國立政治大學資訊科學系
Department of Computer Science
National Chengchi University

碩士論文

Master's Thesis

以型態組合為主的關鍵詞擷取技術在學術寫作字彙上的研究

A Pattern Approach to Keyword Extraction for Academic Writing Vocabulary

研 究 生:邵智捷

指導教授:劉吉軒

中華民國九十九年七月

July 2010

摘要

隨著時間的推移演進，人們瞭解到將知識經驗著作成文獻典籍保存下來供後人研究開發的重要性。時至今日，以英語為主的學術寫作論文成為全世界最主要的研究交流媒介。而對於英語為非母語的研究專家而言，在進行英語學術寫作上常常會遇到用了不適當的字彙或搭配詞導致無法確切的傳達自己的研究成果，或是在表達上過於貧乏的問題，因此英語學術寫作字彙與搭配詞的學習與使用就顯得相當重要。

在本研究中，我們藉由收集大量不同國家以及不同研究領域的學術論文為基礎，建構現實中實際使用的語料庫，並且建立數種詞性標籤型態，使用關鍵詞擷取關鍵詞擷取(Keyword Extraction)技術從中擷取出學術著作中常用的學術寫作字彙候選詞，當作是學術常用寫作字彙之初步結果，隨即將候選詞導入關鍵詞分析的指標形態模型，將候選詞依照指標特徵選出具有代表指標意義的進一步候選詞。

在實驗方面，透過對不同範圍的樣本資料進行篩選，並導入統計上的方法對字彙進行不同領域共通性的分析檢證，再加上輔助篩選的機制後，最後求得名詞和動詞分別在學術寫作中常用的字彙，也以此字彙為基礎，發掘出語料庫中常用的搭配詞組合，提出以英語為外國語的研究學者以及學生在學術寫作上的常用字彙與搭配詞組合作為參考，在學術寫作上能夠提供更多樣性且正確的研究論述的協助。

Abstract

With the evolution over time, people start to know the importance of taking their knowledge and experience into literature texts and preserving them for future research. Until now, academic writing research papers mainly in English become the world's leading communication media all over the world. For those non-native English researchers, they often encounter with the inappropriate vocabularies or collocations which causes them not to pass on their idea accurately or to express their research poorly. As a result, it's very important to know how to learn or to use the correct academic writing in English vocabularies and collocations.

In this study, we constructed the real academic thesis corpus which includes different countries and fields of academic research. The keyword extraction technique based on the several Part-of-Speech tag patterns is used for capturing the common academic writing vocabulary candidates in the academic works to be the initial result of the common vocabulary of academic writing. The candidate words would be introduced to the index analysis model of keyword and be picked out to the further meaningful candidate words according to the index characteristics.

For the experiments, the sample data with different fields would be filtered and the vocabularies on different fields of commonality would be analyzed and verified through statistical methods. Moreover, the auxiliary filter mechanism would also be applied to get the common vocabularies in academic writing with nouns and verbs. Based on these vocabularies, we could discover the common combination with the words in the academic thesis corpus and provide them to the non-native English researchers and students as a reference with the common vocabularies and collocations in academic writing. Hopefully the study could help them to write more rich and correct research papers in the future.

目錄

圖表目錄

第一章

簡介

1.1 背景

歷史一詞是人類自從發明文字以來開始產生的，在此之前的史前時代，人類的生活智慧與經驗教訓都只能依賴口耳相傳而延續下去。當發生了天災人禍時，這些資訊的累積可能就隨著傳承者的死亡而消失，之後又必須一切重新開始。然而文字的出現所刻下的歷史，不但留給下一代生活經驗的基礎，更留給了後人無數智慧結晶演進的脈絡可循。同時從原本記錄媒體的取得不便，使得這些資訊的傳承有著時間與空間上的限制，但進入了資訊時代之後，這些限制迅速的被跨越，人們在隨時隨地都可以讀取同一份文學名著或是研究紀錄，無論是以傳統的紙張形式或是數位化的檔案形式。

隨著文化的進步，人們也意識到知識傳承的重要性，國家或政府藉由學校這個組織將經過知識訓練的人們，透過紀錄著知識的教材以口述或筆記的方式不斷的向之後的世代教育，期待培育出更具有智慧的人材。而正是因為處於這種環繞著大量知識累積的人材與完整的文獻典籍的環境之下，同時在國家也願意大力扶持的立場下，使得學校機構組成的學術界，成為學術研究發展的最佳場所，而其集研究精華於一身的學術著作論文，則是屢屢於世界上改進人們生活的創新產物的重要基石。

處於現在的全球無國界立場，世界上各國都有許多優秀的研究學者，隨時都有可能研發出新的學術理論或是技術更新，期待與更多的學者專家分享，但往往受限於語言的限制無法傳遞給他人理解。另一方面，英美的崛起以及數百年強權的歷史發展結果，英語成為了世界上最通用的語言。無論是貿易通商或是資訊交流都形成了以英語作為主流的趨勢，當然學

術發展也不例外。作為領導地位的美國，在學術上組織了許多重要的學術組織，以英語為主要語言讓全球的知識份子能夠做最大範圍的交流。因此，學術著作以英語為主也就佔了大多數，相對於英語為外國語的研究學者而言，英語學術論文寫作也就成為非常重要的一項專門技術。

為了能在有限度的的文字內清晰的闡明研究內容與所得的成果，學術論文格式上設定了許多的規範。許多學術寫作及書籍都指出，學術寫作本身具有正式(formal)、客觀(impersonal)、精緻(sophisticated)、精確(precise)、簡潔(concise)、專業(specialized)等特色[1]。但對 EFL (English as Foreign Language) 的研究學者或是學生來說，本身缺乏英語本身的語意背景，在字彙的選擇上沒有適當的參考，而導致用字不正確或是意義偏差等問題，進而由字彙所組成的片語等相繼使用了不適合的字可能使語意完全不同，諸如此類問題時常發生。因此做為組成學術論文寫作的最小單位的字彙，能夠如何精準地使用字彙和字彙間組合的而成的搭配詞對於 EFL 的學術作者而言，便成為相當重要的課題。

1.2 研究動機

在資訊發達的現代，專家學者取得研究資訊的管道也從文字期刊移至網際網路上，但對國家研究資源不豐富的研究者而言，即使在網際網路上，可取得的該國語言的研究資源仍有限。在現在研究資源以英語為主的整體環境中，英語為主的研究資源是最為豐富的。從另一角度來說，對於 EFL 背景的研究學者、研究助理、學生以及相關人員，若欲貢獻自己的研究成果給世人得知或是分享給多數人使用，將本國語言的研究論文翻譯成英語版本發表於國際論壇上是最佳的方法。然而鑒於每個研究者本身英語程度不一，產出的論文品質上多少有所差異，而經由他人代為翻譯卻因為每個人不同的文化背景對同樣的文字產生的解釋不同，翻譯後文件容易對作者本身欲闡述的意義失焦，而作者卻無法自行修正，這樣的結果對作者來說無疑是對自己的研究成果打了折扣。

也因為如此，市面上有許多英語寫作專家針對英文論文寫作這部份著作了不少參考書籍，試圖輔助研究者寫出合乎體裁的英語學術論文。但這方面的著書作者往往是英語寫作領域或是某特別領域的專家。以文法來說毋庸置疑可依循專家的建議，但專家選出來的字彙與搭配詞或許因為本身學識豐富，提供的字彙上可能一般不常使用，或是使用的範圍較切進本身研究的領域，有些部份不適用於一般性全領域的學術論文撰寫。故找出頻繁且常用的學術

寫作字彙變成為相當迫切的課題。目的不是要取代專家學者所提供的字彙，而是希望能夠補齊不足之處，進而提供較一般性的常用學術字彙供作者習得，在學術研究上無論是閱讀或寫作都有所助益。

1.3 研究目的與方法

本文希望建構一跨領域的學術論文語料庫，透過自然語言處理的語料庫內容預先處理以及我們採用的形態分析模型，應用在關鍵詞擷取技術的基礎上，建立各領域真實狀況下常用的學術寫作字彙，同時在進而分析建立一般共通領域適用的學術寫作字彙。另外，以這些字彙為基礎，分析出一般常用的搭配詞組合，供 EFL 研究者作為英文學術論文寫作的參考。

語料庫的建構設計理念則是，為了能跨領域交叉分析，我們選擇了 CS(Computer Science)、ELT(English Learning & Teaching)以及 MED(Medical)三個領域的學術論文，MED 領域的選擇是基於其屬於專門集中領域，期望在高度專門領域中以本研究的方法取出的詞彙是由一般性通用的學術寫作字彙組成，故選擇以高度專門領域為主。而在內容選擇的部份，則收集了台灣、日本與美國三個國家的論文作者為主的學術論文，同時為了比較並且強調作者 English as Foreign Language(EFL)作者與 Native Speaker 作者的詞彙使用上的差異，故在美國作者的論文部份，我們以收集博士畢業論文為主，其他部份則是以碩士畢業論文以及期刊論文為中心。

本篇論文是以關鍵詞擷取的方法輔佐以形態分析模型使用進而產生出最終的字彙列表。而關鍵詞擷取研究本身屬於資訊擷取(Information Extraction)的一環，當中有許多文字預先處理部份與資訊擷取相同，包含 Tokenization、Morphological Processing、Syntactic Analysis、Domain Analysis 等預先處理步驟，依研究目的的不同選擇適當的步驟，接著就是屬於關鍵詞擷取技術的範圍。

根據研究[2]指出，在英語中經常同時出現有明確意義的詞彙組合是由三個單字或大於三個單字以上組成的組合，稱為搭配詞(Collocation)，搭配詞中的組合則大多數以名詞和動詞為構成搭配詞意義的核心，其他詞類如介系詞或形容詞副詞等，則是表現關係和修飾其他詞性的角色。在關鍵詞擷取的研究中，由於關鍵詞多數由名詞或是名詞片語(NP, Noun

Phrase)所組成，因此以名詞片語為中心的關鍵詞擷取研究佔了多數。綜合以上的資訊，我們使用了自行定義包含名詞和動詞為主的 PoS Tag Patterns 作為我們關鍵詞擷取的主要擷取型態，並擷取出所有符合 Patterns 的候選詞，作為關鍵詞彙擷取的第一個步驟。

由於語料庫本身的資料龐大，各種領域內的研究主題也相對不小，而 PoS Tag Pattens 在設計上是以搭配詞最小單位為基礎，擷取出的大量 Patterns 必須經由適當頻率的篩選，隨後再將 Patterns 拆為單個字彙的集合，排除非動詞和名詞的其他字彙，而依此兩種詞性分別套用在分析模型上，可求得數種代表不同關鍵詞屬性的指標值。其後以實驗得到的結果，對照於常用於學術教學應用上的學術字彙列表(AWL)[22]，並探討學術寫作應用中的真實狀況下，本文研究結果與學術字彙列表於真實語料庫的分佈狀況，同時歸納出基於研究結果組合而成的學術寫作搭配詞。

1.4 論文架構與貢獻

本篇論文分為五章，第一章說明研究背景、動機、目的及方法。第二章為文獻探討，介紹本篇論文所使用到的相關研究技術的定義與特徵，從語料庫語言學為始，到核心的關鍵詞擷取技術及最後所應用的形態模型等。第三章則是實際語料庫設計以及研究方法的闡述，將分析模型套用在從語料庫以關鍵詞擷取技術得到的成果。第四章為實驗評估與結果討論，將分析模型的各指標實驗結果與作為參考的學術字彙列表在真實語料庫的分佈情形統計性的比較。第五章為結論及未來研究方向，並探討實驗結果可衍生的應用層面。本論文主要貢獻有以下幾點：

I. 藉由真實收集語料庫交叉比對分析而得的學術寫作字彙，不但能補足一般專家著作較缺乏的一般性學術上頻繁使用的字彙，對某些偏重於單一領域或是使用頻率過低的字彙，這些現實中已被應用的字彙能修正其偏差。

II. 由實驗結果所得到的雖然只是常用的學術寫作字彙，但將此資訊重新帶入原本的語料庫，可以得到高頻度且實際使用的搭配詞組合，這些組合結合英語寫作專家所提出的搭配詞相互驗證，除了可信度高之外，也延伸了作者從單一字彙的使用到字彙相互搭配組合的實際應用參照。

III. 語料庫的設計不僅僅只是產出一般性的學術寫作字彙，同時藉由分析各領域間的複合領域字彙使用情形，也能看出即使在同一學術寫作範圍下，各領域之間的學術論文寫作時用字遣詞的差異，這些差異也能提供 EFL 作者未來在寫作上能選擇適合領域的字彙，產出更貼切的論述內容的學習參考。

IV. 除了提供各領域綜合的學術字彙與搭配詞之外，我們也以另一個角度進行分析，分別提出了以 EFL 作者(台灣、日本)以及以英語為母語的 Native Speaker(以下簡稱 NS)，也就是美國作者的常用的搭配詞。提供搭配詞的目的在於，搭配詞的使用較字彙上更為實用也較為繁複，而就搭配詞上的使用狀況可以得知 NS 作者較常用的寫作風格，同時參照自身的搭配詞使用方式，不但可以學習較正確的寫作風格，也可能發現並修正自身潛在的寫作錯誤。

第二章

文獻探討

本篇論文是對真實學術論文語料庫以關鍵詞擷取技術配合指標形態模型分析的方法發掘出一般學術論文寫作常用字彙，而本章節將許多研究技術相關研究文獻提出逐一探討。主要探討內容包含語料庫語言學的特徵與相關應用、關鍵詞擷取技術適用範圍與細節說明，同時介紹以型態分析為主的關鍵詞篩選方法，接著探討字彙與搭配詞的使用在學術寫作上的重要性，並透過英語學習領域專家提出的學術字彙列表跟以資訊技術實驗分析而得的結果相互參照。最後提出本章節的總結。

2.1 語料庫語言學

隨著電腦科技的日漸發達，基於大量計算與統計分析語料庫語言學的相關研究也如雨後春筍般日益崛起，尤其是語料庫語言學對於語言教育與學習的部份有著明顯的增加。本節將對於語料庫語言學的定義、語料庫為主分析研究的特徵，以及在相關領域的應用在此說明。

2.1.1 語料庫以及語料庫語言學的定義與特徵

語料庫語言學是一項透過語料庫以真實發生的範例研究人類所用的自然語言的使用狀況[4]。而語料庫則是一連串由純文字所組成，用來表達其狀態與其多樣性(如口說語言及語

言寫作等)並且可以儲存於電腦內的文字集合[5]。而以語料庫為主的研究隨著各式不同的領域的變化，語料庫為了要能夠符合研究的主題，語料庫的預先設計就顯得格外重要，也因此語料庫設計在檔案尺寸上和表現方式上有著各種的變化[6]。

舉例來說，如探討生態學學術教材為主的語料庫[7]在研究主題是要找出生態學上的學術教材相較於其他學術文章的差異等，在設計上內容選取約 200 篇生態學學術教材文章中不包含每篇文章的第一段以外的所有內文，是為了避免每篇教材的第一段有著概括性的廣泛描述，而脫離以生態學為核心的專門論述。另外一項針對英語學術文章引述(CITATION)的研究[8]，延續了兩個不同原有語料庫的研究，其中一個是由 10 篇跨領域期刊文章所組成，另一個則是由 16 篇博士論文所組成。由上述兩個例子可見，語料庫在內容組成上並無特別的侷限，並不需要完整的保留原有文章的所有內容，而是依照研究目的需要自行定義，相同地在數量選擇上也是如此，只要最後的實驗方法上量足夠於採信即可。

語料庫依據用途建立後，需要經由許多分析的步驟才能達到研究目的。而 Biber, Conrad 與 Reppen[9]提出了以語料庫分析的四項主要特徵，說明如下：

I. 研究者可藉由分析觀察到實際語言文字形態的使用，故此分析是有實證為依歸的。

II. 語料庫分析多半借助 Concordancer 等電腦輔助軟體並可達到 KWIC (Key Word In Context) 顯示的功能，同時也可借助於其他電腦程式進行文法或詞性上的標註和變化指示。

III. 語料庫的內含的語言特徵可同時進行計量性分析與直譯式分析，如使用 Concodancer 可同時顯示”vocabulary“此字彙出現頻率後，進而點選分析此字彙在不同文章中展示的不同意義。

IV. 藉由分析語料庫來模擬探討語言學上的研究問題使得分析本身就是有意義的。

總結來說，語料庫語言學本身由於相關電腦輔助工具的發明，使得語料庫分析不再只是提供規範性的觀點，而能夠提供一種新的描寫性的觀點。

2.1.2 語料庫文字的預先處理與其後續相關應用

本節是以電腦科技的角度出發，探討在資訊科技研究使用語料庫進行研究時經常遇到的預先處理步驟說明，並介紹依照不同的預先處理程度之後可進行的後續研究使用狀況。

上一節曾提到，語料庫設計可依據研究目的需求。相對地，這也代表著語料庫內容的來源可能千變萬化或以各種格式存在。主要的內容來源有兩種，第一種是來自於現有的典籍文獻以及報章雜誌等，這一類的資訊來源共通的問題就是必須將紙本轉換成數位化的檔案，所使用的方式就是經由文件掃描後再經過 OCR(Optical Character Recognition)辨識後文字始能編輯，而依據原始檔案保存或印刷狀況都會影響到文字正確辨識度的多寡，因此辨識完成的文字檔案依狀況仍須人工比對或用電腦字典對照以確保整個語料庫的正確度。

相較於文本各式的內容來源，有許多內容資源都已經數位化成為檔案格式或分佈於網際網路之中，但這些資訊仍然是格式內容不一，必須做過濾格式的預先處理。舉例來說，欲建構以網頁形式為原始內容的語料庫，必須將原始資料中網頁的標籤逐一移除，並且將文字以句子或段落為單位進行分隔，甚至有時也需要解決編碼格式上亂碼的問題。而做完預先處理的文字資料，則是研究方法而存成既定的格式，其中以純文字檔案格式(半結構性)和 XML 資料格式(結構性)為主要格式。

語料庫在建構完成後，除了對語料庫本身進行分析統計外，再經過不同的加工形式(Tokenization、Morphological Processing、Syntactic Analysis、Domain Analysis 等)後，可應用在資訊擷取(包含關鍵詞擷取)、資料探勘、文本探勘、自動翻譯、社會網絡等多種不同的應用分岐。根據王俊弘[14]的研究，建立一個可標記化的語料庫需要八個步驟。然而，根據研究目的的不同預先處理所需的步驟也不同，通用的步驟則包含下列步驟：

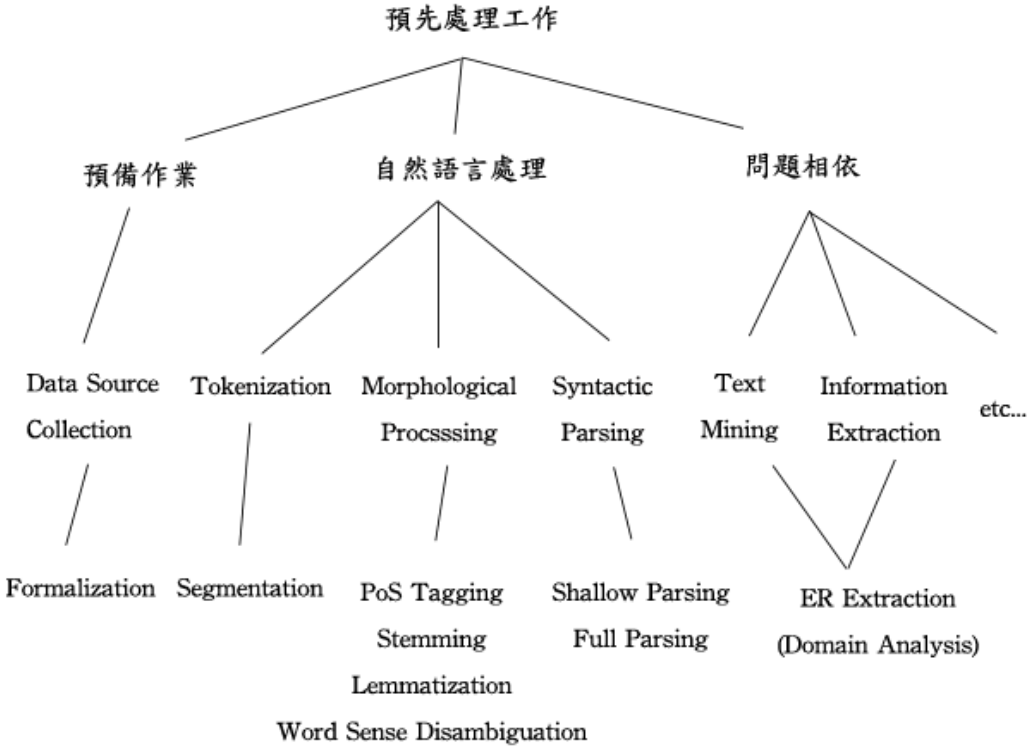

圖 2-1 語料庫與資訊擷取預先處理工作一覽

1. 正規化(Formalization)：由於語料庫內容原始資料來源不盡相同，當中可能包含了標題、副標、圖片與文字格式(如粗體、斜體、底線等)。正規化的目的即為除去這些文件內文以外的其他不需要部份。

2. 斷句(Sentence Segmentation)：資料經過正規化之後的內文，可能還保有原本的形式，文字之間依照段落分隔。一般來說，在自然語言處理中，通常以句子當作一個執行的基本單位，故斷句便是將所有的文字內容依照句點當作區隔其他句子的單位劃分文章內容。而在其他的研究也有依需求將文章內容依片語、段落或是章節區分的狀況[30]。

3. 斷詞(Tokenization)：英語中字與字之間大多與空白分隔，或是依照各種標點符號分隔。斷詞目的在於區分語料庫中最小可供辨識的基本單位"Token"，一般為上述空白或標點區隔的英文單字，也可依據需求將所需的標點符號或特殊符號定義成 token，未被定義成 token 的部份在自然語言處理時會自動被忽略。

4. 詞性標註(Part-of-Speech Tagging)：詞性標註是自然語言處理中最重要的一個步驟，所以後續都分析都以標註後的結果為基礎進行。詞性標註是將句子中每個單

字進行詞性標籤的加註，但有時文章較為複雜也有詞性判斷錯誤的狀況。一般來說，前述步驟在不發生錯誤的狀況下，詞性標註的準確率可達到 95%以上。

總結來說，預先處理的步驟可參照上圖。圖 2.1 為綜合大部份資訊處理相關領域的預先處理步驟，而圖中左半部份也就是上述四個步驟是進行建構語料庫最常見的預先處理步驟，其餘步驟則依研究目的而分別有所不同。

2.2 關鍵詞擷取技術

2.2.1 關鍵詞在學術著作中的定義與特徵

本文曾在第一章時提到，關鍵詞擷取技術是屬於資訊擷取技術的一環，不同於資訊擷取技術的是，關鍵詞擷取技術是將研究範圍縮小集中於對關鍵詞進行擷取的工作。這裡所指的“關鍵詞”本身並不限定於單一字詞，可以由一個單字或是一個片語(數個單字的集合)所構成，而”關鍵詞“一詞，則是依據關鍵詞本身所處的主題範圍有著不同的解釋。在關鍵詞擷取的研究中，關鍵詞的意義是代表與主題領域相關度高，能夠以此一詞作為代表整篇文章或整個領域的詞藻，讓他人能夠看到此關鍵詞便能快速瞭解整篇文章的研究領域或是中心主題，故此詞藻可能就是該文章或領域的專門術語(Terminology)或是較一般性的共通代表辭彙。

同時關鍵詞也具備了一些特性，以在同一篇文章內為例。關鍵詞常有的特性有出現頻率高，或是以同義詞(synonym)或上位字(hyponym)、部份詞(meronym)等形式出現[10]，但都代表同一個意義，或是代表與其它字彙同時出現的比率高[11]等，都是關鍵詞常有的特徵。此外，也有研究指出關鍵詞的組成多數以名詞居多[12]。正因為關鍵詞本身具有這些特性，由這些特徵出發進行關鍵詞擷取的研究也不在少數。而這些研究大致可分為以自然語言處理方法為基礎和以統計分析方法為基礎兩種[13]。將於接下來的章節分別介紹。

2.2.2 基於自然語言處理分析為主的關鍵詞擷取技術

此類的關鍵詞擷取技術，是基於人類學習外來語言的方式，從單字的詞性、字根的變化到片語的構成以至於語意的瞭解，對原始文字資料層層標註，再以電腦分析這些標註，取出有重要代表性意義的詞彙。而本文則以關鍵詞擷取技術常用的標註說明如下：

I. Tokenization(Sentence、Word Segmentation) - 定義原始資料中資訊被處理的最小單位。以英文來說，一般就是以空白分隔的單字視為一個 token，各種特殊標點符號也可被定義為 token。

II. Morphological and Lexical Processing(Part of Speech Tagging 、 Word Sense Disambiguation)

(I) Part of Speech Tagging：詞性標註。舉例來說”design”此字同時可以當作名詞(Noun)以及動詞(Verb)來使用，而”root”這個字可能隨著不同的詞性分別代表不同的意義。詞性標註則是依照前後文已確定的詞性來分別對字彙進行註譯，同時也弭除詞性岐異的問題。

(II) Stemming & Lemmatization：詞幹還原與詞根還原。 Stemming 是將字彙還原成詞幹(root)的形式，而此詞幹可能是完整的單字，也有可能是單字的一部份，而 Lemmatization 則是將大小寫差異、動詞時態、名詞單複數以及形容詞比較級等統一還原成字彙的標準詞根形式[14]。舉例來說，以動詞過去式”waited”為例，Stemming 和 Lemmatization 的結果都同樣為”wait”，但用另外一個動詞過去式”produced”為例時，Stemming 的結果為”produc”而 Lemmatization 的結果為”produce”。

(III)Word Sense Disambiguation：消除歧義。由於詞性標註是以單字為單位，光憑詞性標註有時還是會有無法辨識的問題。而消除歧義則是以相鄰字彙意義與其詞性標籤為基礎，對某些有歧義的字彙判定其意涵，而這方法也是監督式學習的方法(Supervised Learning)。

原始文件經過標註之後，便可進行關鍵詞擷取的步驟。由於關鍵詞多半包含名詞，在 Hulth[15]的研究中，就分別使用了 NP-Chunk、n-gram 以及 PoS Tag Pattern 以名詞為中心的三種方式，來進行初步辭彙取出的單位。NP(Noun Phrase)-Chunk 的概念在於，關鍵詞往往

是具描述性的名詞，這樣的詞是由名詞或是形容詞搭配名詞的組合組成，然而 NP-Chunk 與 NP 的差別在於，NP-Chunk 通常包含的單字量比 NP 少，因為單一 NP-Chunk 無法包含其他 NP-Chunks，而 NP 卻有可能在包含其他較小的 NP(在語法上一個名詞就可以當成 NP，Chunk 則是由名詞組或動詞組等所購成)。

第二種方式 n-gram 方法中的 n，代表的是可變動的正整數，指的是以 n 個單字為一個擷取單位。Hulth 以類似 Turney[16]和 Frank [17] 的研究方法，將所有的 unigram, bigram 以及 trigram 的詞都先擷取出來之後，將這些詞若是頭尾含有 stop word(一些特定的詞，出現頻率相當高而無實質代表意義，因此搜尋引擎等檢索系統並不加以索引，在資訊擷取上也常常被忽略)的詞加以捨棄，最後在對剩下的詞進行 stemming 的處理以提高精確度。最後提出的 PoS Tag Pattern 方法，則是定義各種詞性標籤順序的組合形態，將符合的形態從文中直接抽取出來。Hulth 訂定了共 56 種詞性標籤型態，而最常出現的有 Adjective Noun、Noun Noun 及 Noun 等形態。而無論是使用哪一種方法，選出來的候選詞最後須經過 Machine Learning 的方式，將訓練資料不斷的自我學習才能得到最終的結果。

2.2.3 基於統計分析為主的關鍵詞擷取技術

相對於基於機器理解的方法，以統計分析為主的擷取技術著重於大量統計的資訊如字彙頻率的基礎上。在 Matsuo&Ishizuka 的研究中[11]，作者認為，相較於一般高詞頻的字彙(如 make, kind 等)可能與許多各種字彙在文中公平地共同出現，可能為關鍵詞的高頻詞(如 digital computer, imitation 等)只會與較少特定的字彙共同出現，如此一來，圍關鍵詞的高頻詞與一般的高頻詞在算共同出現比例時就會有所偏差。作者是採用卡方分佈(χ^2- Measure)的統計方法來計算偏差值，而後在對這些可能是關鍵詞的候選詞再進行 Clustering 的分析，以提高方法的可靠度並同時更加凸顯關鍵詞與其它高頻詞的差別。

除了詞頻之外，另一個常用來評估字彙在文章中的重要性的指標 TF-IDF 也常常被用來做關鍵詞擷取的運算。TF 和 IDF 為 Term Frequency 以及 Inverse Document Frequency 的縮寫。TF-IDF 的概念在於，一般來說，詞頻 TF 越高的字可能在該文件的重要性相對較高，但對多份文件構成的資料集來說，在某份文件中詞頻相當高的詞在其他份文件內卻都沒有出現，如此來說，這個詞在整個資料集的分佈頻率便相對降低，因此只用某詞彙的總詞頻來衡量重要性是不夠的，也需要考慮該詞彙在資料集中的文件分佈狀況。IDF 這個指標是代表該

詞彙在資料集之中的逆向文件分佈頻率，兩者相乘便能代表一個詞語普遍重要性的衡量標準。TF-IDF 公式如下[1]：

$$TF_{ij} = \log \frac{n_j}{n_{all}} \quad , \quad IDF_j = \log_2 \frac{N}{df_j}$$

其中 n_j 表示單字 j 在文件 i 的出現次數。 n_{all} 表示文件 i 所有具意義的總詞頻。N 代表所有文件的總數 df_j :代表單字 j 有出現過的文章總數。最後結果為上述兩者的乘積：

$$w_{ij} = TF_{ij} * IDF_J$$

每個詞彙計算所得之 TF-IDF 值則為權重，值越高代表該詞彙在該資料集範圍下的重要性越大。

2.2.4 建立於關鍵詞之上的特徵分析模型

Dutta[18]提出了一個新的觀點，對於關鍵詞(Keyword)一詞有著不同的看法。作者認為，關鍵詞時時刻刻存在於我們的日常生活之中，而關鍵詞本身也有許多不同的定義，在維基詞典 Wiktionary 中：

1) 關鍵詞可能是一串文字裡的任何詞彙。

2) 關鍵詞可以是任何用來參照或連結到其他文字或資訊的詞彙; 它也可以是用來描述文章或書籍主題的詞彙;抑或是在資訊系統內用來代表資料目錄的名稱(在資訊領域之中)。

3) 關鍵詞是代表一個指令或函數的保留字(在程式設計領域內)。

4) 任何在文章中出現次數比平常多的詞彙也可稱為是關鍵詞(語言學領域)。

而作者最終認為，關鍵詞可以用來代表一篇文章主題的精華。在一般搜尋電子格式的資訊系統時，也常用以關鍵詞為主包含關鍵詞比對或是依照主題分類或字母順序瀏覽的方式運作，這也表示了關鍵詞是相當適合作為主題的描述詞。

確定了關鍵詞本身的特性之後，Dutta 設計了對凸顯這些特徵的一套指標模型，其中包含八個關鍵詞特徵指標，每個指標分別代表一種趨勢，此趨勢可說明在當某關鍵詞的一個指標值偏高時，所顯示出此關鍵詞的特徵。在此針對八個指標的詳細定義分別說明如下：

(1) Integrated Visibility Index：以 v(i)表示，定義為 Fr/Nr。此數值越高代表此關鍵詞出現頻繁，且可能為主題中心的、領域共通的或是次要的詞彙。

(2) Momentary Visibility Index：以 m(i)表示，定義為 Fr/Ar。此數值越高代表關鍵詞出現頻繁，但孤立集中。此關鍵詞可作為某一種研究的中心，但對大範圍領域來說卻只能作為其一分支。

(3) Potency Index：以 p(i)表示，定義為 ln(Nr*Fr)。此數值越高代表關鍵詞數量多和分佈率平均，代表領域共通且具高相關度的關鍵詞。

(4) Frequency Density Index：以 d(i)表示，定義為 Fr/J。值越高代表整個文件空間涵蓋率高。

(5) Occupancy Density Index：以 o(i)表示，定義為 Ar/J。值越高代表整個文件時間涵蓋率高。

(6) Keyword Density Index：以 k(i)表示，定義為 Nr/J。 值越高代表高頻率能量。

(7) Stability Index：以 s(i)表示，定義為 (Ar/Amax)*100。實際的分佈狀況與最高可能分佈狀況的比值，值越高代表分佈的高穩定性。

(8) Scattering Index：以 t(i)表示，定義為 Ar/Nr。此數值越高代表關鍵詞在整個主題空間領域是分散的。

在研究主題 S 之下，從年份 y 到(y+l)之間的文章總數為 J，J 隨著 l 的變化有所不同。i 則是由 1 到變數 n 之間的正整數，而 n 則是不同年份 l 之間的分別關鍵詞個數，Fr 為詞彙出現頻率，Ar 為文件分佈頻率，Nr 是 l 年分間的關鍵詞總數。前兩者隨著關鍵詞本身而變化，後者則是根據年份相應值也不同。Amax 則是用來預測某年份內最高的文件分佈頻率，為年份 l 與 Nr 的乘積。

2.3 英語教學相關字彙研究

根據 Nation[19]研究指出，受過教育英語母語使用者大約擁有 20,000 左右的英文字彙量。但對外語學習者來說，想要擁有跟母語使用者同樣的量是相當困難的，即使處於與母語使用者相同的語言文化環境下，由於許多字彙可能是集中於各種不同的環境下才使用的到，要在短時間內都學習到實屬不易，而且大部份的外語學習者較無機會處於母語使用環境下。然而為了協助英語學習者短期內有所進步，學習者可以透過學習一些經由英語教育研究學者整理過的學術常用字彙，進而讀懂大部份的學術教材，達到較顯著的進步成效。

2.3.1 英語教學字彙的定義與特徵

在 Coxhead 與 Nation 的研究中[20]，將英語字彙分成了四個群組。分別為高頻字彙、學術領域通用字彙、學術領域專門字彙以及低頻字彙。分別介紹如下：

- West[21]於約 5,000,000 字的英語語料庫中，建立了一份英語文章中使用頻率最為頻繁的字彙列表 GSL(Genreal Service List)。這份列表由 2,000 個字彙所組成，而 GSL 中的文字在學術文章中使用的字彙量約佔了 80%的涵蓋率(包含重複的字彙)。

- 學術領域通用字彙 AWL(Academic Word List)是由 Coxhead[22]所建立，當中包含了 570 個字彙。AWL 是根據之前的 UWL(University Word List)改進，是從藝術、商業、法律、科學等四大學術領域與其底下共 28 學門的語料庫中選出，不但變得更精確，且經實驗證實，在學術寫作文章組合而成的語料庫中，AWL 能夠達到 8.5%~10%左右的涵蓋率，在商業領域中涵蓋率更可高達 12%。而對於

相同大小非學術文章類的小說語料庫，AWL 的涵蓋率只有 1.4%，也說明 AWL 在學術文章中常被使用。Coxhead 建議教育者應該教導這些學術上頻率高且常被使用的字彙，而 AWL 也被廣泛的使用在學術英語教學上，國內如台大教育視聽館也有關於 AWL 的介紹與教學[23]。

- 學術領域專門字彙指的是來自於各種領域研究分野的字彙，總計約包含 1,000 個字彙，此類的字彙佔學術文章的涵蓋率約只有 5%。

- 低頻字彙則是在出現在多種的探討主題中，也因此字彙較為分散頻率也較低。有的字彙甚至只出現一兩次。

由上述各點可知，在英語教學上，學習者可優先學習 GSL 和 AWL 中的字彙，習得這些字彙後在學術文章的閱讀上，將近 90%以上的字彙都可理解。教育者也應優先教導這些使用率高的字彙，以達良好的教學成效 [26]。

2.3.2 字彙與詞性的組合使用 - 搭配詞

搭配詞(Collocation)一詞的定義，在牛津英語搭配辭典的解釋是，『搭配詞是某種語言字彙合併的方式，以產出自然的口語和文字』[24]。而 Benson[25]等人主張，學習如何將字彙加以組合建構成片語、句子、段落和文章，就是學習搭配詞，而熟習搭配詞就能正確流暢地寫作。也由於搭配詞在使用上並無嚴格的限制，造成 EFL 背景的作者會依照本身所擁有的字彙自行建立搭配詞，導致搭配詞在時態與字彙選擇上的錯誤。

此外，Benson 也提出搭配詞可分成語法搭配(grammatical collocation)和詞義搭配(lexical collocation)兩種。語法搭配是由一個支配詞(dominant word)配上介系詞或不定詞或子句的其他語法結構所組成的片語，而支配詞可以是名詞、形容詞或是動詞;詞義搭配則是由名詞、形容詞、動詞以及副詞所組成，不包含介系詞及其他語法結構。而搭配詞也往往以名詞為中心，名詞本身具備實質意義，也可由形容詞修飾描述，或當作動詞的受詞，更可接介系詞表達關係。而搭配詞也有相關的書籍資源可以參考，如 Oxford collocations dictionary for students of English 或 LTP dictionary of selected collocations 等字典。而國內的研究也有藉由比對雙語語料庫而開發出的一套系統 TANGO[27]，學習者可選擇想學習的字彙，便列出該

字彙的可能的搭配詞組合供學習者使用，這對字彙學習與英語學術寫作上能提供及時且實用的幫助。

在搭配詞的使用上，Verb - Noun 組合的搭配詞是 EFL 學習者最容易犯錯的一種類型，即使對以英語為母語的 Native Speaker 學習者來說也是如此[28]。此外，在搭配詞內容的組成之中，由於搭配詞本身對於時態的限制相當嚴格，因此動詞成了學習者最容易犯錯的字彙 ，故 Verb - Noun 搭配詞也是學習者在學習搭配詞時覺得最為困難的一環。

2.4 本章總結

在英語教學及英語學術寫作的研究上至目前為止已有許多成效，但本文認為仍然有許多值得探究的空間。以學術字彙列表 AWL 為例，AWL 是以改良 UWL 而來，字彙量也從原本的 800 多字精減而至 570 字。然而學術字彙列表本身開發的目的，是希望在英語的教學上能起到顯著的功效而提供給英語學習者的優先學習字彙，因此必須同時考慮到使用頻率以及涵蓋率。但在一般的英語學術寫作中，AWL 則有部份字彙在使用上的機率則相當低，相對的其搭配詞組合也相對減少，如此來說這些字彙在學術寫作上能提供的幫助有限。本研究的目的，便是希望能夠補足在學術寫作應用時學術字彙列表的不足，並提供真實狀況下常用的搭配詞組合，讓 EFL 作者在進行學術寫作時，有更豐富的論述表達方式與正確的寫作風格。

第三章

研究方法

本文前兩章觀察到在學術寫作上現行學術字彙列表的不足，以及期待能提供 EFL 作者在寫作時於學術字彙與搭配詞的使用上有較適當的參考目標。接下來本章將探討以關鍵詞擷取技術輔佐關鍵詞分析模型的使用，抽取出一般性學術共通寫作詞彙的方法。下圖為本章研究方法的流程圖，本章以下各節依照此流程圖之順序進行研究步驟。

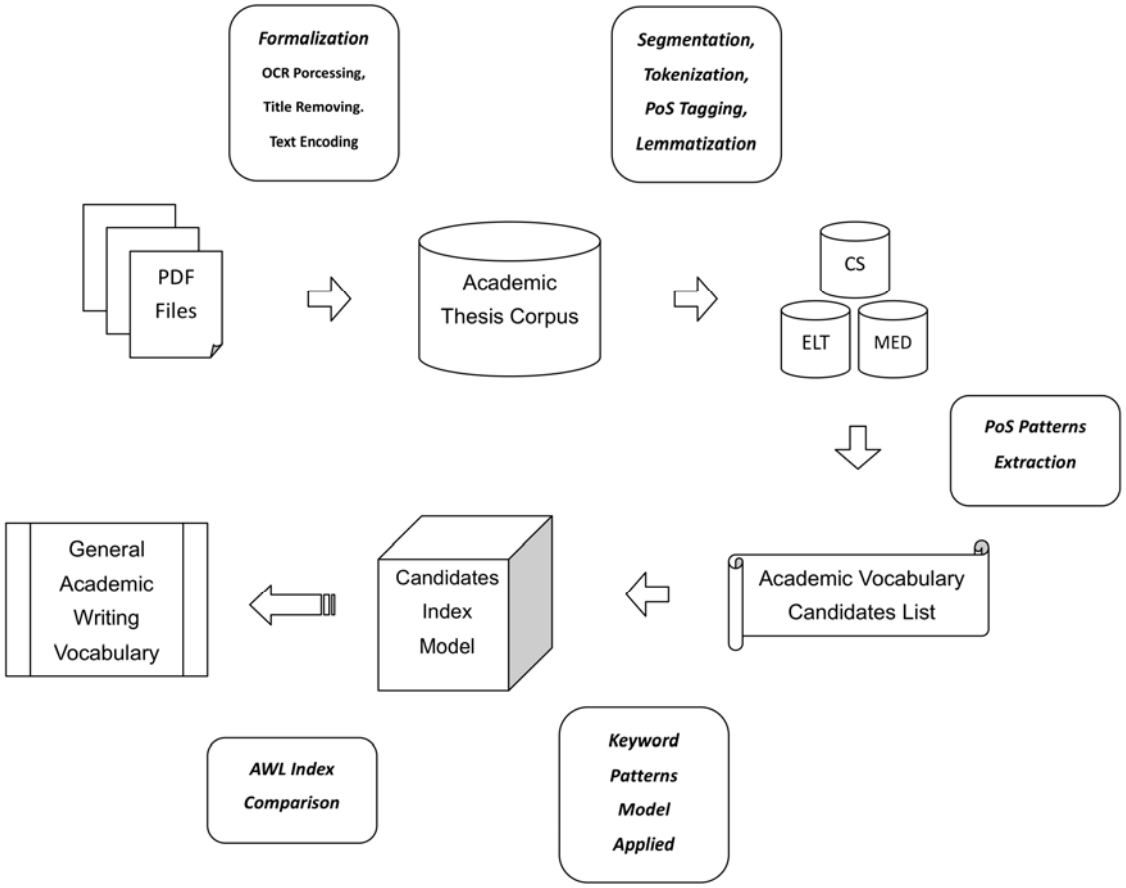

圖 3-1 研究方法之流程架構圖

3.1 語料庫設計

本文的研究目的，除了抽取出跨領域通用的學術寫作字彙外，並且能夠明顯各領域之間字彙使用上的差異。為了達到上述目的，在語料庫設計部份，則分別說明如下：

I. 首先是內容選擇的部份。學術寫作字彙可常見於一般學術課本教材、學術性雜誌、會議期刊論文以及畢業論文等。在內容的編排，教材與雜誌因其讀者眾多而採用較淺顯易懂的描述，字彙與搭配詞分佈也較為鬆散，而會議期刊與學位論文，強調用字精確，論述簡扼，學術寫作字彙分佈集中，故以期刊與學位論文為主。 而資料來源的收集，國內是從各大學所建構之機構典藏與國家圖書館提供的資源下載，國外部份則是以各校圖書館所購置的學術論文資料庫而得，全語料庫由 420 篇文章組成。

II. 其次是跨領域部份的設計。為了強調跨領域的部份，則需至少由三個領域以上，彼此間能夠相互交集印證，得出的結果也較為客觀。領域以本文研究相關的 Computer Science(CS)、English Learning & Teaching(ELT)之外，再加入用詞高度專門術語化的 Medical 領域(MED)，每個領域分別保有 140 篇學術論文，以此比較得出的結果是否為通用性的字彙。

III. 內容組成的部份，以學術論文為中心，其中包括期刊論文、碩士論文和博士論文三種。分別取臺灣、日本及美國三個地區的學術論文。美國學術論文為 NS 作者的代表，並收集其博士論文，藉由大量統計與寫作深度較高的內容構成分析出的結果，能夠作為 EFL 作者的參考。臺灣與日本學術論文的內容，是碩士論文與期刊論文各半，日本部份則是同為 EFL 作者的臺灣之對照。數量上臺灣日本則分別為 120 篇(期刊論文與碩士論文各 60 篇)，美國部份為 180 篇(全為博士論文)。

IV. 資料內容的選取。學術論文經過收集之後，必須將原本的 PDF 檔案格式轉換成 TXT 文字檔，並且將文章中的各種大小標題、圖表及參考文獻等內容移除，只保留摘要以及內文部份。移除文章中的標題目的在於，如 method、

conclusion 等詞彙經常被用在標題之中。若標題不移除此類詞彙的頻率便會偏高，會影響其他詞彙的頻率統計結果，況且此類字彙多已列在 AWL 之中，移除對本研究並無太大的影響。

綜合以上各點可知，語料庫在預先的設計上由三個領域及三個國家共九個集合，目的是藉由兩種維度不同的交叉分析，得到不同性質的分析結果。如圖 3.2 所示，縱向箭頭表示可從單一國家來看各國學術寫作特性，也可結合臺日兩國家(English as Foreign Language, EFL)與美國(Native Speaker, NS)做比較。橫向箭頭則是依照領域來看各領域學術寫作特性，可就單一國家領域探討其特性，也可結合三國家的 CS 領域發掘 CS 領域特有的常用學術字彙。而總和九個象限一起綜合探討，就成為本文研究主題，學術寫作中通用的字彙特性，九個象限中的數字則代表該國該領域下的學術論文數量。

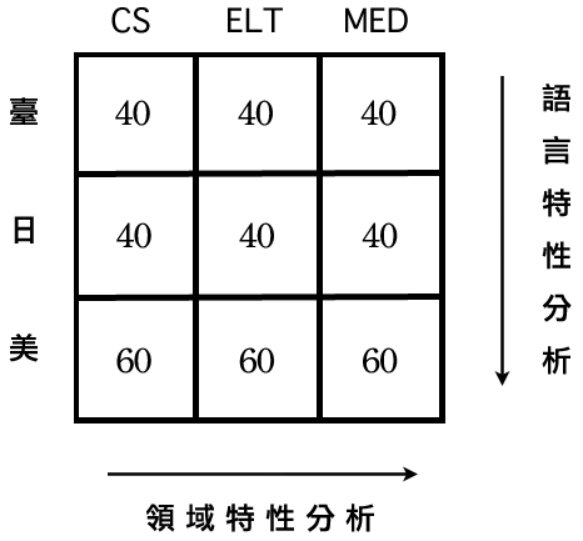

圖 3-2 語料庫結構特性分析

3.2 關鍵詞擷取 - PoS Tag Patterns

從第二章關鍵詞擷取相關文獻研究可知，關鍵詞擷取的最終目的，是從大量資訊之中抽取出可代表此資訊集合的詞彙，故此詞彙是包含主題中心的或是領域共通的資訊。本研究目的則是追求在學術寫作中頻繁使用的字彙，故需先了解此學術寫作字彙的特性，針對其特性而推演出適合的擷取方式，而此學術字彙是在補學術字彙列表(AWL, Academic Word List)之不足，兩者性質類似，因此我們可以從 AWL 的特性進行分析。在 Coxhead 的研究中，強

調 AWL 在學術文章語料庫中可達 10%左右的文件頻率涵蓋率，且 AWL 在學術文章的使用頻率相當高。而搭配詞也是最常用的詞彙組合表現，故學術寫作字彙在搭配詞中也經常出現，而以搭配詞來說，Verb - Noun 組合的搭配詞是 EFL 學習者最長犯錯的一環以及搭配詞以名詞片語為中心的表現方式等，再再都顯示了字彙組成中以動詞和名詞最為頻繁，此兩者也是構成句子意義的主要字彙。

綜合上述論點來看，我們截取關鍵詞應該以動詞和名詞為中心，並且過濾掉低頻率的字彙與低涵蓋率的字彙。為了完成上述任務，我們透過 NLTK[29]設計擷取關鍵詞之演算法。NLTK 由 Bird 與 Loper 所共同開發，是一套開放原始碼以 Python 程式語言為主的各種程式模組與相關的語言學資料檔案(語料庫)，目的是提供自然語言處理與文字分析上的研究與分析使用，並支援多種作業系統平台。本文以 NLTK 為中心設計了一套擷取相關 Patterns 的演算法，將語料庫依照領域分隔，依照文章的句點分隔句子當作擷取的基本單位，每句分別擷取以下四種 patterns：

I. Noun - Verb pairs
II. Verb - Noun pairs
III. Noun - * - Verb patterns
IV. Verb - * - Noun patterns

第一種 pattern 為一般句子表現的方式，由名詞後隨即接動詞而成，是名詞當作主詞的表達方式;第二種 pattern 也是常用的表達方式，名詞前面可接動詞的進行式與過去式用來修飾名詞，或是名詞當作動詞本身的受詞;後兩種則代表的是前兩種 pattern 的廣義表達方式。由於在一般論述的表達上，常會用到許多詞性表達前後文關係，如介系詞、代名詞等，同時形容詞多用來修飾名詞，副詞多用來修飾動詞，因此如想擷取出所有的名詞和動詞相互組合 pattern，必需考慮到這些附屬的詞性出現在兩者之間。然而第三種和第四種 pattern 中間也不能夾雜名詞或動詞出現，如果出現如 NNV 組合的狀況，演算法會將此組合視為第二種 pattern 而只擷取接連出現的 NV pattern。

下圖 3.3 為演算法詳細內容。整個演算法可分為三個部份，首先是語料庫的預先處理。語料庫名稱定義成 AcademicThesisCorpus，對於語料庫內的所有文件，我們以句點分隔的 sentences 作為擷取的最小單位，在迴圈 sentences 中的每個 sent，以一個單字作為一個 token，token 則是自然語言處理程序中的最小可識別單元，所有程序執行都以 token 為基

礎，故接下來的步驟包含詞性標註(PoS Tagging)與詞形還原(Lemmatization)都以單一字彙執行。在此採用詞形還原而不使用詞根還原(Stemming)的原因是，為了正確統計使用字彙的頻率而採取了還原的步驟，但詞根還原會破壞掉字彙本身的原始形式統一還原為字根(root)，在統計上會導致多形式的字彙統計錯誤，與研究目的找出精確的字彙使用上分岐，故採用詞形還原。

Part 1. Corpus Preprocessing

```
from nltk.corpus import AcademicThesisCorpus as atc

for documents in atc.tagged_sents():

     def ie_preprocess(documents):
     sentences = nltk.sent_tokenize(documents)
     sentences = [nltk.word_tokenize(sent) for sent in sentences]
     sentences = [nltk.pos_tag(sent) for sent in sentences]
     sentences = [nltk.stem_lemmatize(sent) for sent in sentences]
```

Part 2. Extraction Grammar Definition

```
grammar = r"""
  CPTP(#1): {<V.*><JJ|IN|RB|PP|DT|TO>?<N.*>}     #2
            {<N.*><TO|IN|DT>?<V.*>}              #3
"""

#1 - CPTP: Custom Pos Tag Patterns
#2 - Consecutive NV pair or N*V PoS Pattern
#3 - Consecutive VN pair or V*N PoS Pattern
```

Part 3. Pattern Recursive Extraction

```
cp = nltk.RegexpParser(grammar.CPTP)

for sentence in atc.tagged_sentences():

        result = cp.parse(sentence)

print result
```

圖 3-3 Custom PoS Tag Patterns Algorithm

第二部份則是關鍵詞擷取的文法定義。文法本身是依照正規表示式(Regular Expression)中定義的方式將詞性標籤進行組合，在此文法名稱為 CPTP。CPTP 下擷取兩種 pattern(#2 與 #3)。#2 所定義的詞性組合，以動詞為首以名詞結尾，無論是何種形態的動詞或名詞皆可。動詞和名詞之中，以一般較常見的形式，副詞(RB)在動詞後修飾動詞、形容詞(JJ)在名詞前

修飾名詞、代名詞(PP)在名詞前修飾主詞或名詞、名詞前出現機率相當高的限定詞(DT)以及介系詞(IN 與 TO)表達關係等，都是可允許出現的詞性，問號(?)表示上述詞性可出現或不出現，因此#2 可分別代表前述第二種與第四種 patterns，#3 代表了第一種與第三種 patterns 的組合。最後的部份則依照定義的文法遞迴式的將整個語料庫中符合文法的 patterns 作為關鍵詞擷取出來。此外，由於擷取是以整個句子作為擷取的基本單位，故有可能某種 pattern 符合#2 的 NV 組合，而其後面的 V 又與之後相連的 N 行程符合#2 的 pattern 狀況，雖說 V 的部份是重複計算其頻率，但後續的分析，會回歸到單個字彙在語料庫象限中的頻率計算，故此方法不會造成誤差。

根據上述演算法擷取出來的 patterns 過濾掉非擷取目的其他詞性，將結果所得之 patterns 列表拆分為單字的集合，同時將名詞與動詞分開處理，並依領域計算其詞頻，此列表為初步研究方法得出之領域別學術寫作字彙的候選詞。我們依據此候選詞列表作為形態分析模型的輸入，同時根據每個分析模型指標的特性，交叉分析得到最終的學術寫作字彙。下表 3. 1 為不同領域的候選詞數量資訊。

候選詞數 \ 領域別	CS	ELT	MED
名詞數量	1104	1622	1689
動詞數量	719	753	709

表 3-1 由 CPTP algorithm 擷取出之各領域學術寫作字彙候選詞

3.3 形態分析模型套用

在本文第 2. 2. 4 節 Dutta[18]的研究中，集結了各種關鍵詞的定義後提出了關鍵詞本身是能代表某個領域或是論述範圍的詞彙，並且設計八種指標，依照每個指標所代表趨勢的不同，關鍵詞的特徵也有所不同。而指標中代表的所有關鍵詞特徵如下所示：

I. 主題中心的

II. 主題共通的

III. 輔助性質的

Dutta 同時強調，關鍵詞本身的屬性在整個主題空間下所代表的意義。就整個研究主題空間來看，Fr 為字彙出現頻率，代表的是關鍵詞在空間上的表現狀態;Ar 則是文件發生頻率，代表著關鍵詞在時間上的表現狀態;而 Nr 為空間內總關鍵詞數量總和，可作為整個研究主題空間中能量分佈的表現。也因此，經由這些代表著不同關鍵詞屬性的變數的計算，引出關鍵詞在研究主題空間的特徵表現。

本研究則以 Dutta 提出的指標分析模型為基礎，針對不同領域下的語料庫候選詞進行指標分析計算。但在基礎假設下本研究與 Dutta 的研究有所不同，必需分析其不同之處，才能進行接下來的步驟。試將基礎上不同之處說明如下所示：

I. Dutta 提出的研究，某些指標是針對從年份 y 開始到總年數 l 之間關鍵詞的變化而計算。但在本研究之中，並無因應年份而進行分類切割，統一假設所有的語料庫文章在同一年份進行實驗，l 則為固定呈正整數 1，因 l 而造成影響的 Amax 與 J 等兩變數有關的的指標計算均無法使用。受到影響的指標有四個，分別為 d(i)、o(i)、k(i)以及 s(i)等，在接下來的實驗中均不使用此四個指標的計算。

II. 在原本的研究之中，Vr 等八項指標定義是基於不同年份下的研究，然而在上一點之中曾提到年份 l 不適用於本研究，也因此相對應的函數值 i 的定義也相對失效，故在此我們將包含變數 i 定義的八個函數指標 v(i)、m(i)、p(i)、d(i)、o(i)、k(i)、s(i)、t(i)，在本研究中重新定義為 Vr、Mr、Pr、Dr、Or、Kr、Sr、Tr 等八個指標，而也由上一點得知 Dr、Or、Kr 及 Sr 等四個指標於接下來實驗不採用。

III. 剩餘四個指標中，唯獨 Tr 本身所代表的趨勢無法判別關鍵詞是屬於上述三種中何種的特徵。Tr 代表每個關鍵詞在整個研究領域的分散程度，並無法藉由 Tr 得知關鍵字本身的特性，於此將 Tr 排除不予計算。

IV. 在原研究內關鍵詞類型可分為三種，Keyword Cluster(三個單字以上)、Keyword-Couple(兩單字構成)與 Single Keyword(單獨字彙)，而其研究中心主軸以

Keyword Cluster 的計算分析為主。而本研究著重於單字彙為中心進行研究，找出學術寫作字彙本身的特徵，也因此由 CPTP 演算法擷取出的詞彙將這些詞彙拆開為以動詞和名詞為主的單字進行分析。

套用分析模型後可算出表 3-1 中各項字彙與其指標數值。在原有指標分析模型下共有三種代表其特性的指標可適用於本研究，分別為 Vr、Mr 以及 Pr。此外，代表字彙重要性的 TF-IDF，在計算時會用到 Ar 與 Fr 作為運算元，可見其重要性。因此我們也將 Ar 與 Fr 納入為本研究中分析模型指標的一部份，如此一來，分析模型便有五個指標可供實驗之用。在下一章中，我們將分別針對每一項指標的特性與趨勢進行實驗分析，同時將結果與適合作為參照目標的 AWL 做比較，擷取出最終的學術寫作字彙。

3.4 本章總結

在本章一開始，我們透過領域分野與語言特性交叉的語料庫設計，實現了對單一語料庫多重分析的可能，並且在關鍵詞擷取的基礎下，採用數種以名詞和動詞為主的類似搭配詞的 PoS Tag Patterns，同時使用 NLTK 作為擷取的軟體工具，將可能為領域共通的各領域關鍵詞候選詞擷取出來。隨後將這些候選詞套用在 Dutta 所提出的關鍵詞指標分析形態模型上，並且依照實際狀況，將許多不必要的指標剔除，只留下適合的指標供後序分析之用。接下來在下一章中，我們將對這些具備有各指標值，也就是具備各指標特性的候選詞作更進一步的篩選，並且透過不同機制的交叉實驗，選出最後符合研究目的的學術寫作字彙與搭配詞。

第四章

實驗分析討論與結果

本章將對第三章研究方法之結果進行實驗分析並討論。4.1 節提出實驗相關測試資料與進行實作之方法。4.2 節則進行實驗結果的討論與分析，並提出學術寫作上常用的字彙列表為本研究的結果。4.3 節則以 4.2 節所得之結果，衍生出在語料庫中之常用搭配詞。4.4 節總結本章。

4.1 實驗資料與實作方法

本文以第三章研究方法所得之各領域之學術字彙候選詞作為實驗資料進行實驗分析，每一候選詞均有 Ar、Fr、Vr、Mr、Pr 五種指標索引值，我們分別對每項指標值個別進行實驗，探討這些候選詞在整個語料庫中之表現與分佈狀況。然而由於語料語本身資料龐大，若直接將候選詞與原始語料庫之字彙對照，便會造成這些候選詞在語料庫之中成為相當稀疏之點狀分佈，故需是先對語料庫本身進行預先篩選。

4.1.1 實驗資料說明

目前整個語料庫原始資料包含 420 篇學術論文共有 79,874 個詞形(word type) 與 7,652,876 個詞次 (word token)，經過 stop word 過濾(共 429 個詞形)的程序之後則為 79,445 個詞形與 3,649,156 個詞次，這些詞次依領域分佈如下表 4-1：

	ALL	CS	ELT	MED
詞次(Token)	3,649,156	1,188,794	1,525,436	934,926
文件數(Documents)	420	140	140	140

表 4-1 AcademicThesisCorpus(ATC)語料庫領域別文件詞次數量分佈

而所有詞次經由詞性標註後，其中總共包含了名詞 55,141 個和動詞 17,233 個詞次。而字彙的頻率分佈大略如下:

	字彙出現頻率								
	> 201	51 ~ 200	21 ~ 50	11 ~ 20	8 ~ 10	6 ~ 7	3 ~ 5	2	1
字彙總量	2,294	3,661	4,089	4,232	2,712	3,056	10,440	10,670	38,291
名詞數量	1,616	2,454	2,927	3,104	2,067	2,205	7,694	7,263	25,810
動詞數量	568	583	622	617	368	427	1,405	1,302	4,746

表 4-2 ATC 語料庫字彙頻率分佈

由上述字彙頻率分佈可知，語料庫的原始字彙裡在頻率 5 之下的佔了 74.3%，而在頻率 10 以下則佔了 81.5%，分佈集中於低頻率字彙。原因是由於語料庫原始資料收集是採各領域隨機收集的方式，故主題較為分散，這也是造成字彙數量多且頻率低的原因。另一方面，我們也預先過濾了 stop word 等佔了原始語料庫 52%以上的一部份高頻率字彙，也是導致字彙頻率集中分佈於低頻率區塊的原因。

在一般關鍵詞擷取的方法中，由於關鍵詞本身具有高頻率的特性，因此為了精確的擷取出關鍵詞同時減少不必要的字彙，會過濾掉部份低頻率的字彙，或是只取統計上頻率較高的部份字彙進行關鍵詞擷取。由 Coxhead[22]提出的學術字彙列表(AWL)中也提到說，學術字彙在一般學術語料庫中所佔的文件頻率涵蓋率約可達 10%，也就是約十篇學術文章中至少有一篇會有學術字彙出現，而本研究中每個領域文章分別為 140 篇，以 Coxhead 的標準來

看，學術字彙在 140 篇文章文件頻率應該是在 14 以上，因此相對的單一學術字彙的詞頻至少也是在 14 以上，然而考慮到實際出現狀況上可能有些許的誤差，在此也稍微放寬篩選的標準，語料庫中詞頻在 10 以下字彙則將其忽略不予計算。故由表 4.2 可得知，詞頻在超過 10 的名詞有 10,101 個，動詞則有 2,390 個。在這些名詞與動詞中，於三個領域皆出現過的名詞有 1,980 個，動詞則是 1,040 個，此資訊將成為我們進行實驗的主要依據，同時整理如下表：

	ALL	CS	ELT	MED	Intersection
名詞數量	10,101	4,112	5,493	4,389	1,980
動詞數量	2,390	1,282	1,669	1,168	1,040

表 4-3 ATC 語料庫領域別動詞名詞數量分佈(出現頻率 Fr > 10)

接著，我們以 AWL 作為數量參照的標準，而取出與學術字彙列表等量以及倍數以上的字彙進行實驗。AWL 一共包含 570 個單字，其中副詞與形容詞佔了 101 個，剩下的 469 個字彙中，由於考慮到有些字彙可同時作為名詞和動詞使用，故在此我們也將此狀況納入考量，因此包含了重複的字彙中，作為名詞的有 331 個而當作動詞的有 213 個。

在上一章中，藉由 PoS Tag Patterns 選出的候選詞(請參照表 3-1)為研究的初步成果。然而這些候選詞，不僅僅是數量過於龐大，而且也非完全是學術寫作常用的字彙。正因如此，這些候選詞仍需要特定的方法來進行精確的實驗才能達到標準，故將這些候選詞套用於指標分析模型，依指標的數值分別對字彙排序。然而在經由指標排名排序過後的字彙，卻無一個適當的比較對象，故於此我們取與其性質相近的 AWL 等量以及 1.5 倍的字彙量兩種不同的數量單位進行實驗比較，較能凸顯出其效果。以名詞來說，AWL 中名詞有 331 個，因此我們取名詞候選詞各指標前 331 個與前 496 個(1.5 倍)，來比較各指標值偏高所取出之字彙與學術字彙列表在語料庫中分佈的情形，同樣地動詞部份也選取各指標前 213 與前 320 個候選詞與學術字彙列表的分佈做比較，並且也將把各領域交集的部份以相同的方式進行實驗。下表總結各項文中所提及之數據如下：

	CS	ELT	MED	Intersection	AWL*1	AWL*1.5
名詞數量	1104	1622	1689	519	331	496
動詞數量	719	753	709	339	213	320
簡稱	S(D)			S(D*)	S(A)	S(A+)

表 4-4 領域別候選詞數量與 AWL 數量統計

4.1.2 實驗方法

接著本節說明進行實驗的方法。首先在此先對欲進行實驗的對象分別進行定義說明，在上表 4-4 中的統計數據分別定義如下：

I. AWL 中所包含的名詞與動詞之字彙列表，目的是用來與等量的各領域動詞與名詞當作 threshold，並且個別對其動詞和名詞跟候選詞中之動詞和名詞做分佈狀況之比較，以下簡稱為 S(A)。

II. 學術字彙中名詞與動詞原始數量的 1.5 倍數量，取其 1.5 倍作為 threshold 是與 S(A)之統計結果提供另一種的標準作參考，只取數量而無實際字彙，簡稱為 S(A+)。

III. 三個領域 CS、ELT 與 MED 分別的候選詞樣本資料的集合， 簡稱為 S(D)，此集合目的是求得分別領域中為學術字彙的詞。

IV. 候選詞中三領域共同交集而成的名詞與動詞之集合，，簡稱為 S(D*)，目的是為了求得各領域共通之學術字彙。

由上述定義可知，進行實驗之內容有 S(D)、S(D*)與 S(A)三種，S(A+)只取其數量而無內容。而 S(A)作為 S(D)與 S(D*)之參照對象，在 S(D)與 S(D*)都有 S(A)與其作分佈的比較。

實驗步驟的第二步則為指標值的計算。在第三章曾經提及到，可當作適當指標計算的有 Ar、Fr、Vr、Mr 與 Pr 五種指標值。在此說明計算方式與範圍：

I. Ar：Ar 為定義範圍內的字彙於文章總數中出現的次數，也就是一般的常見的文件頻率(Document Frequency)。以 S(D)來說，每個領域之 Ar 最多可達 140，在 S(D*)則以整體語料庫來算，最高值為 420。

II. Fr：Fr 為最常見的語料庫統計資訊，即為某字彙在定義範圍文章內的出現頻率總數，同樣地在 S(D)與 S(D*)中有所差異。

III. Vr：Vr 為每個字彙平均出現頻率，定義為 Fr/Nr。Nr 代表的是在不同範圍內之關鍵詞總數，在 S(D)與 S(D*)內也分別隨之變化，Vr 值高代表此字彙為主題中心的、領域共通的或是輔助性質的。然而在第三章中曾經提到，由於 Nr 在本研究中則因為年份 l 固定的關係，Nr 因而成為常數，同時 Vr 所得到的值會類似 Fr。因此在此作調整。各領域候選詞計算出 Vr 後，會依照候選詞 Vr 值由高而低排序，並計算各候選詞之間的 Vr 值差距。當出現 Vr 值差距大於其他候選詞兩倍以上的字彙時，便將此字彙包含 Vr 值高於此字彙的所有字彙先行剔除，去除 Vr 值過高的字彙，使得透過 Vr 值選出的字彙是偏向領域共通而非主題中心的，也因此選出的詞彙與 Fr 所選出的詞彙有所不同。

IV. Mr：Mr 為每篇文章中該字彙的出現頻率，定義為 Fr/Ar。此指標的計算類似 TF-IDF，強調單一範圍下字彙的重要性程度，同樣依 S(D)與 S(D*)而變動，Mr 值高代表此關鍵詞出現頻繁。

V. Pr：Pr 為範圍空間下字彙的分佈密度，定義為 ln(Fr*Nr)。在 S(D)與 S(D*)之中隨之不同，Pr 值高代表此關鍵詞為領域共通的。如同 Vr 值的計算方式，各候選詞也在計算出 Pr 值差距之後，將大於其他部份的候選詞 Pr 差距值兩倍以上的字彙先行剔除，也因此選出的詞彙與 Fr 所選出的詞彙有所不同。

經過上述的定義，對實驗內容 S(D)與 S(D*)計算其指標值後，分別依照上述五個指標值之術值由高至低進行排序，可以得到單一領域 S(D)與共通領域下 S(D*)之字彙列表。

透過計算指標值之結果，實驗的第三個步驟是在 S(D)與 S(D*)的結果中，取出各指標值排序後數值較高的字彙，至於取出的數量則是參考 S(A)與 S(A+)，取兩種數量分別來作比

較。以 S(D)中的 CS 領域部份作為例子，以動詞來說，整份候選詞字彙列表經過指標值計算字彙總數仍為 719 個，接著此份候選詞列表依照每個指標值分別排序之後，取出等同於 S(A)與 S(A+)之兩種數量，也就是五個指標值分別取出排名前 213 個字彙與排名前 330 個字彙，而我們總共可以得到 S(D)中三個領域加上 S(D*)共四份集合中包含動詞和名詞的八份指標序字彙列表， 此結果可說是經過各指標值篩選出之第二階段之候選詞。附錄表一為 CS 領域下動詞取與 AWL 等量數量 S(A)之指標序字彙列表，作為所有指標序字彙列表之代表(因資料量過於龐大，在此取一部份表示)，而我們也取 AWL 作為對照，作為 AWL 中的動詞字彙分佈與各指標選出的動詞字彙分佈之對照。

雖然目前已經得到了進一步篩選的指標序字彙列表，但這些列表仍然無法作為我們最終的結果。其原因在於指標分析模型中的五個指標即使具備了篩選出字彙成為關鍵詞的能力，但根據各指標本身的定義所擷取出的字彙與本研究所追求的一般性學術寫作字彙的定義不盡相同，仍然需要另一種符合研究目的機制對這些指標序字彙列表作更精確的過濾。

學術寫作上共通且經常使用的字彙，依照字面上的定義即為『在學術領域下之各分野都可能出現的字彙，並且字彙頻率本身具有一定水平之上』。同時在 Coxhead 的研究中[22]也提到，學術字彙在學術各領域分野的涵蓋率約為 10 %，也就是代表說這些跨領域之學術字彙在各領域的分佈是均勻分佈，並不會只在單一領域表現突出。基於上述的原則，接下來的實驗步驟則針對指標序字彙列表進行字彙跨領域分佈，也就是單一字彙在不同領域之同質性進行檢證。

在統計學上有許多方法可以檢測出不同集合間的分佈狀況的效果，其中 Mantel[31]在約 50 年前就提出卡方分佈(Chi-Square Distribution)除了可計算樣本資料之間的相關性以外，也適合用來計算這些樣本之間的同質性。在此我們以一個字彙當作一個集合，而該字彙在 CS、ELT 以及 MED 個別領域中的出現頻率作為該集合的樣本資料，藉以算出該字彙於不同領域之間之同質性。卡方分佈的計算公式如下：

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

其中，χ^2 為卡方值，其數值越高表示該樣本間之實際群體分佈與期望值相差甚大，代表該集合之同質性低。O_i 為該集合之樣本值，E_i 則代表所有樣本之間的期望值(在本研究中由於頻率皆為正數，故期望值為三個頻率之平均值)，k 代表著樣本數量，在本研究中 k 值為 3。基於公平原則，語料庫中各領域學術論文乃隨機選取，雖然說每個領域文章中都包含 140 篇學術論文，但從表 4-1 中可以得知，由於文章內容長度不一，每個領域下的詞次相差甚多，三領域詞次數比 CS：ELT：MED 約為 1.27：1.63：1，為了保持學術字彙的均衡出現標準，必需作字彙頻率正規化的處理，於同質性計算時必須將字彙在各領域下之頻率除以各領域之詞次數比後，根據正規化後的頻率計算字彙之同質性數值。

接著取實驗步驟三所得之八份指標序字彙列表進行同質性卡方分佈的計算。由於卡方分佈計算是求出字彙於三領域間之同質性，在進行此計算時假使該字彙並沒有在三個領域中同時出現，該字彙便不予計算，依其分佈歸類於僅出現於兩領域或只出現於單一領域的類別。由表 4-3 來看，以名詞而言，三領域交集之名詞總數共有 1,980 個(動詞有 1,040 個)，而篩選出來的名詞候選詞最多各有 331 個(即為 S(A)，與 AWL 名詞等量)以及 496 個(即 S(A+)，AWL 之 1.5 倍名詞數量)，為了清楚地了解這些候選詞在所有三領域交集名詞中的同質性表現，在此對所有三領域交集的名詞總數和動詞總數的同質性分佈切分為數個區間作為區隔。

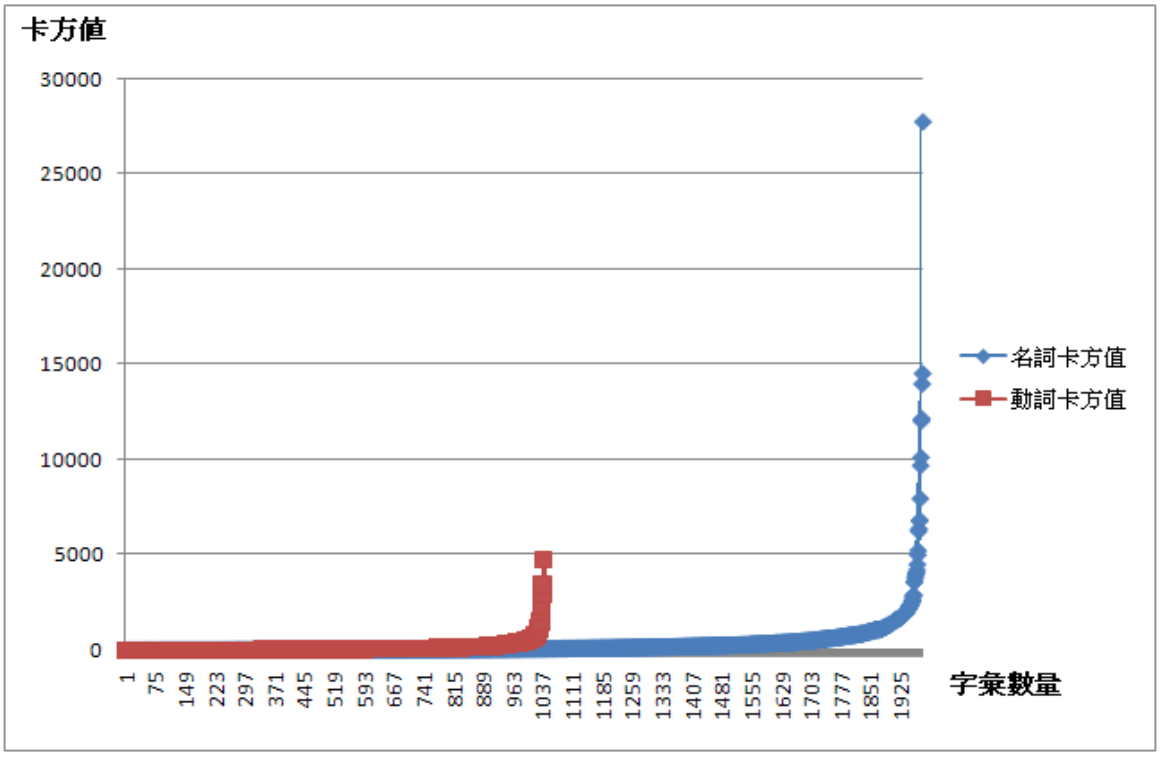

圖 4-1 三領域交集名詞與動詞之卡方值分佈

上圖 4-1 為三領域交集之名詞與動詞依照計算出的卡方值由低而高排序分佈。從圖中可以發現，無論是名詞或是動詞，絕大部份計算出的卡方值都是相當平均的，但部份有些字彙卡方值遠較其他字彙高出許多，若是同時與其它字彙一起區隔，將會導致各區間之卡方值分佈極度不均，對於這些字彙則分隔成另一個區間另外討論。而分隔的條件則以上圖中兩曲線開始急速成長之地方開始，也就是兩曲線之切線斜率急速上升之處，在 1,980 個名詞中為第 1,944 個，其卡方值為 2184(取整數部份)，動詞則是在總數 1,040 中個第 1,028 個，其卡方值為 910。接著我們將過濾掉特高卡方值的部份依其卡方值數值切分成三個等分，並且依照同質性由高而低排列(也就是卡方值由低而高排列)，區分成同質性高(High)、同質性中等(Medium)、同質性低(Low)分佈，再加上一開始分隔出的同質性極低(Very Low)之另一區間，將三領域交集之所有名詞與動詞依照同質性高低之區間分隔開，同時也將附錄表一為代表的第二階段候選詞字彙依四個區間分隔開，下表則為每個區間的同質性數值分佈：

詞性 \ 區間		H_Bound	M_Bound	L_Bound	VL_Bound
名詞	同質性數值	0 ~ 728	728 ~ 1456	1456 ~ 2184	>> 2184
	數量	1,785	118	39	36
動詞	同質性數值	0 ~ 303	303 ~ 606	606 ~ 910	>> 910
	數量	947	62	17	12

表 4-5 三領域交集字彙各區間之同質性數值分佈與包含字彙數量統計

就卡方值計算之結果而言，同質性高的意義代表該集合之下之樣本均與期望值差距不大，用來計算卡方值的樣本為單一字彙於 CS、ELT 與 MED 三領域下之分別頻率值，而期望值則為三頻率值之平均值。如此一來，可能會出現某些字彙的卡方值低，表示其同質性高，但是結果卻是因為三者間頻率很低且頻率值之間的間隔相等之故。為了避免此種狀況，我們將對表一之結果加上頻率篩選的機制，以四種不同的頻率作為 Threshold 來篩選。依照單一領域下(各包含 140 篇文章)學術字彙的可能出現頻率從 1/10~1 之間，分別取 10、35、70、140 作為 Threshold 的基準，也就是期望學術字彙至少分別在十篇、四篇、兩篇甚至於一篇文章中就至少能出現一次作為標準。

而上述所指的頻率是經過正規化程序，也就是字彙在各領域下之原始頻率除以領域間的詞次數比。當字彙於每個領域之出現頻率若沒有同時超過此標準，則將此字彙淘汰不予採用。在此以附錄表一的資料為基礎，同時加入同質性區間分隔與頻率 Threshold 分隔之後的 CS 領域動詞候選詞分佈結果如附表二。

	Ar	Fr	Vr	Mr	Pr
指標趨勢	字彙於文件中分佈平均	字彙使用率高	字彙可能是單一領域、領域共通或輔助性的	字彙於單篇文章出現次數多	字彙為領域共通的

表 4-6 各指標所代表趨勢之特徵

在上一節我們曾經提到各指標代表的趨勢與意義，在此總結如上表 4-6 所示，對於學術上常用之寫作字彙來說，也具備上表中的各項特徵。故同時在五種指標選出來代表字彙中都有出現的字彙，就是我們希望得到的學術領域共通之常用寫作字彙。我們將取附錄表二中五個指標的交集作為代表 CS 領域動詞學術字彙的結果，而此結果也必需滿足 Fr 和 Mr 的特徵，在頻率之 Threshold 上，Threshold 大於 140 的字彙較符合上述條件，因此 Threshold 大於 140 之依各指標並且依照相同區間(同一 Bound 之間進行交集)將表二的字彙作交集，所得到的字彙列表結果如同附錄表三，同時也附上其他領域的動詞之結果字彙列表。

4.2 實驗結果之分析討論

在上一節中，我們對各領域交集之候選詞 S(D)與各領域交集候選詞 S(D*)分別進行了指標值計算，接著依序各指標值選擇與 AWL 等量 S(A)與 1.5 倍數量 SA(+)取出了可作為各指標代表性的字彙後，隨即使用卡方分佈的統計方式驗證這些字彙的同質性，並將字彙同質性分佈切分成三個區間，最後再設立 4 種頻率的門檻檢驗是否為常用字彙的可能性。上述實驗方法的搭配使用，皆代表對於不同的實驗資料與實驗方法，除了確立其本身之正確性之外，同時藉由其結果之交叉分析，也可得知其成效。接下來我們參照實驗所得到的結果，並以討論的方式分析實驗結果。

4.2.1 實驗樣本的差異性

在 4.1.2 節所進行的實驗方法之中，我們分別對各領域分別候選詞 S(D)以及各領域交集候選詞 S(D*)此兩份樣本資料分別進行實驗。以研究目的而言，直接對領域交集之後選詞 S(D*)進行篩選或許就可以達到求得學術寫作上領域共通的字彙，是個很直覺式的方法。但是在本研究中開始對候選詞進行統計時發現(請參照下表)，同時於每個領域都出現之候選詞 S(D*)數量，都只佔了各領域候選詞中不到一半的數量，在名詞部份領域中甚至佔不到三分之一。若是只選擇交集部份來進行實驗，就意味著必需捨棄大量的可能性。

	CS	ELT	MED	Intersection
名詞數量	1104	1622	1689	519
非三領域共通名詞	585	1103	1170	0
動詞數量	719	753	709	339
非三領域共通動詞	380	414	370	0

表 4-7 各領域候選詞與非領域共通候選詞數量統計

再者，考慮到即使是在學術領域這個限定範圍之中，各個領域在寫作上的呈現方式有所不同。舉例來說，在 CS 領域下，network 這個字彙是使用率相當高的字彙，經常出現於研究方法或實驗設計的論述之中，但是在其他兩個領域之中使用率則相對降低，因此就本研究來說 network 的同質性並不高。然而 Coxhead 所提出的 AWL[22]之中，許多如同 network 此類字彙仍然是屬於 AWL 的其中之一。故在一開始就只取 S(D*)作為實驗資料樣本，可能會導致錯失許多領域共通但具有領域集中特性的字彙。

如上述所說，單憑同質性的驗證也可能造成判斷錯誤。由於同質性的計算是以頻率之卡方分佈為作為基準，當某字彙在三領域中之出現頻率都很低時，所得到的同質性相對就非常高，但這種字彙並不符合需求。另一方面，當某字彙在各領域中頻率都很高時，同質性也就相對低落，尤其是如上述字彙 network 之狀況，在三領域中都有一定的出現頻率，但在 CS 領域中頻率卻相較高出許多，而其同質性卻成為候選詞中最低的一群。因此在同質性和

出現頻率的選擇上需取其平衡點，故我們取在頻率上具有一定水準之上的字彙但同質性低的字彙也納入最終結果的考量。

4.2.2 不同實驗樣本之實驗結果

在上一節中，我們以各領域分別候選詞 S(D)以及各領域交集候選詞 S(D*)分別進行實驗，而各領域分別候選詞 S(D)與各領域交集候選詞 S(D*)之間的差別主要在於，S(D)是分別領域下進行指標排序後最後透過指標相互交集得到的結果，而 S(D*)則是一開始進行候選詞交集、指標排序而交集，可說是經過了兩次的交集程序。然而在 S(D)的實驗結果上，由於各領域候選詞在不同指標值表現下差異甚大，最後經由頻率篩選交集後得到的字彙為數較低，不過成為能該領域下較常使用的代表字彙，這是 S(D*)所無法達成的。

另一方面，S(D*)在第一次的交集程序上，萃取出了能夠於多種領域表現良好的字彙，也因此再經過指標排序後第二次交集出的字彙與各指標單獨計算所得之字彙並無太大的差別，這也表示在基於研究目的基礎上，先行對候選詞交集挑選出領域共通的候選詞的效果上，明顯大於直接對各領域候選詞進行指標排序後交集的效果。但在最終結果的選擇上，若是只選擇 S(D*)的實驗結果作為最終的研究成果，只能選擇到同質性高而頻率高的字彙，部份分佈頻率高而同質性低的邊緣化字彙如 network 等將會被忽略，而這些字彙在各領域中佔了相當少數。

而就五個指標分別來討論，指標的表現上也是有所差異。舉例來說，下表節錄在附錄表二中的 CS 領域動詞候選詞在頻率大於 10 下的同質性分佈來說，各指標都是選出前 213 個代表性的候選詞然後分別依照區間進行分隔。但是在總數 213 個動詞之中，可以發現像 Mr 指標只選出在三個領域中具有同質性的動詞一共 78 個。相對地，Ar 指標則可選出 190 個動詞，Pr 指標則可選出 163 個動詞，至於 Fr 和 Vr 的指標表現則是雷同。而在附錄表二中，若是將頻率的 Threshold 提高後，各指標選出的動詞也隨之遞減，但是遞減的程度也是不會影響各指標的表現，因此可以推斷，就指標本身的學術字彙篩選效果上，是 Ar > Pr > Fr ≈Vr > Mr。這效果不僅僅限於 CS 領域動詞底下，在所有的實驗資料都呈現相同的效果。

	H(0~303)	M(303~606)	L(606~910)	VL(>>910)
Ar_(共 190 個)	139	34	10	7
Fr_(共 130 個)	119	8	2	1
Vr_(共 120 個)	113	5	2	0
Mr_(共 78 個)	72	5	1	0
Pr_(共 163 個)	139	19	3	2

節錄附錄表二 頻率大於 10 的動詞候選詞於各指標與同質性區間下之數量分佈

總結本節的內容來說， 我們以各領域分別候選詞 S(D)以及各領域交集候選詞 S(D*)以指標模型和交集的方式混合使用來進行實驗，試圖藉由兩種方法的優點結合而得到最好的結果。但只就單一方法而論，在求得學術共通寫作字彙的前提下，先對各領域候選詞進行交集的處理方式(也就是 S(D*)的資料)，會大於單獨使用指標的效果。而就指標模型本身的方法來說，五個指標的表現則是各有優缺點，當中以 Ar 效果最佳，Mr 的效果則最差。

4.2.3 學術寫作字彙的篩選機制

基於上述的討論與分析，為了達成兩個篩選準則同質性與頻率皆高的表現，並排除同質性高而頻率低的狀況，同時考慮到語料庫各領域的組成以博碩士論文佔多數，在文章字數相當多時學術字彙出現的機率較高，因此我們將作為第二篩選準則之頻率提至最高，在四個頻率的 Threshold(10、35、70、140)之中選擇最高的 140 成為最終的 Threshold，相當於選出的學術字彙在每篇學術文章中至少出現一次。如此一來，可在同質性和頻率皆可兼顧的狀況下，挑選出較為適當的學術寫作字彙。

接著，在實驗一開始挑選各指標的排名較高代表字彙的數量選擇上，於 AWL 等量的 S(A)與 AWL 數量 1.5 倍的 S(A+)兩者的考量上，當候選詞的數量固定時，交集後的字彙總數取決於每個指標選出的字彙的數量。而從附錄表二中的結果來看，Mr 此指標會選出較少的字彙作為此指標的代表字彙，導致在 CS 領域動詞所得到的字彙相當稀少，而指標 Mr 本身代表的特徵為在單一文章中字彙的出現次數較多(請參閱表 4.5)，就本研究以博碩士論文

為主的語料庫來看，Mr 在篩選出學術寫作常用字彙的目的上是必要的，同時為了避免結果選出來的字彙過少如同附錄表三中的 CS 領域動詞一樣，因此我們選擇 S(A+)數量對各指標序字彙列表交集的結果成為我們最終的結果。

最後，考慮到頻率上表現好因同質性而被邊緣化的字彙，我們則取以各領域分別候選詞 S(D)所得之各領域代表性學術寫作常用字彙與各領域交集候選詞 S(D*)得到的領域共通學術性寫作字彙，將此兩份字彙列表取聯集，也就是同時顧及字彙在單一領域和共同領域的表現，並非只考慮共同領域 S(D*)的部份。根據我們的準則選出來的結果，名詞共有 246 個，而動詞有 147 個，如附錄表四所示。同時也將實驗方法開始至篩選出最終結果字彙的流程統一整理，如下所示：

1. 將各領域分別候選詞 S(D)與各領域共通候選詞 S(D*)分別進行五種指標 Ar、Fr、Vr、Mr、Pr 的計算。

2. 分別對 S(D)與 S(D*)的指標值計算結果，選出每個指標由高而低排名於與 AWL1.5 倍數量 S(A+)之前(在名詞中選取前 496 個，動詞中選取前 320 個，可參閱表 4.4)的字彙，作為具有該指標特性的代表字彙。

3. 由上一步驟 S(D)與 S(D*)的得到的字彙，分別計算該字彙的同質性。計算方式為統計學中常用的卡方分佈計算，以字彙在三領域中的頻率作為樣本資料，三個領域為一個集合計算，而各領域中的字彙頻率須經過正規化計算，也就是將原始字彙頻率除以該領域字彙量的步驟，並將各指標代表字彙依照同質性由高而低分佈。

4. 為了更清楚地了解字彙同質性的分佈狀況，同時也希望字彙能夠依照同質性高低被學習者有效利用，從同質性高之字彙開始學習。將各指標代表字彙的動詞與名詞分別依照同質性數值的四個區間(H、M、L、VL)排列。

5. 考慮到同質性本身的不足，再加入頻率(10、35、70、140)的 Threshold 作為門檻，解決頻率低而同質性高的問題。

6. 最後，將 S(D)與 S(D*)中分別依指標和同質性區間排序，同時在過濾掉頻率低於 140 以下的字彙後，取出五個指標中共同出現之字彙，作為兩者的結果，並將此兩者得到的字彙做聯集，得到最終所求的學術寫作常用字彙列表，如附錄表四所示。

4.2.4 基於地域語言特性的學術寫作字彙

於第三章曾經提到，語料庫在設計上為了後續分析之用可分為領域特性以及語言特性兩種維度的語料庫建構方式(請參照圖 3-2)。而在上一節中我們得到了以綜合領域之間的特性為主的學術寫作字彙。同樣地，我們也仿照上一節的方式，擷取出了綜合三個地域語言特性的學術寫作字彙附於附錄表五。綜合兩份學術字彙列表來討論，可以發現兩份字彙中所包含的 AWL 數量都大約將近三成左右，在選擇結果與 AWL 的相交範圍的表現上並無太大差異，如下表 4-8 所示。

	名詞數量	包含 AWL 名詞數	動詞數量	包含 AWL 動詞數
綜合領域特性	246	69	147	41
綜合語言特性	183	47	109	29

表 4-8 兩種學術字彙列表數量與所包含 AWL 數量

然而，就同質性的表現上，由下表 4-9 可以發現，在綜合領域下，無論是名詞或是動詞的卡方值數值分佈都比綜合語言下的字彙大出許多，這是因為在我們的 ATC 語料庫中字彙在不同領域之間出現的頻率相差極大，相較於在不同語言地域之下的表現。舉例來說在附錄表四中同質性極低但頻率高的字彙 learn，明顯地在 ELT 領域下的出現頻率會比在 CS 及 MED 領域下出現頻率高出許多，而在語言特性中得到的學術寫作字彙，ELT 領域的 learn 字彙分散於三個地域國家之中，導致 learn 此字彙分佈較平均。雖然說在得到兩者的結果前有進行正規化的計算，但在字彙在不同領域上的頻率分佈上較不同地域之間差異為大，這同時也表示，在綜合領域所得的學術寫作字彙，卡方值越高時(即代表同質性越低)，代表這些字彙極有可能是某一領域的常用字彙，如附錄表四中的 L Bound 以及 VL Bound 所示，像是 L Bound 中的 function 則是在 CS 領域下的頻率表現較凸出，skill 在 ELT 領域下表現較為凸出等，至於 VL Bound 內下的所有字彙，都屬於各領域下常用字彙，如附表內標註所示。

	高同質性(H)	中同質性(M)	低同質性(L)	同質性極低但頻率高(VL)
名詞(綜合領域)	0~728	728~1456	1456~2184	>>2184
名詞(綜合語言)	0~194	194~388	388~582	>>582
動詞(綜合領域)	0~303	303~606	606~910	>>910
動詞(綜合語言)	0~108	108~216	216~324	>>324

表 4-9 兩種學術字彙列表之字彙卡方值分佈

而就綜合語言特性的學術寫作字彙來說，L Bound 及 VL Bound 內的字彙則是偏向於 ELT 領域的學術寫作字彙。在表 4-1 中可以發現，ELT 領域下的字彙總 token 數，是 CS 領域和 MED 領域下的 1.2 倍及 1.6 倍左右。也就是說，在三個地域語言的分野下，無論是何種分野，ELT 領域的字彙量都是佔較多的，這也導致在計算同質性進行頻率的正規化時，ELT 的字彙還是在頻率表現上較為凸出，這個結果類似在綜合領域的學術寫作字彙，在某方面頻率上表現特別突出時，造成同質性非常低，而成為了具有特別意義的字彙。同樣地，在綜合語言特性下 L Bound 動詞中的 teach 和 read 字彙明顯地屬於 ELT 領域，VL Bound 之內的字彙則更為明顯。

4.3 延伸應用 - 學術搭配詞

在上一節中，在最終的實驗結果下，總共擷取出了綜合領域特性與綜合地域語言特性的學術寫作常用字彙。這些字彙對於一般英語學習者來說，並不是在日常生活中鮮少見到，用在艱深的學術論文表達的單字，而是在學術寫作之中，甚至是一般的其他領域的英語文章中，如新聞或小說等，也常常出現的字彙。然而學術寫作與其他寫作分野之間的差異，主要在於學術寫作上的結構較為嚴謹，字彙與字彙之間的組合的規定也較精簡。為了能讓英語學術寫作的作者能清楚的了解學術寫作上的字彙組合方式，將本研究 ATC 語料庫中經常使用的搭配詞以一般學術寫作中最常使用的搭配詞組合以及不同語言特性下的常用搭配詞整理出來提供給作者作參考。

在附錄表六以及表七之中，分別是兩類不同的搭配詞列表。附錄表六中的內容，是在原有 ATC 語料庫之中，我們將 CS、ELT、MED 三個領域合併起來共 420 篇文章總和的角度，擷取出的常用搭配詞。其中又包含了兩個子表格，附錄表六之一所列出的是，六大類使用頻率相當高的搭配詞，出現的頻率至少都在 200 次以上，而這些搭配詞組合，如 at the same time 或是 as a result of 等，不僅是在學術寫作中經常出現，在一般文章中也經常可見，屬於較基本必須儘快熟習的搭配詞。而在附錄表六之二之中的搭配詞，則是考慮到對於不同常用的學術字彙作為搭配詞的中心時，多種可使用的前後修飾字彙的組合，而附錄表六之二中的搭配詞可分為六大類，在整個語料庫中出現頻率上也有 50 次以上，並且在考量其多樣性之下所產生的結果。

另一方面，在附錄表七中也包含了兩個子表格，分別代表了以英語為外國語(English as Foreign Language, EFL)地域(台灣及日本)常用的搭配詞與以英語為母語(Native Speaker, NS)地域(美國)兩種個別經常使用的搭配詞組合。由於是分開進行統計，故在兩個子表格中的常用頻率都是取出現 50 次以上作為最小門檻，並且標註紅色的部份是與附錄表六中的搭配詞比較，取表六中沒有的組合形態方便比較。就附錄表七之一之中代表 EFL 語言特性的搭配詞來說，像是 in which the 或是 that there be 等都是 EFL 作者常用的句子接續方式。而 depend on 和 correspond to 等也是我們經常看到的字彙使用組合。

相較於附錄表七之一與附錄表六的兩個子表格來說，在 NS 作者常用的搭配詞中，對於同樣的想表達的核心學術字彙上，在表達方式上所修飾使用的字彙不僅是表現方式較為豐富，在用字上使用片語或是專門字彙也較多。舉例來說，以 in the XX method of 此搭配詞來說，XX 在附錄表六之二代表了所有領域作者常用的 effective/suggested，而在附錄表七之一中的 EFL 作者風格 XX 為 this，在 NS 作者風格 XX 為 effective/suggested/proposed/accessing/hybrid，NS 作者表現風格較為多樣，用字也較深入。而如 NS 作者的 handle/discuss/show/cope with/look into/create/point out/avoid problem 使用兩字的片語也較所有作者的 handle/discuss/show/identify problem 多。另外，如 beyond the scope of 或是 study utilize 等一般 EFL 作者不常用的風格在 NS 作者中也隨處可見。因此在附錄表七之二的 NS 作者常用搭配詞，不僅可跟 EFL 作者比較，也可跟更一般性的附錄表六中的常用搭配詞參照，使用者可以視需要學習 NS 作者的寫作風格。然而就英文句子組成而言，搭配詞的表現極其豐富，在此僅僅列出一小部份使用頻率較高的搭配詞組合，若是想詳細了解搭配詞的使用，也有相關資源[24]可以參考。

4.4 本章總結

在本章中， 我們透過了許多實驗方法以及門檻機制的組合搭配，成功地擷取出在學術寫作上常用的字彙，並且以這些字彙作為核心，在本研究所建立的 ATC 語料庫之中，挑選出了四類搭配詞的組合供參考。而這些挑選出的學術寫作字彙，與 AWL 最大的差別在於，AWL 中的字彙，有些部份字彙的組成，基於其在學術語料庫中必需維持一定涵蓋率，導致這些字彙的使用率相當低。而本研究所擷取出的學術寫作字彙，都是選擇使用頻率高為主，但在不同領域下仍然具有一定的表現。

另一方面，就兩份字彙列表選出的名詞與動詞而言，兩者包含的 AWL 數量皆約為 30%，也代表著 70%的字彙可以補足 AWL 字彙所沒有的。這些字彙經由本研究的實驗方法加以驗證，不論是在出現頻率上以及文章的分佈上皆有良好的表現。對學習 AWL 的 EFL 學習者來說，這些字彙擴充了 AWL 以外五成左右的字彙量，提供了更多在學術寫作上可使用的字彙，同時也列出以這些字彙為核心的常用搭配詞，對於有英文學術寫作需要的研究者或是學生在進行學術寫作的同時，成為即時性的學習資源與參考對象。

第五章

結論與未來研究方向

5.1 結論

人類自從有文明開始，便藉著教育將先人的智慧結晶與經驗教訓不斷的延續下去，而後人則藉由閱讀理解這些文獻典籍的記載，在自己所處的時代不斷的改進生活周遭的事物，創造更好的生活環境，這些文獻典籍可說是人類演進的重要基石。時代演進至今，學術界成為實驗研究的重鎮，而以英語為核心的學術論文可說是全世界進行研究交流的重要媒介，在研究學者專家與企業專家間不停的交流。因此英語學術論文的寫作成為踏入世界性研究交流的基本門檻。但是對於以英語為非母語的研究者來說，在進行英語學術寫作時，為了要精確的表達自己的研究成果，字彙的選擇經常成為棘手的問題。另一方面，對英語學習者來說，學術字彙列表(AWL)的提供固然是一項幫助，但 AWL 在總量上固然有限，同時部份字彙在表現上也有較不實用的狀況。因此本研究針對英語學術寫作與英語學習上遭遇之困難，提出了實用性高且能補強 AWL 的一份字彙列表，並且以這些字彙為基礎，萃取出常用的搭配詞使用方式，提供英語學術寫作和英語學習上即時性的協助。

學術寫作本身具有用字精確與描述簡扼的特性，也充分表現在學術寫作字彙上面。而學術寫作字彙在學術領域文章中，無論是出現的頻率以及文章的分佈狀況，都具有良好的表現。因此我們從建立跨領域學術論文語料庫為基礎，結合資訊科技與統計模型的方式，從非英語學術專家的另一個角度，挖掘出屬於學術寫作共通的實用字彙。

英語學術寫作字彙在界定上並無明顯的定義，即使是建立 AWL 的 Coxhead 也是從學術語料庫中經由分析歸類而得。本研究由觀察學術論文句子的組成中最小組合單位搭配詞為基礎，初步排除了搭配詞中屬於關係與修飾屬性的介系詞、形容詞以及副詞，而以構成搭配詞意義的動詞與名詞為主，透過關鍵詞擷取技術取出了搭配詞中最常使用的 PoS Tag

Patterns，隨即將擷取出的 Patterns 分解成動詞與名詞的候選詞，並且依照領域與詞性分開，作為初步篩選可能符合研究目的之結果。

根據擷取出的候選詞與及交集的字彙集合，作為 S(D)和 S(D*)兩種不同的實驗樣本，輸入多指標為主的關鍵詞分析模型進行分析，在呈現不同趨勢的每個指標中，依指標值排序而選出對著於 AWL 等量或倍數的數量並指標值由高至低的字彙，每一份選出的字彙就成為該領域下獨具意義的字彙。而藉由對應到研究目的，將不同指標序字彙再做進一步的交集，可得到該領域下代表的學術常用字彙。

然而在單一領域適用之學術常用字彙並無法適用於所有的學術寫作範圍。為了達到通用性的效果，我們採用統計上常用來計算同質性的方法，以卡方分佈(Chi-Square Measure)對字彙逐一檢驗，將字彙於各領域下的出現頻率作為樣本資料，計算集合為三個領域交集下的單一字彙分佈狀況，當卡方值數值低時，表示字彙在各領域分佈較為平均，其同質性較高。但單獨計算字彙之同質性可能會導致最終的結果字彙偏向在領域之頻率都偏低但同質性高的字彙，故需另外一個輔助性的 Threshold 來修正實驗結果。

最終以在各領域出現頻率大於 140，並且同質性高的 S(D*)候選詞，在選擇 AWL 數量之 1.5 倍後代表各指標意義的字彙列表交集而成的字彙為主，同時為了補足頻率高而同質性低但可能為學術寫作常用字彙，將 S(D)的候選詞也依上述條件選出的結果，與 S(D*)之結果進行聯集而得到最終的字彙列表。其中名詞有 246 個，動詞則有 147 個，這些字彙可作為在學術寫作上與英語學習上 AWL 的補遺，同時也提供以這些字彙為主的常用搭配詞，能讓使用者更快速的學習這些字彙的使用。

5.2 未來研究方向

本研究是以關鍵詞擷取技術配合指標分析模型對多領域學術論文語料庫進行剖析，而在關鍵詞擷取部份是採用 PoS Tag Patterns 作為擷取的目標，取出佔多數的名詞加動詞與動詞加名詞的組合。但英文句子的表現上詞性的組合相當多種，而且在組成上也不限於最少的三字彙搭配詞。基於此兩個因素，『N-gram Patterns』與『多詞性關係組合』可作為我們未來的研究方向。

I. N-gram Patterns：N-gram 為 N 個字彙組成的片斷，其中 N 為正整數，N=1 時稱為 unigram，N=2 時為 bigram，N=3 叫做 trigram，以此類推。在本研究中，N 介於 2 到 3 之間。當 N 變大時，也意味著字彙之間的組合隨之增加，字彙間的關係也隨之複雜。但是透過文法中詞性修飾與組成的分析，可精確的取出以學術字彙為核心的字彙組成片斷。除了 N-gram 外，自然語言處理中針對 Chunks 或是 Noun Phrases 的類似單位都常用於關鍵詞擷取技術的應用上。

II. 多詞性關係組合：英文句子組成中，介系詞主要用於表示與承接其他不同詞性之間的關係。即使是最常用的動詞加名詞的搭陪詞，最後面仍須接上介系詞與後續內容相連，而像是介系詞加名詞加介系詞此類的搭配詞也不在少數。除了介系詞外，副詞常用於修飾動詞，而名詞常用形容詞修飾，這些屬於修飾性質的詞在學術寫作上也經常被使用，如在 AWL 的 570 字組成中，就有 101 個字彙是由這些附屬性的字彙所構成。在加入這些詞性的關係後，不僅能擴充常用學術寫作字彙的數量，且能靈活的使用不同詞性的組合，在學術寫作的進行上更加有所助益。

參考文獻

[1] 郭志華. 學術寫作字彙特色分析. URL: http://ir.lib.nctu.edu.tw/handle/987654321/19252

[2] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. London: Longman.

[3] Chen, C. Y. & Tang, Y. T. (2004). Collocation errors of Taiwanese college students: Oral or written production. In The proceedings of the Eighth International Symposium on English Teaching(pp. 483- 494). Taipei, Taiwan: The Crane Publishing Co.

[4] McEnery T., & Wilson, A. (Eds.). (2001). Corpus linguistics. Edinburgh: Edinburgh University Press.

[5] Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. English for Specific Purposes, Vol. 25, 235-256.

[6] Biber, D. (1998). Variation across speech and writing. Cambridge: Cambridge University Press.

[7] Conrad, C. M. (1996). Investigating Academic Text With Corpus-Based Techniques: An Example From Biology. Linguistics and Education 8, pp. 299-326.

[8] Thompson, P., & Tribble, C. (2001). Looking at Citations: Using Corpora in English for Academic Purposes. Language Learning & Technology, Vol.5, Num. 3 pp. 91-105.

[9] Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

[10] Ercan, G., & Cicekli, I. (2007). Using Lexical Chains for Keyword Extraction. Information Processing & Management, Vol.43, Issue 6, pp. 1705-1714.

[11] Matsuo, Y., Ishizuka, M. (2003). Keyword Exraction from a Single Document using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools. World Scientific Publishing Company.

[12] Giarlo, M. J. (2005). A Comparative Analysis of Keyword Extraction Techniques. Rutgers, The State University of New Jersey.

[13] 魏智強. (2006). 自動化問答系統之研製. 私立中華大學資訊工程研究所碩士論文.民國九十五年八月.

[14] 王俊弘, 劉昭麟, 高照明. (2003). 電腦輔助英文字彙出題系統之研究. 2003 人工智慧,模糊系統及灰色系統聯合研討會論文集.

[15] Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, July, 2003, pp. 216-223.

[16] Turney T. D. (2000). Learning algorithms for keyphrase extraction. Information Retrieval, 2(4):303–336.

[17] Frank E., Paynter G. W., Witten I. H. (1999). Domain-specific keyphrase extraction. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'99), pages 668–673, Stockholm, Sweden.

[18] Dutta, B., Majumder K. & Sen, B. K. (2009). An analytical model for investigation of some characteristics of the keywords of the subject fermi liquid: a case study. Annals of Library and Information Studies, Vol. 56, December 2009, pp. 273-290

[19] Nation, P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

[20] Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), Research perspectives on English for academic purpose (pp.252-267). Cambridge: Cambridge University Press.

[21] West, M. (1953). A general service list of English words. London: Longmans, Green.

[22] Coxhead, A. (2000). The Academic Word List: A Corpus-based Word List for Academic Purposes. TESOL quarterly, 2000.

[23] 台大教育視聽館 Academic Vocabulary, URL : http://efreeway.avcenter.ntu.edu.tw/freeway/postgraduates/vocab/vocab_index.html

[24] 廖柏森. (2008). 英文研究論文寫作 - 搭配詞指引 : 眾文圖書.

[25] Benson, M., Benson, E., & Ilson, R. (2007). The BBI dictionary of English word combinations. 台北 : 書林.

[26] 黃茹玉. (2007). 探討應用語言學期刊論文中學術字彙之使用. 國立清華大學外國語文學系碩士班外語教學組碩士論文. 民國九十六年六月.

[27] Chuang, T. C., Jian, J. J., Chang, Y. C. & Chang, S. C. (2005). Collocational Translation Memory Extraction Based on Statistical and Linguistic Information. Computational Linguistics and Chinese Language Processing Vol. 10, No. 3, September 2005, pp. 329-346.

[28] Nesselhauf, N (2003). The use of collocations by advanced learners of English and some implications for teaching. Applied Linguistics, 24, 223- 242.

[29] Bird, S. (2006) .The Natural Language Toolkit, Proceedings of the COLING/ACL on Interactive presentation sessions table of contents 2006. Sydney, Australia. pp.69 - 72

[30] Lucas, N., Cremilleux, B. & Turmel, L. (2003). Signalling well-written academic articles in an English corpus by text mining techniques. *Proceedings Corpus Linguistics* 2003. pp. 465-474.

[31] Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association, Vol. 58, No. 303. pp. 690-700

附錄表一　CS 領域動詞候選詞之各指標代表性字彙(前 213 個)

1. 各指標中選出前 213 個數量 S(A)作為代表，而 AWL 共 213 個動詞中，只有 193 出現在語料庫內。

2. 表格前面部份表示各字彙在各領域的分佈狀況，在總數 213 個候選詞中，三代表候選詞在三個領域中都有出現，字彙以綠色表示。二領域出現候選詞以藍色表示，只在 CS 領域出現之候選詞以紅色表示。

領域出現數	Ar_ (共 213 個)	Fr_ (共 213 個	Vr_ (共 213 個	Mr_ (共 213 個	Pr_ (共 213 個	AWL (共 193 個
三	204	157	156	111	182	180
二	6	23	20	27	16	10
一	3	33	37	75	15	3
依各指標排序選出之高指標值候選詞字彙列表						
	complete	blind	blind	blind	blind	access
	vary	act	act	foresee	act	achieve
	address	burden	sign	burden	complete	affect
	close	complete	complete	complete	root	analyze
	type	root	sum	address	address	approach
	sum	address	address	charge	close	assume
	reach	charge	charge	trace	trace	benefit
	lower	trace	trace	type	type	bias
	purpose	type	list	fire	sum	clarify
	list	sum	root	sign	promise	code
	code	promise	promise	advise	survey	conclude
	occur	survey	survey	lower	sign	consist
	research	sign	burden	cease	force	construct

	step	force	force	bridge	lower	contact
	parallel	lower	lower	intervene	list	contrast
	cover	list	type	code	balance	contribute
	repeat	balance	balance	visit	code	cooperate
	conclude	code	code	invert	occur	coordinate
	form	research	research	reserve	research	create
	view	visit	project	parallel	substitute	demonstrate
	separate	substitute	substitute	decay	step	denote
	benefit	step	step	subdivide	parallel	derive
	employ	reserve	reserve	click	cover	design
	direct	parallel	travel	attack	form	detect
	function	dot	dot	view	attack	devote
	rate	form	pertain	separate	view	discriminate
	summarize	click	click	escape	project	dominate
	produce	attack	attack	waste	constrain	emerge
	fail	view	view	transfer	separate	enable
	study	project	visit	regress	benefit	encounter
	divide	constrain	constrain	travel	employ	ensure
	index	separate	separate	function	exhibit	establish
	limit	waste	waste	trust	transfer	estimate
	maintain	benefit	benefit	pertain	function	evaluate
	link	exhibit	exhibit	rate	pertain	exclude
	access	transfer	transfer	strike	rate	exhibit
	correct	travel	parallel	boost	drive	extract
	leave	function	function	float	summarize	facilitate
	position	trust	trust	delay	fail	focus
	object	pertain	form	score	study	function
	remain	rate	rate	contact	divide	generate
	transform	drive	drive	analyse	yield	identify
	experiment	study	study	duplicate	count	illustrate
	space	yield	yield	link	index	image
	carry	index	index	access	limit	imply
	attempt	limit	limit	correct	delay	impose
	ensure	delay	delay	plan	score	incorporate

	signal	score	score	fight	duplicate	induce
	integrate	contact	contact	smooth	link	input
	cross	analyse	analyse	position	access	involve
	approach	duplicate	duplicate	supervise	correct	isolate
	target	link	link	counter	plan	label
	replace	access	access	experiment	smooth	link
	provide	correct	correct	grant	position	maintain
	receive	plan	plan	space	locate	maximize
	analyze	smooth	smooth	segment	demand	mediate
	follow	position	position	exchange	transform	minimize
	equal	demand	demand	contract	experiment	modify
	avoid	counter	counter	slide	picture	obtain
	continue	experiment	experiment	feed	tie	participate
	lead	picture	picture	signal	space	pose
	support	tie	tie	display	bias	precede
	contrast	space	space	inscribe	segment	predict
	derive	bias	bias	probe	attempt	process
	approximate	segment	segment	approach	exchange	project
	meet	attempt	attempt	target	miss	promote
	establish	exchange	exchange	equal	plot	publish
	prove	contract	contract	support	sort	range
	contribute	slide	slide	manifest	signal	register
	block	plot	plot	approximate	display	rely
	introduce	sort	sort	block	integrate	remove
	develop	signal	signal	bend	cross	require
	incorporate	display	display	request	trade	research
	grow	inscribe	inscribe	register	approach	restrict
	size	trade	trade	incorporate	target	reveal
	pass	approach	approach	belong	notice	reverse
	image	target	target	load	fall	revise
	input	notice	notice	size	provide	select
	adopt	fall	fall	image	analyze	shift
	control	equal	equal	input	follow	site
	illustrate	support	support	control	equal	survey

	fix	judge	judge	design	lead	target
	design	contrast	contrast	change	support	utilize
	change	aim	aim	test	judge	accumulate
	test	manifest	manifest	shape	contrast	aggregate
	suppose	approximate	approximate	subject	aim	attribute
	subject	span	span	rank	manifest	automate
	expect	guide	guide	site	approximate	challenge
	record	block	block	care	meet	comment
	involve	request	request	train	span	contract
	ignore	track	track	play	guide	diminish
	connect	register	register	weight	block	display
	decide	incorporate	incorporate	interview	request	eliminate
	eliminate	belong	belong	scan	track	exceed
	start	load	load	enable	register	output
	improve	size	size	release	incorporate	phase
	speed	image	image	comment	join	proceed
	affect	input	input	figure	belong	release
	return	review	review	watch	load	resolve
	hold	control	control	query	size	simulate
	accord	design	design	attribute	pass	substitute
	play	change	change	question	image	sum
	weight	test	test	tune	input	suspend
	observe	shape	shape	schedule	review	terminate
	express	subject	subject	sound	control	trace
	enable	rank	rank	model	fix	transfer
	verify	record	record	base	design	advocate
	identify	site	site	measure	change	aid
	consist	care	care	process	test	conceive
	deal	speed	speed	estimate	shape	conform
	figure	return	return	suggest	suppose	dispose
	note	train	train	fear	subject	impact
	regard	accord	accord	doubt	rank	invest
	depend	play	play	label	record	pursue
	prevent	weight	weight	log	adjust	restore

	check	interview	interview	power	site	tape
	investigate	scan	scan	shear	care	violate
	capture	enable	enable	trail	start	assist
	search	release	release	profit	speed	chart
	demonstrate	comment	comment	overlie	return	conduct
	modify	deal	deal	host	train	conflict
	serve	figure	figure	tip	accord	debate
	tend	note	note	strip	play	decline
	evaluate	regard	regard	mail	weight	differentiate
	desire	check	check	wire	scan	expose
	refer	watch	watch	defect	enable	grade
	concern	search	search	corrupt	release	implicate
	account	query	query	coach	deal	interpret
	extend	attribute	attribute	scramble	figure	monitor
	draw	desire	desire	blow	note	persist
	question	concern	concern	kick	regard	reject
	relate	account	account	cheat	check	seek
	combine	outperform	outperform	shake	investigate	stress
	simplify	question	question	bid	capture	survive
	sense	sense	sense	ship	search	sustain
	bind	bind	bind	allot	query	undertake
	remove	range	range	love	attribute	accommodate
	range	hope	hope	orient	desire	adapt
	move	challenge	challenge	overbid	concern	compensate
	challenge	mark	mark	slim	account	constrain
	mention	tune	tune	abort	outperform	evolve
	add	guarantee	guarantee	clang	draw	format
	minimize	schedule	schedule	whitewash	question	grant
	operate	sound	sound	spar	sense	guarantee
	discuss	focus	focus	whistle	bind	interact
	focus	rest	rest	skim	range	justify
	represent	model	model	slack	hope	manipulate
	extract	claim	claim	animate	challenge	parallel
	assign	base	base	smart	mention	quote

	rest	measure	measure	prime	mark	react
	exist	answer	answer	anneal	tune	schedule
	model	offer	offer	dangle	guarantee	bond
	create	process	process	welch	schedule	confer
	satisfy	estimate	estimate	brush	sound	consent
	base	store	store	personify	focus	cycle
	construct	suggest	suggest	wedge	represent	deduce
	result	increase	power	counsel	assign	draft
	call	power	log	blast	rest	feature
	optimize	log	label	chase	model	fund
	calculate	label	box	reprint	claim	institute
	select	box	mind	subsidize	base	issue
	measure	mind	fear	cull	result	layer
	explain	fear	output	relinquish	call	panel
	determine	output	host	flatter	measure	purchase
	handle	host	stage	dwarf	determine	recover
	examine	stage	phase	deflect	handle	regulate
	include	phase	trail	negative	examine	style
	match	trail	mail	sift	include	transport
	implement	mail	profit	forge	match	allocate
	correspond	profit	wire	patch	implement	converse
	describe	wire	strip	shot	correspond	credit
	build	strip	corrupt	brake	describe	deviate
	assume	corrupt	copy	slice	answer	exploit
	apply	bid	format	speech	offer	index
	require	coach	kick	bundle	define	invoke
	begin	light	inspect	grab	process	offset
	choose	orient	abort	pace	estimate	random
	define	copy	negative	tile	store	route
	process	format	smart	mesh	decrease	trend
	compare	kick	prime	pitch	suggest	unify
	estimate	abort	overbid	flood	compute	amend
	distribute	negative	animate	trap	increase	appreciate
	achieve	smart	slim	sprite	power	aspect

	reduce	prime	anneal	hook	log	assemble
	bound	overbid	slack	conceal	label	aware
	denote	animate	route	pad	box	corporate
	obtain	slim	speech	parse	mind	data
	set	anneal	patch	crease	bid	discrete
	solve	slack	stream	route	light	diverse
	store	route	proof	misbehave	corrupt	equate
	decrease	speech	broadcast	credit	copy	erode
	perform	patch	clock	browse	orient	explicit
	suggest	stream	array	curate	format	finite
	update	proof	mesh	download	bite	
	send	broadcast	skim	tag	output	
	generate	clock	parse	lift	stage	
	run	array	slice	hurt	phase	
	propose	mesh	tag	dispatch	host	
	associate	skim	credit	flag	negative	
	compute	parse	download	clock	smart	
	increase	slice	prune	broadcast	abort	
	understand	tag	clip	encrypt	prime	
	learn	credit	slope	cascade	route	
	write	download	sift	clip	proof	
	power	prune	pitch	slope	stream	
	respect	clip	flag	prune	array	
	organize	slope	tile	relay	speech	
	output	sift	shot	stream	anneal	
	stage	pitch	browse	conquer	animate	
	bite	flag	bid	spike	broadcast	
	degree	tile	coach	array	clock	
	random	shot	light	warp	slack	
	negative	browse	orient	proof	patch	

附錄表二　CS 領域動詞第二階段候選詞於不同頻率(正規化後)Threshold 下之同質性分佈

說明：

1. 下表為 CS 領域動詞經由各指標值選出代表的前 213 個動詞後，剔除無法計算同質性的動詞，並且依照區間分隔的 CS 領域動詞數量分佈。

Threshold : Frequency > 10，頻率經正規化後大於 10 的動詞總數共有 687 個				
	H Bound (卡方值 0~303)	M Bound(卡方值 303~606)	L Bound(卡方值 606~910)	VL Bound(卡方值>>910)
Ar_(共 190 個)	139	34	10	7
Fr_(共 130 個)	119	8	2	1
Vr_(共 120 個)	113	5	2	0
Mr_(共 78 個)	72	5	1	0
Pr_(共 163 個)	139	19	3	2
AWL_V(共 140 個)	124	15	1	0
Threshold : Frequency > 35，頻率經正規化後大於 35 的動詞總數共有 389 個				
	H Bound (卡方值 0~303)	M Bound(卡方值 303~606)	L Bound(卡方值 606~910)	VL Bound(卡方值>>910)
Ar_(共 174 個)	124	34	9	7
Fr_(共 86 個)	75	8	2	1
Vr_(共 87 個)	73	11	3	0
Mr_(共 42 個)	36	6	0	0
Pr_(共 126 個)	102	19	3	2
AWL_V(共 87 個)	74	12	1	0
Threshold : Frequency > 70，頻率經正規化後大於 70 的動詞總數共有 270 個				

	H Bound (卡方值 0~303)	M Bound(卡方值 303~606)	L Bound(卡方值 606~910)	VL Bound(卡方值>>910)
Ar_(共 153 個)	108	31	8	6
Fr_(共 55 個)	46	7	1	1
Vr_(共 55 個)	50	3	2	1
Mr_(共 26 個)	20	5	1	0
Pr_(共 82 個)	72	17	2	1
AWL_V(共 55 個)	44	10	1	0
Threshold : Frequency > 140，頻率經正規化後大於 140 的動詞總數共有 160 個				
	H Bound (卡方值 0~303)	M Bound(卡方值 303~606)	L Bound(卡方值 606~910)	VL Bound(卡方值>>910)
Ar_(共 120 個)	86	23	5	6
Fr_(共 30 個)	25	3	1	1
Vr_(共 28 個)	24	2	1	1
Mr_(共 15 個)	12	2	1	0
Pr_(共 57 個)	45	10	1	1
AWL_V(共 35 個)	27	7	1	0

附錄表三 各領域動詞依指標交集而得的領域學術字彙列表

Frequency > 140，依分別指標且同一區間交集，字彙依同質性由高而低排列	
三領域交集候選詞 S(D*)實驗所得結果 (共 124 個動詞)	act, complete, vary, address, enter, reach, occur, conclude, form, view, employ, exhibit, produce, fail, study, limit, maintain, predict, leave, characterize, utilize, remain, lose, attempt, provide, receive, analyze, follow, explore, avoid, continue, lead, support, derive, meet, prove, contribute, imply, introduce, develop, grow, perceive, control, illustrate, fix, design, change, test, expect, record, involve, pay, start, improve, affect, hold, recognize, accord, play, observe, express, identify, consist, note, regard, depend, investigate, capture, demonstrate, collect, modify, serve, tend, reveal, evaluate, refer, concern, extend, draw, relate, combine, move, mention, add, discuss, focus, reflect, respond, represent, assign, exist, create, base, construct, result, call, select, measure, explain, determine, examine, implement, describe, assume, offer, apply, require, begin, choose, define, compare, achieve, reduce, obtain, set, perform, suggest, generate, run, feel, propose, increase, understand, learn, write
CS 領域候選詞實驗所得結果 (共 12 個動詞)	address, code, link, support, control, design, change, test, play, base, measure, suggest
ELT 領域候選詞實驗所得結果 (共 116 個動詞)	act, complete, vary, address, occur, conclude, form, view, employ, produce, fail, study, limit, maintain, link, predict, leave, characterize, remain, lose, carry, attempt, provide, receive, analyze, follow, explore, continue, lead, support, derive, meet, establish, contribute, imply, introduce, develop, grow, perceive, control, illustrate, design, change, test, expect, involve, start, improve, affect, hold, recognize, accord, play, observe, express, identify, note, regard, depend, investigate, demonstrate, collect, serve, tend, reveal, evaluate, refer, concern, extend, draw, relate, move, mention, add, conduct, discuss, focus, interpret, reflect, respond, represent, assign, share, exist, create, base, construct, result, call, select, measure, explain, determine, mean, examine, describe, assume, offer, apply, treat, require, begin, choose, define, compare, set, perform, suggest, feel, propose, associate, report, increase, understand, learn, write

MED 領域候選詞實驗所得結果 (共 97 個動詞)	complete, vary, address, enter, occur, cover, conclude, view, employ, produce, fail, study, limit, maintain, predict, leave, remain, provide, receive, analyze, follow, explore, continue, lead, support, derive, establish, contribute, introduce, develop, grow, control, design, change, test, expect, involve, pay, improve, affect, accord, observe, express, identify, note, regard, demonstrate, collect, serve, tend, evaluate, refer, concern, relate, combine, move, mention, add, conduct, discuss, focus, reflect, represent, exist, create, base, construct, result, select, measure, explain, determine, examine, implement, describe, assume, offer, apply, treat, require, begin, choose, define, compare, achieve, reduce, obtain, set, perform, suggest, generate, run, propose, report, increase, understand, write

附錄表四 最終選出之學術寫作上常用之字彙(綜合領域)

符號說明：

1. 字彙順序依照同質性由低而高排序
2. H 為同質性，F 為出現頻率。
3. 標註紅色之字彙為與 AWL 重複之字彙。

名詞部份	
High H (同質性數值 由 0~728)	methodology, degree, setting, confidence, series, manner, organization, support, addition, type, respect, program, comparison, transfer, stage, effectiveness, exception, line, principle, contrast, possibility, event, movement, contribution, frequency, definition, failure, demand, element, issue, limitation, variety, dimension, dissertation, strength, attempt, conclusion, position, paper, average, influence, body, evaluation, tool, author, basis, correlation, hand, relation, percentage, assumption, phase, collection, hypothesis, history, choice, effort, variation, expectation, person, unit, category, framework, importance, presence, technology, observation, reason, device, improvement, detail, classification, procedure, source, perception, direction, range, access, project, score, advantage, impact, center, base, product, scale, property, view, description, focus, difference, condition, reference, response, target, purpose, criterion, surface, change, interaction, amount, mode, version, background, field, test, percent, train, home, characteristic, length, table, measure, structure, feedback, instance, parent, situation, quality, error, concept, combination, nature, relationship, power, list, peer, interpretation, result, card, benefit, period, approach, expression, sequence, resource, aspect, material, selection, difficulty, decision, job, analysis, argument, session, plan, pattern, literature, constraint, research, participant, people, attention, evidence, component, theory, researcher, perspective, term, process, service, opportunity, world, goal, requirement, distance, ability, role, sense, technique, size, team, computer, account, item
Medium H (同質性數值	chapter, domain, rule, subject, idea, level, increase, pair, implementation, individual, step, development, experiment, variable, solution, sample, rate,

由 728~1456)	content, policy, task, activity, outcome, context, experience, distribution, behavior, practice, representation, form, effect, control, frame, information, factor, mechanism, trial, environment, performance, design, strategy, code, data, community
Low H (同質性數值由 1456~2184)	input, object, image, method, time, instruction, section, function, space, skill, population, knowledge, feature, action
Very Low H with High F (同質性數值遠大於 2184)	site(MED), figure(CS), network(CS), class(ELT), question(ELT), application(CS), set(CS), model(CS), system(CS), child (ELT), study (ELT), word (ELT), agent(CS), student (ELT)
動詞部份	
High H (同質性數值由 0~303)	act, complete, vary, address, enter, reach, list, code, occur, cover, repeat, conclude, form, view, distinguish, classify, employ, exhibit, summarize, produce, fail, study, divide, limit, maintain, link, predict, leave, characterize, utilize, remain, lose, carry, attempt, ensure, display, replace, provide, receive, analyze, follow, explore, avoid, continue, lead, support, derive, meet, establish, prove, contribute, imply, introduce, develop, grow, perceive, control, illustrate, fix, design, change, test, expect, record, involve, pay, start, improve, affect, hold, recognize, accord, play, observe, express, identify, consist, note, regard, depend, investigate, capture, demonstrate, collect, modify, serve, tend, reveal, evaluate, refer, concern, extend, draw, relate, combine, move, mention, add, conduct, discuss, focus, interpret, reflect, respond, represent, assign, share, exist
Medium H (同質性數值由 303~606)	create, base, construct, result, call, select, measure, explain, determine, mean, examine, include, implement, describe, assume, offer, apply, treat, require, begin, choose, define, compare, achieve, reduce, obtain
Low H (同質性數值由 606~910)	set, perform, suggest, generate, run, feel
Very low H, with high F (同質性數值遠大於 910)	propose(CS), associate(MED), report(MED), increase(MED), understand(ELT), learn(ELT), write(ELT)

附錄表五 最終選出之學術寫作上常用之字彙(綜合語言特性)

符號說明：

1. 字彙順序依照同質性由低而高排序
2. H 為同質性，F 為出現頻率。
3. 標註紅色之字彙為與 AWL 重複之字彙。

名詞部份	
High H (同質性數值由 0~194)	aspect, feedback, image, answer, improvement, thesis, element, approach, treatment, factor, difference, link, direction, content, table, task, issue, product, interaction, matrix, protein, people, function, understand, reference, term, technique, category, region, age, performance, video, time, computer, concept, interpretation, analysis, comparison, phase, role, purpose, hypothesis, verb, course, development, person, definition, speaker, situation, measure, period, position, process, topic, line, correlation, step, ratio, list, node, procedure, idea, background, domain, recognition, rule, scale, frame, session, control, section, form, frequency, event, program, attention, support, solution, observation, context, rate, example, change, relationship, input, reason, effect, sequence, code, component, material, theory, skill, text, error, condition, class, experiment, character, practice, figure, activity, length, lesson, interview, mechanism, network, population, parent, data, distribution, property, level, speech, sample, experience, value, type, finding, vector, cost, story, goal, size, unit, weight, score, difficulty, space, stage, ability, language, design, feature, world, path, information, behavior, pattern, object, question, parameter, pair, train, algorithm, subject, response, communication, result, action, structure, discussion, mean, environment, research, chapter, set, item, variable, distance
Medium H (同質性數值由 194~388)	study, knowledge, target, sentence, learn, system, instruction, relation, participant, user, cell

Low H (同質性數值 由 388~582)	child, researcher, model, test, strategy, learner
Very Low H with High F (同質性數值 遠大於 582)	method(ELT), expression(ELT), teacher(ELT), student(ELT), vocabulary(ELT), word(ELT)
動詞部份	
High H (同質性數值 由 0~108)	examine, process, assess, involve, combine, speak, contribute, affect, extend, avoid, discuss, hold, refer, suggest, form, achieve, derive, feel, explain, receive, report, understand, apply, design, concern, recognize, collect, illustrate, compute, measure, treat, estimate, reflect, lead, tell, limit, study, focus, occur, follow, compare, share, cause, construct, call, increase, change, improve, relate, include, consider, set, analyze, generate, perform, demonstrate, regard, add, identify, represent, require, solve, means, reduce, contain, produce, exist, employ, expect, look, consist, start, choose, introduce, determine, support, select, move, reveal, note, express, result, appear, define, conduct, play, assume, begin, develop, create, investigate, mention, observe, indicate, evaluate
Medium H (同質性數值 由 108~216)	associate, accord, obtain, calculate, provide, base, describe, help, allow
Low H (同質性數值 由 216~324)	teach, read
Very low H, with high F (同質性數值 遠大於 324)	propose(ELT), write(ELT), learn(ELT)

附錄表六 學術寫作上字彙之常用搭配詞(整體)

詞性說明：

1. 本附表資料為整個 ATC 語料庫 420 篇文章之統計結果，依照使用頻率及多樣性分成兩個子表格。
2. S. (主詞)，為搭配詞核心字彙。
3. O. (受詞)，以名詞居多。
4. DT (定冠詞)，如 a / the。
5. 其他為 Noun(名詞)、Verb(動詞)、Prep.(介系詞)、Adj.(形容詞)、Adv.(副詞)、Pronoun(代名詞)

附錄表六之一　　高頻率搭配詞組合(出現頻率大於 200)

<table>
<tr><td colspan="4">[1.] Prep. + (DT) + (Adj.) + S.(Noun) + Prep.</td></tr>
<tr><td>as</td><td>as</td><td>as</td><td>as</td></tr>
<tr><td colspan="4">[2.] V.(be) + S.(V-ed) or Adj. + Prep.</td></tr>
<tr><td rowspan="3">be</td><td>related/associated/required/</td><td colspan="2">to/with</td></tr>
<tr><td>likely</td><td colspan="2" rowspan="2">to</td></tr>
<tr><td>due</td></tr>
<tr><td colspan="4">[3.] Prep. + (DT or Pronoun) + (Adj.) + S.(Noun) + (Prep.)</td></tr>
<tr><td rowspan="2">in</td><td rowspan="2">the/this/other</td><td>terms/context/field/development/course/form</td><td>of</td></tr>
<tr><td colspan="2">words/section/case/study/way</td></tr>
<tr><td>of</td><td>the</td><td>previous</td><td>study</td></tr>
<tr><td>at</td><td>the</td><td>same</td><td>time</td></tr>
</table>

		begin	of
for	the	purpose	of
		most	part
[4.] Prep. + (DT or Pronoun) + S.(Noun or Verb) + Prep.			
in	in		in
as	as		as
as	as		as
with	with		with
[5.] DT + S.(Noun) + Prep.			
the	purpose/majority/difference/performance/time/means/relationship/case/context/number/result/role/absence/importance/remainder/		of/between
[6.] Others			
seems	seems		seems
showed	showed		showed
be	be		be
because	because		because
more	more		more
most	most		most
whether	whether		whether
average	average		average
played	played	played	played

附錄表六之二　　多樣性搭配詞組合 (出現頻率大於 50)

[1.] Prep. + (DT) + S.(Noun) + Prep.			
	on/with/as/by/from	result	of
[2.] (Prep. + DT) + Adj. + (DT) + S.(Noun) + (Prep.+ O.)			
in the	broad/great/superior/ sufficient/growing	knowledge	of
	pilot/previous/few/several/recent/current	study	
	wide/large/broad/huge/rich/ abundant	variety + of	way/reason/language
	top/null/major/alternative	hypothesis	
	quantitative/statistical/objective/accurate/full/precise	analysis	
	big/obvious/main/tiny/ statistical	difference	
	effective/suggested	method	
	primary/general/original/ specific/final	aim	of
	comprehensive/intensive/ systematic/academic/ ongoing/present/empirical	research	
[3.] V. + (DT) + S.(Noun) + (Prep.+ O.)			
	construct/test/support/ propose/modify	hypothesis	
	make/perform/provide/ require/conduct	analysis	
	formulate/give/adopt/follow /suggest/change	definition	
	answer/ reply to/deal with/solve/address	question	
	find/advance/explain/form	cause	
	handle/discuss/ show/identify	problem	
[4.] S.(Noun) + V. + (DT + O.)			

	data	show/provide/imply/prove/ indicate/support	
	result	indicate/show/suggest/ present/reveal/summarize	
	study	examine/demonstrate/aim/pr ovide/show/suggest/investig ate/use/conduct/reveal/find	
[5.] (N.) + S.(Verb) + Adv.			
	increase	significantly/slowly/greatly/ gradually	
	compare	directly/statistically/well to/with	
	contrast	clearly/strongly against/to/with	
[6.] (N.) + S.(Verb) + Prep.			
	focus	on/of/upon	
	vary	form...to.../with/in/ among/according to	
	agree	with/to/upon	
[7.] N. + (Prep.) + S.(Noun)			
	degree/area/pattern/amount	of + difference	
	set/huge amount/ interpretation/source	of + data	
	research/test/laboratory/ survey	instrument	
[8.] S.(Verb or Noun) + N.			
	research	center/group/site/subject/res ult/problem/tool/procedure	
	yield	conclusion/outcome/result/b enefit/relationship/evidence	

附錄表七 學術寫作上字彙之常用搭配詞(依語言特性)

詞性說明：

1. 本附表資料為整個 ATC 語料庫 420 篇文章之統計結果，附錄表七之一為 EFL 作者(台灣及日本)之常用搭配詞，附錄表七之二則為 Native Speaker 作者(美國) 之常用搭配詞。
2. S. (主詞)，為搭配詞核心字彙 ； O. (受詞)，以名詞居多 ； DT (定冠詞)，如 a / the。
3. 下列兩表中標註紅色部份為附錄表六中並未出現的搭陪詞使用方式。
4. 其他為 Noun(名詞)、Verb(動詞)、Prep.(介系詞)、Adj.(形容詞)、Adv.(副詞)、Pronoun(代名詞)。

附錄表七之一　　EFL 作者常用搭配詞組合(出現頻率大於 50)

[1.] S.(Noun) + Verb + that		
result	indicates	that
table	shows	
finding	suggests	
[2.] V.(be) + S.(V-ed) or Adj. + Prep.		
be	based	on
	defined	as
	regard	
	applied	to
	related	
	necessary	

	due		
[3.] V. + Prep.. + DT(the)			
correspond	to		the
depend/focus	on		
[4.] Prep. + (DT or Pronoun) + (Adj.) + S.(Noun) + (Prep.)			
in	the/this	method	of
	the	paper	
		present/following/first	target/subject
[5.] Prep. + (DT or Pronoun) + S.(Noun or Verb) + Prep.			
in	order/addition		to
as	well		as
as	showed		in
with	respect		to
[6.] DT + S.(Noun) + Prep.			
the	number/result/effect/development/importance/relationship/performance/effectiveness/purpose/process		of/between
[7.] DT(the) + Noun or Adj. or V-ed+ S.(Noun)			
the	target		language
	proposed		algorithm/method
	vocabulary	learning	strategy
	short	term	memory
	self	efficacy	belief
[8.] Others			
in	which		the
seems	to		be

showed	in		table
a	set/number		of
because	of		the
it	be		possible
average	waiting		time
that	there		be
one/all/most	of	the	participant/target/subject
on	the	other	hand

附錄表七之二　　NS 作者常用搭配詞組合 (出現頻率大於 50)

[1.] Prep. + (DT) + S.(Noun) + Prep.			
on/with/as/by/from	(a/the)	set/number/variety/function/ result/series/total/range/list/lots	of
[2.] (Prep. + DT) + Adj. + (DT) + S.(Noun) + (Prep.+ O.)			
(in the)	broad/great/superior/ sufficient/growing/ common/scientific/prior	knowledge	of
	previous/several/recent/ current/present/social	study	
	wide/rich	variety + of	way/reason
	top/null	hypothesis	
	objective/full/precise/ detail/meta	analysis	

	big/obvious/main/statistical/ significant/only	difference	
	effective/suggested/proposed/ accessing/hybrid	method	
	primary/general	aim	of
	comprehensive/ further/additional	research	
[3.] V. + (DT) + S.(Noun) + (Prep.+ O.)			
	construct/propose/ draw/retain/adopt	hypothesis	
	provide/require/ carry out/need	analysis	
	give/adopt/suggest/ fulfill/meet/alter	definition	
	answer/deal with/ solve/address/ work out/comprehend/evade	question	
	find/attribute	cause	
	handle/discuss/show/ cope with/look into/ create/point out/avoid	problem	
	verify/cite/request/ignore	evidence	
	make/base on/ hold/have/challenge	assumption	
[4.] S.(Noun) + V. + (DT + O.)			
	evidence	demonstrate/confirm/ contradict/support	
	research	estimate/offer/disclose/ prove/yield	
	information	provide/contain/include	
	data	show/provide/imply	
	result	indicate/show/suggest/reveal	

	study	examine/demonstrate/show/ suggest/reveal/ produce/look at/utilize/last	
[5.] (N.) + S.(Verb) + Adv.			
	compare	directly/with	
	contrast	clearly to/with	
	vary	systematically	
[6.] (N.) + (be) + S.(Verb or Noun or Adj.) + Prep.			
(be)	focus	on/of/upon	
	vary	with/in/according to	
	agree	with/to/upon	
	account	for	
	compare	to/with	
	consist	of/in/with	
	explain	to/about/in	
	aim	at/for	
	serve	as	
	refer/seem	to	
	design/apply/expect/ require/relate	to/with	
	participate	in	
	due	to	
	consistent	with	
	difficult/possible/similar	to	
[7.] N. + (Prep.) + S.(Noun)			
	degree/area/significance/ amount/magnitude	of	difference
	interpretation/source/role/nature/level/problem/quality/size		data

	/form/concept/goal/state		
	kind/basis/probability/lack/ rest/purpose/cost/impact/ distribution/content/benefit/ idea/version/value/type		experiment
[8.] S.(Verb or Noun) + N.			
	research	center/group/site/subject/result/ problem/tool/procedure	
	yield	conclusion/outcome/result/ benefit/relationship/evidence	
	apply	theory/knowledge/ formula/standard	
	contribute	perspective/information/ paper/time	
[9.] Prep. + DT + S.(Adj. or Noun) + Prep.			
as	a	result	
in		absence/presence/form	
within		context	
with		exception	of
at	the	begin/end/time	
beyond		scope	
by		end	
from		point	
[10.] Others			
on	the	other	hand
at	the	same	time
for	the	first	time
with	the	increasing	task