漢語學術詞表提取

摘要：自 1953 年至今，在英語學習領域有關於詞表提取的相關研究已經形成了相當成熟的體系，尤其是針對於學術語料庫的學術詞表提取方法已形成了兩個流派，並針對不同學科也產生了許多成果。但類似的研究在漢語領域卻顯得略有不足，例如在詞彙學習的過程中我們並不會產生專門學習學術詞彙的概念，以及現有的相關研究相比英文領域比較落後，研究成果較為少見。但事實上，學術詞彙作為學術領域中十分重要的一部分，每一位學術研究人員都應當進行一定程度的學習，因此需要一份新學術漢語詞表來填補這一空白。我們根據 2021-2022 年中文社會科學引文索引（CSSCI）收集了 581 種權威中文期刊收錄的論文元資訊，構建了一個上億字元的學術語料庫，並以使用頻次、跨文本分佈、詞語均勻分佈值三個參數作為特徵與篩選條件，並在先前學術界研究的基礎上對使用特徵進行了進一步的改進，希望能夠得到在學術語料中“更重要”的詞彙詞表。我們最後應用計算語言學方法提取出了一個基於 2019-2022 年全部期刊論文摘要資料提取得到的通用學術漢語詞表，以及基於法學、管理學、教育學、語言學四個學科領域的最早自 1954 年以來至 2022 年發表論文的摘要資料組成的子語料庫提取得到四個特殊用途學術漢語的學科專用詞表，並在這五個詞表的基礎上使用漢語停用詞表（stop words）對重複的詞彙進行了排除，得到了刪除停用詞和不刪除停用詞兩個版本的結果，最後從語料庫覆蓋率、詞彙排序與重複詞分析三個方向與現有通用詞表對比進行了驗證。我們認為新學術漢語詞表的創新點不僅僅是學術漢語領域學術語料庫規模的提高以及通用學術詞表和多

個學科專用詞表的提出，同時也體現在包含英文的國際學術詞彙研究領域對詞語均勻分佈值這一指標改進演算法的應用。我們相信新學術漢語詞表的建設在詞彙研究、語料庫語言學、英語教學等多個領域都能給予相當的幫助和參考。

關鍵字：語料庫；學術詞彙；詞表；特徵提取

Extraction of Chinese Academic Vocabulary

Abstract: Since 1953, the related research on word list development has formed a fairly mature system in the field of English study. Especially for academic word list development based on academic corpus, two schools of academic word list development method have been formed, and many achievements have been made for different disciplines. However, similar studies in the field of Chinese seem to be slightly inadequate. For example, in the process of vocabulary learning, we will not realize to learning academic vocabulary specially, and the existing related studies are relatively backward compared with the English field, with relatively few results. In fact, as an important part of the academic field, academic vocabulary should be studied to a certain degree by every academic researcher. Therefore, a new Chinese academic vocabulary is needed to fill this gap. Based on the Chinese Social Sciences Citation Index (CSSCI) from 2021 to 2022, we collected meta-information of 581 authoritative Chinese journals, constructed an academic corpus with hundreds of millions of characters, used Frequency, Range, and Dispersion as the characteristics and screening conditions, and further improved the usage characteristics on the basis of previous academic research. We hope to find "more important" words based on academic corpus. Finally, we used computational linguistics method to developed a Chinese general academic word list based on the abstract data of all journal papers from 2019 to 2022, and four discipline word list which is Law, Management, Pedagogy and Linguistics based on sub-corpus from 1954 to 2022. And we remove repetitive words by Chinese stop words on these five word-list. The results of deleting the stop words and not deleting the stop words are obtained. Then,

the results are compared with the existing general word lists in three directions: the coverage of the corpus, the sorting of words and the analysis of repetitive words. We think that the innovative of the new Chinese academic word list is not only the improvement of the size of the academic corpus as well as the development of the general academic word list and the word list for many disciplines, but also the application of the improved algorithm to the index of word distribution value in the research field of international academic vocabulary. We believe that the construction of the new Chinese academic word list can provide considerable help and reference in many fields such as vocabulary research, corpus linguistics and English teaching.

Key words: Corpus; Academic Word; Word List; Feature Extraction

目錄

圖表目錄

1 緒論

1.1 研究背景

隨著電腦時代的不斷發展，各個研究領域都不同程度的因此而產生相應的改變，而語料庫語言學也正是在這股時代的浪潮中出現並不斷發展的。迄今為止，語料庫語言學已經應用在語言學領域的方方面面，同時因為其可以將定量分析與定性分析相結合的研究優勢也深受研究人員的歡迎，因此產生了許多優秀的研究成果。而詞典編撰和詞表提取方向，可以算是語料庫語言學研究應用最早，最重要的方向之一。

所謂語料庫，根據(杨惠中 & 卫乃兴, 2002)的定義，指的是“一個由大量的語言實際使用的資訊組成的，專供語言研究、分析和描述的語言資料庫，在電腦網路技術和資訊技術快速發展的現代社會，語料庫主要指經科學取樣和加工的大規模電子文本庫”。同時現如今有關語料庫的研究已經深入到諸多方面，例如詞典編撰(Rundell 等, 2009; 朱冬生, 2009)、語義標注(林丽, 2013)、句法分析(荀恩东等, 2016)等，同時其不限於語言學領域，在電腦領域像情感分析(Alnawas & Arıcı, 2018)、機器翻譯(柏晓静 等, 2002)，文本生成(Brow 等, 2020)等方向，語料庫都是其不可或缺的重要組成部分。例如近期火熱的 chatgpt，其底層原理語言模型就是在一個原大小 45TB 資料處理後大小為 570GB 文本的語料庫基礎上訓練得來的(Brown 等., 2020)。可以說只要包含自然語言處理的相關知識的領域裡，都能見到語料庫的身影。現在全球範圍內多種語言都已經產生了各自的語料庫，像英語(Kennedy & Ooi, 1998)、漢語(荀恩东等, 2016; 詹卫东等, 2019; 周强, 2004)、日語(Lee & Kesuke, 2017)、土耳其語(Ozkan, 2013)等等，並且隨著相關研究

的不斷發展，其規模不斷擴大，定義愈發完善，分類與格式也呈現出不斷精細化的趨勢。

而詞典編撰、詞表提取作為在語言學領域語料庫應用最重要的方向之一，也產生非常多優秀的成果。在英語研究領域關於詞表的研製自二十世紀五十年代就開始了，且隨著語料庫知識的引入以及電腦算力的不斷增長，其研製過程與評價指標也隨之不斷的變化，從早期 MichaelWest (1953)的通用英語詞表（The General Service List，GSL），到第一個在英美高校被廣泛使用，由 Xue & Nation (1984)研製的大學單詞表（The University Word List，UWL），再到在學術界產生了極大反響的 Coxhead (2000)學術單詞表（Academic Word List，AWL）。到了近代，詞表研製領域的研究成果在 Coxhead 詞表研製方法的影響下顯示出專用性更強，研究目的更加細緻的特點，產生了針對不同語言，不同模態，不同學科，不同目的的各種詞表，同時使用的自然語料庫規模也在不斷擴大，詞表的評價指標也更加完善。和英語領域的詞表提取研究相比，中文領域的詞表研究較為落後，雖然也有相對應的成果出現，但從研製方法而言仍存在一定的差距。

1.2 研究問題

1. 構建人文社科類的學術詞表需要什麼樣的語料？

2. 採用什麼方法來構建學術類的學術語料庫？

3. 從哪幾個角度研究驗證詞表的有效性？

4. 人文社科學術詞表和現代漢語通用詞表的差異有哪些？

1.3 研究意義和目的

基於中文學術語料庫的中文學術詞表提取研究不僅有助於中文研究緊跟國際學術詞彙研究步伐，同時在各個應用領域也有廣泛的應用前景。基於語料庫的學術詞表提取研究在英文領域已經產生了相當多具有影響力的研究成果，並且在英語教學、特別用途英語等領域都產生了一定的應用。相比而言，中文有關學術詞彙的相關研究則略顯滯後，現有的中文詞表研究中，尚未存在一個被廣泛認可的學術詞彙表，學術領域大部分仍在使用通用詞表，在學術詞彙這一細分領域研究略顯不足。現有有關中文學術詞彙詞表提取的研究中，在語料庫規模和詞彙篩選特徵兩個方面和英文領域相比都仍存在一定差距。

因此我們綜合現有的研究成果，自主構建了一個上億字元的學術語料庫，在進行資料處理後，採用多個特徵提取的方法篩選出我們認為在學術語料中最重要的詞彙，從一個更加資料化，精細化的視角出發，提取出新的學術漢語詞彙表。而我們除了通用學術漢語詞表外，選擇建立的法學、管理學、教育學、語言學四個學科專用學術詞表都為人文社科類的學科的原因主要有兩點。第一是因為現有的漢語學科專用學術詞表成果大部分為人文社科類學科例如語言學、經貿學等。因此對於漢語研究我們是應用我們詞表提取流程進行學科專用詞表提取，起到對現有研究成果起到一個完善作用。第二是鑒於收據學術語料資料的困難程度，我們只完成了法學、管理學、教育學、語言學這四個學科所有核心期刊自電子版發刊時間起至 2022 年 8 月全部文獻資料的收集。我們研究首先希望的是能創立一個較為標準的研究範式，並將其應用在通用以及學科專用學術詞表的創建這一方式填補中文領域在這一部分的研究空白，推動學術詞彙研究的進一步發展。並且我們參考英文學術詞表的相關應用情況認為，雖然漢語學術詞

表這一研究成果對於漢語作為第二外語的學習者們而言難度有些大，並不能代替 HSK 等常用漢語詞彙表在對外漢語教學領域的作用，但其對於想要在對漢語進一步學習或者想要成為漢語老師的學習者而言仍然是非常優秀的補充材料，具備一定的應用前景。

1.4 研究設計

本實驗綜合了現有的基於語料庫語言學提取學術詞表的研究成果，在學術詞彙研究領域創建了一個大資料學術語料庫，並改進了提取學術詞彙表的實驗設計，首先我們將 2021-2022 年中文社會科學引文索引（CSSCI）作為我們學術語料的來源收集、整理、清洗最後得到了一個總字元量超過一億字元的漢語學術語料庫，接著使用 pku_seg 代碼包進行分詞、詞彙特徵提取、篩選得到五個不同目的的學術詞彙表、最後進一步進行詞彙表去除停用詞、現有通用詞表對比等驗證工作。具體實驗流程如下：

1. 創建學術語料庫：首先我們確定了我們學術語料的來源，中文社會科學引文索引（CSSCI）是目前中文學術界公認的最重要，最權威，應用最為廣泛的期刊索引表，並且學術期刊論文一直是學術語料最重要的來源之一。同時我們認為它具備了整個中文學術語料不同類別科目的原始權重資訊，因此我們可以避免考慮到收集到不同學科大量學術語料後的子語料庫平衡問題。整個 CSSCI 包含了 583 種期刊，以及臺灣期刊 30 種，報紙理論版 2 種，並將其分為了中國文學,人文經濟地理,體育學,冷門絕學,歷史學,哲學,圖書館、情報與文獻學,外國文學,宗教學,心理學,政治學,教育學,新聞學與傳播學, 民族學與文化學,法學,社會學,管理學,經濟學,統計學,綜合性社會科學,綜合性高校學報,考古學,自然資源與環境科學,藝術學,語言學,馬克思主義理論,高校社科學報一共 27 個種類。我們以中文知網（https://www.cnki.net/）作為期刊的語料的詳細資訊來源，收集

了其中部分期刊自有電子版期刊資料出版以來發表的全部論文摘要語料構成了我們的漢語學術語料庫。

2. 建立學術漢語詞彙表：收集到的資料都為非結構化資料，因此我們需要對收集到期刊和論文的元資訊進行整理，清洗，最後得到僅剩全部摘要資料。接著我們使用目前較為先進的由北京大學語言計算與機器學習研究組研製的北大開源中文分詞工具包 pku_seg 來對摘要語料資料進行分詞處理。接著對經過分詞得到的資料進行頻次、跨文本分佈、詞語均勻分佈值三個特徵的提取，接著綜合現有的文獻從這三個維度來進行詞彙篩選，根據不同語料得到我們不同使用目的的學術漢語詞彙表。

3. 停用詞去除與詞表驗證：在得到我們新建立的幾個學術詞表後，我們進行了一個自然語言處理常用的停用詞去除步驟，將百度停用詞表、哈工大停用詞表和四川大學智慧實驗室停用詞庫綜合使用，得到了學術詞表去除停用詞與不去除停用詞的兩個版本。接著我們從詞表語料覆蓋率、詞彙排序與重複詞分析三個方向與通用詞表對比來對我們的學術詞彙表進行了進一步分析。

1.5 小結

本文主要根據現有的研究背景，自主構建了一個大型學術語料庫，在進行分類與資料處理後，進一步改進了現有的詞彙篩選特徵，提取出新的學術詞彙表，並對詞彙表進行了驗證。

本文主共分五章，主要研究流程圖如圖 1-1 所示：

第一章緒論，主要簡述了有關學術詞表提取的研究背景，研究成果的目的和意義以及研究的大致實驗設計。

第二章文獻綜述，主要介紹了語料庫與學術詞表提取的歷史與相關研究，顯示了中英文兩種語言在學術詞表提取領域的差距。

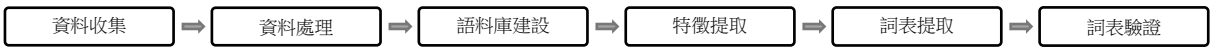
第三章語料庫建設，主要介紹了自主建設學術語料庫的詳細構成例如子語料庫情況、期刊數、摘要數等，以及相關的組建過程。

第四章詞表提取，主要介紹了使用如何進行分詞處理，以及頻次，跨文本分佈，詞語均勻分佈值三個指標的意義以及相關篩選標準，演算法的選擇細節。

第五章結果與討論，首先進行了去除停用詞表處理得到了我們的全部學術詞表成果，接著我們將學術詞表與現有的常用通用詞表進行了對比來進一步分析。

第六章結語，對我們的實驗流程與實驗結果進行了簡述，接著討論了整個實驗的創新點與不足之處，指出了未來進一步研究的方向。

圖 1-1 實驗流程圖



2 文獻綜述

2.1 語料庫研究

2.1.1 國外語料庫研究

自十七世紀開始英語語言學家們受經驗主義影響就開始關注真實語料在語言學研究的作用，而後美國結構主義語言學即實證主義和行為主義思潮進一步推動了經驗主義在語言研究的地位，於是包含大量真實語料的語料庫隨之出現，但再接著是喬姆斯基轉換生成語法理論的興起，語料庫研究進入低谷期，但隨著電腦算力的不斷發展，其相關研究又進入了新的高峰期，現如今語料庫研究已經深入語言學及其應用研究的方方面面，成為語言學研究中不可缺少的重要部分。(何中清 & 彭宣维, 2011)

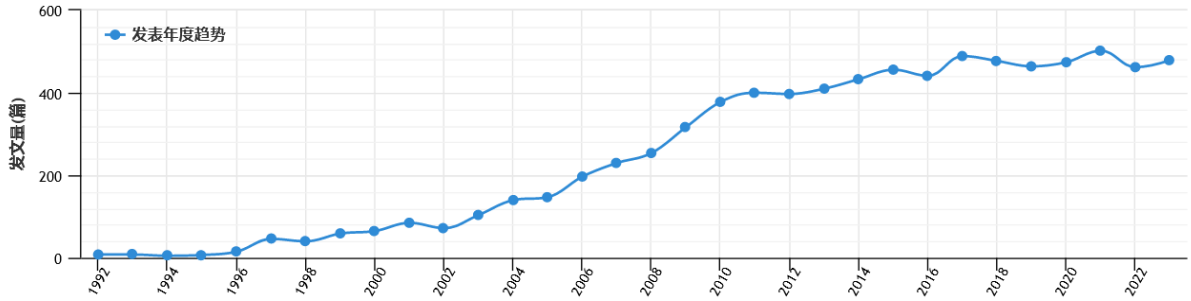
早期語料庫和現代語料庫的差異是巨大的，在二十世紀五十年代以前的早期語料庫製作基本依靠人工進行篩選，記錄，匯總製作效率低，週期長，在體諒有限的同時手動記錄對人體的健康來說都有著相當大的負擔。(冯志伟, 2002)到了 20 世紀六十至七十年代出現了第一代電子語料庫：BROWN、LOB 和 LLC 三大經典語料庫(Kennedy & Ooi, 1998)。第一次使用電腦儲存語料庫，大大簡化了語料庫使用、保存和移動的步驟。同時還對語料庫進行了標注，為電腦與語料庫交互的進一步發展奠定了基礎。到了 20 世紀 80 至 90 年代，電腦儲存成本進一步下降，運算處理能力加強，同時隨著光學字元辨識（OCR）技術的發展，大批量轉化處理文本成為現實，語料庫規模開始擴大、同時對語料保存的標準也開始逐漸統一。(冯跃进 & 潘璠, 1998)突出的像 CO-Build、BNC、ICE 等語料庫，語料庫規模達到上億字元(Kennedy 1998)，此時的語料庫以通用語料為主，語料庫內部進行了不同類型語料的分類、且保存格式一致方便對其進行進一步處理。

現如今語料庫的相關研究更加成熟，新一代語料庫，體量更大，分類更成熟，且已經應用到語言研究的方方面面。隨著電腦的進一步發展，現在英文領域語料庫規模大都達到數億甚至數十億級別，如英國國家語料庫“British National Corpus”(BNC)、當代美國英語語料庫“the Corpus of Contemporary American English”(COCA)。分類也更加細緻，例如區分了不同地域英語使用的基於網頁全球英語語料庫“Corpus of Global Web-Based English”（GloWbE），以及專門提供歷時性研究的美國歷史英語語料庫“The Corpus of Historical American English“（COHA），同時他們都將資料儲存在伺服器，提供網頁版本的對話模式並集成在www.english-corpora.org，為全世界廣大英語語言研究者提供更方便查詢和使用語料庫的方式。語料庫研究也已經深入語言研究的各個領域，除了線上大型語料庫為對詞彙使用，地區英語差異研究的學者提供了大量語料，雙語語料庫的出現在機器翻譯領域也產生了許多成果，以及詞典建立也許要語料庫研究方法的參與，包括牛津，科林斯，朗文，劍橋等多個詞典等等(Ozkan, 2013)。除了語料庫本身，語料庫的發展還促進了產生了許多相關軟體和處理方法成果的誕生，像對著名的antconc 軟體系列，由 Lancaster 大學研製的#LancsBox，以及各種對語料庫進行加工的技術例如自動切分、自動標注等演算法，具體的例如 CLAWS 演算法，HMM（隱瑪律可夫模型），神經網路等(冯志伟, 2002)。同時在發展的還有關於語料庫相關理論例如語料庫本體論的研究，學界對於如何正確認識語料庫語言學區分成了基於語料庫（Corpus-based）與語料庫驅動（Corpous-driven）兩個流派(Tognini-Bonelli & Elena, 2001)。可以說，語料庫語言學是英語語言學研究領域不可分割的一個部分，“毫不誇張地說，語料庫及語料庫研究在過去幾十年裡給語言研究以及語言應用研究帶來了一場革命性的變化”(Hunston, 2003)。

2.1.2 國內語料庫研究

而對於中文領域的語料庫研究而言，學術領域認為最早應用語料庫思想的研究成果是著名教育學家陳鶴琴先生于 1925 年編撰而成的《語體文應用字彙》，雖然當時還沒有語料庫的概念，但陳先生通過統計方法收集了總計六十萬字的語料創建了按漢字在語料中通過絕對頻次排序的“字數次數對照表”。(冯志伟, 2002)雖然中文語料庫的起始出現時間很早，但隨著時間推移中文語料庫的發展與應用較英文而言比較緩慢。慶倖的是，近年來語料庫的相關研究逐漸火熱，我們在中國知網資料庫中以“語料庫”作為關鍵字，“核心期刊”“cssci”“cscd”作為篩選條件，檢索出了 7672 條結果，按年劃分的發表論文數量視覺化結果如圖 2-1 所示。

圖 2-1 國內語料庫文獻年度分佈



鑒於語料庫應用的普遍性，語料庫研究的相關成果也顯示出多領域，應用廣的特點。例如在語言教學領域，由基於漢語或英語語料庫的輔助課堂教學(娄宝翠, 2020; 郑艳群, 2013)，利用雙語語料庫的翻譯課堂教學(韩露等, 2017)。在詞典編撰以及詞表提取領域，基於大規模語料庫可以為編撰者提供具體詞彙查找、詞頻表、搭配檢索、句法框架等相關知識的實證案例(荀恩东等, 2016)，而這些成果都可以應用在不同的學科中。還有應用於情感分析領域情感詞典的創建(饶洋辉等, 2014)、以及在其他領域例如微博、

新聞、汽車評論都有利用語料庫進行文本分析的範例(苗祥等, 2014)。再有在資訊檢索和語言對比翻譯研究中的應用等等(孙东云, 2018)。(黃水清 & 王东波, 2021)

語料庫的應用如此廣泛，在中文領域自然也出現了非常多的語料庫，各大院校基本都創建了自己研究的相關語料庫。我們介紹了目前比較具有代表性的幾個語料庫。首先是北京語言大學語料庫中心創建的 BCC 語料庫（BCC 語料庫 (blcu.edu.cn)），語料庫以漢語為主，還兼顧了英語、法語、西班牙語、土耳其語等語言，總規模達數百億字，同時還創立了線上檢索系統提供給大家線上使用(荀恩东等, 2016)。還有北京大學的 CCL 語料庫（CCL 語料庫檢索系統（網路版） (pku.edu.cn)），包含了古代漢語，現代漢語以及漢英句對其平行語料，其中現代漢語大約六億字元，古代漢語大學兩億字元，包括了從周代到民國的相關語料，以及不方便按照朝代劃分的雜類語料例如大藏經、全唐詩，道藏等(詹卫东等, 2019)。以及清華漢語樹庫，其中語料都經過了自動斷句、標注，形成了有完整句法結構的句法樹庫語料庫(羅振聲, 1996)。除此之外，還有很多像國家語委現代漢語通用平衡語料庫(黃水清 & 王东波, 2021)、中國科學院漢英平行語料庫(薛松, 2003)、漢語仲介語語料庫(张宝林, 2019)、HSK 動態作文語料庫、少數民族語料庫(斯日古楞, 2010)也值得一提。

2.2 學術詞表提取研究

2.2.1 學術詞彙

根據 Wang、Paul (2004)的定義詞彙可以分成四類：首先是以 GSL（general service vocabulary）為代表的高頻詞彙，例如 measure, small, wealth 等；接著是學術詞彙(academic vocabulary)，例如 assist, economy, complex 等，再到一般與某一特定領域聯繫緊密的技術詞彙(technical vocabulary)，例如 semantically, psychologically 等，最後是在文本

中較少出現的低頻詞彙，例如 obliteration, panjandrum 等。早在 1934 年，研究者們就已經注意到有一些詞彙在不同學科的學術文本中均有出現，而這一類在當時被稱之為“子技術詞彙”或“半技術詞彙”的詞在現在被深入研究並統稱為“學術詞彙”(Coxhead, 2000; Dresher, 1934; Farrell, 1990; Flowerdew, 1993; Wang & Paul, 2004; Yang, 1986)。同時學術詞彙知識一直被認為是學術閱讀能力中不可或缺的重要組成部分(Biemiller, 1999; Corson, 1997; William 等, 2012),其與學術成就，經濟機會等直接相關(Gardner & Davies, 2014; Goldenberg, 2008; Ippolito 等, 2008; Jacobs, 2008)。在中文領域也存在類似的定義，高增霞、刘福英 (2016)指出學術漢語是訓練學生用漢語從事專業學習和學術活動的漢語教學，可分為通用學術用途漢語和特殊學術用途漢語，其中通用學術用途漢語的詞彙研究就與英語中學術詞彙的研究相類似。(王笑然 & 王佶旻, 2022)

2.2.2 英語學術詞表

詞典編撰、詞表提取作為詞彙研究中最重要的應用之一，而以學術文本為研究物件，學術詞彙為中心的詞表提取在英文領域也已經產生了非常多優秀的研究成果。(Coxhead, 2000; Gardner & Davies, 2014)同時隨著電腦技術的不斷發展，進行學術詞表研究的方法也不斷更新，其語料庫規模和判斷詞彙時採用的評價指標也不斷進步，在詞彙研究領域展現出新的活力。早期的研究例如 MichaelWest (1953)在 1953 年開發的通用英語詞表 GSL（The General Service List，GSL），從一個五百萬詞的語料庫得出，包含了 2000 個詞族；第一個被學術界廣泛應用的詞表是由 Xue＆Nation (1984)在 1984 年開發的大學單詞表 UWL（University Word List），當代最具影響力的由 Coxhead (2000)開發的學術詞彙表 AWL（Academic Word List）就是在 UWL 的基礎上得出的。而 Gardner 和 Davies 研製新的學術詞彙表 AVL（Academic Vocabulary List）與 AWL 在是否排除通用學術詞表

和是使用詞族（word family）還是詞元（lemma）兩個方面的不同形成了對比，也引起了學界的關注。現如今學術詞表研究領域基於 AWL 和 AVL 兩種方法形成了兩個流派，兩者都產生了非常豐富的研究成果(Chen & Ge, 2007; Coxhead 等, 2010; Dang & Webb, 2014; Hsu, 2018; Hyland & Tse, 2007; Lei & Liu, 2016; Li & Qian, 2010; Liu & Han, 2015; Martinez 等, 2009; Munoz, 2015; Valipouri & Nassaji, 2013; Wang 等, 2008; Ward, 2009; Yang, 2015)。

由 Coxhead 在 2000 年開發的 AWL 自發表以來就被廣泛使用，在詞彙研究、語料庫語言學、英語教學（TESOL）、學術英語（EAP）以及特別用途英語（ESP）等領域都有一定影響(Coxhead, 2016)，根據 2023 年 web of science 的統計結果，文章已經被引 1200 次，在學術詞彙領域中被引量最高。AWL 是在一個 350 萬詞的書面學術文本語料庫的基礎上，根據專用性（specialized occurrence）：所選單詞必須在 GSL 常用詞表外；廣泛分佈（range）：單詞在子語料庫中的分佈情況；頻次（frequency）三個篩選條件得到的，包含 570 個詞族的學術詞表，覆蓋了語料庫中大約 10%的詞彙量。作為與語料庫結合的非常優秀的研究成果，AWL 的相關研究主要集中在兩個方面，一方面是是其本身的實證性研究，例如在英語口語、醫學、農學、理工科及社會學、工程學、應用語言學、金融學等學術文本的基礎上使用 AWL 進行分析(Chen & Ge, 2007; Dang & Webb, 2014; Hyland & Tse, 2007; Li & Qian, 2010; Martinez 等, 2009; Munoz, 2015; Ward, 2009)。另一方面是在 AWL 的基礎上進一步開發其他詞表，例如學術英語科技詞表(Coxhead & Hirsh, 2007)和中學英語詞表(Coxhead, 2016)等。除此之外，與 AWL 相關的語料庫處理工具的開發也是 AWL 如此熱門的原因質疑，比如例如 Tom Cobb 開發的 Compleat Lexical Tutor (http://www.lextutor.ca/)與 Heatley, Nation, Coxhead's (2002) Range Programme（downloadable from http:// www.victoria.ac.nz/lals/about/staff/paul-nation）都為學術研究者們開發與分析自己的語料庫提供了方便(Coxhead, 2016)。

當然作為當時新出現的詞表，AWL 也接受了許多批評與討論。例如(Hyland & Tse, 2007)就提出同一詞族的單詞的不同變形在不同的情況下具有不同的意義，並且在不同學科領域下其差異會被進一步放大，因此對我們是否需要一個統一的核心學術詞彙表提出質疑。Hyland 在質疑核心學術詞彙表的同時也強調了不同學科領域更加細化學術詞彙表開發的必要性，因此出現許多使用 AWL 的方法開發不同學科學術詞彙表的成果，醫學、化學、環境科學、中醫(Hsu, 2018; Liu & Han, 2015; Valipouri & Nassaji, 2013; Wang 等, 2008)等。此外(Gardner & Davies, 2014)指出 AWL 排除了 GSL 中的所有詞彙，因此 GSL 中很多高頻的學術詞彙也同時被排除了，例如 company、market、exchange 等詞彙本身都具有重要的學術意義。

為瞭解決上述的問題，後續的研究者做出了諸多方法的嘗試以及多方面的改進。例如 Fraser （2007，2009）開發的藥理學詞表 PWL（Pharmacology Word List）將 GSL，AWL 與專業詞彙結合在一起，創建了 2000 個最常用詞族詞表達到了 90%的覆蓋率(Hsu, 2018)。Gardner 與 Davies's 更是做了諸多改進，一是使用詞元而不是詞族組成詞表，二是不排除 GSL 中具有學術意義的詞彙，三是引入 Julliand dispersion 指標來衡量詞彙在語料庫中的均勻程度(Gardner & Davies, 2014)。其中後者的影響更大，獲得了學界的諸多認可，包括 Coxhead 本人也高度評價了 AVL 的研究成果，認為其在一個大語料庫的前提下完善了詞表提取的原則(Coxhead, 2016)。AVL 是由 4 億兩千五百萬詞的 COCA 語料庫中一億兩千萬詞的學術子語料庫中得到的，使用了四個篩選條件：（1）頻次比率（ratio）：所選單詞在學術文本和非學術文本中的頻次必須達到一定差異。（2）廣泛分佈（range）：單詞在子學科語料庫的分佈情況。（3）均勻分佈（dispersion）：單詞的均勻分佈指數（Julliand D）。（4）學科指數（discipline measure）：單詞在某學科的出現頻次不能過高。最後得到的 AVL 包含 3015 個詞元，覆蓋了大約 14%的學術文本。為了和

AWL 形成對比，Gardner 與 Davies's 還取了 AVL 中最高的 570 個詞族與 AWL 進行對比，發現 AVL 的覆蓋率是 AWL 的近兩倍。在 AVL 的影響下還產生了許多優秀的研究，例如(Lei & Liu, 2016)就採用類似的方法，開發了醫學英語詞彙表 MAVL（Medical academic Vocabulary List），還有 Durrant (2016)在大學寫作中驗證了 AVL 的可用性，Wingrove (2017)使用 AVL 對 TED 講座的學術特性進行了測量等。除此之外，對於指標也有學者進行了進一步的研究，Gries (2008)在語料庫規模不斷擴大的時代背景下對 Julliand Dispersion 指標的有效性提出了質疑，並提出了新的 Gries' DP 值演算法，並被 (Biber 等, 2016)等進行了驗證。可見有關學術詞表提取研究的方法論正不斷朝著更精細、更準確的方向不斷發展。

2.2.1 漢語學術詞表

與英語領域較為豐富的研究成果相比，漢語領域有關學術詞表的成果則比較鮮見與滯後。雖然陳鶴琴先生進行字數頻次研究的時間相當早，但之後的研究進度較為緩慢。在 Coxhead 的 AWL 出現之後，國內學術詞彙的有關研究更多是以英文作為研究主題，像 Wang 等 (2008)參考 AWL 的方法創建的醫學學術詞彙表（EMP）；Coxhead“學術詞彙表”的適用性研究(吳瑾 & 王同順, 2007)；學術詞彙表（ＡＷＬ）的研究進展與啟示(吳璟, 2010)；專門用途英語學術詞表創建研究——以航海英語為例(赵志刚, 2015)；由(Lei & Liu, 2016)創建的醫學英語詞彙表(Medical Academic Vocabulary List，MAVL)等。

而有關漢語的研究則相對比較少見，直到近些年在自然語言處理領域在中文分詞方面的效果不斷提升，語言學學者們也有相當一部分學習了相關的知識並具備了較高的電腦水準，同時電腦能力的不斷提高也增強了我們處理大量語料庫的能力，這才產生了一些相關的學術成果。

像薛蕾 (2017)收集了對外漢語、國際漢語教育學科研究領域中最具代表性的的四本權威期刊《世界漢語教學》、《漢語學習》、《語言文字應用》中的對外漢語板塊以及《語言教學與研究》中近三年來最新研究成果，構成了 250 萬字規模的漢語語言學論文語料庫。使用 CAJ 閱讀器將 PDF 格式轉換為 TXT 格式使電腦能夠讀取，接著使用 Python2.x 專業分詞軟體，但並未詳細描述是什麼軟體，接著利用 Wordsmith 詞頻統計軟體對學術詞彙的出現頻次和分佈特點進行了統計，最後以出現頻次 10 次作為篩選條件，得到了一個總詞數為 5187 詞的《漢語語言學學術詞表》。再將其與《現代漢語常用詞表》和《漢語水準考試詞彙與漢字等級大綱》進行了對比從高頻詞在詞表中所占的比例體現出新詞表的優越性。

朱明玉 (2020)收集了漢語言文學、法學、國際經濟與貿易、軟體工程四個專業的教材各一本、碩博論文各 10 篇，建立了總字數為 340 萬字的學術漢語語篇語料庫。接著運用 ROST 資料計算工具進行分詞處理和詞頻統計，採用頻次和跨文本分佈作為特徵，生成了一個按頻次高低排序的詞彙表，首先截取了詞表中出現頻次 50 詞以上的詞彙得到《學術漢語總詞彙表》，接著將基於四個專業語料分別建立的《漢語言文學專業學術漢語詞彙表》、《法學專業學術漢語詞彙表》、《國際經濟與貿易專業學術漢語詞彙表》、《軟體工程專業學術漢語詞彙表》截取了出現頻次不低於五次的詞彙，最後將四個專業詞表共有的詞彙表與截取頻次 50 詞的《學術漢語總詞彙表》進行合併，使用 ROST CM6 軟體的詞頻分析功能，得到一個按頻次高低排序的詞彙表，以兩次作為篩選條件得到總詞數為 1026 的《通用學術漢語詞表》。最後將其與《HSK 詞彙等級大綱》進行了對比並分析了詞彙的重合情況和分級特點。

(劉華 & 李曉源, 2022)以中醫漢語類教材、中醫專業類教材、中醫網站三大語料來源，建設中醫漢語語料庫；利用詞語聚類演算法和圖式語義場理論，形成中醫漢語內

部主題作為語料來源，自建了中醫漢語語料庫，雖然沒有注明語料庫大小，但根據圖式語義場理論採用了比較複雜的詞語聚類演算法，得到了一個中醫漢語主題詞表。

(王笑然 & 王佶旻, 2022)的最新研究則採用了更為與時俱進的方法，收集了經貿類專業本科生（或本科留學生）前三個學期的五本專業基礎課教材以及經濟學大類下五種綜合性核心期刊中的５０篇論文分成六個子語料庫作為語料來源，自建了一個總字數為２００萬字左右的經貿類學術語料庫。接著採用 Python 的 pkuseg 中文分詞庫進行分詞處理工作，而後以跨文本分佈、專用性、詞頻和詞語離散度作為選詞標準。其中跨文本分佈指出所選詞語至少要出現在六個子語料庫中的三個，共有 6039 個詞種；專用性則指出參考 Coxhead（2000）的研究，要排除來華留學預科教育用的《基礎漢語常用詞彙表》種包含的詞彙，保留了 4299 個詞種。詞頻則根據已有研究中最低指標在 25-30 次每百萬詞，因此選擇在 120 萬詞的語料庫種保存出現 30 詞及以上的詞語，剩餘 1616 個詞種。最後詞語離散度也就是我們所說的詞語均勻分佈值採用了學界使用最多的 Julliand Dispersion 值，根據已有研究種採用的最低閾值 0.3（Oakes & Farrow 2007）、0.5（Paquot 2007）、0.8（Gardner & Davies 2014），根據自建語料庫中的資料特徵，選擇了比較低的 0.3 取值閾限。就得到的詞表包含 1454 個詞語，構成了經貿類學術詞表的初表。最後再經過人工篩選後得到最終保留了 1349 個詞語，整合後共 1336 個詞條的經貿類學術漢語總詞表。在結果討論中作者討論了詞表詞彙的分佈情況，並從詞表總體覆蓋率角度進行了驗證，指出詞表對經貿類學術語料庫的覆蓋率為 19%，和先前研究比展現了優越性，接著與來華留學生預科教育所使用的《基礎漢語常用詞彙表》和《經貿漢語教學大綱》兩個詞表進行了對比。

我們注意到即使是最新的研究成果也只是簡單的使用了常用的幾個指標，尤其是對於詞語均勻分佈值的使用實際上是參照了英文領域中常見的取值範圍再根據自己的

實驗結果進行選擇。但我們注意到學界已經對相關指標進行了非常詳細的分析，例如 Biber 等(2016)等人對 Julliand 演算法進行了詳細的分析，指出在語料庫早期研究中 Julliand dispersion 確實是一個非常實用且有效的演算法，這是由於早期子語料庫數量較少。但他們對 Julliand‘D 演算法仔細研究後發現，隨著子語料庫數量趨於無限大，Julliand’D 值趨近於無窮，因此反映在現代語料庫的結果就是，隨著現代語料庫規模的增大，子語料庫數量的增多，Julliand’D 值的有效範圍為會呈現災難性降低，他們設計了兩個實驗來對此結論進行進一步驗證。同時我們的實驗結果也同樣證明瞭這一結論，而作為 Julliand‘D 值的替代，Gries’ DP (Gries 2008)是一個很好的方法，Gries 在論文中對現有均勻分佈值的方法進行了討論，然後提出了基於實際分佈值和理想均勻分佈值的差異這一思想而開發的 Gries’ DP，能有效避免隨著子語料庫數量增大，Dispersion 有效區域隨之降低的問題，根據其原理，Gries’ DP 值越接近於零的詞語，其分佈越均勻。我們在詞表研製的過程中首次使用了 Gries’DP 值作為詞語均勻分佈值的篩選標準，這在整個學術詞表提取研究中包括英文和中文領域都是創新性的，對詞語均勻分佈值指標的進行了一定程度的改進。

2.3 小結

本章節從語料庫發展與基於語料庫的學術詞表提取兩個角度進行了相關文獻的介紹，尤其在學術詞表提取這一研究方向體現了國內漢語研究與英語研究的差異，並對學術詞表提取的步驟和指標進行了簡要介紹。

3 語料庫建設

3.1 語料庫資料收集

我們首先確定了我們的學術語料來源，學術論文是學術語料最重要的來源之一，而摘要又是公開資料中學術語料最豐富，同時也是整個文章中最精煉、最總結性的部分。因此我們參照 2021-2022 年中文社會科學引文索引（CSSCI，https://cssrac.nju.edu.cn/cpzx/zwshkxywsy/20210425/i198393.html）來源期刊目錄，CSSCI 是中文學術界公認的最權威且全面的期刊目錄之一，其中包含了 583 種期刊以及 30 種臺灣期刊和 2 種報紙理論版，並將其分為了中國文學,人文經濟地理,體育學,冷門絕學,歷史學,哲學,圖書館、情報與文獻學,外國文學,宗教學,心理學,政治學,教育學,新聞學與傳播學,民族學與文化學,法學,社會學,管理學,經濟學,統計學,綜合性社會科學,綜合性高校學報,考古學,自然資源與環境科學,藝術學,語言學,馬克思主義理論,高校社科學報一共 27 個子學科種類。語料庫以中國知網（https://www.cnki.net/）作為主要的資料來源，除少數幾種無法搜索到外，收集了全部期刊收錄文獻的摘要資料。

整個過程分成了兩個部分，第一部分我們收集了 2020-2022 年 8 月的全部種類期刊發表的論文元資訊，其中包含中國文學期刊 18 種，人文經濟地理 12 種，體育學 11 種，冷門絕學 5 種，歷史學 28 種，哲學 14 種，圖書館、情報與文獻學 20 種，外國文學 6 種，宗教學 3 種，心理學 7 種，政治學 39 種，教育學 37 種，新聞學與傳播學 17 種，民族學與文化學 15 種，法學 24 種，社會學 12 種，管理學 36 種，經濟學 70 種，統計學 4 種，綜合性社會科學 47 種，綜合性高校學報 63 種，考古學 7 種，自然資源與環境科學 6 種，藝術學 24 種，語言學 25 種，馬克思主義理論 20 種，高校社科學報 11 種，一共 581 種期刊。經過統計，我們一共收集了 212246 條記錄，提取其摘要後包含英文

和符號等字元的總語料庫大小 54570006 字元，僅包含中文的語料數量達 48594232 字，詳見表 3-1。我們將其作為提取通用學術詞表的語料庫基礎，同時鑒於 CSSCI 期刊的權威性與全面性，我們認為其同樣時間段內不同子學科期刊的數量可以代表各自學科在學術語料中的權重，因此我們不需要再將子語料庫剪裁以保證子語料庫的平衡性。

表 3-1 通用語料庫概況

學科	期刊數	時間跨度	記錄總條數	摘要條數	字數				
					英文字元	數字	中文	特殊字元	總字元數
法學	24	2020-2022.8	5812	5738	4820	11758	1548337	137585	1,702,500
高校社科學報	11	2020-2022.8	2663	2629	5911	9622	747578	64082	827,193
管理學	36	2020-2022.8	15708	15097	60232	65808	4048251	323055	4,497,346
教育學	37	2020-2022.8	16718	16114	31199	37141	3637477	335936	4,041,753
經濟學	70	2020-2022.8	20840	20449	68631	112410	5555655	450018	6,186,714
考古學	7	2020-2022.8	1586	1567	3640	12195	275648	30420	321,903
冷門絕學	5	2020-2022.8	1021	1010	1053	4417	179560	22599	207,629
歷史學	28	2020-2022.8	6011	5851	7023	25845	1250787	139546	1,423,201
馬克思主義理論	20	2020-2022.8	9363	9177	4174	22677	1952707	181138	2,160,696
民族學與文化學	15	2020-2022.8	5576	5459	6954	12603	1330378	123024	1,472,959
人文經濟地理	12	2020-2022.8	5670	5620	31237	51623	1652597	157607	1,893,064
社會學	12	2020-2022.8	2272	2237	5043	12390	582406	52338	652,177
體育學	11	2020-2022.8	3531	3505	37726	27752	994271	106031	1,165,780
統計學	4	2020-2022.8	3506	3502	27875	21842	820684	66391	936,792
圖書館、情報與文獻學	20	2020-2022.8	9699	9596	64822	33437	2292520	230902	2,621,681
外國文學	6	2020-2022.8	1337	1312	1395	3113	278014	27499	310,021
心理學	7	2020-2022.8	2710	2703	24099	21463	639027	63860	748,449
新聞學與傳播學	17	2020-2022.8	10459	8673	20530	22116	1774104	171129	1,987,879
藝術學	24	2020-2022.8	13605	11555	25137	34381	2023854	224116	2,307,488
語言學	25	2020-2022.8	5200	5117	33102	16647	991263	109914	1,150,926
哲學	14	2020-2022.8	4717	4690	9503	8220	1040708	111710	1,170,141
政治學	39	2020-2022.8	7536	7432	9647	21406	1940833	177230	2,149,116
中國文學	18	2020-2022.8	7631	6714	13709	26139	1264214	164930	1,468,992
自然資源與環境科學	6	2020-2022.8	3599	3590	36669	59836	1212457	128803	1,437,765
宗教學	3	2020-2022.8	1194	1147	2190	2840	226644	25265	256,939
綜合性高校學報	63	2020-2022.8	18472	18188	21753	41045	4527956	428638	5019392
綜合性社會科學	47	2020-2022.8	25810	25166	47845	42120	5806302	555243	6,451,510
總計	581	2020-2022.8	212246	203838	605919	760846	48594232	4609009	54570006

第二部分我們目的是建立法學、管理學、教育學和語言學四個學科的專用學科詞表，同樣參照 2021-2022 中文社會科學引文索引（CSSCI）來源期刊目錄，但我們此次收集了各自學科核心期刊自存在電子記錄以來直至我們收集語料之時的全部文獻，同時記錄了更加詳細的相關元資訊，創建了四個學科各自的學科語料庫。同時統計了不同學科內不同期刊的年份記錄、含英文字元數、中文字數、記錄總條數、摘要條數等資訊，詳見表 3-2。並將其作為創建四個學科專用詞表的語料庫資料基礎。各個學科語料庫不同期刊子語料庫的概況見附件。

表 3-2 學科語料庫概況

學科	記錄總條數	摘要條數	詞例數		詞種數		中文字數
			含外文	排除外文	含外文	排除外文	
法學	96563	88909	10806450	9208942	123721	110039	16302192
管理學	208398	193271	22083448	18959500	192138	139586	33944617
教育學	229090	204955	22090179	18687949	200101	156783	33385011
語言學	84297	77637	8432448	6763304	210365	148803	11673779

最後我們匯總兩部分得到的學術語料庫規模超過一億字元，並使用其進行學術詞表提取研究，與以往研究對比可知，在語料庫大小的角度看，我們將語料庫規模提升了兩個數量級，且創建了目前我們已知最大規模用作學術目的的學術語料庫。

3.2 語料庫資料處理

語料資料收集的過程中我們還需要進行許多的資料處理工作，第一部分通用學術語料庫我們收集了文獻的期刊名稱、ISSN 國際刊號、CN 期刊號、文獻標題、作者、摘要六種中繼資料，接著按不同子學科進行分類，最後將摘要資料提取出來作為我們進一步處理的主要語料。第二部分學科學術語料庫我們除了第一部分的幾種中繼資料外還收集了像關鍵字、基金、DOI、分類號等一共 23 種中繼資料，同樣將摘要資料提

取出來作為我們後一步處理的主要語料。接著我們進行了分詞、中文詞彙篩選、簡繁轉換三個步驟，完成了我們進行詞表提取前的全部資料處理工作。

3.2.1 分詞

分詞我們使用了由北京大學語言計算與機器學習研究組研製的北大開源中文分詞工具包 pku_seg，主要使用 python 語言進行實現。Pku_seg 屬於中文分詞領域較為新穎的研究成果，其分詞效果較 jieba 和 THULAC 兩種分詞都更好（詳見：https://github.com/lancopku/pkuseg-python），由於學術語料所包含學科眾多，知識比較複雜，內容非常豐富，因此我們使用了預設模型來對語料進行分詞處理。

3.2.1 篩選中文詞彙

對分詞處理得到的結果，我們需要將其中非中文的部分排除，僅留下包含我們所需學術漢語詞彙的部分，像數位元元元、空格、英文、特殊字元都需要被排除。我們通過編寫函數來實現，其主要原理是根據 Unicode 編碼原則，由於我們所有資料的儲存編碼方式都是 utf-8，而 Unicode 漢字編碼的範圍是\u4E00-\u9FA5，因此我們將分詞結果進行測試，將屬於漢字編碼範圍內的詞進行保留，其他的則排除掉。

3.2.1 簡繁轉換

因為我們發現我們收集的語料在絕大部分簡體中仍存在少量繁體記錄，例如在外語教學與研究期刊前期的論文就都是以繁體出現。而在後續詞彙統計的過程中簡體和繁體同樣的分詞結果會被視作兩個不同的詞語，因此我們需要進行簡繁轉換避免資料因此受到影響。我們的簡繁轉換同樣通過 python 語言使用 opcc 代碼包實現，對我們的經過前面所有步驟的分詞結果進行處理，檢測其中所有的繁體詞彙並將其轉換為簡體。

3.3 小結

本章我們詳細描述了我們創建語料庫的過程，分成了構建語料庫和語料庫資料處理兩個部分。在語料庫構建部分介紹了我們語料庫的構成，並展示了不同部分統計的詳細結果。接著我們介紹了我們進行語料庫資料處理的主要步驟，分成了分詞、中文詞彙篩選、簡繁轉換三個步驟並解釋了相應的實現過程。

4 詞表創建

在語料庫資料的相關處理完成後我們獲得了組成通用語料庫以及法學、管理學、教育學、經濟學四個專用語料庫的所有學術詞彙。經過對學術詞表提取領域的研究現狀研究，同時我們認為常用漢語常用詞彙中很多詞彙具備相當高的學術意義，因此我們選擇了類似 AVL 的方法，並未選擇專用性選詞標準排除所得詞表中與漢語常用詞彙重合的部分，而是選擇了詞頻、跨文本分佈、詞語均勻分佈值三個特徵作為選詞標準得到我們的學術詞表。值得一提的是，我們還參考了特徵研究領域的研究成果，創新性的選用 Gries’ DP 值作為我們詞語均勻分佈值評價指標，而不是常用的 Juilland’ s D 值，避免了 Juilland’ s D 值在語料庫規模增大時的出現的問題。

4.1 通用學術詞表提取

4.1.1 所使用的語料庫資料

使用我們在語料庫構建過程中第一部分收集自建的通用學術語料庫，因為我們希望子語料庫分學科更加明確，因此刪除了原語料庫中學科分類不夠清晰三個部分的語料，也就是高校社科學報、綜合性社會科學、綜合性高校學報三個子語料庫資料，剩餘 24 個子學科種類，460 種期刊從 2020 年至我們收集時間點收錄文獻的全部元資訊（詳見表 3-1），其中摘要資料作為詞表提取的基礎語料庫。所使用語料中排除空格的全部字元數為 42,271,911，其中含英文字元 530410，數位元元元 668059，中文字元 37512396，特殊字元 3561046。

經過第一步分詞後所得的分詞結果中包含了英文，特殊符號等字元的分詞結果，經過統計其中包含詞例數 24538554，詞種數 311207。接著進行第二步篩選中文詞彙步

驟，經過篩選後統計其中包含詞例數 20683200，詞種數 251268。最後進行簡繁轉換步驟，將篩選中文詞彙步驟的結果進行簡繁轉換，並將結果和轉換的繁體詞彙分別進行保存，所得結果經過統計詞例數不變，詞種數 247314，有 3,954 個繁體詞種被排除，最後所得中文分詞結果共含中文字元數量 37328346。詳細詞表見附件。

4.1.2 頻次篩選

在進行詞表提取相關研究中，頻次是應用最早，最為廣泛的篩選條件。在現有的研究中，Coxhead（2000）研製的 AWL 在 350 萬詞的語言庫中選擇了出現頻次大於 100 詞的詞彙；Jing Wang(2008)在百萬詞醫學語料庫中選擇出 30 次以上的詞彙；Gardner & Davies(2014)研製的 AVL 設定詞彙在學術文本中的頻次是非學術文本的 1.5 倍，而 COCA 語料庫學術文本的占比是 28%，因此 Gardner 認為這樣選擇詞彙在學術文本中的占比至少 42%；薛蕾（2017）在 250 萬子的語料庫中收集了出現頻次 10 次以上的詞彙；朱明玉（2020）設定的 50 次；王笑然，佶旻（2022）設定的的 30 次/百萬詞等。綜合來看，我們選擇了其中要求較高的 30 次/百萬詞，作為我們頻次篩選的標準，而我們通用語料庫中一共包含 20683200 個詞例，因此我們選擇出現次數大於 620 次的詞彙。最後得到了一個包含 3155 個詞種，最多出現的“的”字出現 1390874 次，出現 621 次的詞彙有“科學化”、“制度性”、“刊發”、“崩盤”四個。詳細詞表見附件。

4.1.3 跨文本分佈篩選

接著進行我們跨文本分佈特徵的篩選，這也是學界進行詞表提取的常用特徵之一，Coxhead（2000）設置詞彙必須出現在 28 子學科領域中的 15 個，且在語料庫的四個主要部分中每個部分至少出現 10 次；Gardner & Davies(2014)則設置詞彙至少出現在九個子學科中的七個；Wang 等 (2008)指出所選詞彙要出現在至少一半子語料庫中，使用了一

半子料庫作為標準；中文方面朱明玉（2020）設定詞彙在全部語料庫中出現；王笑然，王佶旻（2022）則選擇詞彙出現在語料庫的一半以上。我們首先將上一步篩選得到的3155 個單詞在子語料庫中出現的數量和分佈進行了統計，經過觀察我們設定我們選擇的詞彙必須出現在 24 個子學科中的 20 個以上這一較高的標準，篩選後詞表保留了 2648 個詞彙，捨棄了“審判”、“利潤“、“就業”這些出現在了 20 個子學科語料中，比較具有鮮明學科特性的詞彙。詳細詞表見附件。

我們以漢語通用學術詞表排名第 277 位的“框架”一詞為例，將其在語料運用中的實例以及實例的關鍵元資訊加以展示。

詞語“框架”實例 1：

期刊來源：《電子政務》

發表時間：2020.12

題目：試論網路資訊內容治理主體構成及其行動轉型

作者：周毅

例句：文章以習近平總書記的重要講話為依據,分析了網路資訊內容治理的多主體參與框架及其治理行動任務。

詞語“框架”實例 2：

期刊來源：《電子政務》

發表時間：2020.12

題目：“互聯網+政務服務”優化了營商環境嗎？——基於 31 省的模糊集定性比較分析

作者：廖福崇

例句：提出了行政負擔的組態分析框架,以 31 省的"放管服"改革為研究物件,採用定性比較分析的方法,探討"互聯網+政務服務"對營商環境的影響和內在機制。

4.1.4 詞語均勻分佈值篩選

下一步是進行詞彙詞語均勻分佈值的計算，Julliand Dispersion 演算法是學術詞表提取領域中最常用的用來衡量詞語在語料庫中給分佈是否均勻的指標，Gardner、Davies (2014)，Liu、Han (2015)等人都使用了 Julliand’D 來衡量詞語均勻分佈值，但 Biber 等(2016)在文章中詳細分析了 Julliand‘D 值的有關問題，通過分析 Julliand‘D 值的原始演算法，指出 Julliadn’D 值的有效範圍會隨著語料庫數量的增大而減小。而隨著電腦處理能力的增加，語料庫規模的增加，更多子語料庫的數量可以包含更多的資訊，更好的反映單詞料庫中的分佈情況，而 Julliand’D 演算法與這個趨勢背道而馳。因此他提到了由 Stefan Gries (2008)提出的 Gries’ DP 演算法，使用詞彙實際分佈情況和理想均勻分佈的差異來計算詞語均勻分佈值，有效避免了 Julliand‘D 演算法的問題。因此我們在將語料庫分成等大的 1000 個子語料庫的基礎上，使用 Gries’ DP 演算法來計算詞彙的均勻分佈值，這也是我們實驗設計的創新點之一。

同時在我們實驗過程中也驗證了 Julliand‘D 演算法的缺陷，當將語料庫劃分未 10 個子語料庫時，其 Julliand‘D 的最低有效值為 0.3，而如果將語料庫劃分為 1000 子語料庫，最低有效值就變為了 0.7，有效範圍發生了嚴重下降。而我們使用 Gries’ DP 演算法在 1000 子語料庫基礎上取得的有效範圍值為 0.06-0.89。接著我們經過觀察詞表，選定了 0.8 作為我們詞語均勻分佈值的篩選標準，捨棄了像“文學“，”雜誌社“，”體育“這一類分佈不均勻的詞彙。經過篩選後，詞表保留了 2593 個詞種。詳細詞表見附件。

4.1.2 去除停用詞

接著我們在獲得詞表中去除在百度停用詞表、哈工大停用詞表、四川大學機器智慧實驗室停用詞庫三個停用詞表中出現過的全部停用詞，一共 429 個詞彙。所得到去除停用詞版本的通用學術詞表包含 2164 個詞彙。詳細詞表見附件。

4.2 法學專用學術詞表提取

4.2.1 所使用的語料庫資料

所使用語料庫為我們自建的法學學術語料庫，將我們收集的第二部分 24 種核心期刊自電子版發刊以來到 2022 年我們收集的時間點為止法學文獻的全部元資訊（詳見表 3-2）的摘要資料作為詞表提取的基礎語料庫。所使用語料中排除空格的全部字元數為 18,177,377，其中含英文字元 96766，數位元元元 167801，中文字元 16373508，特殊字元 1539302。

經過第一步分詞後所得的分詞結果中包含了英文，特殊符號等字元的分詞結果，經過統計其中包含詞例數 10806450，詞種數 123721。接著進行第二步篩選中文詞彙步驟，經過篩選後統計其中包含詞例數 9208942，詞種數 110039。最後進行簡繁轉換步驟，將篩選中文詞彙步驟的結果進行簡繁轉換，並將結果和轉換的繁體詞彙分別進行保存，所得結果經過統計詞例數不變仍是 9208942 因為簡繁轉換不會減少詞例數量，詞種數 109343，說明共由 696 個繁體詞種被排除，最後所得中文分詞結果共含中文字元數量 16302192。詳細詞表見附件。

4.2.2 頻次篩選

根據上文文獻綜述與先前研究的經驗，我們對頻次指標仍是你以 30 詞/百萬詞作為篩選條件，因語料庫包含詞例 9208942，因此我們將出現次數在 276 詞以上的詞彙保存在我們的詞表中，此時我們的詞表包含 3020 個詞種。詳細詞表見附件。

4.2.3 跨文本分佈篩選

同樣根據先前的研究經驗，整個法學學術語料庫中包含了 24 種期刊，這就構成了 24 個子語料庫，我們統計了頻次篩選後詞表中 3020 個詞彙每個詞種在多少個子語料庫出現，以及在各子語料庫中分別出現了多少次。接著我們發現即使最少的詞彙也在子語料庫的十個部分中出現，且只有“月刊“和”頁碼“兩個詞彙，其他 3018 個詞彙都出現在了法學學術語料庫中的一半以上，已經具備先前研究中被收錄的條件，這說明經過我們第一步頻次篩選的詞彙在跨文本分佈特徵方面也具有相當好的表現。經過研究我們選擇去除了出現在 18 以及 18 個子語料庫以下的詞彙，其中包括”月刊“，”“頁碼”，“引文”，“位址”等 18 個法學意義較少的詞彙，保留了出現在 19 個子語料庫中的像“搶劫罪”，“犯罪分子”，“公證”等詞彙。並將原詞表和篩選後詞表都進行了保存，所得詞表包含 3002 個詞種。詳細詞表見附件。

我們以漢語法學學術詞表排名第 315 位的“審判”一詞為例，將其在語料運用中的實例以及實例的關鍵元資訊加以展示。

詞語“審判”實例 1：

期刊來源：《比較法研究》

發表時間：2022.5

題目：《刑法》第 88 條“不受追訴期限的限制”研究

作者：阮齊林

例句：極端終止說在終止說的基礎上提出認定“逃避偵查或者審判”“八不論”主張，極端擴張《刑法》第 88 條的適用，此說明顯不妥。

詞語“審判”實例 2：

期刊來源：《比較法研究》

發表時間：2022.4

題目：論民事糾紛相對性解決原則

作者：張衛平

例句：以這一原則對民事訴訟實踐活動進行審視，可以發現人們在解決民事糾紛時，往往沒有顧忌這一原則，糾紛一次性解決、穿透式審判等實踐活動都可能存在跨界越邊的情形。

4.2.4 詞語均勻分佈值篩選

我們依然根據 Gries’ DP 演算法的特點，將整個語料庫分為 1000 個大小相等的子語料庫，並統計上一步處理後得到詞表的 3002 個詞種在 1000 子語料庫中的分佈情況，並憑此計算 Gries’ DP 值與 Julliand‘D 值並分別保存。經過對結果的觀察，法學 Julliand‘D 值的有效區域在 0.75-0.99，我們使用的 Gries’ DP 值範圍則在 0.03-0.93 之間，避免了 Julliand‘D 值的缺陷，但在由於一部分法學學術意義較為豐富詞彙在詞表均勻分佈值排序中並靠末尾，如“撤訴”：0.909、“洗錢”：0.92，“信訪”：0.904 等，這是由於詞語的分佈情況決定的，兩種演算法都會展現出這個問題，而我們不希望將這些

詞彙在篩選過程中被排除，因此我們並未在這一步中對我們的詞表做出裁剪，並得到我們最終的法學學科專用學術詞表，其中包含 3002 個詞種。詳細詞表見附件。

4.2.5 去除停用詞

我們在獲得詞表中去除在三個停用詞表的重複詞彙，一共 533 個詞彙。所得到去除停用詞版本的通用學術詞表包含 2500 個詞彙。詳細詞表見附件。

4.3 管理學專用學術詞表提取

4.3.1 所使用的語料庫資料

資料使用所有期刊自電子版發刊以來至 2022 年我們收集語料為止的所有資料，其中包含 36 種期刊，所使用語料中排除空格的全部字元數為 38108902，其中含英文字元 541471，數位元元元 564938，中文字元 34130535，特殊字元 2871958。

經過第一步分詞後所得的分詞結果中包含了英文，特殊符號等字元的分詞結果，經過統計其中包含詞例數 22083448，詞種數 192138。接著進行第二步篩選中文詞彙步驟，經過篩選後統計其中包含詞例數 18959500，詞種數 139701。最後進行簡繁轉換步驟，將篩選中文詞彙步驟的結果進行簡繁轉換，並將結果和轉換的繁體詞彙分別進行保存，所得結果經過統計詞例數不變仍是 18959500 詞種數 139586，說明共由 115 個繁體詞種被排除，最後所得中文分詞結果共含中文字元數量 33944617。詳細詞表見附件。

4.3.2 頻次篩選

根據上文文獻綜述與先前研究的經驗，我們對頻次指標仍是你以 30 詞/百萬詞作為篩選條件，因語料庫包含詞例 18959500，因此我們將出現次數在 569 詞以上的詞彙保存在我們的詞表中，此時我們的詞表包含 2748 個詞種。詳細詞表見附件。

4.3.3 跨文本分佈篩選

經過統計，頻次篩選後詞彙中在 36 個子語料庫出現 18 次以下的詞彙僅有兩個，“科學學”與“運算元”且分佈較為集中在在數個語料庫內，說明我們頻次篩選的結果較好。同時我們決定出現在 28 個語料庫及以下的詞彙加以排除，去除了像“商品經濟”，“分析師”，“審計師”等一些專業性較強的詞彙，以及像“厭惡”，“追趕”，“拼湊”等不具備太大學術屬性的詞彙。最後得到了包含 2668 個詞種的詞彙表。詳細詞表見附件。

我們以漢語管理學學術詞表排名第 263 位的“投資”一詞為例，將其在語料運用中的實例以及實例的關鍵元資訊加以展示。

詞語“投資”實例 1：

期刊來源：《管理評論》

發表時間：2021.12

題目：地區差異會影響金融與實體之間的相互轉化程度嗎?——基於我國製造業上市公司的微觀視角

作者：徐立;吳文鋒;

例句：實體經營與金融投資的有機結合被認為是實現我國製造業由大到強崛起的一個重要途徑。

詞語“投資”實例 2：

期刊來源：《管理評論》

發表時間：2021.12

題目：大氣污染與環境規制對企業庫存的影響機制研究——基於廣東省工業企業資料

作者：王雪清;劉勇;

例句：本文從系統學視角出發,構建了大氣污染與環境規制對工業企業庫存產生效用的系統動力學模型,研究減排政策、工業污染治理投資、交通基礎設施投資以及科技投資的變化對廣東省工業企業庫存和工業經濟的影響效應。

4.3.4 詞語均勻分佈值篩選

接著進行在上述成果的基礎上僅性詞語均勻分佈值篩選，同樣在將語料庫均分為 1000 個子語料庫的基礎上計算 Griess_dp 值，得到結果有效範圍在 0.05-0.87 間。而在均勻度較差的詞彙中存在“政務”，“電商”，“經濟帶”等我們不希望被去除的詞彙，因此不對詞表進行截取，最後得到了一個包含 2668 個詞種的管理學專用學術詞表。詳細詞表見附件。

4.3.5 去除停用詞

我們在獲得詞表中去除在三個停用詞表的重複詞彙，一共 446 個詞彙。所得到去除停用詞版本的通用學術詞表包含 2222 個詞彙。詳細詞表見附件。

4.4 教育學專用學術詞表提取

4.4.1 所使用的語料庫資料

所使用語料庫同樣為我們自建的教育學學術語料庫，將我們收集的第二部分和第一部分資料組合在一起，相當於收集了 37 種教育學種核心期刊從發刊以來至我們收集資料時間為止收錄的所有文獻中繼資料，並使用其中的摘要資料作為語料庫的主體。所使用語料中排除空格的全部字元數為 37552186，其中含英文字元 397899，數位元元元 481962，中文字元 33468688，特殊字元 3203637。

包含英文，特殊字元的分詞結果經過統計，包含詞例 22090179，詞種 200101。排除所有非中文後，剩餘中文詞例 18687949，詞種 157032 個。簡繁轉換後詞例數依舊不變，詞種數剩餘 156783，排除了 249 個繁體詞彙。最後所得中文分詞結果總詞表共含中文字元數量 33385011，並將完整詞典進行保存。詳細詞表見附件。

4.4.2 頻次篩選

根據上文文獻綜述與先前研究的經驗，我們對頻次指標仍是你以 30 詞/百萬詞作為篩選條件，因語料庫包含詞例 18687949，因此我們將出現次數在 560 詞以上的詞彙保存在我們的詞表中，此時我們的詞表包含 2771 個詞種，出現次數 561 次的詞彙有“美的”，“援助”，“民國”三個詞彙。詳細詞表見附件。

4.4.3 跨文本分佈篩選

接著對詞表進行跨文本分佈特徵計算，在 37 種不同核心期刊中，出現在 19 次與 19 次以下的詞彙僅有三個“欺凌”，“鋼琴”與“聲樂” ，這說明我們頻次條件篩選的

結果比較理想，而經過觀察，我們決定保留出現在 28 與 28 次以上子語料庫種的詞彙，因為出現 27 次的“工學院”以及 27 次以下的詞彙例如“節目”，“舞蹈”，“動畫”等詞彙不如跨文本分佈值為 28 的“民辦教育”，“作文”，“課件”等更具教育意義，最後我們在詞表中保留了 2749 個詞彙。詳細詞表見附件。

我們以漢語教育學學學術詞表排名第 222 位的“教學”一詞為例，將其在語料運用中的實例以及實例的關鍵元資訊加以展示。

詞語“教學”實例 1：

期刊來源：《比較教育研究》

發表時間：2022.6

題目：資助、雇傭和教育：美國博士研究生助教制度功能之爭

作者：楊雪芬;李子江;

例句：博士研究生助教的身份之爭，表面上是助教和大學管理者對助教教學任務性質的觀念爭鋒，實際上卻隱含著助教維護教學權益和大學管理者節約教學成本之間的利益博弈。

詞語“教學”實例 2：

期刊來源：《比較教育研究》

發表時間：2022.5

題目：世界一流教育學院使命陳述研究——基於 39 所世界一流教育學院使命文本的分析

作者：韋鳳彩;陽榮威;

例句：它們的使命陳述具有共同的表現特徵，包括重視教學與科研的卓越發展，努力打造全球教育的標杆；擁有培養傑出人才的旨趣，致力於培養行業的領軍人物；以推

動知識創新進步為己任，追求引領的國際地位；秉承公平公正的服務理念，豐富教育對社會的貢獻；注重對外交流與合作，不斷提升自身地位和影響力；尊重多樣性與多元化，營造相容並包的環境。

4.4.4 詞語均勻分佈值篩選

進一步進行詞語均勻分佈值進行計算，同樣將語料庫劃分為 1000 個子語料庫，計算詞彙的 Griess_dp 值，有效範圍在 0.04-0.89 之間，同時我們發現在 Griess_dp 較高的詞彙中同樣包含像“函授”，“產教”等教育學意義豐富的學術詞彙，而我們不想這些詞彙被排除，因此我們保留了整個詞表，得到了我們最終的語言學學科專用學術詞表，共包含 2749 個詞種。詳細詞表見附件。

4.4.5 去除停用詞

我們在獲得詞表中去除在三個停用詞表的重複詞彙，一共 462 個詞彙。所得到去除停用詞版本的通用學術詞表包含 2287 個詞彙。詳細詞表見附件。

4.5 語言學專用學術詞表提取

4.5.1 所使用的語料庫資料

資料使用的語料庫為語言學 25 種期刊自發刊以來收錄的全部文獻的摘要資料，經過統計語料庫大小刪除空格後為 14,237,095 個字元，其中包含英文字元 670647，數位元元元 241180，中文 11844207，特殊字元 1481061。

包含英文，特殊字元的分詞結果經過統計，包含詞例 8432448，詞種 210365。排除所有非中文後，剩餘中文詞例 6763304，詞種 150018 個。簡繁轉換步驟後詞例數不變，

詞種數剩餘 148803，排除了 1,215 個繁體詞彙，和其他三個學科專用詞表相比，繁體詞彙較多。最後所得中文分詞結果總詞表共含中文字元數量 11673779。詳細詞表見附件。

4.5.2 頻次篩選

根據上文文獻綜述與先前研究的經驗，我們對頻次指標仍是你以 30 詞/百萬詞作為篩選條件，因語料庫包含詞例 6763304，因此我們將出現次數在 202 詞以上的詞彙保存在我們的詞表中，此時我們的詞表包含 3046 個詞種，在出現次數 203 次的詞彙中包含了“熟語”，“詞句”等具備語言學意義的的詞彙，也有“創辦”，“並存”等具備一定學術意義的詞彙。詳細詞表見附件。

4.5.3 跨文本分佈篩選

在對詞表進行跨文本分佈特徵計算後，我們決定將出現在 20 子語料庫，即出現在二十種期刊以上的詞彙保留，對於一些出現期刊數較少的，或者集中出現在某些期刊中的詞彙加以排除，例如“電化教育”，“苗語”，“莎士比亞”等。最後剩餘 2764 個詞種。詳細詞表見附件。

4.5.4 詞語均勻分佈值篩選

接著對詞語均勻分佈值進行計算，同樣將語料庫劃分為 1000 個子語料庫，計算詞彙的 Griess_dp 值，有效範圍在 0.04-0.92 之間，同時我們發現在 Griess_dp 較高的詞彙中同樣包含像“音標”，“屈折”等語言學意義豐富的學術詞彙，而我們不想這些詞彙被排除，因此我們保留了整個詞表，得到了我們最終的語言學學科專用學術詞表，共包含 2764 個詞種。詳細詞表見附件。

我們以漢語語言學學學術詞表排名第 2761 位的“位移”一詞為例，將其在語料運用中的實例以及實例的關鍵元資訊加以展示。

詞語“位移”實例 1：

期刊來源：《當代語言學》

發表時間：2020.4

題目：漢語 A’-型依存結構在最簡方案下的句法推導

作者：潘俊楠

例句：而重建效應只與語缺、複指代詞和零代詞等受 A’-約束成分的內部結構有關,與位移無關。本文證實孤島效應和重建效應產生的根本原因以及出現的句法環境不同。

詞語“位移”實例 2：

期刊來源：《外語電化教學》

發表時間：2020.10

題目：國內英語專業知識體系的系譜學考察

作者：張和龍

例句：本文採用系譜學的研究方法,考察清末、民國與新中國 70 年英語專業知識體系的歷史嬗變,梳理文學學科從中心向邊緣位移的演變過程,探討中國社會變遷、教育政策沿革以及課程設置變化對知識體系建構的影響,並對專業知識體系的重構提出思考。

4.5.5 去除停用詞

我們在獲得詞表中去除在三個停用詞表的重複詞彙，一共 533 個詞彙。所得到去除停用詞版本的通用學術詞表包含 2231 個詞彙。詳細詞表見附件。

4.6 小結

本章我們使用我們構建的上億字元大型語料庫，創建了一個學術語料通用學術詞表與四個學科專用學術詞表，並對其進行了自然語言處理中的停用詞去除操作，最後得到去除停用詞和不去除停用詞兩個版本一共十個詞表。使用了頻次，跨文本分佈、詞語均勻分佈值三個指標，並根據詞表的實際情況設定了篩選條件，接下來我們會對結果與現有常用詞表進行對比驗證。

5 結果與討論

本章中我們將實驗得到的十個學術詞表與現有比較權威與常用的常用詞頻表進行對比，分別是 BCC 漢語常用詞頻表，BCC 漢語詞頻表與國家語委現代漢語語料庫詞頻表。從學術詞表在語料庫中的覆蓋率，學術詞表與常用詞表重合詞彙，重合詞彙語料庫覆蓋率，不重合詞彙語料庫覆蓋率，在常用詞表中從前往後取和學術詞表相同數量詞匯計算語料庫覆蓋率以及重合詞彙在學術詞表順序和不在對比常用詞表中的順序這幾個角度來驗證所得詞表可用性以及和現有常用詞表的差異。

5.1 通用學術詞表

首先將通用詞表與 BCC 漢語常用詞詞頻表進行比較，BCC 漢語常用詞詞頻表是近年來最具影響力的語料庫研究成果 BCC 語料庫的常用詞彙資源之一，其包含 57681 個詞種。我們主要通過 python 語言的 pandas 代碼庫來實現兩者的比較，統計結果顯示兩者之間重合的詞彙數量為 2370，不重合的詞彙數量為 233，也就是說重合詞表占通用學術詞表的比例為 0.914，不重合詞表中具有代表性的詞彙有“認知“，”有效性“，”特質“，”互聯網“等。同時整個通用詞表占通用漢語學術語料庫的比例為 0.78，重合詞表的詞彙占整個語料庫比例為 0.75，不重合詞彙僅占語料庫比例的 3%。這意味著包含 57681 個詞種 BCC 漢語常用詞詞頻表能覆蓋 90%的通用漢語學術詞表，同時剩餘的 10%佔據了 3%的語料庫詞彙。接著我們在常用詞表中從前往後截取和通學術詞表相同數量的詞種來計算語料庫覆蓋率，在這裡是 2593 個詞種，結果顯示 BCC 漢語常用詞詞頻表中最常用的前 2593 個詞種占通用漢語學術語料庫中的 0.497，和我們得到的通用學術詞表的 0.78 相比差異明顯，其中常用詞詞頻表中的”不能”,”都有”,”中央委員”等一共 38 個詞彙詞彙不包含在我們的語料庫中，同時我們還對比了重合詞彙在學術詞

表順序和不在對比常用詞表中的順序的變化，差異更加明顯，後兩項數值差異大的原因主要是 BCC 漢語常用詞詞頻表將單個漢字不作為詞彙，但通用學術詞表和 BCC 常用詞頻表同樣數量的詞覆蓋率的差異同樣能說明通用學術詞表在語料中更具代表性。

接著將通用詞表與 BCC 漢語詞頻表中的科技詞表進行對比，科技詞表包含 503691 個詞種。結果顯示兩者之間重合的詞彙數量為 2474，不重合的詞彙數量為 119，重合詞表占通用學術詞表的比例為 0.95，不重合詞表中具有代表性的詞彙有“維度“，”大資料“，”高等教育“，”弱勢“等。重合詞表的比例和 BCC 漢語常用詞詞頻表相比更大，說明我們構建的通用學術詞表與提取 BCC 漢語詞頻表中的科技詞表所使用的科技語料的重合度更高。而我們在科技詞表中從前往後截取和通學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為 0.67，科技詞表截取部分中不包含在我們語料庫中的內容大都是特殊符號與英文，不但能說明我們構建的通用詞表能更明顯覆蓋所使用的學術語料，同時還表明了 BCC 漢語詞頻表的設計過程中沒有對特殊字元進行資料清洗。重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化顯示有大約一半詞彙的相對順序發生了前移，前 100 個頻次的詞彙中發生前移的有 27 個，三個順序不變，其他全部後移，變化比較明顯的例如”治理“，”基於“，”學生“，”提升“，”構建“等詞彙，說明 BCC 漢語詞頻表中的科技詞表的詞頻排序與我們的詞表仍有一定的區別，主要體現在上述類似的學術詞彙在詞頻表中的位置。

同時在通用詞表與 BCC 漢語詞頻表中的全部詞表進行對比後展現了和科技此表對比類似的結果，在重合詞表部分完全一致，而全部詞表包含 1818649 個詞種。說明在我們構建的學術詞表不存在報刊、博客、微博、文學詞表中有但科技詞表中沒有的詞種，充分體現了詞表的學術特性。

最後是將通用詞表與國家語委現代漢語語料庫詞頻表進行對比，國家語委現代漢語語料庫詞頻收錄了出現次數 50 次以上的詞彙，一共 14629 個詞種。統計後發現與通用詞表的重合詞彙數為 2230，不重合的詞彙數量為 363，重合詞表占通用學術詞表的比例為0.86，不重合詞表中具有代表性的詞彙有“認同“，”研發“，”論壇“，”解析“等。說明通用詞表與國家語委現代漢語語料庫詞頻表和上兩個相比差異稍大，漢語語料庫詞頻表缺失了大約 15%的通用學術詞彙。在現代漢語語料庫詞頻表中從前往後截取和通學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為0.60，和通用學術詞表的覆蓋率 0.78 相比，表示同樣數量的高頻詞彙覆蓋的語料庫內容有較大差距。同時現代漢語語料庫詞頻表截取的詞種中有 44 個未在語料庫中出現，例如”中國共產黨“，”社會主義建設“，“並不是”等，我們認為這是由於分詞所使用方法的效果不同導致的。在重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化顯示有 40%的詞彙順序發生了後移，且基本是處於學術詞表前半部分的詞彙，而學術詞表後半部分的詞彙順序則普遍出現了上升，我們認為這是由於我們詞表前半部分的詞彙學術性更強，在學術語料中出現的更多，而現代漢語語料庫詞頻表所統計的多是在通用語料中更常用的詞彙，因此兩者相比較出現了學術詞表後半部分更具通用性特點的詞彙上移的現象。

5.2 法學專用學術詞表

與通用詞表進行詞表比較的過程類似，將我們所得到的法學專用學術詞表與我們現有的三個常用詞表在法學專用學術語料庫中進行對比。首先將法學專用詞表與 BCC 漢語常用詞詞頻表進行比較，統計結果顯示在包含 3002 個詞種的法學專用學術詞表與 BCC 常用詞詞頻表兩者之間重合的詞彙數量為 2648，不重合的詞彙數量為 354，重合詞

表占法學學術詞表的比例為0.88，不重合詞表中具有代表性的詞彙有“刑事訴訟法“，”民法典“，”合同法“，”法學院“等具有明顯法學意義的詞彙，體現了詞表的法學專用性。同時整個法學詞表占法學學術語料庫的比例為0.85，重合詞表的詞彙占整個語料庫比例為0.80，不重合詞彙即屬於法學詞表但不屬於常用詞詞頻表的詞彙占語料庫比例的 5%。這意味著包含 57681 個詞種 BCC 漢語常用詞詞頻表能覆蓋 88%的法學學術詞表，同時剩餘的 12%詞彙佔據了 5%的語料庫詞彙。接著我們在常用詞表中從前往後截取和法學學術詞表相同數量的詞種來計算語料庫覆蓋率，在這裡是 3002 個詞種，結果顯示 BCC 漢語常用詞詞頻表中最常用的前 3002 個詞種占通用漢語法學學術語料庫中的0.53，和我們得到的法學學術詞表的 0.85 相比差異明顯，其中常用詞詞頻表中的”不能”,” 司令員”,” 生產資料”等一共 62 個詞彙詞彙不包含在我們的語料庫中，同時我們還對比了重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化，但因為 BCC 漢語常用詞詞頻表將單個漢字不作為詞彙因此差異較大，但法學學術詞表與 BCC 常用詞頻表同樣數量的詞彙覆蓋率的差異同樣能說明法學學術詞表在法學語料中顯然更具代表性。

接著將法學專用學術詞表與 BCC 漢語詞頻表中的科技詞表進行對比，科技詞表包含 503691 個詞種。結果顯示兩者之間重合的詞彙數量為 2805，不重合的詞彙數量為 197，重合詞表占法學學術詞表的比例為 0.93，不重合詞表中具有代表性的詞彙有“民法典“，”物權法“，”刑法學“，”檢察權“等。重合詞表的比例和 BCC 漢語常用詞詞頻表相比更大，說明我們構建的法學學術詞表與提取 BCC 漢語詞頻表中的科技詞表所使用的科技語料的重合度更高。而我們在科技詞表中從前往後截取和法學學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為0.69，科技詞表截取部分中不包含在我們語料庫中的部分包括 242 個詞種，其中依舊包含大量特殊符號與英文，表明了 BCC 漢語

詞頻表的設計過程中沒有對特殊字元進行資料清洗，以及我們構建的法學專用學術詞表對學術漢語語料有更好的覆蓋率。重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化中顯示像“法律”，“制度“，“司法”，“犯罪”，“我國”，“立法”等法律意義豐富的詞彙其頻次順序都發生了明顯上移，一些常見詞的順序有所下降，充分體現了我們法學專用詞表的法學特性。

同時在法學詞表與 BCC 漢語詞頻表中的全部詞表進行對比後展現了和科技此表對比類似的結果，在重合詞表部分完全一致。說明在我們構建的學術詞表不存在報刊、博客、微博、文學詞表中有但科技詞表中沒有的詞種，充分體現了詞表的學術特性。

最後是將通用詞表與國家語委現代漢語語料庫詞頻表進行對比，統計後發現與通用詞表的重合詞彙數為 2406，不重合的詞彙數量為 596，重合詞表占法學學術詞表的比例為0.80，不重合詞表中具有代表性的詞彙有“法治“，”人權“，”物權“，”法系“等具有法學意義的詞彙，說明國家語委現代漢語語料庫詞頻表的 14629 個詞種中缺少了許多具備法學意義的詞彙，體現了我們構建的法學專用詞表與漢語語料庫詞頻表的區別，體現出法學專用學術詞表的優越性。以及法學詞表與國家語委現代漢語語料庫詞頻表重合詞表占說明法學詞表的比例和上兩個相比差異稍大，表明現代漢語語料庫詞頻表與我們構建的法學詞表差異與前兩個詞表相比較大，表中缺失了大約 20%的法學學術詞彙。在現代漢語語料庫詞頻表中從前往後截取和通學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為 0.65，和法學學術詞表的覆蓋率 0.85 相比，表示同樣數量的高頻詞彙覆蓋的語料庫內容有較大差距。同時現代漢語語料庫詞頻表截取的詞種中有 92 個未在語料庫中出現，例如”中國共產黨“，”人民群眾“，“曉得”等，我們認為這是由於分詞所使用方法的語料不同導致的，我們的法學語料庫中應該不包含現代漢語語料庫詞頻表中使用的一些政治和口語化的語料。在重合詞彙在學術詞表順

序和不在對比常用詞表中的順序的變化中同樣顯示出法學意義的詞彙排序上移，常見詞順序下降的情況，與 BCC 漢語詞頻表中的科技詞表的對比情況相似。

5.3 管理學專用學術詞表

我們管理學專用學術詞表與三個常用詞表在管理學學專用學術語料庫中進行對比。首先將管理學專用詞表與 BCC 漢語常用詞詞頻表進行比較，統計結果顯示在包含 2668 個詞種的管理學專用學術詞表與 BCC 常用詞詞頻表兩者之間重合的詞彙數量為 2398，不重合的詞彙數量為 270，重合詞表占管理學學術詞表的比例為 0.9，不重合詞表中具有代表性的詞彙有“中小企業“，”地方政府“，” 市場經濟“，”管理者“等具有明顯管理學意義的詞彙。同時整個管理學詞表占管理學學術語料庫的比例為0.87，重合詞表的詞彙占整個語料庫比例為0.83，不重合詞彙即屬於管理學詞表但不屬於常用詞詞頻表的詞彙占語料庫比例的 4%。這意味著包含 57681 個詞種 BCC 漢語常用詞詞頻表能覆蓋 90%的管理學學術詞表，同時剩餘的 10%詞彙佔據了 4%的語料庫詞彙。接著我們在常用詞表中從前往後截取和管理學學術詞表相同數量的詞種來計算語料庫覆蓋率，在這裡是 2668 個詞種，結果顯示 BCC 漢語常用詞詞頻表中最常用的前 2668 個詞種占管理學學術語料庫中的 0.56，和我們得到的管理學學術詞表的 0.87 相比差異明顯，其中常用詞詞頻表中的”不能”,” 兩國”,” 有了”等一共 43 個詞彙詞彙不包含在我們的語料庫中，同時我們還對比了重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化，同樣因為 BCC 漢語常用詞詞頻表將單個漢字不作為詞彙因此差異較大，但管理學學術詞表與 BCC 常用詞頻表同樣數量的詞彙覆蓋率的差異同樣能說明管理學學術詞表在管理學語料中顯然更具代表性。

接著將管理學專用學術詞表與 BCC 漢語詞頻表中的科技詞表與全部詞詞表進行對比，科技詞表包含 503691 個詞種，全部詞表包含 1818649 個詞種。結果顯示兩者之間重合的詞彙數量為 2526 與 2527，不重合的詞彙數量為 142 與 141，兩者之間的區別僅為“創新型”這一詞彙出現在全部詞表而在科技詞表中未出現，因此總體來看，我們的詞語與 BCC 科技詞表仍有較大的相似性。重合詞表占管理學學術詞表的比例為 0.95，不重合詞表中具有代表性的詞彙有“供應鏈“，”供應商“，”大數據“，”物流“等。重合詞表的比例和 BCC 漢語常用詞詞頻表相比更大，說明我們構建的管理學學術詞表與提取 BCC 漢語詞頻表中的科技詞表所使用的科技語料的重合度更高。而我們在科技詞表中從前往後截取和管理學學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為 0.75。科技詞表截取部分中不包含在我們語料庫中的部分包括 152 個詞種，於之前結果類似，以及我們構建的管理學專用學術詞表對學術漢語語料有更好的覆蓋率。重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化中顯示像“企業”，“發展“，“創新”，“模型”，“績效”，“戰略”等詞彙其頻次順序都發生了明顯上移，說明瞭這些詞彙具有較高的管理學意義，而一些常見詞的順序有所下降，充分體現了我們管理學專用詞表的管理學特性。

最後是將管理學專用學術詞表與國家語委現代漢語語料庫詞頻表進行對比，統計後發現與通用詞表的重合詞彙數為 2230，不重合的詞彙數量為 438，重合詞表占管理學學術詞表的比例為 0.84，不重合詞表中具有代表性的詞彙有“績效“，”審計“，”研發“，”演算法“等具有管理學意義以及學術意義的詞彙，說明國家語委現代漢語語料庫詞頻表的 14629 個詞種中缺少相應的學術詞彙，體現了我們構建的管理學專用詞表與漢語語料庫詞頻表的區別，體現出管理學專用學術詞表的優越性。以及管理學詞表與國家語委現代漢語語料庫詞頻表重合詞表占管理學詞表的比例和上兩個相比差異稍

大，表明現代漢語語料庫詞頻表與我們構建的管理學詞表差異與前兩個詞表相比較大，表中缺失了大約 16%的管理學詞彙或高頻學術詞彙。在現代漢語語料庫詞頻表中從前往後截取和管理學學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為0.66，和管理學學術詞表的覆蓋率 0.87 相比，表示同樣數量的高頻詞彙覆蓋的語料庫內容有較大差距。同時現代漢語語料庫詞頻表截取的詞種中有 56 個未在語料庫中出現，例如”中國共產黨“，”人民群眾“，“不能”等，我們認為這是由於分詞所使用方法的語料不同導致的。在重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化中同樣顯示出管理學意義的詞彙排序上移，常見詞順序下降的情況，與 BCC 漢語詞頻表中的科技詞表的對比情況相似。

5.4 教育學專用學術詞表

使用教育學專用學術詞表與三個常用詞表在教育學學專用學術語料庫中進行對比。首先將教育學專用詞表與 BCC 漢語常用詞詞頻表進行比較，統計結果顯示在包含 2749 個詞種的教育學專用學術詞表與 BCC 常用詞詞頻表兩者之間重合的詞彙數量為 2476，不重合的詞彙數量為 273，重合詞表占教育學學術詞表的比例為 0.90，不重合詞表中具有代表性的詞彙有“資訊化“，”中小學“，”學習者“，”高職“等具有明顯教育學意義的詞彙。同時整個教育學詞表占教育學學術語料庫的比例為0.87，重合詞表的詞彙占整個語料庫比例為0.73，不重合詞彙即屬於教育學詞表但不屬於常用詞詞頻表的詞彙占語料庫比例的 4%。這意味著包含 57681 個詞種 BCC 漢語常用詞詞頻表能覆蓋 90%的教育學學術詞表，同時剩餘的 10%詞彙佔據了 4%的語料庫詞彙。接著我們在常用詞表中從前往後截取和教育學學術詞表相同數量的詞種來計算語料庫覆蓋率，在這裡是 2749 個詞種，結果顯示 BCC 漢語常用詞詞頻表中最常用的前 2749 個詞種占教育學學術

語料庫中的 0.57，和我們得到的教育學學術詞表的 0.87 相比差異明顯，其中常用詞詞頻表中有 44 個詞彙詞彙不包含在我們的語料庫中，同時我們還對比了重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化，同樣因為 BCC 漢語常用詞詞頻表將單個漢字不作為詞彙因此差異較大，但教育學學術詞表與 BCC 常用詞頻表同樣數量的詞彙覆蓋率的差異同樣能說明教育學學術詞表在教育學語料中顯然更具代表性。

接著將教育學專用學術詞表與 BCC 漢語詞頻表中的科技詞表與全部詞詞表進行對比，科技詞表包含 503691 個詞種，全部詞表包含 1818649 個詞種。結果顯示兩者之間 BCC 科技詞表重合詞彙的數量為 2595，全部詞表為 2596，僅有“創新型”一個詞彙產生區別，可見這個詞彙不僅在作為我們語料庫中的學術詞彙，在 BCC 語料庫的其他部分例如微博、新聞語料中也有出現，詞種數相差如此多但僅有出現一個詞彙的差別，足以說明 BCC 語料庫的科技詞表與我們的學術詞表更具有相似性。重合詞表占教育學學術詞表的比例為0.94，不重合詞表中具有代表性的詞彙有“高等教育“，”高等學校“，”國家教委“，”一“等。重合詞表的比例和 BCC 漢語常用詞詞頻表相比更大，說明我們構建的教育學學術詞表與提取 BCC 漢語詞頻表中的科技詞表所使用的科技語料的重合度更高。而我們在科技詞表中從前往後截取和教育學學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為0.74。科技詞表截取部分中不包含在我們語料庫中的部分包括 155 個詞種，不包含的內容與之前結果類似，我們構建的教育學專用學術詞表對教育學語料有更好的覆蓋率。重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化中顯示像“‘‘語言”’，“教學“，“教師”，“大學”，“學習”，“課程”等詞彙其頻次順序都發生了明顯上移，其中順序最高“教育”從科技詞表中的排序 70 名上升至第二名，“課程”這一詞彙上升了 771 個順位，說明瞭這些詞彙具有較高的語

言學意義，而一些常見詞的順序有所下降，充分體現了我們教育學專用詞表的教育學特性。

最後是將教育學專用學術詞表與國家語委現代漢語語料庫詞頻表進行對比，統計後發現與通用詞表的重合詞彙數為 2310，不重合的詞彙數量為 439，重合詞表占教育學學術詞表的比例為0.84，不重合詞表中具有代表性的詞彙有“師範大學“，”國家教委“，”互聯網“，”教學法“等具有教育學意義以及學術意義的詞彙，說明國家語委現代漢語語料庫詞頻表的 14629 個詞種中缺少相應的學術詞彙，體現了我們構建的教育學專用詞表與漢語語料庫詞頻表的區別，體現出教育學專用學術詞表在教育學領域的優越性。以及教育學詞表與國家語委現代漢語語料庫詞頻表重合詞表占教育學詞表的比例和上兩個相比差異稍大，表明現代漢語語料庫詞頻表與我們構建的教育學詞表差異與前兩個詞表與前兩個相比較大，表中缺失了大約 16%的教育學詞彙或高頻學術詞彙。在現代漢語語料庫詞頻表中從前往後截取和教育學學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為 0.68，和教育學學術詞表的覆蓋率 0.87 相比，表示同樣數量的高頻詞彙覆蓋的語料庫內容有較大差距。同時現代漢語語料庫詞頻表截取的詞種中有 49 個未在語料庫中出現，我們認為這是由於分詞所使用方法的語料不同導致的。在重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化中同樣顯示出教育學意義的詞彙排序上移，常見詞順序下降的情況，與 BCC 漢語詞頻表中的科技詞表的對比情況相似。

5.5 語言學學專用學術詞表

使用語言學專用學術詞表與三個常用詞表在語言學學專用學術語料庫中進行對比。首先將語言學專用詞表與 BCC 漢語常用詞詞頻表進行比較，統計結果顯示在包含 2764

個詞種的語言學專用學術詞表與 BCC 常用詞詞頻表兩者之間重合的詞彙數量為 2532，不重合的詞彙數量為 232，重合詞表占語言學學術詞表的比例為 0.92，不重合詞表中具有代表性的詞彙有“語料庫“，”語法化“，”語言學家“，”語體者“等具有明顯語言學意義的詞彙。同時整個語言學詞表占語言學學術語料庫的比例為0.80，重合詞表的詞彙占整個語料庫比例為0.76，不重合詞彙即屬於語言學詞表但不屬於常用詞詞頻表的詞彙占語料庫比例的 4%。這意味著包含 57681 個詞種 BCC 漢語常用詞詞頻表能覆蓋92%的語言學學術詞表，同時剩餘的 8%詞彙佔據了 4%的語料庫詞彙。接著我們在常用詞表中從前往後截取和語言學學術詞表相同數量的詞種來計算語料庫覆蓋率，在這裡是 2764 個詞種，結果顯示 BCC 漢語常用詞詞頻表中最常用的前 2764 個詞種占語言學學術語料庫中的 0.43，和我們得到的語言學學術詞表的 0.80 相比差異明顯，其中常用詞詞頻表中有 68 個詞彙詞彙不包含在我們的語料庫中，同時我們還對比了重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化，同樣因為 BCC 漢語常用詞詞頻表將單個漢字不作為詞彙因此差異較大，但語言學學術詞表與 BCC 常用詞頻表同樣數量的詞彙覆蓋率的差異同樣能說明語言學學術詞表在語言學語料中顯然更具代表性。

接著將語言學專用學術詞表與 BCC 漢語詞頻表中的科技詞表與全部詞詞表進行對比，科技詞表包含 503691 個詞種，全部詞表包含 1818649 個詞種。結果顯示兩者之間重合的詞彙數量完全相同，重合詞彙數量都為 2608，足以說明 BCC 語料庫的科技詞表與我們的學術詞表具有驚人的相似性。重合詞表占語言學學術詞表的比例為0.94，不重合詞表中具有代表性的詞彙有“語篇“，”語用學“，”印書館“，”構詞“等。重合詞表的比例和 BCC 漢語常用詞詞頻表相比更大，說明我們構建的語言學學術詞表與提取 BCC 漢語詞頻表中的科技詞表所使用的科技語料的重合度更高。而我們在科技詞表中從前往後截取和語言學學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為0.64。

科技詞表截取部分中不包含在我們語料庫中的部分包括 193 個詞種，與之前結果類似，以及我們構建的語言學專用學術詞表對學術漢語語料有更好的覆蓋率。重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化中顯示像“語言”，“研究“，“翻譯”，“教學”，“漢語”，“英語”等詞彙其頻次順序都發生了明顯上移，說明瞭這些詞彙具有較高的語言學意義，而一些常見詞的順序有所下降，充分體現了我們語言學專用詞表的語言學特性。

最後是將語言學專用學術詞表與國家語委現代漢語語料庫詞頻表進行對比，統計後發現與通用詞表的重合詞彙數為 2363，不重合的詞彙數量為 401，重合詞表占語言學學術詞表的比例為 0.85，不重合詞表中具有代表性的詞彙有“修辭學“，”語用“，”語篇“，”語料庫“等具有語言學意義以及學術意義的詞彙，說明國家語委現代漢語語料庫詞頻表的 14629 個詞種中缺少相應的學術詞彙，體現了我們構建的語言學專用詞表與漢語語料庫詞頻表的區別，體現出語言學專用學術詞表在語言學領域的優越性。以及語言學詞表與國家語委現代漢語語料庫詞頻表重合詞表占語言學詞表的比例和上兩個相比差異稍大，表明現代漢語語料庫詞頻表與我們構建的語言學詞表差異與前兩個詞表與前兩個相比較大，表中缺失了大約 15%的語言學詞彙或高頻學術詞彙。在現代漢語語料庫詞頻表中從前往後截取和語言學學術詞表相同數量的詞種來計算語料庫覆蓋率的結果為 0.62，和語言學學術詞表的覆蓋率 0.80 相比，表示同樣數量的高頻詞彙覆蓋的語料庫內容有較大差距。同時現代漢語語料庫詞頻表截取的詞種中有 51 個未在語料庫中出現，我們認為這是由於分詞所使用方法的語料不同導致的。在重合詞彙在學術詞表順序和不在對比常用詞表中的順序的變化中同樣顯示出語言學意義的詞彙排序上移，常見詞順序下降的情況，與 BCC 漢語詞頻表中的科技詞表的對比情況相似。

6 結語

本次實驗我們基於 2021-2022 中文社會科學引文索引（CSSCI）來源期刊目錄自主構建了一個上億字元的學術語料庫，並使用語料庫進行漢語學術詞表提取方面的研究。我們綜合了相關領域的現有研究成果，決定使用頻率、跨文本分佈、詞語均勻分佈值三個特徵進行篩選，最後得到了一個通用學術漢語詞表與法學、管理學、教育學、語言學四個專用學科學術漢語詞彙表，並參考自然語言處理的方法對其進行了去除停用詞處理，最後得到了一共十個詞表的研究成果。

接著我們將不同的詞表與 BCC 漢語常用詞詞頻表、BCC 漢語詞頻表以及國家語委現代漢語語料庫詞頻表三個詞表進行了對比，結果顯示雖然我們詞表的詞彙很大程度上包括在了上述通用詞表中，但這是由於詞表的規模不同導致的，在同樣數目出現頻次最多的詞彙進行對比，通用詞表和我們構建的學術詞表在學術語料庫中的覆蓋率差距在 20%-30%之間，展現了學術語料與通用語料在詞彙使用上的差異，同時說明瞭我們構建的學術詞表在學術語料中相較於傳統通用詞表的優越性。在分析具體詞表詞彙的過程中，我們也發現不同學科在詞彙使用方便也存在這明顯的不同，屬於各個學科常用的詞彙在不同學科專用詞表的頻率排序變化非常顯著。這也說明瞭我們提取的學科專用詞表的有效性。

綜合來說，本實驗設計的創新點體現在三個部分。第一，我們將學術界常用學術詞表提取的語料庫大小提升了兩個數量級，由百萬級提升到上億字。大型語料庫更充分的包含了學術詞彙所代表的資訊，而充分利用語料庫中的詞彙資訊，也就意味著我們所得到的詞表更加可信，更加科學。第二，我們在分詞階段所使用的 pku_seg 代碼庫是近年來自然語言處理領域的優秀研究成果，與其他分詞方法例如軟體分詞、人工檢

查的方法相比，無論在準確性還是效率都體現了一定的優勢。第三，在測量詞語均勻分佈值部分，我們經過研究並沒有採用常用的 Julliand Disperson 作為我們分佈值的篩選條件，而是創新性的使用 Griess_dp 來進行詞語均勻分佈值的衡量，避免了 Julliand ‘D 值在面對大型語料庫中有效範圍災難性下降的問題。本實驗吸取了國際領域學術此表提取的豐富經驗，同時加入了自主研究的創新點，最終形成了一個基於大型學術預料庫進行學術詞表提取的研究流程，屬於語料庫語言學研究成果。

當然，實驗也存在一些缺陷。首先，語料庫的實驗資料我們採用的使用學術期刊摘要部分的全部語料，雖然摘要作為學術文獻的濃縮，具備鮮明的學術特性，但同樣由於其概括性強的特點，可能與正文部分的語料存在一定的差異。未來進行改進的方向之一就是擴充語料庫中學術文本的種類，將期刊正文或者教材文本加入語料庫的內容中，接著將其匯總進行分析，期以得到更優秀的詞表成果。第二，由於時間原因以及收集語料的難度，我們創建通用學術詞表使用的語料僅是 2020 至 2022 年 8 月的期刊資訊，以及學科專用詞表僅完成了法學、管理學、教育學、語言學四個部分，下一步的研究要繼續擴充學術語料庫的內容，完成其他學科的歷史期刊的語料收集工作。

總的來說，本實驗構建了基於學術語料庫進行學術詞表提取的研究流程，推動了漢語學術詞表提取的研究進展，根據多特徵提取的學術詞表成果在包括詞彙研究、學科教育等領域都有豐富的應用前景。

參考文獻

Alnawas, A., & Arıcı, N. (2018). The Corpus Based Approach to Sentiment Analysis in Modern Standard Arabic and Arabic Dialects: A Literature Review. *Journal of Polytechnic*.

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non)utility of Juilland's D to measure lexical dispersion in large corpora [Article]. *International Journal of Corpus Linguistics*, *21*(4), 439-464. https://doi.org/10.1075/ijcl.21.4.01bib

Biemiller, A. (1999). Language and reading success.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., & Amodei, D. (2020). Language Models are Few-Shot Learners.

Chen, Q., & Ge, G. C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs) [Article]. *English for Specific Purposes*, *26*(4), 502-514. https://doi.org/10.1016/j.esp.2007.04.003

Corson, D. (1997). The Learning and Use of Academic English Words. *Language Learning*, *47*(4), 671-718. https://doi.org/10.1111/0023-8333.00025

Coxhead, A. (2000). A new academic word list [Article]. *Tesol Quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Coxhead, A. (2016). Reflecting on Coxhead (2000), "A New Academic Word List" [Article]. *Tesol Quarterly*, *50*(1), 181-185. https://doi.org/10.1002/tesq.287

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list [Article]. *Revue Francaise De Linguistique Appliquee*, *12*(2), 65-78. <Go to ISI>://WOS:000261303800005

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why Might Secondary Science Textbooks be Difficult to Read? *New Zealand Studies in Applied Linguistics*.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English [Article]. *English for Specific Purposes*, *33*, 66-76. https://doi.org/10.1016/j.esp.2013.08.001

Dresher, R. (1934). Training in Mathematics Vocabulary. *Educational Research Bulletin*, *13*(8), 201-204. https://doi.org/10.2307/1470596

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? [Article]. *English for Specific Purposes*, *43*, 49-61. https://doi.org/10.1016/j.esp.2016.01.004

Farrell, P. (1990). Vocabulary in ESP: A Lexical Analysis of the English of Electronics and a Study of Semi-Technical Vocabulary. CLCS Occasional Paper No. 25. *Electronics*, 87.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System*.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List [Article]. *Applied Linguistics*, *35*(3), 305-327. https://doi.org/10.1093/applin/amt015

Goldenberg, C. (2008). Teaching English language learners: what the research does — and does not — say. *American Educator*. https://doi.org/ http://www.aft.org/pdfs/americaneducator/summer2008/goldenberg.pdf

Gries, S. T. (2008). Dispersions and Adjusted Frequencies in Corpora. *International Journal of Corpus Linguistics*, *13*(4), 403-437. https://doi.org/10.1075/ijcl.13.4.02gri

Hsu, W. H. (2018). The most frequent BNC/COCA mid- and low-frequency word families in English-medium traditional Chinese medicine (TCM) textbooks [Article]. *English for Specific Purposes*, *51*, 98-110. https://doi.org/10.1016/j.esp.2018.04.001

Hunston, S. (2003). Corpora in Applied Linguistics. *Elt Journal*, *57*(4), págs. 416-420.

Hyland, K., & Tse, P. (2007). Is there an "Academic vocabulary"? [Article]. *Tesol Quarterly*, *41*(2), 235-253. https://doi.org/10.2307/40264352

Ippolito, J., Steele, J. L., & Samson, J. F. (2008). Introduction: Why Adolescent Literacy Matters Now. *Harvard educational review*, *78*(1), 1-6. https://doi.org/10.1111/j.1465-3435.2007.00338.x

Jacobs, V. (2008). Adolescent Literacy: Putting the Crisis in Context. *Harvard educational review*, *78*(1), 7-39. https://doi.org/10.17763/haer.78.1.c577751kq7803857

Kennedy, G., & Ooi, V. (1998). An Introduction to Corpus Linguistics.

Lee, S.-H., & Kesuke, I. (2017). Native Speaker of the Japanese Speech Style Shift to Personal Relationships: Using Corpus of Spoken Japanese by Basic Transcription System for Japanese [対人関係に応じた日本語母語話者のスタイル切り換え - 『btsj による日本語話し言葉コーパス』を用いて -] [research-article]. *Journal of North-east Asian Cultures*, *1*(50), 275-288. https://doi.org/10.17949/jneac.1.50.201703.015

Lei, L., & Liu, D. L. (2016). A new medical academic word list: A corpus-based study with enhanced methodology [Article]. *Journal of English for Academic Purposes*, *22*, 42-53. https://doi.org/10.1016/j.jeap.2016.01.008

Li, Y. Y., & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus [Article]. *System*, *38*(3), 402-411. https://doi.org/10.1016/j.system.2010.06.015

Liu, J., & Han, L. N. (2015). A corpus-based environmental academic word list building and its validity test [Article]. *English for Specific Purposes*, *39*, 1-11. https://doi.org/10.1016/j.esp.2015.03.001

Martinez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study [Article]. *English for Specific Purposes*, *28*(3), 183-198. https://doi.org/10.1016/j.esp.2009.04.003

MichaelWest. (1953). *A General Service List Of English Words*. Longman Group Limited.

Munoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study [Article]. *English for Specific Purposes*, *39*, 26-44. https://doi.org/10.1016/j.esp.2015.04.001

Ozkan, B. (2013). Corpus Based Dictionary of Turkey Turkish's Lexicon: Method and Application. *Bilig - Turk DunyasI Sosyal Bilimler Dergisi*(66), 149-178.

Rundell, M., 夏立新, & 朱冬生. (2009). 语料库词典学的最新发展和未来趋势(上)——语料库数据在学习词典中的显性应用. *辞书研究*(3), 8.

Tognini-Bonelli, & Elena. (2001). Corpus Linguistics at Work. *Computational Linguistics*, *28*(4), 583-583.

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes*, *12*(4), 248-263. https://doi.org/https://doi.org/10.1016/j.jeap.2013.07.001

Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a Medical Academic Word List [Editorial Material]. *English for Specific Purposes*, *27*(4), 442-458. https://doi.org/10.1016/j.esp.2008.05.003

Wang, M., & Paul, N. (2004). Word Meaning in Academic English: Homography in the Academic Word List. *Applied Linguistics*(3), 3. https://doi.org/10.1093/applin/25.3.291

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates [Article]. *English for Specific Purposes*, *28*(3), 170-182. https://doi.org/10.1016/j.esp.2009.04.001

William, Nagy, Dianna, & Townsend. (2012). Words as Tools: Learning Academic Vocabulary as Language Acquisition. *Reading Research Quarterly*, *47*(1), 91-108. https://doi.org/10.1002/rrq.011

Wingrove, P. (2017). How suitable are TED talks for academic listening? [Article]. *Journal of English for Academic Purposes*, *30*, 79-95. https://doi.org/10.1016/j.jeap.2017.10.010

Xue & Nation. (1984). A university word list. *LanguageLearning and Communication*, *2*, 215-229.

Yang, H. (1986). A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts: (An Interim Report). *Literary and Lingu Computing*(2), 2. https://doi.org/10.1093/llc/1.2.93

Yang, M. N. (2015). A nursing academic word list [Article]. *English for Specific Purposes*, *37*, 27-38. https://doi.org/10.1016/j.esp.2014.05.003

柏晓静, 常宝宝, 詹卫东, & 吴拥华. (2002). 构建大规模的汉英双语平行语料库. 机器翻译研究进展——2002年全国机器翻译研讨会论文集,

冯跃进, & 潘璠. (1998). 语料库语言学的最新动态及未来发展趋势. *山东外语教学*(4), 5. https://doi.org/CNKI:SUN:SDWY.0.1998-04-001

冯志伟. (2002). 中国语料库研究的历史与现状. *Journal of Chinese Language and Computing*, *11*, 127~136.

高增霞, & 刘福英. (2016). 论学术汉语在对外汉语教学中的重要性. *云南师范大学学报：对外汉语教学与研究版*, *14*(2), 8. https://doi.org/CNKI:SUN:YNJX.0.2016-02-009

韩露, 余静, 吴虹, & 余亚微. (2017). 汉英双语平行语料库在高职英语教学中的应用研究——以中医双语翻译人才培养为例. *职教论坛*(8), 5. https://doi.org/CNKI:SUN:ZJLT.0.2017-08-018

何中清, & 彭宣维. (2011). 英语语料库研究综述:回顾,现状与展望. *外语教学*(1), 6.

黄水清, & 王东波. (2021). 国内语料库研究综述. *信息资源管理学报*, *11*(3), 15.

林丽. (2013). 试析框架语义标注在新闻事件抽取中的应用——以越南语军事新闻为例. *山西大学学报：自然科学版*, *36*(4), 7.

劉華, & 李曉源. (2022). 基于語料庫的中醫漢語主題詞表構建. *華文教學與研究*(02), 77-85. https://doi.org/10.16131/j.cnki.cn44-1669/g4.2022.02.012.

娄宝翠. (2020). 语料库在研究生学术英语教学中的应用探索. *学位与研究生教育*(7), 6. https://doi.org/10.16750/j.adge.2020.07.008

羅振聲. (1996). 清華大學 TH 大型通用漢語語料庫系統的研制. *清華大學學報(哲學社會科學版)*(01), 94-98. https://doi.org/CNKI:SUN:QHDZ.0.1996-01-015

苗祥, 刘业政, & 孙春华. (2014). 领域同义特征词的统计规律及其在情感分析上的应用研究. *计算机应用研究*, *31*(11), 4. https://doi.org/10.3969/j.issn.1001-3695.2014.11.030

饶洋辉, 李青, 刘文印, & 李晶晶. (2014). 公众文本之情感词典研究进展. *中国科学:信息科学*, *44*(7), 825-835. https://doi.org/10.1360/N112013-00167

斯日古楞. (2010). *《現代蒙古語語料庫管理平臺》建設* 內蒙古大學].

孙东云. (2018). BCC 汉语语料库在英汉翻译教学中的应用. *外语教学理论与实践*(3), 9. https://doi.org/CNKI:SUN:GWJX.0.2018-03-011

王笑然, & 王佶旻. (2022). 經貿類本科專業學術漢語詞表研究. *語言教學與研究*(04), 9-19. https://doi.org/CNKI:SUN:YYJX.0.2022-04-002.

吴瑾, & 王同顺. (2007). Coxhead"学术词汇表"的适用性研究. *外语教学理论与实践*(2), 28-33. https://doi.org/10.3969/j.issn.1674-1234.2007.02.005

吴璟. (2010). 学术词汇表(AWL)的研究进展与启示. *云南农业大学学报：社会科学版*, *4*(1), 6. https://doi.org/CNKI:SUN:YNNS.0.2010-01-018

薛蕾. (2017). *基于漢語語言學論文語料庫的學術漢語詞匯析取及特征研究* [碩士, 云南師范大學].

薛松. (2003). *汉英平行语料库中名词短语对齐算法的研究* 中国科学院软件研究所].

荀恩东, 饶高琦, 肖晓悦, & 臧娇娇. (2016). 大数据背景下 BCC 语料库的研制. *语料库语言学*(1), 18.

杨惠中, & 卫乃兴. (2002). *语料库语言学导论*. 语料库语言学导论.

詹卫东, 郭锐, 常宝宝, 谌贻荣, & 陈龙. (2019). 北京大学 CCL 语料库的研制. *语料库语言学*(1), 17.

张宝林. (2019). 从 1.0 到 2.0——汉语中介语语料库的建设与发展. *国际汉语教学研究*(4), 12. https://doi.org/CNKI:SUN:HJXY.0.2019-04-017

赵志刚. (2015). 专门用途英语学术词表创建研究——以航海英语为例. *重庆交通大学学报：社会科学版*, *15*(6), 5. https://doi.org/CNKI:SUN:CQJS.0.2015-06-033

郑艳群. (2013). 语料库技术在汉语教学中的应用透视. *语言文字应用*(1), 8. https://doi.org/CNKI:SUN:YYYY.0.2013-01-027

周强. (2004). 汉语句法树库标注体系. *中文信息学报*, *18*(4), 2-9.

朱冬生, M. R. 夏. (2009). 语料库词典学的最新发展和未来趋势(下)——语料库数据在学习词典中的显性应用. *辞书研究*.
朱明玉. (2020). *通用學術漢語詞表研究* [碩士, 云南大學].