者整理詞表的過程中，專注於異體字及同義異形詞的專業判斷工作，一方面降低詞表整理的複雜度，一方面可避免聯想的遺漏，在提高詞表編輯效能的同時也提高詞表的品質。

4.1.6 構詞率與組字力的運算

掌握收字多、字義明顯字力強的部件，例如：「其」在溝通使用上比較偏向書面語，所以，在詞語表中編在比較後面的級別。但，「其」是許多常用字的部件，如：「期」、「基」、「斯」、「旗」等，先學「其」對於學習包含該部件的字具有降低記憶負擔的效果。另外，在組字力的計算上，不能只考慮部件可組字的多寡，還要考慮所組成字的語料庫頻率，以及該字本身的頻率。因為有些部件雖能組合出很多字，但卻多是低頻字，或雖然是很

多高頻字的部件，但該字本身很少單獨使用，例如：「酉」字。故在組字力的計算上，語料庫的字頻是重要的依據。字的構詞率在進行字的分級也是很重要的參考依據，許多字的語料庫頻率雖不高，但有相當高的構詞率，例如：「醬」的字頻雖不高，卻是衍生性非常高的詞尾字。透過語料庫的統計，可了解每個漢字的構詞率（黃淑齡、白明弘、吳鑑城、李詩敏、林慶隆 2018）。

4.2 語料庫應用系統開發

語料庫的應用必須依賴語料庫分析工具提供檢索及分析功能，我們為將華語文語料庫應用於建置華語文標準體系之漢字、詞語與語法點分級標準等應用，我們已經研發多套語料庫應用系統，且未來也將持續增強其功能（吳鑑城、白明弘、林慶隆 2019）。

4.2.1 語料庫索引典系統²

語料庫索引典語料內容包括圖書語料、新聞語料、口語語料等，本系統在標準體系建置應用上，可以統計漢字、詞語及語法點在語料庫中的使用頻率、提供基礎詞語例句編寫依據等；在詞典編撰上，可以提供詞彙的語法及語義等訊息。本文所研發的語料庫索引典系統採用 CWB (Christ 1994)及 CQPweb (Hardie 2012)技術。

語料庫索引典系統包含了下列功能：

1. 基本查詢表示式：基本查詢表示式提供了萬用字元查詢及詞類查詢，詞類集合以中研院平衡語料庫所定義的 46 個詞類標記。
2. CQP 查詢表示式：CQP 查詢表示式是系統底層的查詢法，提供詞彙各欄位更精確及更複雜的複合式條件查詢。華語文分級標準建置的語法點頻率即是以 CQP 查詢表示式為主要的查詢方法。此外，我們也為中文詞彙的特性提供了特有的查詢方式，例如：重疊詞的查詢功能。
3. 排序：查詢的結果依照前後文排序。
4. 過濾：依查詢結果進一步設定條件，選擇保留或濾除一些結果。
5. 搭配詞統計：提供最常和關鍵詞一起搭配出現的詞彙。
6. 語料庫隨機取樣：為節省分析時間，從查詢的結果中隨機取出部分樣本分析。

² 語料庫索引典系統（<https://coct.naer.edu.tw/cqpweb/>）。

7. 詞語分布統計：統計關鍵詞在不同主題的語料中出現的比例，用來觀察詞彙分布的特性。

語料庫索引典系統屬於多用途的語言分析工具，在華語文教學上的應用非常廣。可能的應用包括：（1）詞語頻率的比較，例如：比較「腳踏車」與「自行車」的使用頻率；（2）語義的考察，例如：「了」一共有幾種不同的語義；（3）語法結構觀察：例如觀察「一...就...」。（4）搭配詞分析及統計：可用於強化詞彙的運用、語言的學習；（5）近義詞用法的區辨：例如，透過搭配詞的比較，以了解「感覺」、「覺得」及「感受」在用法上的差異。

4.2.2 華英雙語索引典系統³

華英雙語索引典系統最主要的技術包括雙語文句自動對應及雙語詞彙對應技術(Ma 2006; Koehn et al. 2007)。華英雙語索引典系統主要功能由三大部分所構成：1. 雙語互譯例句並陳功能、2. 關鍵詞語翻譯建議功能以及 3. 常用搭配詞建議功能。

1. 雙語互譯例句並陳功能是將包含關鍵詞的例句同時列出。以「教育」一詞為例，系統會找出語料庫中所有包含「教育」的中文例句以及互為翻譯的英文例句（如圖 1）。同時系統會將「教育」從中文例句中標出，以及標出英文相對應的翻譯。
2. 關鍵詞語翻譯建議功能是從語料庫中自動找出英文的翻譯，同樣以「教育」為例，系統自動從雙語語料庫中找出並依照次數排序 education (2403)、educational (640)、educated (107)、school (83)、teaching (65)、training (61) 等和「教育」相對應的翻譯及對應次數（如圖 2）。
3. 常用搭配詞建議功能提供常用搭配詞的使用建議，包括形容詞搭配，如：「高等教育」，名詞搭配如「大學教育」等；動詞搭配，如：「受教育」等（如圖 3）。

³ 華英雙語索引典系統 (<https://coct.naer.edu.tw/bc>)



圖 1：中英雙語互譯例句並陳功能



圖 2：關鍵詞語翻譯建議功能

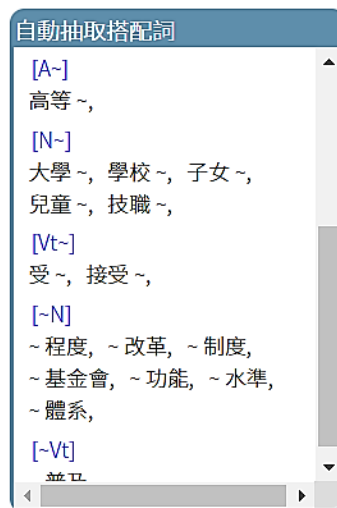


圖 3：常用搭配詞建議功能

華英雙語索引典系統在華語文教學上可以有許多種不同的用法。首先，可以應用在華語文寫作教學上。本系統目前可接受中文或英文的查詢詞，學

生可以透過比較熟悉的英文，找到正確的華語詞彙、片語及搭配等，讓學習者在寫作使用的語法及搭配更接近華語母語者的用法。同時，在翻譯的建議上，可以提供數組不同的翻譯以及頻率，讓學習者可以替換不同的詞語，不會一直使用相同的詞語，也可避免使用太艱澀的詞語。

其次，可以應用在華語文教材的開發。以「採取」一詞為例，雙語索引點提供英文常用翻譯 **adopt**，可以方便教材編輯者編寫「採取」的解釋。同時提供「~ 措施」、「~ 態度」、「~ 方式」、「~ 政策」、「~ 手段」等常用搭配，讓學習者能夠掌握「採取」的用法。除此之外，雙語索引點還提供常用搭配的例句，以及例句的翻譯。透過中英例句，一方面方便編輯者編寫詞彙或搭配的使用例句，另一方面，也可以用來整理華語的語法。例如，「採取...態度」這一組搭配詞，中間可以包含「保留」、「審慎」、「肯定」、「開放」、「支持」及「妥協」等種種不同的態度。使得教材的編寫，更加方便與有彈性。

4.2.3 華語中介語索引典系統⁴

華語中介語索引典系統開發的技術採用語料庫管理引擎包括 **CWB** (Christ 1994)及 **CQPweb** (Hardie 2012)技術。目的在於提供研究者探討中介語現象。在華語文教學上可以提供研究者了解華語學習者的困難。透過語言現象的檢索與比對，分析語言學習者在學習第二語言時所可能遭遇的困難。研究者可以統計中介語語言現象發生在不同母語背景的學習者的分布，藉以了解中介語現象的發生是否與母語遷移有關，或是與不同的語言能力、學習時間、性別、居住地、使用教材等有關，藉以了解困難發生的可能原因。例如：張莉萍（2017）透過中介語語料庫的分析以了解華語關係子句對學習者的困難之處，同時比較不同母語者的困難性質，以了解困難的形成與母語之間的關係。

4.2.4 語義場關聯詞查詢系統⁵

語義場關聯詞查詢系統以自然語言處理技術及機器學習中的詞彙內嵌技術(Mikolov et al. 2013)，自動從華語文語料庫中，計算詞彙之間的相似度。語義場是在同一個語義系統中，在共時條件下，若干個具有共同義素的義位

⁴ 華語中介語索引典系統 (<https://coct.naer.edu.tw/cqpweb/>)。

⁵ 語義場關聯詞查詢系統 (<https://coct.naer.edu.tw/word2vec/>)。

聚合起來的聚合體。例如：從動詞語義場下手可以觀察到整組動詞的共同處，像是一致的論元結構、相同的搭配關係，以抽離出一組互相牽制的語意屬性（張麗麗、陳克健、黃居仁 2000）。因此，具相似結構之近義詞、反義詞可能歸屬於同一語義場。過去的近義詞編輯都是詞典編輯專家依據各人經驗提供近義詞訊息。然而，專家經驗過度依賴個人語感，容易流於主觀與偏狹（李詩敏、白明弘、吳鑑城、黃淑齡、林慶隆 2016）。

由於詞彙的語義場近似程度是以語料庫為基礎自動計算語義向量，所以語料庫所收錄的文本特性將影響到詞彙相似性。例如：在工商新聞為主的報紙文本中，「蘋果」一詞的使用多指「蘋果公司」，語義場關聯詞系統所學習的語義自然也傾向於「蘋果公司」。因此，語義場關聯詞中多包含電腦公司，例如：惠普、戴爾、微軟、英特爾等。使用者在本系統中可以選擇不同的語料庫，以探討在不同語料庫為基礎之下的語義場關聯詞。

語義場關聯詞系統在華語文教學上有許多應用。例如，可以應用在華語文寫作教學上，透過近義詞的提供，學習者可以選擇使用不同的詞彙來表達意思。除了可以避免重複相同的語詞造成呆板的文句外，還可以增加學習所認識的詞彙。另外，也可以應用在教材或詞典的近義詞編寫以及辨析上。

4.2.5 教材分級檢索系統

教材分級檢索系統研發的技術應用 CWB (Christ 1994)及 CQPweb (Hardie 2012)技術。目的方面為讓使用者能根據需求選擇適合閱讀程度之教材，另一方面則是為分級標準分析不同等級的漢字、詞語及語法點的特徵。使用者可輸入欲查詢的詞彙，並選擇所想要的教材等級及教材名稱進行文句之檢索。檢索的結果也可以顯示 CEFR 的等級分布。圖 4 為「畢業」一詞出現在各等級教材中的分布情況，可以發現，「畢業」主要出現在 B1、B2 及 C1 三個等級的教材之中，從中可以了解，一般教材將「畢業」一詞歸在中等等級教材中。圖 5 則是「畢業」一詞出現在各類教材中的分布情況，從中可以了解，哪些教材中較常使用這個詞彙。

教材分級的分析除了用來了解詞彙的等級之外，也可以用來分析漢字及語法點的分級，在標準體系的建置及教材編輯的參考上，有很大的應用空間。

另外，這套檢索系統也可以應用來分析教材中搭配詞的使用。圖 6 為「畢業」一詞在教材中的搭配詞使用情況，如果和一般語料庫比較可以發現，在教材中「畢業」相對較少與「於」搭配，這可能是因為教材編輯者認為「於」

這個詞的用法比較困難，但也有可能是教材編輯者不自覺的遺漏。如此利用教材語料庫系統和一般語料庫系統的比較，可以分析現有教材語料庫的一些特性。

Distribution breakdown for query "畢業": this query returned 38 matches in 26 different texts				
Categories:	General information ▼	Show as:	Distribution table ▼	
Category for crosstabs:	No crosstabs ▼		Show distribution ▼	Go!
Based on classification: CEFR分級				
Category [↓]	Words in category	Hits in category	Dispersion (no. files with 1+ hits)	Frequency [↓] per million words in category
A1	0	0	0 out of 117	0
A2	0	1	1 out of 306	0
B1	0	6	5 out of 365	0
B2	0	10	8 out of 700	0
C1	0	13	9 out of 298	0
C2	0	8	3 out of 33	0
Total:	0	38	26 out of 1819	0

圖 4：「畢業」在不同等級教材中的分布

Category [↓]	Hits in category	Dispersion (no. files with 1+ hits)
僑教雙週刊_字的故事	<u>1</u>	1 out of 108
初中華文	<u>2</u>	2 out of 84
實用中文讀寫	<u>2</u>	2 out of 26
實用視聽華語	<u>5</u>	4 out of 73
Hello 華語	<u>1</u>	1 out of 60
民間故事	<u>1</u>	1 out of 48
海華文庫	<u>1</u>	1 out of 333
當代中文課程	<u>7</u>	5 out of 84
迷你廣播劇	<u>7</u>	2 out of 12
遠東生活華語	<u>1</u>	1 out of 52
高中華文	<u>1</u>	1 out of 35
中文讀本	<u>3</u>	3 out of 83
中華民國的故事	<u>6</u>	2 out of 48
Total:	38	26 out of 1819

圖 5：「畢業」在各教材中的出現情況

There are 176 different words in your collocation database for "[word="畢業"%c]". (Your query "畢業" returned 38 matches in 26 different texts) [0.056 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	以後	479	0.035	7	7	61.913
2	後	413	0.03	5	3	42.028
3	證書	2	0	2	2	39.721
4	分手	6	0	2	2	30.933
5	典禮	13	0.001	2	1	27.258
6	時	656	0.047	2	1	11.18
7	的	21,143	1.524	5	5	5.274
8	找	453	0.033	1	1	4.939
9	了	9,460	0.682	3	3	4.402
10	兩	875	0.063	1	1	3.68
11	那	1,169	0.084	1	1	3.141

圖 6：「畢業」在教材中使用的搭配詞彙

4.2.6 例句編輯輔助系統

教科書中使用太難的詞彙會降低課文內容的可讀性，進而影響學生吸收新知的效率(Lively and Pressey 1923)。因為語言學習者所認識的詞彙量有限，過多的新詞會造成文句理解的困難。同樣的，在辭典的釋義及例句編輯也必須要考慮到學生在語文方面的能力。對辭典或教材的編輯者來說，要考慮學生的語言能力是一個極大的挑戰。首先編輯者在編寫釋義或例句時，必須掌握語言能力分級的知識，哪些字、詞、語法是目前學習者所熟悉的，才能發現其中難以閱讀及理解的部分。其次，編輯者除了必須熟悉解釋義及例句編寫的方法，或必須顧及同義詞代換的知識，才能將釋義或例句中太難的詞語及語法修改成合適的內容。

例句編輯輔助系統所使用的技術包括中文自動分詞與標記的技術，目標為提供編輯者詞語的分級知識，當編輯者將編好的釋義、例句或文章輸入到系統時，系統即時進行斷詞及標記，將每個詞語標出分級標準的等級。例句編輯輔助系統在華語文教學的應用上，主要應用在教材課文以及例句的編寫，幫助編輯者可以了解課文內容的整體難易度，以了解課文的難度是否與設定的學習者相符。而個別語詞難度的級別標記，使編輯者直接看出哪些詞

語的難度較高，並針對太難的詞彙進行替換。除了應用在教材課文以及例句的編寫外，本系統也可適用於在詞典釋義以及例句的編寫。

4.2.7 語料庫覆蓋率統計系統⁶

語料庫覆蓋率統計系統依據語料庫覆蓋率公式，提供線上互動式字表及詞表的覆蓋率計算，訊息涵蓋圖書、報紙、口語、教材等數種語料的累計覆蓋率，以提供字詞表的分級評估。一般的詞頻訊息可以提供詞語是否常用的訊息，但是無法得知這些詞彙是否夠用，覆蓋率則提供了更明確的訊息。覆蓋率訊息可以和能力指標配合，成為語言標準體系的制定標準。每一個等級的學習者該具備多少的詞彙量都能有一個客觀的標準。

4.2.8 作文錯別字自動批改系統⁷

作文錯別字的批改是一件耗費人力的工作，然而，語言學習者用錯別字的機率很高，在沒有人批改的情況下，很難改正錯字的問題。作文錯別字自動批改系統必須使用語言模型來判斷文章內容是否合理。而此語言模型的建立必須使用數量龐大，而幾乎沒有錯別字的高品質語料庫來建置。國教院所建置的臺灣華語文語料庫，在品質與數量上，皆符合語言模型的需求。因此，作文錯別字自動批改系統採用臺灣華語文語料庫中的書面語語料作為語言模型之語料來源，並應用 Hsieh、Bai、Huang 與 Chen (2015) 所提出的理論為基礎所開發，開發的目的為提供學習者一個錯別字自動批改的工具以減少錯別字，經過實際的驗證，錯別字批改的正確率可以達到 78%。

5. 結論與建議

華語文語料庫在華語文教育的應用包括華語文漢字、詞語及語法點分級標準的研發建置，及整合應用語料庫、語料庫科技及華語文分級標準研發了多套華語文教學應用系統。

在華語文分級標準的研發，我們應用華語文語料庫書面語、口語及華英雙語語料的詞頻、覆蓋率、分布均勻度、類詞綴、語義場關聯詞、構詞率及組字力等的統計分析結果，輔以學者專家和資深華語文教師諮詢，完成 1-7 級 3,100 個漢字、14,267 個詞語和 73 個詞綴的分級，及 1-5 級 520 個語法點

⁶ 語料庫覆蓋率統計系統 (<https://coct.naer.edu.tw/tools>)。

⁷ 作文錯別字自動批改系統 (<https://coct.naer.edu.tw/spcheck>)。

的分級。

在華語文教學應用系統的研發，建置了語料庫索引典系統、華英雙語索引典系統、華語中介語索引典系統、語義場關聯詞查詢系統、教材分級檢索系統、例句編輯輔助系統、語料庫覆蓋率統計系統及作文錯別字自動批改系統。而且，建置華語文語料庫與標準體系整合應用系統整合入口（網址為 <https://coct.naer.edu.tw/>），提供更方便的應用這些系統在華語文教學及研究。

目前，國內學者使用這些系統的相關研究，有國內碩博士論文 4 本，期刊論文 3 篇。未來，亦可應用在編撰華語文教材、文法教材、學習者辭典、易混淆字詞表，也可應用在華文例句及試題自動產生、華文拼字及文法檢查、易讀性自動分級、華文作文評分、E-learning 學習平台等華語文教學及學習工具。同時，計畫所建置的臺灣華語文語料庫及華語文漢字、詞語及語法點分級標準也可運用在人工智慧開發及應用上，成為新的華語文教育資源（柯華葳等 2013, 2014；林慶隆、柯華葳、吳鑑城、白明弘、陳茹玲 2019）。

最後，建議持續將華語文語料庫應用在通用詞頻表的建置、基礎詞彙表的建構、商務及觀光等專業領域詞彙蒐集研究、華語文搭配詞結構分析、及詞類標記對應之研究等的研究發展，並且研發各類教學應用系統。

誌謝：本文係教育部補助國家教育研究院華語文八年計畫「建置應用語料庫及標準體系」102 年～107 年工作計畫期末報告之部分研究成果改寫。參與人員包括歷年主持人：許添明、林慶隆、柯華葳、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍、郭工賓；專家學者及教師：方麗娜、王萸芳、李明懿、周美慧、屈承熹、信世昌、張麗麗、曹逢甫、陳立元、陳克健、黃沛榮、劉昭麟、劉顯親、蔡宜妮、鄧守信、鄭錦全、戴浩一、謝佳玲、謝舒凱、蘇以文、孫懿芬、張黛琪、陳懷萱、楊尤媛、盧翠英（分別依姓氏筆畫排列）；研究人員：吳鑑城、白明弘、陳茹玲、李詩敏、吳欣儒；計畫參與人員：黃淑齡、丁彥平、張玳維、余昱瑩、陳威佑、林佳錡、張洪瑄。本文作者感謝教育部經費補助與其他參與人員的貢獻。另外，感謝蔡旻穎小姐協助文章校訂。

引用文獻

Al-Yahya, Maha, Luluh Aldhubayi, and Sawsan Al-Malak. 2014. A pattern-based approach to semantic relation extraction using a seed ontology. Paper

- presented at *the 2014 IEEE International Conference on Semantic Computing*, June 16, 2014. CA: Newport Beach.
- Alshawhi, Hiyan, and David Carter. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics* 20.4: 635-648.
- Bai, Ming-hong, Jian-cheng Wu, Ying-ni Chien, Shu-ling Huang, and Ching-lung Lin. 2016. A study on dispersion measures for core vocabulary compilation. *Computational Linguistics and Chinese Language Processing* 21.2: 1-18.
- Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI). *Computer Studies in the Humanities and Verbal Behavior* 3.2: 61-65.
- Christ, Oliver. 1994. A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX '94 3rd Conference on Computational Lexicography and Text Research*, 23-32. Budapest, Hungary.
- Church, Kenneth W., and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16.1: 22-29.
- Curran, James R., and Marc Moens. 2002. Improvements in automatic thesaurus extraction. *Proceeding of the ACL-02 Workshop on Unsupervised Lexical Acquisition-Volume 9*, 59-66. Philadelphia, PA.
- Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41.6: 391-407.
- Grishman, Ralph, and John Sterling. 1994. Generalizing automatically generated selectional patterns. *Proceeding of the 15th Conference on Computational Linguistics-Volume 2*, 742-747. Kyoto, Japan.
- Hardie, Andrew. 2012. CQPweb - Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17.3: 380-409.
- Hsieh, Yu-ming, Ming-hong Bai, Shu-ling Huang, and Keh-jiann Chen. 2015. Correcting Chinese spelling errors with word lattice decoding. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 14.4: 18.
- Huang, Chu-ren, Hua-rui Zhang, and Shi-wen Yu. 2005. On predicting and

- verifying a basic lexicon: Proposals inspired by distributional consistency. *POLA Forever: Festschrift in Honor of Professor William SY. Wang on His 70th Birthday*, eds. by Dah-an Ho, and Ovid J. L. Tzeng, 57-69. Taipei: Institute of Linguistics, Academia Sinica.
- Juilland, Alphonse, and Eugenio Chang-Rodríguez. 1964. *Frequency Dictionary of Spanish Words*. The Hague: Mouton.
- Kilgarriff, Adam, and David Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography. *Proceedings of the Workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation"* 39th ACL & 10th EACL, 32-38. Toulouse, France.
- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 171-180. Prague, Czech Republic.
- Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL-1998)*, 768-774. Montreal, Canada.
- Lively, Bertha A., and Sidney L. Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision* 9: 389-398.
- Luhn, Hans P. 1966. Keyword-in-context index for technical literature (KWIC Index). *Readings in Automatic Language Processing*, ed. by David G. Hays, 159-167. New York: American Elsevier.
- Ma, Xiaoyi. 2006. Champollion: A robust parallel text sentence aligner. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 489-492. Genova, Italy.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of the*

- International Conference on Learning Representations (ICLR)*. Scottsdale, Arizona, USA.
- Pecina, Pavel. 2005. An extensive empirical study of collocation extraction methods. *Proceedings of the ACL Student Research Workshop*, 13-18. Ann Arbor, Michigan
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. Doha, QA.
- Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de Linguistique Appliquée* 1: 103.
- Ruge, Gerda. 1992. Experiments on linguistically-based term associations. *Information Processing & Management* 28.3: 317-332.
- Salton, Gerard. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.
- Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19.1: 143-177.
- Summers, Della. 1996. Corpus lexicography - The importance of representativeness in relation to frequency. *Longman Language Review* 3: 6-9.
- Turian, Joseph, Lev Ratinov, and Yoshua Bengio 2010. Word representations: A simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384-394. Uppsala, Sweden.
- van der Plas, Lonneke, and Gosse Bouma. 2004. Syntactic contexts for finding semantically related words. *Proceedings of the Meeting of Computational Linguistics in the Netherlands (CLIN)*. Leiden, Netherlands.
- van der Plas, Lonneke, and Jörg Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 866-873. Sydney, Australia.
- Verlinde, Serge, and Thierry Selva. 2009. Corpus-based versus intuition-based

- lexicography: Defining a word list for a French learners' dictionary. *Proceedings of the Corpus Linguistics 2001 conference*, 594-598. Lancaster, UK: Lancaster University.
- Wang, Chang, Liang-liang Cao, and Bo-wen Zhou. 2015. Medical synonym extraction with concept space models. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 989-995. Buenos Aires, Argentina.
- Wang, Wenbo, Christopher Thomas, Amit Sheth, and Victor Chan. 2010. Pattern-based synonym and antonym extraction. *Proceedings of the 48th Annual Southeast Regional Conference*, 64. New York, NY: ACM.
- 尹海良。2007。《現代漢語類詞綴研究》。山東：山東大學博士論文。[Yin, Hai-liang. 2007. *Study on the Quasi-affix of Modern Chinese*. Shandong: Shandong University Ph. D. dissertation.]
- 李詩敏、白明弘、吳鑑城、黃淑齡、林慶隆。2016。〈中文近義詞的偵測與判別〉，《第 28 屆自然語言與語音處理研討會論文集》，342-351。[Li, Shih-min, Ming-hong Bai, Jian-cheng Wu, Shu-ling Huang, and Ching-lung Lin. 2016. Detection and discrimination of Chinese near-synonyms. *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing ROCLING XXVIII (2016)*, 342-351.]
- 呂叔湘。1979。《漢語語法分析問題》。北京：商務印書館。[Lu, Shu-xiang. 1979. *Problems in the grammatical analysis of Chinese*. Beijing: The Commercial Press.]
- 吳鑑城、白明弘、林慶隆。2019。〈臺灣華語文語料庫建構與應用〉，《第二屆中研語言學論壇》。臺北：中央研究院語言學研究所。[Wu, Jian-cheng, Ming-hong Bai, and Ching-lung Lin. 2019. Construction and applications of the corpus of contemporary Taiwanese Mandarin. *The 2nd ILAS Annual Linguistics Forum*. Taipei: Institute of Linguistics, Academia Sinica.]
- 沈光浩。2014。〈漢語新興類詞綴研究綜述〉，《襄陽職業技術學院學報》，第 13 卷第 2 期，77-79。[Shen, Guang-hao. 2014. Review of Chinese new quasi-affixes research. *Journal of Xiangyang Vocational and Technical College* 13.2: 77-79.]
- 林慶隆、柯華蕙、吳鑑城、白明弘、陳茹玲。2019。《建置應用語料庫及標準

體系》期末研究報告。國家教育研究院研究計畫成果報告。新北市：國家教育研究院。[Ching-lung Lin, Hwa-wei Ko, Jian-cheng Wu, Ming-hong Bai, and Ju-ling Chen. 2019. *The Research Report of The Construction and Application of Mandarin Chinese Corpus and Standard Systems*. New Taipei City: National Academy for Educational Research.]

柯華葳、方麗娜、林慶隆、信世昌、范信賢、高照明、張俊盛、張郁雯、陳浩然、蔡雅薰。2013。《華語文八年計畫「建置應用語料庫及標準體系」102 年工作計畫期末報告》。臺北：國家教育研究院。[Ko, Hwa-wei, Li-na Fang, Ching-lung Lin, Shih-chang Hsin, Hsin-hsien Fan, Zhao-ming Gao, Jason S. Chang, Yu-wen Chang, Howard Hao-jan Chen, and Ya-hsun Tsai. 2013. *The 2013 Research Report of The 8-Year Project of Construction and Application of Mandarin Chinese Corpus and Standard Systems*. Taipei: National Academy for Educational Research.]

柯華葳、林慶隆、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍、范信賢。2014。《華語文八年計畫「建置應用語料庫及標準體系」103 年工作計畫期末報告》。臺北：國家教育研究院。[Ko, Hwa-wei, Ching-lung Lin, Jason S. Chang, Howard Hao-jan Chen, Zhao-ming Gao, Ya-hsun Tsai, Yu-wen Chang, Po-hsi Chen, Li-ping Chang, and Hsin-hsien Fan. 2014. *The 2014 Research Report of The 8-Year Project of Construction and Application of Mandarin Chinese Corpus and Standard Systems*. Taipei: National Academy for Educational Research.]

柯華葳、林慶隆、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍。2015。《華語文八年計畫「建置應用語料庫及標準體系」104 年工作計畫期末報告》。臺北：國家教育研究院。[Ko, Hwa-wei, Ching-lung Lin, Jason S. Chang, Howard Hao-jan Chen, Zhao-ming Gao, Ya-hsun Tsai, Yu-wen Chang, Po-hsi Chen, and Li-ping Chang. 2015. *The 2015 Research Report of The 8-Year Project of Construction and Application of Mandarin Chinese Corpus and Standard Systems*. Taipei: National Academy for Educational Research.]

柯華葳、林慶隆、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍。2016。《華語文八年計畫「建置應用語料庫及標準體系」105 年工作計畫期末報告》。臺北：國家教育研究院。[Ko, Hwa-wei, Ching-lung Lin,

- Jason S. Chang, Howard Hao-jan Chen, Zhao-ming Gao, Ya-hsun Tsai, Yu-wen Chang, Po-hsi Chen, and Li-ping Chang. 2016. *The 2016 Research Report of The 8-Year Project of Construction and Application of Mandarin Chinese Corpus and Standard Systems*. Taipei: National Academy for Educational Research.]
- 許添明、林慶隆、柯華葳、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍。2017。《華語文八年計畫「建置應用語料庫及標準體系」106 年工作計畫期末報告》。臺北：國家教育研究院。[Sheu, Tian-ming, Ching-lung Lin, Hwa-wei Ko, Jason S. Chang, Howard Hao-jan Chen, Zhao-ming Gao, Ya-hsun Tsai, Yu-wen Chang, Po-hsi Chen, and Li-ping Chang. 2017. *The 2017 Research Report of The 8-Year Project of Construction and Application of Mandarin Chinese Corpus and Standard Systems*. Taipei: National Academy for Educational Research.]
- 郭工賓、林慶隆、柯華葳、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍。2018。《華語文八年計畫「建置應用語料庫及標準體系」107 年工作計畫期末報告》。臺北：國家教育研究院。[Kau, Gung-bin, Ching-lung Lin, Hwa-wei Ko, Jason S. Chang, Howard Hao-jan Chen, Zhao-ming Gao, Ya-hsun Tsai, Yu-wen Chang, Po-hsi Chen, and Li-ping Chang. 2018. *The 2018 Research Report of The 8-Year Project of Construction and Application of Mandarin Chinese Corpus and Standard Systems*. Taipei: National Academy for Educational Research.]
- 張郁雯。2003。〈詞彙分級研究〉，《華語文能力測驗編製：研究與實務》，柯華葳（主編），83-102。臺北：遠流。[Chang, Yu-wen. 2003. The study of vocabulary grading. *Chinese Language Proficiency Test Preparation: Research and Practice*, ed. by Hwa-wei Ko, 83-102. Taipei: Yuan Liou.]
- 張莉萍。2017。〈TOCFL 學習者語料庫的偏誤標記〉，《語料庫與華語教學》，陳浩然（主編），159-196。臺北：高等教育出版社。[Chang, Li-ping. 2017. TOCFL xue xi zhe yu liao ku de pian wu biao ji. *Corpus and Teaching Chinese as a Second Language*, ed. by Howard Hao-jan Chen, 159-196. Taipei: Higher Education Publishing.]
- 張莉萍、陳鳳儀。2006。〈華語詞彙分級初探〉，《第六屆漢語詞彙語義學研討會》。新加坡：中文與東方語文信息處理學會。[Chang, Li-ping, and Feng-yi

- Chen. 2006. A preliminary approach to grading vocabulary of Chinese as second language. *Proceeding of 6th Chinese Lexical Semantics Workshop (CLSW-6)*. Singapore: Chinese and Oriental Languages Information Processing Society.]
- 張麗麗、陳克健、黃居仁。2000。〈漢語動詞詞彙語意分析：表達模式與研究方法〉，《中文計算語言學期刊》，第 5 卷第 1 期，1-18。[Chang, Li-li, Keh-jian Chen, and Chu-ren Huang. 2000. A lexical-semantic analysis of Mandarin Chinese verbs: Representation and methodology. *International Journal of Computational Linguistics and Chinese Language Processing* 5.1: 1-18.]
- 湯廷池。2014。《華語詞法研究入門(上)》。臺北：致良。[Tang, Ting-chi. 2014. *An Introduction to Linguistic Analysis of Chinese Vol. I*. Taipei: Jhih Liang Press.]
- 黃淑齡、白明弘、吳鑑城、李詩敏、林慶隆。2018。〈國家教育研究院華語字表與其他字表比較研究〉，《華語文教學研究》，第 15 卷第 3 期，85-126。[Huang, Shu-ling, Ming-hong Bai, Jian-cheng Wu, Shih-min Li, and Ching-lung Lin. 2018. A comparative study of the NAER Chinese character list and other Chinese character lists. *Journal of Chinese Language Teaching* 15.3: 85-126.]
- 鄭昭明。1997。〈漢語水平考試的定位、編製及「字彙」與「詞彙」使用的問題〉，《華文世界》，第 85 期，42-47。[Cheng, Chao-ming. 1997. Han yu shui ping kao shi de ding wei, bian zhi ji “zi hui” yu “ci hui” shi yong de wen ti. *The World of Chinese Language* 85: 42-47.]

[審查：2019.7.1 修改：2019.7.29 接受：2019.8.29]

華語文教學研究

吳鑑城

Jian-Cheng WU

10644 臺北市和平東路一段 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

wujc@mail.naer.edu.tw

白明弘

Ming-Hong BAI

10644 臺北市和平東路一段 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

林慶隆

Ching-Lung LIN

10644 臺北市和平東路一段 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

cllin@mail.naer.edu.tw

Applying the Corpus of Contemporary Taiwanese Mandarin in Teaching Chinese as a Second Language

Jian-Cheng WU Ming-Hong BAI Ching-Lung LIN

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

Abstract

The main reason for the National Academy for Educational Research to construct the Corpus of Contemporary Taiwanese Mandarin (COCT) is to make sure a comprehensive applications for Teaching Chinese as a Second Language (TCSL). The COCT includes corpora taken from written language, spoken language, bilingual Chinese-English and Chinese learners' interlanguage. This paper aims to explore the application of the COCT in establishing difficulty levels of Chinese characters, words, and grammar for TCSL, and the development of corpus techniques in TCSL with standard system integration.

After conducting statistical analyses of lexical frequency, coverage, distribution uniformity, affixes, semantic-field-related words, character and word formation rates from the COCT, as well as consulting with experts and senior TCSL teachers, the researchers have been able to establish a standard for the classification of Chinese characters, words, and grammatical patterns. Furthermore, a NAER concordance system, a Semantic-field-related word query system, a Writing typos automatic correction system and an Example sentences editing-assistance system were completed by integrating the standard system and corpus techniques.

Finally, this paper puts forward some suggestions for the future use of the COCT in the construction of a common-word frequency table, a basic vocabulary table, and the analysis of the Chinese collocation structure.

Keywords: COCT, corpus of contemporary Taiwanese Mandarin, teaching Chinese as a second language, teaching Chinese as a second language classification standards