以對比語料庫為本之「詞語搭配」研究[1]

盧慧娟/ Hui-Chuan Lu
成功大學外文系 副教授
Department of Foreign Languages & Literature, Cheng Kung University

【摘要】

本研究主要以史密斯詞彙分析工具中詞彙頻率表、關係詞與索引三功能來分析「成功大學西班牙語學習者語料庫」(CATE-NCKU)中 81 篇三年級學生作文書寫形式中常用詞彙和詞語搭配組合的模式類型與分佈傾向，並進一步和西班牙語自然語料(CLE)做異同的對比分析和檢定，根據研究結果做出教學建議，以作為教學者教案設計之參考。

【Abstract】

This research is based on the three main functions, Word List, Key Words and Concord in WordSmith Tools to analyze the pattern and distribution tendency of the frequently used vocabulary and collocation on the writing of 81 junior compositions in CATE-NCKU (Taiwanese Learner Corpus of Spanish-National Cheng Kung University). Furthermore, this paper compares and examines the differences and similarities with data compiled in the natural language corpus CLE (Spanish Corpus), and makes suggestions based on the research result to provide the reference of teaching design for related fields.

【關鍵詞】

語料庫、詞語搭配、索引、西班牙語學習

【Key words】

corpus, collocation, concordance, Spanish learning

[1] 感謝國科會專題研究計畫(NSC 94-2411-H-006-016)之經費補助；感謝論文審查者所提出的寶貴修改建議。

前言

本論文在語料庫研究法的基礎上，以台灣西語學習者之作文與彙整的西班牙語自然語言爲語料的研究對象，延續過去相關領域、主題的研究，在電腦詞彙分析軟體工具不同功能的輔助下，剖析量化的結果，並進而推論出概化且具代表性的結論。

在語料庫一系列的研究、建構計畫中，我們首先以西班牙語爲母語者建構「西班牙語自然語言語料庫」(CLE: Corpus de Lengua Española)，建構的原因在於現有的西班牙皇家學院(RAE: Real Academia Española)之「現代西班牙語語料庫」(CREA: Corpus de Referencia del Español Actual)雖然有豐富的主題和語料量來源，但對於特殊目的的探討範疇與限定的分析功能則無法滿足進一步更深入的研究需求。未來延伸的語料庫建構計畫還包括以翻譯爲本質的「台灣西漢翻譯平行語料庫」(CPTEC: Corpus Paralelo de Traducción del Español al Chino)。另一方面，除了收集以西班牙語爲母語者的自然語料外，我們則以建構「台灣西語學習者語料庫」(CATE: Corpus de Aprendices Taiwaneses de Español)的母庫爲最終目標。在其初始化階段，我們則分別建構出 13659 字的「成功大學外文系西語組三年級」(CATE-NCKU-3) 和 11277 字的「成功大學外文系西語組」(CATE-NCKU) 作爲一系列子語料庫的基礎。後續的建構計畫還包括「外文系西語組學習者語料庫」(CATE-DLE)和「台灣西文系學習者語料庫」(CATE-DE) 等的子語料庫。預期台灣西語語料庫母庫整體架構的完成將有助於未來理論與應用語言學領域系統化的研究發展。本論文以上述檔案系統格式化的語料庫爲語言分析的來源，結合詞彙分析軟體工具的統計功能應用，來研究詞彙與其延伸之詞語搭配的主題。

近二十年來詞彙教學於第二語教學中的地位漸被重視，其中的詞語搭配、慣用語和短語並列爲固定語彙學(fraseología)研究中的三大分支。根據 LLBA, MLA 和 ERIC 三個語言學資料庫的查詢結果，最早一篇有關「詞語搭配」的文獻記載是發表於 1962 年以字典爲研究對象的研究成果。而近 40 年來詞語搭配研究成果成長速度的倍增[2]則反映出研究趨勢逐漸形成與其地位漸受重視。另一方面，若與每年新增約 17000 筆語言學領域的研究論文成長總篇數比較，則顯示出在詞語

[2] 相關研究文獻的篇數分別從1965年至1974年間的7篇、1975年至1984年間的57篇、1985年至1994年間的123篇，以及最近10年內(1995年到2004年)的240篇論文。

搭配的領域中仍有極大的研究發展空間。

首先，「詞語搭配」(collocation)一詞最早從 Firth (1957)開始界定，具體地說，如西文中的例子 tomar una decisión (Corpas Pastor 1996:20)或 vocero del gobierno (Bolshakov & Miranda-Jiménez 2004:248)等。該詞在其字面與概念定義是指兩個或兩個以上個別的詞彙成分 X 和 Y (以及 Z...)通常會共同出現在句中某種程度距離範圍內的位置，且彼此間有一種互相搭配、依存的狀態與關係，這種關係只存在於 X 和 Y 而不存在於 X 和 A，或是 X 和其他成分 B 之間。本論文從現有環境的可利用資源中，以自然語料庫和學習者語料庫的語言資料爲基礎，關鍵字與詞群在句中的搭配爲語料的研究對象，結合統計學相關性檢定，進行學習者語料庫語詞使用傾向的分析和對比西語語詞結構分析的研究工作。

2. 文獻回顧

本論文的參考文獻涵蓋自然和學習者兩方面不同語料來源，以及包括語料庫和詞語搭配兩大領域研究成果的相關論著。首先，本研究選擇在語料庫的架構中探討詞語搭配的主題乃是由於語料庫之特性及其學理和應用上的研究價值。結合語料庫的自然語言研究是爲了要探視語言的真實面，Woolard 於 Lewis et al. (2000)指出我們可能會發現使用文法規則所創造出來的語言並不存在的事實(林&吳 2002:24)。在實證功能上，Biber et al. (1994:167)則指出使用以語料庫爲本的分析法可以對自然言談的龐大語料進行語言使用模式的實證分析；Kennedy (1998:7)也指出以語料庫爲基礎的研究有助於語言學的描述與分析。此外，語料結合統計法和電腦自動化工具，更可擴大研究分析的範圍與獲取結論的客觀性，如黃居仁(1997)、陳克健(1997)、陳浩然(2000)等文獻所呈現之研究成果。

在以學習者語料爲本和有關學習者語料庫的研究文獻方面，黃麗儀(2000:1)指出透過對語言學習者語料庫的研究分析，可以充分瞭解、掌握語言學習者全面性的語言系統。Chuang (1996:6)也指出以語料庫爲本，透過學習者書寫文本的量化分析結果在語言習得的研究中佔有舉足輕重的地位。有別於傳統對學習成果的分析，學習者語料庫能有效提供一個較具系統性的分析結果。台灣目前現有的一個較大型且帶標記的英語學習者語料庫是由 Shih 所建構，內容涵蓋中山大學、東吳大學和台灣大學約 2000 篇英語學習者的作文 (Shih 2000a:91)。本論文在參

考學習者語料庫的架構經驗與模式中，特別著重語言差異可能衍生的問題，以做爲我們擴建台灣西語學習者語料庫的後續參考。有了學習者語料庫作爲分析基礎，學習成效的分析結果將更具代表性，而學習成效的研究也更具意義。

另一方面，在詞語搭配的研究上，Hargreves 在 Lewis et al. (2000)指出詞語搭配與日遽增的重要性，由越來越多的英語測驗單位認爲該把它包含在考試命題評量的項目中，以達到客觀評量學習者語言程度的事實得以看出（林 & 吳 2002:139)。在詞語搭配主題的研究需求性和學習技能差異上，Shei & Pain (2001:4) 亦曾指出詞語搭配是台灣長榮管理學院學生學習外語時兩大有待加強的課題之一；而 Shei (2002:139)也進一步指出詞語搭配在語言表達方面的困難遠超過理解方面。本論文在詞語搭配的研究主題上，也將以學習者的書寫表達結果爲語料的研究對象，以進行分析。此外，如從語料庫語言學和外語教學結合的現有文獻論著中觀察，則不難發現由於研究主題特性的關係，在構詞方面的分析遠較其他層次的語言學知識的探討來得廣泛，且討論範圍也多與詞語搭配或共現性等功能有關。我們除了根據語料庫研究法中重要的分類(POS: 詞類)爲基本原則外，並參考 Corpas Pastor (1992: 67-75)等西班牙語組合模式的範例以做爲本論文所欲分析之詞語搭配類型的基礎。此外，利用學習者語料庫做爲詞語搭配分析對象的文獻如 Shih (2000b)藉由台灣學習者英語語料庫和大英國家語料庫兩個語料庫的對照，探討台灣英語學習者詞語搭配能力缺乏的問題。本論文即建立在對比學習語和自然語的基礎上進行研究。

在語詞成分間差異的研究方法與相關性統計法的應用上，Biber et al. (1996: 115-116)指出語料庫語言學是透過語言結構的用法、頻率與語境之關係確認，分析出彼此的相關性模式。在詞語搭配方面，從統計導向法的層面討論，在結合語料庫和詞語搭配的研究上，統計機率性提供了研究的附加價值與意義，「詞語搭配」的定義由統計結果的穩定性和顯著差異性可做出較爲科學性的界定。針對本論文的研究主體，我們將採取以互見訊息制式化，如 Shei (2002:148)等所界定的係數（大於 3）來做詞語搭配相關的詞彙分析探討。

最後，在過去一系列語料庫研究課題中，Lu (2005:17)對成功大學外文系三年級 60 篇作文的研究中發現：學習者使用的詞彙和「西班牙皇家學院現代西語語料庫」的自然語料中，兩者前 100 筆高頻資料的比較結果顯示，未達顯著的檢定結果顯示兩者間有同質性。對於同質性的程度探討，Lu & Wu (2005:11-12)進一步比較成功大學外文系一、二、三年級共 70 篇學習者作文語料和皇家學院現

代西班牙語語料庫的整體語料，彼此間有 34%的相似度。另外，在差異性方面，由 WordSmith 軟體工具的「關鍵字」功能所搜尋出達顯著差異的關鍵性詞彙約計有 100 個。有關詞語搭配方面的研究，Lu & Lin (2004:204)針對成功大學外文系三年級 24 篇作文的錯誤分析中發現「動詞＋介系詞」的類型組合是最難學習的。另外，Lu (2005:25)的研究結果則顯示「名詞＋形容詞/形容詞＋名詞」是學習者使用得較多的詞語搭配類型。在本論文中，我們將以過去研究的結果爲基礎，結合更豐富的學習者語料量、主題導向控制的自然語料，以及軟體工具中更嚴謹、精確的功能應用，對研究主題作延續性、更深入的剖析探討。

3. 研究

3.1. 研究目的與問題

本論文之研究目的在於結合語料庫大量語料及統計法之功能檢視，分析學習者語言和自然語言之詞彙和詞語搭配特性之異同。在理論方面，根據台灣西語學習者語料庫來進行詞彙和詞語搭配使用之異同對比分析的研究結果；並以語料庫研究法檢驗、歸納通則與發現特例。

我們的研究問題是：台灣外文系第二外語西語組的學習者，以書寫技能所評估的西班牙語學習成果與西班牙之西班牙語自然語間有何異同？其異同大小程度如何？根據研究問題，所設定的零假設是：學習語和自然語兩者間沒有差異存在。

透過研究問題的探討，我們希望得以瞭解西班牙語學習成果與自然語言間的距離，以提供日後教材規劃之參考。藉由結合台灣西語學習者語料庫詞彙和詞語搭配使用分析探討的研究成果，並以特定主題且涵蓋性廣泛的語料庫爲基礎進行量化之語言對比分析，以瞭解自然語料和學習者語料間普遍語言現象的異同。未來如能有效配合教學法的執行，教學者將因對學習者慣用傾向的掌握、對語言本身的瞭解和分析以及透過兩者之比對，而有助於台灣西語教學品質與學習成效的提昇。

3.2. 研究對象

本論文研究對象的範圍設定在特定主題中，對字彙單詞與搭配詞使用的分析探討，語料的來源主要有兩類型的語料：第一、由成功大學外文系西語組三年級 81 篇作文所建構而成的學習者語料庫（CATE-NCKU-3: Corpus de Aprendices Taiwaneses de Español-National Cheng Kung University-3）；第二、由西班牙皇家學院現代西語語料庫所篩選而來、代表自然語料的西班牙語語料庫(CLE: Corpus de Lengua Española)。

代表學習者西語習得語料的是台灣西語學習者語料庫母庫(CATE)中，成功大學外文系西語組子庫(NCKU)裡 2005 年三年級 27 位學習西文時數達 180 小時左右的參與者，每人 3 篇，分別涵蓋 fiesta, huracán 和 SIDA 三主題，共計 81 篇的作文所收集、彙整而成的語料庫，總字數爲 13659，三主題個別的字數分別爲 3825, 4765, 5069，各佔 28%, 35%和 37%。

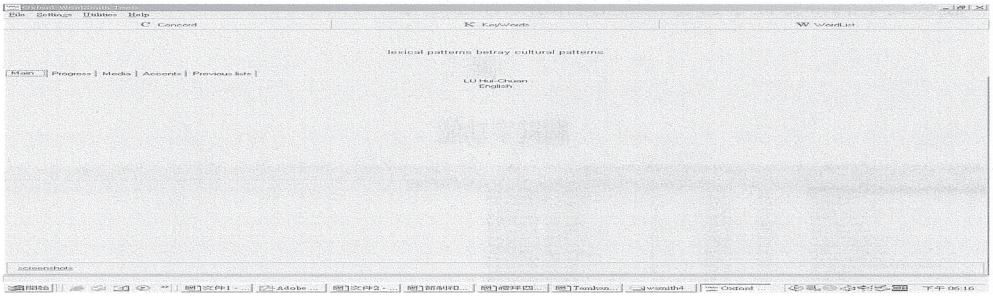
另一方面，代表西班牙語自然語料的西班牙語語料庫(CLE)篩選自西班牙皇家學院(RAE: Real Academia Española)的現代西語語料庫(CREA: Corpus de Referencia del Español Actual)。本論文捨棄現有語料庫而自行建構自然語言語料庫的原因主要是由於主題導向、比例分配和查詢功能的三大考量。選取的原則則是以學習者語料庫的三個主題爲關鍵字，依照上述三主題所佔比例，搜尋到近 8 倍字數量的語料，作爲以西班牙語爲母語者的自然語言語料，字數量分別爲 30438, 38048 和 40222 字，總字數爲 108708 字。

3.3. 研究方法

我們利用 Mike Scott 所設計的史密斯詞彙分析工具(WordSmith Tools) 4.0 版（見圖一）中詞彙表(Word List)、關鍵字(Key Words)和索引(Concord)三個主要的功能，以學習者語料庫和自然語料庫之語料爲文本，進行兩語料庫間單詞和詞語搭配的對比分析。

圖一

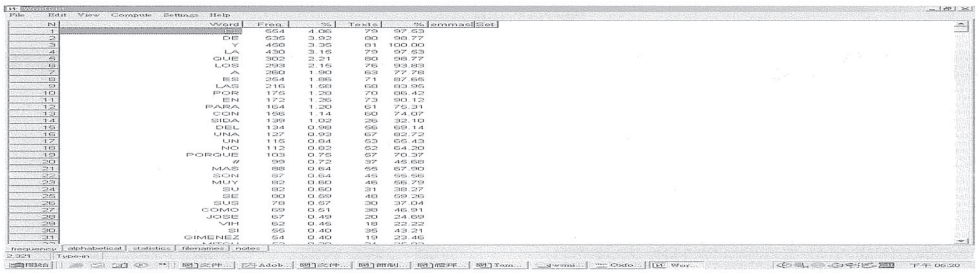
史密斯詞彙分析工具



首先，透過詞彙表（見圖二），我們可以瞭解到個別語料庫中詞彙的類型數(type)，每一個類型字出現的次數和在整個語料庫中所佔的百分比，以及詞彙出現頻率的高低排序。更進一步，我們可以分析高頻詞彙的詞類特性等。此外，我們藉由對比兩個語料庫所分別得出的詞彙表中特定筆數的高頻字來計算兩個語料庫間的相似度。

圖二

詞彙表功能



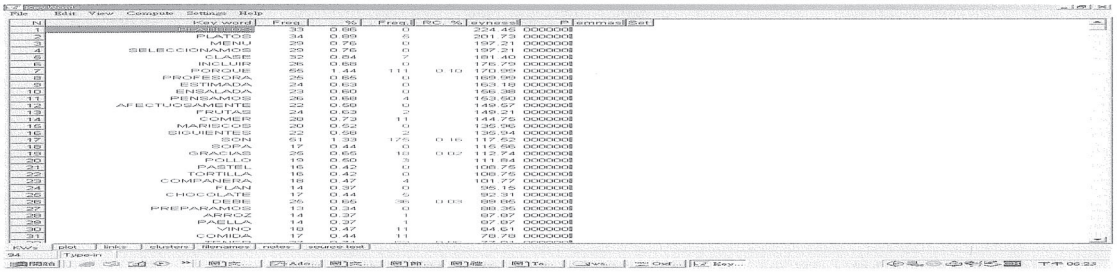
接著，透過關鍵字（見圖三）的功能，以學習者語料庫的詞彙表爲基礎，自然語料庫的詞彙表爲參考的資料庫，透過卡方檢定(P 值設定爲小於、等於0.05)，來獲得兩個語料庫中的關鍵字[3]。其關鍵字主要分兩類：正關鍵性與負關鍵性。正關鍵性的字是指對比自然語言的西班牙語語料庫，在學習者語料庫不正常高頻率出現的字；而負關鍵性則正好相反：是反常低頻率出現的字。我們藉助上述特質來瞭解、分析兩語料庫的差異情形。事實上，不管是正關鍵或負關鍵性，

[3] 此處「關鍵字」的功能有別於一般觀念中所謂的「重要」或高頻的詞彙，其最根本的意義在於透過統計檢定的結果，凸顯兩個被比較的語料庫間統計上的顯著差異。

這些詞彙都是值得教學者特別注意的，如何去拉近兩語料庫間詞彙使用的距離以達到彼此間的對應，是教學上應該努力的方向。詞彙表和關鍵字的功能讓研究者觀察到單詞本身在學習者和母語者間使用的異同。

圖三

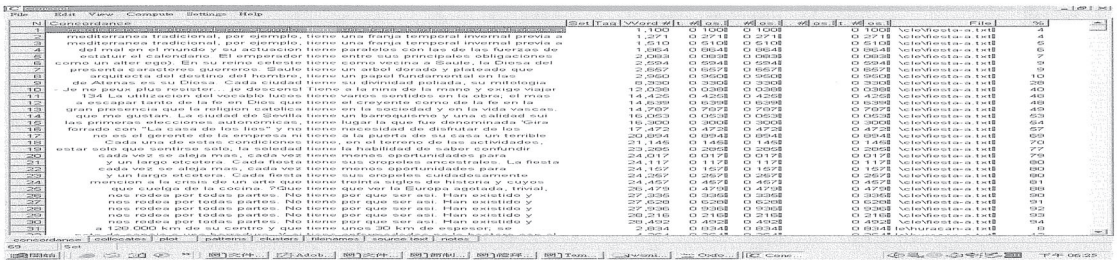
關鍵字功能



最後，以關鍵字表爲基礎，透過索引（見圖四）功能中搭配、模組和群組的不同子功能來分別搜尋兩個語料庫語詞間搭配的分佈傾向，並對比觀察其異同大小。

圖四

索引功能



3.4. 研究結果與討論

3.4.1. 詞彙表

詞彙表的研究結果顯示兩個語料庫不管是在所佔百分比或排序上都不是100%的雷同，但同異程度究竟如何？首先，我們觀察到：雖然兩個語料庫的語料量大小有 7 倍的差距，但仍可由某些詞彙在整個語料庫所佔的比例或頻率排序的階層中發現相當程度上的相似性，如表一所示。

表一

學習者語料和自然語料之詞彙比較

詞彙	El	La	De	Los	Que
學習者語料	188 (4.9%)	178 (4.7%)	149 (3.9%)	140 (3.7%)	94 (2.5%)
排序	1	2	3	4	5
自然語料	3560 (3.2%)	4286 (3.9%)	6903 (6.3%)	1936 (1.8%)	3235 (2.9%)
排序	3	2	1	8	5

在表一中，我們可以觀察到所列出的 5 筆資料中，雖然排序和所佔百分比皆有所不同，但其中有 4 筆是重複的。此外，我們也觀察到頻率較高的筆數多偏屬於閉鎖性的功能性詞彙。我們也注意到：儘管彼此語料量的多寡有異，但在整個語料庫中所佔的比例相仿。例如：在兩個語料庫中排名 3 和 5 的詞彙大概分別佔 3%和 2%左右。如進一步比較詞彙表中前 10 筆的資料：學習者語料前 10 筆的詞彙包括 el, y, la, de, los, que, es, las, con 和 porque；而自然語料前面 10 筆的詞彙爲 de, la, el, en, que, y, a, los, se 和 las，兩語料庫彼此間有 60%的相似度。接著，如比較最前面 50 筆和 100 筆的資料中，則發現分別有 20 筆(佔 40%)和 35 筆（佔 35%）的詞彙重複出現在兩個語料庫中。如果我們繼續比較，則發現學習者語料中前 150 名高頻率的詞彙有 49 筆（佔 32.6%）與自然語料是重複的，再增加至前 200 筆時，兩語料庫間則有 32.5%的相似度。從上述的對比中，我們注意到學習者語料和自然語料相似度的百分比雷同。如此的結果也與 Lu & Wu (2005)中與西班牙皇家學院自然語料相比的結果類似（有 34%的相似度）。上述的研究結果以及比較的基礎源自於語料的實證，有別於傳統教學者之主觀認定。

另外，我們也進一步分析兩個語料庫前 150 筆詞彙的詞性，其結果如表二所示。

表二

兩語料庫前 150 筆高頻詞彙之詞性比較

語料庫	名詞	形容詞	介系詞	動詞	連接詞
學習者語料	57 (38%)	28 (18.7%)	6 (4%)	20 (13.3%)	6 (4%)
自然語料	43 (28.6%)	22 (14.7%)	12 (8%)	15 (10%)	8 (5.3%)

由表二，我們可以觀察到在學習者語料中，名詞、形容詞和動詞屬開放性詞彙所佔的百分比高於自然語料，但屬於閉鎖性詞彙的介系詞和連接詞在學習者語料庫所佔的比例則低於自然語料庫。如此的對比結果讓我們對兩類型語料的差異有進一步的瞭解。

透過上述的比較，其結果皆顯示在提升詞彙相似度以縮短與自然真實語料的距離上，教學者和學習者都還有很大的努力空間。至於哪些是兩個語料庫差異度極高的詞彙，則透過下面關鍵字的功能來進行分析。

3.4.2. 關鍵字

藉由史密斯詞彙分析軟體中關鍵字的功能，我們可以瞭解到學習者和自然兩語料庫間詞彙使用的差異。研究結果顯示關鍵字表共有 149 字，其中有 destruyó, fritas, leche, el, ofrecer 和 que 共 6 筆資料的 P 值大於 0.05，未達顯著水準，故不納入討論。在其他 143 筆呈顯著關係性的關鍵字中，有 136 筆是屬於正關鍵性，例：nombre, pueden, para, deliciosos 等；有 7 筆是呈負關鍵性的，包括 era, había, ha, entre, se, de 和 en。表三列舉關鍵字表中的 7 筆資料說明如下。

表三

兩語料庫之關鍵字

關鍵字	學習者語料庫頻率（%）	自然語料庫頻率（%）	關鍵性	P 值
Es	254 (1.86%)	664 (0.6%)	194.79	0.000
Porque	103 (0.75%)	111 (0.1%)	184.55	0.000
Familia	52 (0.38%)	28 (0.03%)	132.55	0.000
Urgente	7 (0.05%)	0 (0%)	30.89	0.002
Era	4 (0.03%)	202 (0.18%)	-25.35	0.047
Se	80 (0.59%)	1760 (1.59%)	-106.73	0.000
De	535 (3.92%)	6903 (6.25%)	-131.38	0.000

表三中，前面四個關鍵字 es, porque, familia 和 urgente 是屬於正關鍵性（關鍵性之係數爲正）之列，其關鍵性分別爲：194.79, 184.55, 132.55 和 30.89，統計法卡方相關性檢定的結果，其 P 值分別爲：0.000, 0.000, 0.000 和 0.002。上述結果顯示：和自然語料庫相比較（出現頻率所佔百分比分別爲 0.6%, 0.1%, 0.03% 和 0%），學習者語料庫中該四字的使用頻率是異常地高（出現頻率所佔百分比分別爲 1.86%, 0.75%, 0.38%和 0.05%）。也就是說：學習者語料庫過度重複使用了正關鍵性所包含的這一類詞彙。

此外，表三的後三個字 era, se 和 de 是屬於負關鍵性（關鍵性之係數爲負）的字，其關鍵性分別爲：-25.35, -106.73 和-131.38，卡方檢定的 P 值分別爲：0.047, 0.000 和 0.000。亦即：兩個語料庫相比較的結果顯示，學習者語料庫中的使用頻率過低(0.03%對 0.18%, 0.59%對 1.59%和 3.92%對 6.25%)。這意謂著：爲了讓習得的語言成果更接近自然語言，負關鍵性的字是教學者應該強調、學習者應該加強使用的。

再者，對照表一，我們比較 de 這個詞彙在學習者和自然兩個語料庫中個別所佔比例與排序：在兩個語料庫中，de 都是在個別的語料庫中屬於高頻（分別佔 3.9%和 6.3%）且排序甚前（排名分別爲 3 和 1）。另一方面，在表三中，de

是一個負關鍵性的詞彙。換句話說，雖然 de 的使用在學習者語料庫已佔有相當高的比例，但和自然語料相比較仍是偏低的，這也是教學者與學習者在未來需要共同努力的。

接著，我們將針對代表著兩語料庫間的差異，佔學習者語料庫類型總數（2321 個類型字）的 6%、且已達顯著水準的 143 筆關鍵字進行分類與討論。首先，根據詞性分類，負關鍵性的 7 個字，era, ha 和 había 3 個字的詞性是動詞，兩個過去式和一個完成式；en, de 和 entre 3 個是介系詞，se 是代名詞，都是屬於閉鎖式且較繁複的詞彙等級。這幾個詞彙是學習時需要特別被強調與教導的。另一方面，從正關鍵性的 136 個詞彙的詞性分析，其分佈情形如表四所示。

表四

正關鍵性 136 個詞彙之詞性分佈

詞性	名詞	代名詞	形容詞	介系詞	動詞	副詞	連接詞
個數	74	4	20	1	31	4	2
(%)	(54.4%)	(2.9%)	(14.7%)	(0.7%)	(22.8%)	(2.9%)	(1.4%)

從表四，我們可以觀察到正關鍵性的詞彙主要集中在名詞(54.4%)，依次爲動詞(22.8%)和形容詞(14.7%)三大詞類。名詞中除 6 筆專有名詞(SIDA 和 VIH 等)外，其餘的普通名詞的單數型多於複數型，包括 familia, clase, relación, plan, dinero, responsabilidad, casas 等。34 個動詞中有 19 個現在式(es, voy, debe, tiene, gusta 等)，8 個原形動詞(ayudar, comer, tener, usar 等)，而過去式只有一個(causó)；20 筆形容詞，包括 importante, sexual, rica, muchos, urgentes, terrible 等；4 筆代名詞中主要是人稱代名詞(usted, ellos 等)；副詞共有 4 個，分別是 afectuosamente, muy, primero 和 finalmente；連接詞有兩個，分別是 porque 和 y；介系詞只有 para 1 個。這些正關鍵性詞彙的共同特點是簡易導向：多屬於初學者所習得的字彙，且簡多於繁，易多於難。和自然語料相比，這些詞彙有著反常的偏高使用率，要趨近自然語料，則需提昇繁與難的詞彙類型，特別是負關鍵性類型中代詞、過去式或完成式時態的動詞和介系詞片語等層次之詞彙量的使用。

3.4.3. 搭配詞

最後，以關鍵字表中的詞彙爲根本，以索引功能中搭配和模組的子功能查

詢其左右兩邊 5 字內之可能搭配語詞，表五包含一正關鍵性關鍵字 sexuales 和一負關鍵性關鍵字 en 的查詢結果。

表五

索引左右 5 字內之可能搭配於兩語料庫之比較

正關鍵性	左 5	左 4	左 3	左 2	左 1	索引	右 1	右 2	右 3	右 4	右 5
學習	sangre	y	no	tiene	Relaciones	sexuales	con	una	persona	infectada	X
自然	se	higiene	de	los	Relaciones		y	controles	médicos	adecuados	y
負關鍵性											
學習	gracias	por	tener	una	Fiesta	en	la	clase	mi	compañero	y
自然	de	de	de	de	Que		el	de	de	de	de

表五顯示，正關鍵字的 sexuales 在學習者和自然語料中，左邊最鄰近的搭配詞都是 relaciones，但再往左的搭配詞則有所不同：學習者語料由近而遠分別是 tiene, no, y 和 sangre；而自然語料則分別是 los, de, higiene 和 se。再看索引詞右邊的搭配情形，在學習者語料分別是 con, una, persona 和 infectada；而在自然語料則是 y, controles, médicos, adecuados 和 y。彼此的不同代表著學習者語料和自然語料間的差異，要縮短兩者的距離，教學者在瞭解學習者的詞語搭配使用後，必須要以自然語爲學習目標，以自然語的搭配模式作爲未來教學內容設計的根本。

另一方面，在負關鍵性的關鍵字 en 的搭配情形中，學習者語料的模式左邊由遠至近是 gracias, por, tener, una 和 fiesta，而右邊由近至遠分別是 la, clase, mi, compañero 和 y。但在自然語料中，我們則發現左右兩邊 10 個搭配詞中，介系詞 de 佔了 8 個。這說明 en 這個關鍵字的結構用法比學習者慣用的模式複雜甚多，是教學者與學習者要投注更多心力才能夠達到拉近習得語和自然語之間的距離，以落實提昇學習者詞語搭配之能力。

最後，以索引中群組的子功能分析比較學習語和自然語所觀察到的詞語搭配異同。

（一）動詞：(1) es: 在學習者語料庫中有相當比例的 es muy importante 和 lo importante es que 的群組；另一方面，在自然語中除 es importante 外，還可觀察到同樣是無人稱結構的 es posible que 群組。此外，自然語料中由 es importante 延伸的群組中出現最爲頻繁的是 es importante reservar。(2) debe: 在自然語料中，有相當高頻的 se debe a 搭配詞的組合，這是學習者語料中所沒有的。(3) tener: 在學習者語料庫中出現最頻繁的是 gracias por tener una fiesta en；而在自然語料庫中最高頻的是 tener en cuenta 的搭配詞。(4) ayudar: 兩個語料庫僅有的共通點是 ayudar a「動詞＋介係詞」的搭配組合，學習者需要進一步學習此搭配詞所延伸的更大群組。另外，在 ayuda 的索引查詢結果中，雖然兩個語料庫也都有 ayuda a 的詞語搭配，但對比之下，在學習者語料庫中此詞組的出現頻率則遠高過於自然語料庫中的出現頻率。(5) comer: 在兩個語料庫中都出現 de comer「介系詞＋動詞」的搭配詞組合，但在自然語料庫中還出現了另一組頻率頗高的搭配詞 a comer。

（二）名詞：(1) víctimas: 我們發現在學習者語料庫中沒有特別的詞語搭配，但在自然語料中 principales víctimas 的詞語搭配群組的出現頻率頗高。(2) comidas: 在自然語料中有 comidas rituales 的詞語搭配，是學習者語料中所沒有的。

（三）形容詞：sexuales: 兩個語料庫的共通點是 relaciones sexuales 詞組，而這樣詞語搭配的出現頻率也相當高。進一步延伸，在搭配的動詞方面，自然語料多使用 mantener，但學習者則較常使用 tener，這是在自然語料庫中所觀察不到的。

（四）副詞：muy: 兩個語料庫的共通點是有爲數不少的 es muy + 形容詞的搭配詞群組。

（五）連接詞：porque: 在自然語料中出現了爲數不少 bien porque 的詞組搭配，這是學習者語料中所未見的，也是學習者需要學習的。

總結以上的對比分析結果，若在學習者和自然語料庫中皆是高頻出現的搭配詞，則顯示學習者所使用的語彙是較貼近自然語言的；反之，如果在學習者語言中出現頻率高，但在自然語言中卻沒有此現象，則表示學習者需要特別學習以接近自然語言的使用模式。

4. 結論

基於台灣西班牙語搭配詞相關主題研究成果的匱乏，本論文以母語者之自然語料庫和非母語之學習者語料庫之語料和統計方法爲主要的研究基礎，融合現有語料分析之工具、功能與特色，進行差異性與關係性之檢驗，探討西語詞彙和詞語搭配之類型與使用分佈的情形。整體而言，在語言的研究中，利用語料庫爲詞語搭配的搜尋工具不僅可以便利語言分析與研究的過程，其結果也更具客觀性。

透過學習者語料和自然語料之研究比較，其結果顯示在提昇兩語料庫之詞彙相似度和拉近彼此距離的目標上，教學者與學習者都還需要付出更多的努力。另一方面，在兩語料的差異性上，學習者語料和真實的自然語料相比較的結果顯示，學習者語料庫中詞彙使用率偏高的是屬於簡易的字彙類型。因此，如想要接近自然語料，則需擴增繁與難的詞彙類型，特別是像學習者語料中使用率偏低的負關鍵性類型中像代名詞 se、過去式動詞 era, había 和完成式時態動詞 ha 的類型，以及由介系詞 de, en 或 entre 所組成的片語等類似層次的詞彙使用量。在詞語搭配方面，針對在學習者語言中高頻出現，而在自然語言中屬於低頻或甚至沒有出現的字彙，教學者需要特別注重、教導，以接近自然語言的使用模式。兩類語料彼此間的詞彙和詞語搭配差異代表著學習者的習得語言和自然語料間有落差，教學者在瞭解學習者的使用傾向後，必須要以自然語的模式作爲學習的最終目標，遵循其頻率高低做教導先後的排序依據，並搭配適當模組做教材內容的完整規劃。

本論文以語料庫爲基礎，其重要性有二：第一、在學術理論上，語料庫語言學的研究議題從 40 年前開始，其語言學理論和語言習得方面的應用在近 30 年漸被探討，但結合統計學並應用至台灣西語學習者爲對象之研究與台灣學習者語料庫之建立則未見，因此其成果可作爲目前仍未起步的西漢平行語料庫和西漢機器翻譯的重要參考基礎。第二、詞語搭配的正確選擇與表達是中、進階西語學習者成功掌握西語不可或缺的一環，本論文透過語料庫探究學習語和自然語語言本身，系統性的歸納以協助學習者有效提昇程度，以期在語言教學的實際應用上發揮功能。此外，其成果將可提供西文學術與教學界學習者語料庫之資源共享，以及科技與統計化研究結果應用於外語教學研究等實際經驗之參考依據。

本論文在語料庫結合詞語搭配主題的延伸研究上，在未來希望藉由「台灣西語學習者語料庫」逐年擴增的語料量和來源的多樣化讓研究結果更具普遍與代表性。此外，除針對詞語搭配中不容忽視的語言文化層面做進一步的探究外，同時也希望從台灣第一外語英語和第二外語西語的詞語搭配學習成果進行比較，以對學習之可能變因做更深入之剖析。

本論文於 2006 年 10 月 11 日通過審查。

參考書目

陳浩然 (2000)，〈建構英語爲外語學習者的語料庫與第二語詞彙習得研究〉，《行政院國家科學委員會專題研究計畫摘要》(NSC89-2411-H019-005)。

陳克健 (1997)，〈以詞彙爲中心、語料庫爲本的國語語法與語意研究--國語構詞律及詞組結構的自動抽取與分析(I)〉，《行政院國家科學委員會專題研究計畫摘要》(NSC87-2411-H001-041-M7)。

黃麗儀 (2000)，〈學習者語料庫與台灣大學生之英語口語教學〉，《行政院國家科學委員會專題研究計畫摘要》(NSC89-2411-H004-039)。

黃居仁 (1997)，〈現代漢語功能詞之用法與分佈研究--詞彙句法特性之建立(I)〉，《行政院國家科學委員會專題研究計畫摘要》(NSC87-2411-H001-040-M7)。

Biber, D., S. Conrad & R. Reppen. (1994) "Corpus-Based Approaches to Issues in Applied Linguistics" *Applied Linguistics*, 15.2: 169-89.

---. (1996) "Corpus-Based Investigations of Language Use" *Annual Review of Applied Linguistics*, 16: 115-36.

Bolshakov, I. A. & S. Miranda-Jiménez. (2004) "A Small System Storing Spanish Collocations" *Computational Linguistics and Intelligent Text Processing: 5th International Conference, CICLing 2004 Seoul, Korea. February 15-21, 2004 Proceedings*, Ed. A. Gelbukh. Seoul: Springer-Verlag Heidelberg, 248-252.

Chuang, Y. (1996) *Corpus Analysis of the Vocabulary in the Junior and Senior High School Students' English Textbooks and Writings in Taiwan*, Taipei: Crane.

Corpas Pastor, G.. (1992) "Tratamiento de las Colocaciones del Tipo A+S/S+A en Diccionarios Bilingües y Monolingües (Español-Inglés)" *EURALEX´90. Actas del IV Congreso Internacional*, 331-340.

---. (1996) *Manual de Fraselogía Española*, Madrid: Gredos.

Firth, J.R. (1957) "The Technique of Semantics" In *Papers in Linguistics 1934-1951*, London: Oxford University Press.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*, New York: Addison Wesley Longman.

Lewis, M. et al. (2000) *Teaching Collocation: Further Development in the Lexical Approach（英語詞語搭配教學）*, Hove: Language Teaching Publications, 2000, 中譯本：林文婷&吳佩詩譯，台北：文鶴。

Lu, Hui-Chuan & Lieu-Tsun Lin. (2004) "Estudio de Colocación: Aplicación de Corpus en la Enseñanza" *Séptimo Congreso de Didáctica del Español en la República de China*, Taipei: Tamkang University, 198-214.

Lu, Hui-Chuan. (2005) "Análisis, Mediante Corpus, de la Colocación Utilizada por los Aprendices" 87th AATSP Annual Conference, New York, U.S.A.

Lu, Hui-Chuan & Szu-Chia Wu. (2005) "Estudio del Léxico a Partir del Contraste de Dos Corpus: CATE y CLE" 4th International Contrastive Linguistics Conference. Santiago de Compostela, Spain.

Shei, C.-C. and H. Pain. (2001) "Learning a Foreign Language Through Machine Translation: Focusing on Sentence Stems and Collocations" AI-ED Workshop on CALL: Implementing Intelligent Language Tutoring Systems, www.swan.ac.uk/cals/staff/shei/publication/MTandEFL.htm [2005/11/25]

Shei, C.-C. (2002) "On the Issue of Collocation in Chinese-to-English Translation" *Journal of Chang Rong* 5.2: 135-149.
http://www.swan.ac.uk/cals/staff/shei/publication/collocation_translation.htm

Shih, Hsue-Hueh. (2000a) "Compiling Taiwanese Learner Corpus of English" *International Journal of Computational Linguistics & Chinese Language Processing*. 5.2: 89-102.

Shih, Hsue-Hueh. (2000b) "Collocation Deficiency in a Learner Corpus of English" The Pacific Asia Conference on Language Information and Computation (PACLIC 14), Tokyo: Logic-Linguistic Society of Japan, 281-288.

CREA de RAE: www.rae.es