

香港語文教學語篇 難易度測定：

基於漢字測試語料庫的研究

彭志全 博士 / Dr. PANG Chi-chuen



香港語文教學語篇難易度測定：
基於漢字測試語料庫的研究

彭志全

紅出版（青森文化）

2026 年 10 月

Assessment of Text Difficulty in Hong Kong Language Education:
A Study Based on a Chinese Character Test Corpus

by
Pang Chi Chuen

Red Publish (Green Forest)

October 2026

出版說明

本書由作者於 2025 年 9 月提交的香港教育大學教育博士論文修訂而成。

摘要

在信息科技高速發展的時代，語言教學領域愈來愈重視善用網絡資源與人工智能（AI）技術，提升教學質量與學習效能。近年來，無論中國內地、臺灣，還是香港，皆積極推動語文教學智能化，特別是在自動化閱讀難易度分級、個性化教材推薦等方面取得突破。AI 及自然語言處理（NLP）不僅能自動分析大量文本的詞彙、語法和篇章結構，更可結合學生學習數據，為教師推薦最適切、最具可讀性的學習材料，推動分級閱讀與針對性教學。本研究據此建立適用於初小至初中（P1-S3）的九級字表。

本文立足香港語文教學環境，結合語料庫語言學理論與近年 AI 技術成果，綜合參考國內外語料庫建設經驗，自主研發漢字語料庫，並制定《九級字表》（九級＝小一至小六＋初中一至三），建立量化文本篇章難易度的標準。為方便教師應用及提高教學效率，筆者設計並開發「九級字表分析」網上平台，利用 AI 自動分級功能，為每篇教材快速評定難度，生成科學、客觀的難易度數據。教師可據此整理分級教材，編製校本課程，並利用學習語料庫提升教學針對性，兼顧學生多樣學習需要。

在校本課程推行過程中，AI 語料庫平台亦可協助教師建立評估語料庫，方便針對學生能力進行分級評核，並自動生成閱讀分析報告，為教學決策與學生學習提供數據支持。進一步地，教師還可據「評估語料庫」設計前測、後測，定期追蹤學生閱讀能力發展，並將結果反饋到學習語料庫，完善教材編排與教學策略。

通過對語料庫建構規律與 AI 技術的探索，本文確立了科學系統的教學框架，有助推動香港漢語語言學習語料庫發展和教學改革，為未來語文教育提供重要的理論參考與實踐示範。希望該語料庫平台可成為教師教材編寫、課堂教學、校本設計和智能評估的重要幫手，真正實現減輕教師負擔、提升學生語文素養的目標。

關鍵詞：語料庫語言學、人工智能（AI）、九級字表、分級教學／分級閱讀、校本課程

Abstract

In an era of rapid information technology development, language education has increasingly emphasized the integration of online resources and artificial intelligence (AI) technologies to enhance teaching quality and learning effectiveness. In recent years, Mainland China, Taiwan, and Hong Kong have all been actively advancing the digitalization of language teaching, especially achieving breakthroughs in automatic readability assessment and personalized material recommendation. AI and natural language processing (NLP) can automatically analyze a large number of texts for vocabulary, syntax, and discourse structure, and, by integrating student learning data, recommend the most appropriate and readable materials for teachers—thus facilitating tiered reading and targeted instruction.

This study, grounded in the context of Hong Kong language education and drawing on corpus linguistics theory and recent AI advancements, references domestic and international corpus-building experiences to independently develop a Chinese character corpus and a “Nine-Level Character List” (Nine Levels = Primary 1 to Primary 6 + Secondary 1 to Secondary 3 [P1–P6 + S1–S3]) as a quantitative standard for text difficulty. To facilitate practical application and improve teaching efficiency, a web-based “Nine-Level Character List Analysis” platform has been developed, utilizing AI-driven automated grading functions to efficiently evaluate the difficulty of teaching materials and generate scientific and objective data. Teachers can use this to organize graded materials, design school-based curricula, and leverage

the learning corpus to enhance teaching precision and cater to diverse learner needs.

In the implementation of school-based curricula, the AI corpus platform can also assist teachers in establishing assessment corpora, enabling targeted, level-based student assessments and automatically generating reading analysis reports to inform teaching decisions and student learning. Furthermore, teachers can design pre- and post-tests based on the assessment corpus to regularly monitor student reading development, with results feeding back into the learning corpus to refine materials and teaching strategies.

By exploring the principles of corpus construction and AI technology, this study establishes a scientific and systematic instructional framework, helping to promote the development of Chinese learning corpora and educational reform in Hong Kong. This provides an important theoretical reference and practical model for future language education. It is hoped that this corpus platform will become an indispensable tool for teachers in classroom instruction, teaching material development, school-based curriculum design, and intelligent assessment, thereby reducing teachers' workload and improving students' language proficiency.

Keywords: Corpus Linguistics, Artificial Intelligence (AI), Nine-Level Character List, Tiered Instruction / Graded Reading, School-Based Curriculum

致謝聲明

本研究能順利完成，承蒙多位良師益友的支持與啟發，特此致以誠摯謝意。

首先，衷心感謝陳學超老師、謝家浩老師、張連航老師及廖先老師的悉心指導與鼓勵。諸位老師在我學術旅程中的不同階段均給予我寶貴的啟發與堅定的支持，無論在研究方向的确立、理論架構的建構，還是寫作過程中的精心點撥，皆使我獲益良多。沒有諸位的栽培與信任，本研究難以成形。

同時感謝香港教育大學研究院提供完善的學術資源與行政支援，使我能專心投入研究。特別鳴謝在研究初期協助建立語料庫基礎的中學同事葉榮昌老師，以及於我任職香港大學期間擔任研究助理、協助資料整理和建立網上平台的楊曦彤小姐，二位的專業協助與無私支持，對本研究的順利推展貢獻良多。亦感謝參與研究的學校、教師與學生，讓本研究得以順利推展。

最後，衷心感謝家人無條件的包容與支持，使我能無後顧之憂地完成這段漫長的學術旅程。

彭志全

2025 年 9 月

目錄

出版說明	v
摘要	vi
Abstract	viii
致謝聲明	x
縮寫表	xv
列表清單	xvii
第一章 緒論	18
1.1 研究背景與問題	18
1.2 研究目的與意義	20
1.2.1 香港的語言狀況及「兩文三語」政策	20
1.2.2 香港的學校語言課程設置	20
1.2.3 反思與小結	22
1.3 研究問題與假設	23
1.3.1 利用語料庫推動漢語文教學	23
1.3.2 建立學習語料庫幫助學生有效學習漢語	23
1.4 研究方案	24
1.4.1 研究目的	24
1.4.2 研究方法	25
1.4.3 創新之處和研究難點	27
第二章 文獻綜述	30
2.1 語料庫設計的概況	30
2.1.1 語料庫建設的意義	30

2.1.2 語料庫建立整體水準提升的需要	30
2.1.3 語文教學建立的需要	32
2.2 語言學習語料庫源流	33
2.2.1 相關概念	33
2.2.2 語言學習語料庫的定義	34
2.2.3 語言學習語料庫的相關數據檢索功能	36
2.2.4 語言學習語料庫類別	37
2.2.5 語言學習語料庫拓展史	39
2.2.6 漢語文課程語言學習語料庫和漢語文教學	44
2.3 語言學習語料庫建設現狀（國外－內地－香港）	45
2.3.1 國外研究現狀分析	45
2.3.2 內地研究現狀分析	49
2.3.3 香港研究現狀分析	65
2.4 閱讀難度模型相關理論	68
2.4.1 閱讀理論基礎：簡單閱讀觀與 Lexile 框架	69
2.4.2 詞彙與字級難度分級模型	70
2.4.3 香港現有字表比較	72
第三章 研究方法	74
3.1 語文教學相關語言學習語料庫的設計	74
3.1.1 語言學習語料庫的建庫目的	74
3.1.2 語料樣例的容量	76
3.1.3 語料樣例的多樣化	77
3.1.4 語料樣例的處理	80
3.2 閱讀理論研究	80
3.2.1 「語言輸入假設」及「圖式理論」	80

3.2.2 文本難易度研究	82
3.3 語文教學相關語言學習語料庫的開發	87
3.3.1 統計的依據	87
3.3.2 採集語料	92
3.3.3 研究詞表的編製	93
3.4 研究詞表與漢語水準考試詞表的比較分析	97
3.4.1 術語界定	105
第四章 香港語文學習語料庫的應用	108
4.1 語文學習語料庫在教學中的應用	108
4.1.1 字本位的選擇	108
4.1.2 語料庫教學任務和內容	111
4.1.3 小學階段學習語料庫和教學應用	111
4.1.4 初中階段學習語料庫的創建和教學應用	123
4.1.5 創建更多學習語料庫的可行性	142
4.2 語文課程評估語料庫在教學中的應用	146
4.2.1 校本評估的擬卷探究	146
4.2.2 「全港性系統評估」的擬卷探究	156
4.2.3 學校校內考試與公開考試的相互影響	166
4.3 學習語料庫與評估語料庫在教學中的配合	167
4.3.1 特徵選取	167
4.3.2 演算法：基於語料統計的字表切分與可讀性評估	168
4.3.3 教師回饋機制	170
4.3.4 界面設計邏輯	171
4.3.5 小結	172

第五章 討論與結論	174
5.1 理論與技術創新	174
5.1.1 分級方法比較：本研究模型與兩套大型語言模型的特點 與應用考量	176
5.2 教育啟示	177
5.3 研究局限與改善	179
5.3.1 未來展望：界面升級構想與色彩輔助設計	181
5.4 跨域發展	182
5.5 學術意義	183
5.6 社會價值	184
5.7 建議	185
5.7.1 動態更新策略：長期追蹤與優化	185
5.7.2 教師回饋平台建構	186
5.7.3 系統開放與 GitHub 構想	188
5.8 研究結論	190
參考文獻	192
附錄 A. 學習語料庫小學字表	206
附錄 B. 教育局中學中國語文學習參考篇章目錄（初中階段）	228
網上附錄目錄	255

縮寫表

縮寫	中文全稱	英文全稱
EDB	教育局	Education Bureau
TSA	全港性系統評估	Territory-wide System Assessment
HSK	漢語水平考試	Hanyu Shuiping Kaoshi
NLP	自然語言處理	Natural Language Processing
AI	人工智能	Artificial Intelligence
LIVAC	漢語多元語言對比語料庫	Linguistic Variations in Chinese Communities
QEF	優質教育基金	Quality Education Fund
QF	資歷架構	Qualifications Framework
CDC	課程發展議會	Curriculum Development Council
CRIE	中文閱讀理解引擎	Chinese Reading Index Engine
HKUST	香港科技大學	Hong Kong University of Science and Technology
CUHK	香港中文大學	The Chinese University of Hong Kong
BNC	英國國家語料庫	British National Corpus
CHILDES	幼兒語言資料交換系統	Child Language Data Exchange System
API	應用程式界面	Application Programming Interface
HKC	香港中文句子與篇章閱讀語料庫	Hong Kong Corpus of Chinese Sentence and Passage Reading
CUCHILD	香港粵語兒童語言資料庫	Cantonese Child Language Corpus
HKDSE	香港中學文憑考試	Hong Kong Diploma of Secondary Education Examination
LLM	大型語言模型	Large Language Model

GPT	生成式預訓練模型	Generative Pre-trained Transformer
BERT	雙向編碼器表示模型	Bidirectional Encoder Representations from Transformers
OCR	光學字符識別	Optical Character Recognition
ELRA	歐洲語言資源協會	European Language Resources Association
ACL	計算語言學學會	Association for Computational Linguistics
HKBU	香港浸會大學	Hong Kong Baptist University
HKPU	香港理工大學	The Hong Kong Polytechnic University
UNESCO	聯合國教科文組織	United Nations Educational, Scientific and Cultural Organization

列表清單

- 表 1： 國內語料庫建設一覽表
- 表 2： 影響閱讀教材難度的變數
- 表 3： 研究詞表的構成
- 表 4： 《香港小學學習字詞表》及《中學中國語文學習參考篇章》
數量
- 表 5： HSK（詞彙等級大綱）詞彙量分布
- 表 6： HSK 詞表與研究詞表不重複的狀況
- 表 7： 內地特有的詞語
- 表 8： 內地、香港表達存在差異的詞語
- 表 9： 舊詞的遺留
- 表 10： 兩地使用頻率差異大的詞語
- 表 11： 為適應學習者需要而加入的詞語
- 表 12： 據聯想性原則添加的詞語
- 表 13： 由於語料選擇差異而出現的假高頻詞
- 表 14： 研究詞表未涵蓋的真高頻詞
- 表 15： 研究詞表與甲乙級詞語的相關度
- 表 16： 學習語料庫中學字表
- 表 17： 啟思初中篇目
- 表 18： 初中評估篇章不在九級字表之文字分析

第一章 緒論

1.1 研究背景與問題

自 2004 年以來，香港語文教育經歷了重大變革，課程改革由單一著重閱讀、寫作，逐步擴展至同時強調閱讀、書寫、聆聽和口語四大範疇¹。雖然課改的本意在於提升學生的整體語文能力，實踐上卻未必達致預期成效，部分學生的語文水平甚至出現下滑。改革的最大特點，在於教學模式創新，譬如取消指定的篇章、推動主題式與跨範疇教學，以及將文化元素融入語文學習。

以往教學多依賴指定範文²學習，老師按課文授課，學生只要熟讀即能應試。如今，所有文本均可作教學材料，課堂學過的篇章在考試未必會出現，考核重點轉向學生的語文應用能力。這一變化帶來了「廣泛閱讀」的理念，教材選文多元，學生面對的文章難度及風格變化大增。

然而，缺乏必讀篇章使得教材選取主要依賴教科書編輯，是否適合學生程度、是否因難度過高或過低影響學生學習動機，都值得深入反思。有效閱讀是語文能力培養的基礎，若文章難度未能配合學生實際水平，容易導致學習挫敗。作為一線教師，筆者親歷其困難，因此嘗試運用語料庫理論，評估及分級文本難易度，以促進教學突破，改善本地語文教學現狀。

1 自 2000 年起，香港課程發展議會陸續發佈多份「學會學習」課程指引，目的是為學校構建一套更為廣泛、均衡並以學生為中心的課程架構。這套課程將原本各學科歸納為八個主要「學習領域」，其中一項為中國語文。其後，2004 年 10 月，課程發展議會攜手考試及評核局，詳細說明並規劃了中國語文課程的內容。

2 2004 年以前，初中（三年共學習 36 篇）與高中（兩年共 26 篇）各有指定範文學習，考試時按照範文出題。

語料庫建構的同時，筆者特別關注「字」的難易度分級，思考哪些漢字應優先學習，不同學者對漢字學習順序的理論差異，成為設計本地教學語料庫的重要依據。香港在篇章難易度評估方面起步較晚，經過多番嘗試，初步建立了線上分析平台，希望協助教師、學生快速篩選合適的閱讀材料。相較於內地與臺灣，香港的語文教育改革尤其受獨特語言環境影響，推動難度分級教學更見迫切。

在此基礎上，本文計劃結合現有香港中小學語文教學實況、應用漢語語料庫及人工智能（以下簡稱 AI）最新技術，探討語料庫建構及教學實踐，尋找以教師為主導、學生為中心、教學與學習緊密結合的新途徑。語料庫除資訊傳遞外，更協助培養學生自主學習、批判思維與協作精神，推動知識與技能並重。

近年來，隨著 AI 及自然語言處理（NLP）技術快速進步，語文教育與語料庫研究迎來新的發展。AI 可協助分析大量文本的難度，並在結合學習數據後提供分級與選材參考。現時中文可讀性分析工具（如 CRUE）已能綜合詞彙、句法、語義及篇章等多層次特徵，提供文本難度分析與分級資訊（Sung et al., 2016）。在此基礎上，人機協同亦成為語文教育科技的重要發展方向：教師可參考系統的難易度分析結果選材，並結合學生實際表現持續調整教學。本研究正回應這一脈絡，融合 AI 語料庫技術與本地語文教學實需，建立動態、科學的篇章難易度分析平台，補足傳統選文主觀性偏強的不足，提升本地語文教育針對性和公平性，為『兩文三語』政策下分級閱讀教學提供重要依據。

1.2 研究目的與意義

1.2.1 香港的語言狀況及「兩文三語」政策

香港特區政府實施「兩文三語」（Bi-literacy and Tri-lingualism）語言政策，培養學生的中、英文的寫作閱讀能力，以及英語、普通話和粵語之口語技能。

鴉片戰爭後，香港在英國 154 年的統治下成為一個雙語社會（Bilingual Society），以英語為官方語言。不只企業和政府部門使用英語作為工作語言，部分中學和高等學府也將英語作為主要的教學語言。但是大多數人日常的溝通語言依舊以廣東話為主，因而出現雙語並存的情況。在殖民時期，英語被視為高等語言，粵語等方言則被視為低級語言。2001 年教育評議會的調查顯示 97% 以上的香港人口是華人，95% 能說粵語。「能使用粵語的人口約佔總人口數的 95%。因此，香港語言問題的討論，並不是在本地居民如何交流的問題，而是單純涉及採用什麼語言與外界交流」³，所以香港語言問題主要是外部溝通語言的選擇，而不是內部族群溝通的問題。由於普遍使用粵語，要融入主流需要學習粵語，但書面語主要仍是標準中文，這對推行「兩文三語」政策形成一定困難。

1.2.2 香港的學校語言課程設置

1997 年之後，教育當局發布了〈中學教學語言指引〉，當中語文政策明確指出「兩文三語」是政府語文教育的總目標，希望學生能夠同時精通中文和英文的書寫能力，並能流利運用粵語、普通話及英語進行溝通。香港特區政府在語文教育政策方面，致力推動學生具備「兩文三語」的能力，藉此加強語文學習的基礎和保障學生的語文發展。

3 周中天（2003）。〈英語啟蒙教育下東亞地區語言應用與教學之比較〉。發表於《東亞地區語文使用與教學現況學術會議論文集》。

中國政府沒有在語文策略上對香港設限，因此香港社會普遍認為小學不應全面採用英語教學，允許粵語在小學作為教學用語。而中學方面，則按語言授課，劃分為中文與英文中學。周中天教授指出，內地於 1949 年解放之後，英文中學在香港較受歡迎，即使香港中文大學已成立，也無法遏制中文中學的衰落。1970 年有英文中學 229 間，中文中學 114 間，但到 1988 年變為英文中學 343 間，中文中學 57 間。

黃坤堯教授表示，香港的小學現在每週起碼都有一堂普通話課，然而許多學校對此不甚重視，以致學生在離開學校後，還需自學中文找工作。但香港的語言教育策略討論的重點仍在中學教學語言，特別是是否採用英語授課。之所以聚焦在中學，是因為中學和小學的銜接問題。大多幼稚園和小學以中文教學為主，都設英語科。教育局對中文教學學校提供支援，如鼓勵出版中文課本，少數中文班可增聘英語教師等。如果在中學階段實施中文教學，則會推行「學習環境增潤語文研究」，允許學校在課堂上使用英語進行跨學科授課，以提高學生使用英語的機率⁴。

原本香港政府教育局為推動中文教學，只允許部分中學使用英語教學。但因為家長擔心子女升讀大學時無法適應全英文授課，希望子女入讀英文中學，導致中文中學嚴重流失學生。加上家長老師強烈反對，政府被迫對中學教學語言採取放任政策。1997 年回歸後，教育界認為教育要配合時代發展，因此制定了中文教學和新學制⁵政策，許多

4 黃坤堯（2001）。〈香港語文教育省思〉。載於鄧仕樑主編，《香港語言教育反思》，頁 7-14。香港：中文大學出版社。

5 所謂「三三四學制」，是香港教育制度革新的關鍵里程碑。過往在殖民地時期，中學教育沿襲英國體制，分為三年初中、兩年高中、兩年預科以及三年大學，即「三二二三」結構。回歸以後，教育局提出新學制改革，將中學分為三年初中及三年高中，大學則為四年制。根據相關諮詢報告，這一新學制自 2009 年 9 月於中學層面開始實施，至 2012 年大學全面轉為四年課程。改革過渡期內（2009 至 2012 年），新舊學制曾並行運作。

原英文中學被改為中文中學，校方雖不情願但只能接受。2007 年教改檢討發現英語成績下降，教育局恢復學校語言自主權。為提高英語，許多學校遂自行決定採用英語教學。

黃坤堯教授曾經指出，雙語政策已在學校普及，但學生的中英文口頭及書面表達能力普遍有待提升。中文上，同學日常使用廣東話，轉換為書寫語言時遭遇困難；英文則不是母語，除個別學生外，英語程度普遍偏弱。隨著香港回歸，雖然英語地位穩固，但普通話影響力擴大，與中國國際地位提升有關；同時香港社會仍以粵語為主，形成三語鼎立局面。因此，如何提高「兩文三語」的相互支撐，仍是政府和教育局亟需研究的課題⁶。

1.2.3 反思與小結

在香港特區政府和教育局推行「兩文三語」政策，以及學校的積極努力下，這二十年來，學生獲得了更多使用英語和普通話的機會。但是從效能來看，與期望仍有差距。同學可以使用中文（粵語）交流、閱報、讀書、在網絡發表言論，但缺乏正規的語文操練，普遍的語文能力難提升至從事創作文學、研究學術、法理辯論和傳道等高層次活動的程度。在粵語和英語雙重影響下，學生語文能力普遍下降。

此外，在欠缺語言環境的情況下，學生的普通話聽力和口語能力，以至英語的讀、寫、聽、講技能也不如預期理想。在高速發展和周邊國家競爭激烈的環境下，語文政策失衡所造成的負面影響會很嚴重，這需要政府和教育機構高度重視並進行研究。

近日社會熱烈討論香港是否能保持優越性，與周邊地區和國家展開競爭。在社會形成「語言政策會影響社會競爭力」共識的大前提下，

6 同註 4。

筆者構想從改進和提升語文教學入手，改變香港語文教育的面貌。香港配套設施良好，並獲得中央政府全力支援，如果能好好運用資訊科技的優勢，開發完善語料庫支援語文教學，將對推動漢語文教學非常重要。

1.3 研究問題與假設

1.3.1 利用語料庫推動漢語文教學

以往的教學過程論多集中於探討教學步驟的理想模式，較少關注實際教學中所發生的現象。有學者認為，「教學的推進是一個分階段設計和培育人才的歷程，強調師生雙方共同參與，並在互動中持續促進能力成長」⁷。筆者以香港語文教學為研究起點，將漢語語料庫作為核心，理論上結合語言學原理、電子計算機軟件開發及數據庫技術，並融會語文教學的理論與方法。這樣的結合能更全面地認識和應用語料庫。在香港語文教學的實踐中，語料庫的建立有效促進了師生的整體實踐與能力提升。

1.3.2 建立學習語料庫幫助學生有效學習漢語

本研究旨在根據香港中小學語文教學的實際情境，結合漢語語料數據的最新研究進展，探討語料庫在課程設計和教學提升中的實際應用。透過此研究，希望為本地語文教學尋找新的突破點，建立一套由教師主導、以學生為核心、強調教學互動的學習語料庫。這一平台不僅有助學生有系統地掌握語文知識，更著重培養自主學習能力，包括主動查閱工具書、預習與複習、擬題、歸納、反思和質疑等學習習慣，

7 陳伯璋（1999）。〈從九年一貫課程教師角色的再定位談師資培育因應之道〉。見《迎向千禧年——新世紀的教育展望國際學術研討會論文集》（國立中正大學教育學院編，頁12-18）。高雄：麗文出版社。

亦促進師生間的團隊協作，使學生在知識和技能兩方面同步成長。

李筱菊教授認為，當前教育理念已經由過去以教師為核心，逐步傾向以學生需求為出發點，教學實踐上亦日益強調個人化學習的重要性。她進一步闡述，個人化學習不僅要求學生能主動參與學習過程，更強調他們有能力靈活安排和調整自己的學習計劃，包括選擇學習的時間、進度、內容順序、學習方式、策略運用及學習分量等⁸。筆者透過引入電子計算機及語料庫工具，為學生提供便捷平台，方便他們及時了解自己的學習狀況，從而促進養成主動學習的良好習慣。在語料庫支持下，語文課堂強調自主探究，現今教師不再只是知識的傳遞者，更是引導學生探索與學習的推動者。這與傳統單向灌輸的教學模式有根本區別。個性化語料庫建構初期，教師需要積極監督與指導，既要掌握學生學習重點，又需具備解難能力；而當學生逐步建立知識、技能與自主學習習慣後，教師則可逐漸放手，促使學生獨立發展。

藉助語料庫平台，香港語文教師能靈活整合教科書、選用讀本、網上資源或新聞文章，進行有效的文本分級與分類，確保所選材料更貼近學生程度。同時，教師也可運用語料庫工具設計分級考核，靈活調整教學進度與評估策略，令學生在學習與評核兩方面皆能獲得更合適的支持，提升整體學習效能。

1.4 研究方案

1.4.1 研究目的

在教育技術快速發展的背景下，構建語料庫並應用於語文教學，已成為當前學界亟需關注的重要課題。從教育技術角度切入，語料庫研究應強調與教學現場的緊密結合，不能僅停留於傳統應用語言學對

8 李筱菊（2001）。《語言測試的科學與藝術》。長沙：湖南教育出版社。

語言規律和語言結構的探討，而更需要聚焦於課堂教學中的整體性與系統性應用。設計與運用語料庫時，必須融入整體教學設計，配合現代學習理論作為指導，讓語料庫在提升語文教學效能方面發揮關鍵作用。

本研究立足於現有漢語語料庫的學術基礎，目標在於探索學習語料庫的建構路徑及其實踐模式，並以提升語文教學成效為最終追求。通過系統分析語料庫建構的理論與操作流程，逐步建立起適合語文教學的學習資料庫框架，繼而推動漢語學習語料庫與語文教學改革的融合。這一過程亦為未來語文教學提供理論依據和具體操作指南，促進教學創新，從而建立一個適用於初小至初中（P1-S3）的九級字表。

本課題以文獻梳理、理論探討及前人經驗為基礎，聚焦於漢語學習語料庫的建設與語文教學改革的內在規律。透過對現有案例的歸納與分析，深入研究如何將語言學習語料庫有效應用於語文教學，提出實用的新方法，並力求在實踐層面加以驗證與推廣。未來，語料庫的建構與其在語文教學推廣上的應用，將會是持續深化的研究方向，並帶動相關領域的長遠發展。

1.4.2 研究方法

本研究在廣泛梳理相關文獻、借鑒前人研究成果以及現有漢語資料庫的基礎上，聚焦於學習語料庫的搭建和語文教學革新，重點分析兩者之間的內在規律。研究重點在於探討如何在語文教學過程中，有效設計和應用語言學習語料庫，並提出具體的構建策略。通過理論分析與實踐驗證，力求證明這套學習語料庫方案具有可行性與推廣價值，為未來語文教學創新提供經驗和參考。

筆者從事語文教學多年，熟悉漢語教學的各類課型，掌握不同層次的學生特徵和教學特徵；多年來，在漢語教學方面有較為扎實的經驗和原理積累，這為本研究打下了較好的基礎。此外，本選題的確定通過了較長時間、較充分的醞釀和準備。早期確定選題，筆者在研究重點和研究內容的構造等方面，作了多次調整，然而，研究對象和目的始終沒變，體現了筆者對本選題思考的穩定性，著重於探討語料庫是否能有效推廣並應用到語文教學上。其次，探索語料庫的進程及意義，了解其獨特性，並思考如何使用及制定學習語料庫。筆者探討內地語料庫的現況，試圖藉此參考其建構原則，甚至嘗試可否借用以作為本地語料庫的研究。及後收集和歸整內地、臺灣和香港相關的研究材料，並進行梳理和分析，關注相關研究成果，持續收集、歸納和整理相關材料，敘述語料庫歷史演變及特性，為自定語料庫建立良好的規模，因此積累了較為多樣的文獻材料。

然後，闡述語料庫建構的基礎，以及如何利用閱讀理論以建構學習語料庫。再進一步，敘述如何建立學習語料庫——需要把資料逐一分析並合併，建立小學分級字表；再利用香港教育署建議中學文本篇章（初中）600篇、啟思出版社《初中中文》篇章，並再揀選兩家（例如朗文和教圖）出版社的教科書進行比較研究，以提升研究的準確性和廣泛性，整理並制定《九級分級字表》。稍後，組織工作團隊，按照本研究所設計的文本篇章分級概念，完成「九級分級字表網上分析平台」。為達到字表的準確度與兼容性標準，額外採用並分析新聞報紙語料（《明報》從2022年12月4日持續至12月10日共7日內港聞版、《星島日報》從2022年12月4日持續至12月10日內港聞版）及莫言小說《豐乳肥臀》電子版以補充字表字數，初步建立學習語料庫。

最後，敘述如何建立評估語料庫——利用全港性系統評估（TSA）中國語文科閱讀考卷（2006 至 23 年度）的文本篇章資料和筆者曾任教學校的閱讀評估考卷，進行《九級字表》分析——初步發現對編訂評估機制有極大幫助，評估與學習可以互相配合。

總的來說，筆者在高校從事漢語教學工作多年，有較為理想的教學實踐基地，有利於相關研究的順利進行。以上所有都為本研究打下了較好的基礎。

1.4.3 創新之處和研究難點

本研究認為，以香港中國語文教學和語料庫建設研究為啟發點，著眼探究語料庫建設和語文教學規則性問題，分析如何在語文教學中有效運用語料庫，並在理論和實踐上全面論證這一理念在教學中的可行性和有效性。期望提出將學習語料庫融入語文教學的新觀念，並結合實際操作與積極引導，建立適用的學習語料庫和完善語文教學原理。

筆者根據前人的研究，廣泛蒐集語料，篩選合用的語料進行分析，並替語料作進一步的分類，初步釐定語料的級別。再加上自行研究的成果，制定《中小學九級字表》，依據這個表量度文本篇章的難易度。據難易度的概念，組成團隊研發搭建「九級字表分析網上平台」⁹。及後，利用字表工具，進行文本分析，按文本篇章進行級別分類，產生難易度數據。根據難易度數據結果，建立「學習語料庫」，將文本篇章進行整理分級，規劃難易度分級標準，編製校本課程。把文本初步分為初小文本篇章、高小文本篇章、初中文本篇章三類。再者，建立

9 網址：<https://ckc.eduhk.hk/temp/corpus>

「評估語料庫」，幫助學校設計分層評核制度，用以評估學生的閱讀能力，並可產出分析報告。另一方面，也可以根據「評估語料庫」設計預測評估，評估同學閱讀語文的技能，以測試結果作為「學習語料庫」之依據，幫助妥善處理校本課程。

本研究主要以「字」為單位作基本測試依據，取材以香港的教材為主，設計程式及整理數據技術不斷地提高，這幾年愈發進步。當然，仍有許多內容需不斷調整更新。例如，分割詞組時，在資源匱乏的情況下，不能說做得很完美，但仍需不斷改善。此後進一步的研究將以「詞組」為分析單位，擬制定「分級詞彙分析表」，並嘗試爭取研究資源。

本研究採用了「以單字為基礎、詞彙作延伸」的分階段策略。首先，漢字（「字」）作為最小書寫單位，具有良好的跨語境通用性，例如跨簡繁、跨地區，且不存在斷詞模糊性問題。本研究因此以單字頻率與習得年級為依據建構適用於初小至初中（P1-S3）的九級字表，提供一個穩定的分級框架。之後再於此基礎之上，逐步納入「詞彙」層級作補充；具體而言，未來會視學生理解情境與篇章結構需要，加入多字詞的難度分級邏輯，形成「雙層分級」機制。這樣的策略兼顧了實用性與擴展性：初期使用單字分級已能滿足主流教學需求，後續融合詞彙難度則提升系統精確性與語境解讀能力，進而完整反映閱讀複雜性。

文本篇章難易度分析，只按字釐定難易度，本身仍有許多可改善空間。文本篇章的難易度，往往不能單就字詞分析來確定，還須就文意深淺細分文本篇章的難易度，這樣的結果將更為客觀。目前，先以字為基本單位制定難易度，再按內容意思的深淺進行難易度分類。

字表的難易度研究依據的是前人的成果與筆者整理的香港初中文本篇章資料，這個原則是依據現時內地、香港及臺灣的相關研究，分析並比較漢字的使用頻率與難易程度，重點檢視教育署《小學常用字表》所列 2600 字的學習次序和使用情況是否合理。研究同時從文字學角度切入，剖析部分常用字的字源及字形演變過程，為教師解釋字形提供理論依據。所參考的資料包括內地、香港、臺灣的漢字字頻及難易度研究、識字教學的著作，還有關於漢字字形的各類型文獻及考古發現等多個來源。

第二章 文獻綜述

2.1 語料庫設計的概況

2.1.1 語料庫建設的意義

隨著網絡化信息時期的來臨，電腦應用技術得到快速發展，特別是最近兩年人工智能的快速發展，使語文教學再次達到另一境地。香港面臨二十一世紀最新一輪課程改革，為語文教學帶來新的動力和發展契機，在對課程的要求、教學模式的發展與教學內容的探究方面，都有很大的拓展空間。尤其是發展中的語文課程標準提到：「鼓勵擴展語文學習與應用的範疇，強調跨學科內容結合與現代科技手段的融合，讓學生透過不同任務與方法的交疊，提升視野與學習效率，並積累符合現代社會需求的語文實踐能力」¹⁰；「語文課程設計需因應學校、地區及學生的多元化特性作出靈活調整，以及隨著地區變遷與資訊化發展持續自主調整和擴展。與時俱進地推動語文教育內容和方式的創新，是提升課程質素的重要方向」¹¹。研究者在教育理念的指引下，要勇於探索和創新，嘗試將涉及語言學、概率數據統計學、電腦科學和文獻學等多學科的科學技術，融入到語料庫建設，融合到語文教學當中。

2.1.2 語料庫建立整體水準提升的需要

語料庫涵蓋學習者的書面語及口語兩種類型，這兩類語料庫均與一般的學習語料庫存在一定區別。憑藉電腦數據化處理，外語學

10 讀寫算編輯部（2015）。《讀寫算（教育導刊）》，2015（8）。

11 教育部（2002）。《全日制義務教育語文課程標準》，北京師範大學出版社。

習者操作電腦所產出的語言文本可建成數據庫，進而成為中介語（Interlanguage）¹²。在西方國家，根據數據統計，目前已經建成或正在建設中的學習者專用英語書面語語料庫共有八個，口語語料庫則有兩個，例如比利時的 International Corpus of Learner English¹³（ICLE）、瑞典烏普薩拉（Uppsala）大學建立的 Uppsala Student English Corpus¹⁴（USE）、美國 Michigan English Learner Database¹⁵（MELD）、英國的商業性學習者語言學習語料庫 London-Lund Corpus¹⁶（LLC）和 Cambridge Learner Corpus¹⁷（CLC）等¹⁸。

中國學習語料庫的建立和研究，從上個世紀九十年代中期開始。近幾年來，拓展運算速度激增，已達到國際領先水準。中國語言學習語料庫普遍分為兩類，一類是中國學習者英語語言學習語料庫，另一類是西方國家學習者，學習漢語中介語的語料庫。專為內地學生設計的英語學習語料庫，現時已建成或正興建的共有六個¹⁹。然而，專門

12 Quirk, R. (1960). Towards a description of English usage. *Transactions of the Philological Society*, 59, 40–61.

13 <https://www.abo.fi/en/english-language-and-literature-research/icle-the-international-corpus-of-learner-english/>

14 <https://www.uu.se/en/departement/english/research/english-linguistics/electronic-resource-projects/uppsala-student-english-corpus-use>

15 https://www.mischooldata.org/english-learner-student-growth?Common_Locations=1-S,8436,1029,68~2-A,0,0,0&Common_SchoolYear=21&Common_Grade=Grade03&Common_LocationIncludeComparison=False&Portal_InquiryDisplayType=Snapshot&Common_EarlyChildhoodReportCategory=AllStudents&Common_EthnicityList=AllEthnicities

16 <https://varieng.helsinki.fi/CoRD/corpora/LLC/>

17 https://www.cambridge.org/elt/corpus/learner_corpus2.htm

18 Wallis, S. (2007). Markup, data retrieval, and experimental research. In A. Meurman-Solin & A. A. Nurmi (Eds.), *Annotating Variation and Change* (Electronic Publication). Helsinki: VARIENG, University of Helsinki.

19 Sankoff, D., & Sankoff, G. (1973). Approaches to sampling and computational analysis for grammatical variation research. In R. Darnell (Ed.), *Canadian Languages in Their Social Context* (pp. 7–64). Edmonton: Linguistic Research Inc.

面向漢語為母語的學習者所建立的學習語料庫，其拓展前景並不樂觀，核心原因在於，此類語料庫的數據庫類別單一，嚴重依託口語語料庫。這便要求語料庫設計和開發者建立龐大的與語文教學相關的語料庫，但相應的建設難度也因此提升。

2.1.3 語文教學建立的需要

當前，語料庫資源工作人員在搜索和歸整語言數據化文本核心時，主要憑藉各類電子資源搜尋引擎或相關網站，例如百度、谷歌、雅虎或搜狗等；注音則憑藉 MS Office 辦公軟體中的 Word 所自帶的注音功能來進行。為生字尋找相關的字、詞、句則是憑藉工作人員的個人創作，或是借助電子詞典等查詢錄入。

然而，上文中提到的語言學習語料庫，在製作資源的操作形式、方法和工具運用方面都存在缺陷，尤其在語料數據化、文本資源的電子相關數據檢索方面都有所不足。假如使用的電子資源搜索引擎是通用的，則只能依據所需語言材料中所涵蓋的文字來搜索相聯繫的網頁，而不可以根據文章具體的主題或是體裁來搜索與當前正在處理課文的主題密切相關、體裁相同或相似的數據化文本閱讀材料，更不用說僅僅依據出處、標題或作者來搜索需要的特定作品。因為電子搜索引擎並不對網頁屬性進行標注，所以採取電子搜索引擎，去查找語料閱讀資源的查全率普遍較高，還會回饋眾多的網頁信息，然而資源準確率卻是很低的。如果在回饋海量信息中，再進行人工找尋，便會費時費力，嚴重影響工作效率。

語料資源製作過程中存在的問題，必然會制約跨越式設計、開發和製作試驗的進展。因此，這是亟待解決的問題。然而，要解決這些問題，就需要建立語文教學相關的語言學習語料庫。這個語言學習語

料庫，大多以文本篇章為基本單位，收集並總結歸整了適合小學生進行閱讀的各類體裁的文章。語料庫力求在思想和知識基礎上，進一步提升趣味性，全面覆蓋名家大作，可以反映當前漢語拓展領域的前沿成果；同時，各篇章的語言、組織和屬性，例如體裁、語體、作者、標題、主題、字數或出處，甚至引用文獻等，都應當標注完備，這樣就全都可以支持模糊查詢形式，甚至能進行全文相關數據檢索。更進一步，要求大幅度提升數據化文本素材的精準率，使該數據化文本具備客觀的可擴展和可編輯性能。最後應當以互聯網絡（Web）服務的形式發佈出來，以便各語言庫工作人員使用。

語言學習語料庫這個概念，換句話說就是自然語言電子數據處理運行系統的重要組成部分。在電子語言教學體系設計、開發和研究過程中，電腦對語言的數據化文本處理，已由「錄入」－「儲存」－「全文相關數據檢索」等傳統工作模式，轉向「自動化分類」－「智慧相關數據檢索」－「信息提取」－「機器翻譯」等寬領域、深層次的研究模式。而相關電腦技術的進步，尤其是現代漢語數據化信息處理領域的迅速推進，為語文教學的深層次、利用電腦進行輔助的研究，開闢了全新思路，創造了有利條件。基於要創建屬於香港漢語的語料庫，需要參考過去語料庫源流以設置獨特的體系，以下將就過往的發展源流作一綜述。

2.2 語言學習語料庫源流

2.2.1 相關概念

語言學習語料庫（Language Learning Corpus）是以實際語言運用案例為基礎所進行的語言學研究。藉助語料庫，研究者可以從語法和句法層面分析自然語言，同時探究其與其他語言之間的聯繫。早期這

類分析多靠人工處理，現今則主要依靠電腦技術自動化運算。語料庫語言學者普遍認為，只有依據新近蒐集的語料及自然語境進行語言分析²⁰，並盡量減少外在實驗干擾，才能獲得最真實、精確的結果。

2.2.2 語言學習語料庫的定義

一般是指為語言數據分析與研究所蒐集、整理並以電子方式保存的多樣語言數據資源，這些語料大多選自自然產生的書面語或口語標準樣本，能反映特定教學語言或語言變體的代表性特徵。只要經過精心篩選和標註，並具備一定規模，語言學習語料庫便可較為客觀地呈現實際教學語言的運用與實踐情況。學習者及語文教師可藉此觀察並理解語言結構與規律，進行系統性的剖析與探究。如今，該項語料庫不僅是語言學基礎理論探究的重要工具，也是應用語言學、語言工程與課程設計不可缺少的資源。

廣義來說，語言學習語料庫是以隨機抽樣方法蒐集並整合成的，具有代表性的真實語言資料集合，反映語言實際應用的情境。而語料庫語言學則是一門運用這些語料資源來進行語言研究和知識探索的學科。

語言學習語料庫於語言學研究中的運用，可追溯至十九世紀末。當時的研究方法多以人工檢索相關數據為主。隨著電腦與數據存儲技術不斷發展，現代語言學習語料庫語言學逐漸興起，大大提升了語料收集、加工、儲存和檢索的效率。二十世紀三四十年代，語料庫語言學曾經經歷過快速發展。到了資訊產業迅速擴張、電腦普及的時代，語言研究開始向前沿方向突破，積極嘗試創新研究模式，逐步擺脫傳統分析方法的局限。

20 Sinclair, J. (1992). The automatic analysis of corpora. In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82* (pp. 379-397). Berlin: Mouton de Gruyter.

也正因為上述背景，現代語言學習語料庫語言學逐漸發展成形。過去語言研究主要採用兩種方式，一為「內省法」，另一為「實驗法」²¹。「內省法」就是研究者需憑藉自己對語言的直覺和自身的智慧，依靠主觀的臆斷進行語言撰寫。這其實就是學術領域內的閉門造車，因此被視為藏在沙發裡的語言學。至於「實驗法」就是利用個別或者少數語病患者的日常話語語料，進行分析和研究，它關注的僅僅是研究人員利用這些語言來做什麼，並不適用於研究和解釋普通人使用的語言。因此該研究模式又被稱之為藏在診療室裡的語言學。前研究模式缺少對語言自然性和真實性的探討，而後者缺少廣泛研究的代表性；語言學習語料庫裡的語言學，恰好可以彌補這兩種研究模式固有的缺陷，為語言研究提供了前沿的研究思路，可以促使語言學更好地為語言學習實踐提供服務，而最顯著的便是其對語言教學的貢獻。

語言學習語料庫主要的特徵，首先是內容和構造性，「是有目的地就語料推行的整合，並不是語言材料的任意堆積，這便要求語言學習語料庫具備合理的組織內容和構造」²²。在語言學習語料庫建立的初期，就應該制定明確的語言材料選材原則。其次是具備代表性，語言學習語料庫在應用領域中，應當具有足夠的典型性和代表性，如此才能保障基於語言學習語料庫得出的具體知識具備相對普遍性和較高的信息資源連結完備性。還需要有可機讀性，這是語言學習語料庫必然的要求，即以文本形式儲存，並可由電腦讀取和處理的語料整合。只有如此的語言學習語料庫，才能滿足計算機網絡技術下的現代化語言學學術研究的需求。

最後當然還需要有一定規模。有些學者收集和歸納一些文字卡片

21 何安平 (1999)。〈語料庫相關研究的範疇與技術方法評析〉。《外國語》，2，10-17。

22 Fung, C. W. (2002). A review on the development and current situation of corpus studies in China. *Journal of Chinese Language and Computing*, 11(2), 127-136. Singapore.

和影像資料，都不可被視為現代意義上的語言學習語料庫。語料的集合要想成為語言學習語料庫，還應當形成一定程度的研究、設計和開發規模。統計學相關原理表明，如果樣例的數量比例未能達到一定規模，據此得出的結論就失去了可靠性，語言學習語料庫也因此失去研究、設計和開發的代表性。

2.2.3 語言學習語料庫的相關數據檢索功能

不少國內外學者都曾對語言學習語料庫的數據檢索功能展開深入研究。以胡明揚教授為例，他認為一個完善的語言學習語料庫檢索系統，應能讓使用者方便地查詢多項語言信息，包括全文總字數與字頻、依據詞類大小分類提取詞類清單、各類詞彙的出現頻率與總數、字詞類別及其統計數據、特定詞彙的篩選、單字與詞語的語境關聯，以及不同句型、句類的數量統計和語法結構分析等²³。

楊惠中教授則指出，數據檢索工具的基礎應用和操作功能，還應涵蓋自動生成詞彙表、資料彙整、語言材料及結構分析、語用情境規劃、詞彙搭配關係、關鍵詞索引、詞語組合統計、主題詞萃取、聯想詞的數據整理與再組合、詞群的統計處理，以及詞彙圖譜的設計與統計等多項功能²⁴。

數據檢索功能設計，在數據庫建構初期，便應當有一個清晰概念和設計方案，讓使用者應用時簡便，而又可明確掌握數據資訊，作為教學上的重要參考憑證。

23 Hu, M. Y. (1992). Introduction to the Survey Corpus of English Usage and Other Corpora. *Foreign Linguistics*, (4), 37-42.

24 楊惠中（2002）。Corpus Linguistics: An Introduction（中文版《語料庫語言學導論》）。上海：上海外語教育出版社。

2.2.4 語言學習語料庫類別

對該項數據庫的類別，筆者根據不同的研究結果和區別標準，採取了不同的分類辦法。本研究以漢語語言學習語料庫語言學文獻研究為參考，把代表語言研究基礎的語言學習語料庫進行分類：一是百萬詞級及以下，各種小規模的語言學習語料庫，其中包含 Brown University Standard Corpus of Present-Day Edited American English²⁵ (BROWN)、Lancaster-Oslo/Bergen Corpus of British English (LOB)²⁶、上海交通大學的科技英語語料庫²⁷ (JDEST) 等個體自建類別的數據化文本語言學習語料庫；二是千萬至上億詞級，規模相對較大的語言學習語料庫，比如 a Component of the Collins COBUILD Bank of English²⁸、British National Corpus²⁹ (BNC) 等；三是用作專門用途的語言學習語料庫，比如以口語資料為主的學習語料庫、針對學習者收集的語料資料庫、雙語互動的語言學習語料庫等；四是標注了加工形式和類別的語言學習語料庫，如帶有詞類分解附碼、句法分解附碼、錯誤信息分解附碼的語言學習語料庫³⁰。

從上文可以看出，此分類標準是交叉性的——在標注了加工信息的小規模語言學習語料庫中，也會同時出現小規模的、以專門用途為特徵的語言學習語料庫。依照語料庫性質，可以將它們分為「原始語

25 <https://varieng.helsinki.fi/CoRD/corpora/BROWN/#:~:text=The%20Brown%20Corpus%20was%20the,and%20the%20Brown%20MARC%20Form>

26 <https://varieng.helsinki.fi/CoRD/corpora/LOB/>

27 <https://www.lib.sjtu.edu.cn/f/main/index.shtml>

28 <https://www.english-corpora.org/coca/help/compare-wbo.asp>

29 <http://www.natcorp.ox.ac.uk/>

30 參見何安平（2004）出版的《語料庫在外語教育領域的應用：理論與實踐》。廣州：廣東高等教育出版社。

言」和「標注性語言」³¹。

「原始語言」又被視為生語言學習語料庫，一般指依據還未經過標注、正在等待處理的原始語言語料組成的語言學習語料庫。然而也需要對數據化文本進行清檢，以保障語料的機讀模式化。原始語言學習語料庫，可以廣泛服務於語言研究領域，涉及句法、詞彙和語篇等各層次的語言研究領域。

至於「標注性語言」，指的是通過對數據化的文本釐清，依照研究的具體用途和特定目的，對它開展附碼或分割操作後，所形成的語言學習語料庫。經過標注後，不但保障了數據化文本信息提取的精準性，還使語言學習語料庫的使用更加方便和快捷。

有學者根據語言學習語料庫的核心功能，把語料庫細分成「學習者」、「平行」以及「口語」等不同類型³²。

「學習者」主要用於透過語料庫方法深入觀察及理解學習者在真實語境下的語言表現，並將相關分析結果應用於外語教學實踐。這類語料庫的建構，不僅為語言學者開拓了嶄新的研究領域，更為語言教學提供了大量語言運用例證，幫助教師和研究者克服分析難題。建立學習者語料庫的目標，不應僅限於錯誤辨析或診斷式教學，更應進一步發展為對學習者語言特點進行系統性、全面性的對比研究。

至於「平行」，則是進行語言對比和翻譯性研究的重要資源。這類資料對推動語言對比、翻譯和雙語詞典編纂，以及翻譯技術開發等領域均具深遠影響。學習者可以利用平行語料庫主動查找語法、詞彙及語義上的對應關係，進一步認識母語和外語之間的差異，從而提升

31 參見謝應光（1996）在《外語教學與研究》第三期所發表的有關語料庫語言學與外語教學的文章（頁29）。

32 參見王建新於《外語教學與研究》第三期（1996年）發表的〈介紹當代三個英語語料庫〉一文（頁37）。

語言學習與翻譯能力。平行語料庫的發展與應用，正反映國際語言教學和翻譯研究新趨勢。

此外，「口語」在當代語言學與教學中亦日益受到重視。許多學者及教師認為，與書面語相比，口語數據更能反映語言習得的本質過程。透過語料庫技術，學習者和研究人員可收集、整理自然口語語料，並分析其話語特徵，為口語教學和語言習得研究提供實證基礎。

綜上所述，研究人員可以借鑒這幾類語料庫的建設的經驗，取長補短，打造適合學生學習的漢語學習語料庫。

2.2.5 語言學習語料庫拓展史

雖然早在 1747 年，英語詞典的創始人 Johnson 就率先開始大規模收集和歸整英語語料，用於詞典的編纂。但是，現代意義上的語言學習語料庫，卻是由 Quirk 等人於二十世紀五十年代末期在倫敦大學創建，最初被稱為「英語用法調查」。這一語言學習語料庫，廣泛收集和歸納整理了當代社會慣用的英語語言材料，達到了近百萬英語詞量，隨後還在持續擴充。隨著電腦技術飛速發展，更大規模的語言學習語料庫也相繼出現。例如 Lancaster-Oslo/Bergen Corpus of British English (LOB)、朗文－蘭卡斯特、倫敦－隆德以及英國國家語料庫等著名語料庫資源，其收錄的詞彙總量通常已達數億詞級別，數據規模極為龐大³³。

自 Survey of English Usage³⁴ (SEU) 創立以來，語料庫語言學進入新世代。隨後學界陸續建立了具有代表性的語料庫，如布朗語料

33 Wallis, S., 與 Nelson, G. (2001)。〈以語法標註語料為基礎的知識發掘〉。Data Mining and Knowledge Discovery，第 5 期，頁 307-340。

34 <https://www.ucl.ac.uk/arts-humanities/english/research/survey-english-usage>

庫（Brown University Standard Corpus of Present-Day Edited American English）、倫敦－隆德語料庫（London–Lund Corpus）以及蘭卡斯特－奧斯陸語料庫（Lancaster–Oslo/Bergen Corpus），標誌語料庫語言學的蓬勃發展。

二十世紀八十年代，許多國家在高校中附設了受國家科研機構支持或聯合建立的大規模語言學習語料庫，比如英國伯明翰大學就曾建立 Birmingham Collection of English Texts³⁵（BCET）語料庫。而法國的國家科學研究中心，跟美國芝加哥大學共同合作，建立了法語語言學習語料庫 Trésor de la langue française³⁶（TLF）。赫爾辛基大學則建立了世界範圍內的歷史語言學習語料庫——赫爾辛基歷史英語語言學習語料庫 The Helsinki Corpus of English Texts³⁷（通常簡稱 HC）。進入九十年代，語料庫的建立更實現了大步跨越。1994 年英國國家語言學習語料庫 British National Corpus（BNC）收集和歸納整理的英語語料，就已經達到將近一億的詞量。而大批語料庫建立項目，也迅速地被提上日程³⁸。

二十世紀八十年代後期，語言學習語料庫語言學在世界範圍內的蓬勃拓展，推動內地加快建立這門學科，而且越來越受到國家的重視。內地比較多的語言研究學者，積極投身於學習語料庫的建立和使用研究。在 1982 年，黃人傑和楊惠中兩位學者領銜於上海交通大學設計並建立了上海交通大學的科技英語語料庫（JDEST），其容量為一百萬詞次；1993 年前後，香港科技大學語言中心，編製了電腦科學

35 <https://www.birmingham.ac.uk/research/centres-institutes/centre-for-corpus-research>

36 <http://atilf.atilf.fr/tlf.htm>

37 <https://helsinki.corpus.arts.gla.ac.uk/display.py?what=index>

38 Dukes, K., Atwell, E., & Habash, N. (2011)。〈通過協作方式對古蘭經阿拉伯語進行句法標註〉。Language Resources and Evaluation Journal。

英語語言學習語料庫；1996年，廣州外國語學院建立了內地學生交際英語語言學習語料庫；2003年，內地應用語言學界楊惠中及桂詩春兩位教授曾主導並落實的書面英語中國英語學習者語料庫³⁹，標誌著內地對於英語語言學習語料庫的建立由此走向高潮。

二十世紀九十年代以來，電腦應用水準有了極大的提高，網絡應用的普及、電子手帳的湧現、無紙化辦公的逐漸實現，使收集和歸納整理電子數據化文本語料，開始變得非常輕鬆和便捷。伴隨著電腦硬件技術的高速拓展、電腦記憶體容量的擴展、電腦中央處理器的速率有效提升，以及光電掃描儀和其他信息辨識軟件精度的快速提升，電腦的信息索引和檢索技術迅速發展，極大提高了對大規模語言學習語料庫的搜索和處理運算速度。這些新狀況和現象的出現，為研發與建立超大型語言和監控語言的學習語料庫，提供了可靠的物質和技術基礎。

1990年，COBUILD語料庫負責人John Sinclair宣布將建成一個龐大的語料庫——Bank of English（COBUILD語料庫的一部分）。此監控語料庫（Monitor Corpus）自創建以來不斷更新，包含約6.5億詞的現代英語用法，涵蓋書面語與口語素材，標誌著語料庫語言學進入新里程碑⁴⁰。2000年10月，此語言學習語料庫有4.15億詞次，並在持續增長。超大型語言學習語料庫能為基於語言學習的語言學和自然語言的數據化處理研究，提供更寬泛、多樣和可行的知識。況且，因為這些語料規模龐大，設計者就有相對較大的空間去規劃語言學習語料庫的內容和構造，因而減少了語言學習語料庫設計時的障礙。因為每

39 楊惠中等（2005）發表了基於CLEC（Chinese Learner English Corpus，中國學習者英語語料庫）的英語學習者語料研究，書名為Corpus-based Analysis of Chinese Learner English，由上海外語教育出版社出版。

40 Garside, R., Leech, G., & McEnery, A. M. (1997). Corpus annotation. Published by Longman, London.

種類別的語料數量比較多，在語料採集時，語料的典型性與代表性的要求也可相應減輕，因而減少了語料採集的負擔。當樣例達到一定規模時，概率數據統計規律往往趨於平衡。因此，超大型語言學習語料庫，可降低對人工平衡設計的依賴，並提升語料代表性。

然而，建立超大型語言學習語料庫對電腦技術的要求比建立普通的語言學習語料庫更多更高。首先，必須要有高品質的自動化標注軟件。對超大型語言學習語料庫的標注，倘若採取軟件標注加人工校對的方法進行，基本上是不大可行的。其次，超大型語言學習語料庫對電腦相關數據檢索技術也提出了更多和更高的要求。相關數據檢索軟件，需要有更快的運算速度，並需提升檢索的完整性與準確性，才能達到更佳的搜尋效果。

COBUILD 語言學習語料庫，還是一種監控語言學習的工具。此語料庫融合了動態的元素，新數據化文本被持續地融合，憑藉軟件分析、辨識和搜索，不斷發現新詞和詞彙的新用法。當前，國際互聯網上，COBUILD 語言學習語料庫每週都會向網絡電子郵件客戶輸送一份 Word Watch 郵件，提醒用戶一些關於社會用語的最前沿的動態演變⁴¹。

全球化的英語監控語言學習語料庫（The Global English Monitor Corpus⁴²）項目在 2001 年終啟動。這是一個長期項目，預計往後將會持續取得一定成果。它的研究目的是憑藉此語言學習語料庫，匯總語言術語和專有名詞在當前時期的學術意義，報告不同國家英語用法的動態演變新趨勢；同時，用當前的數據和以往作比較，匯總世界新形勢⁴³。

41 參看 <https://collins.co.uk/pages/elt-cobuild-reference-the-history-of-cobuild>

42 <https://www.english-corpora.org/>

43 參看 <https://www.birmingham.ac.uk/research/activity/corpus/index.aspx>

監控語言學習語料庫需要具備及時更新和自動化監控的能力，它需要持續自動地淘汰過時的語料，並選擇時效性強的新語料。網絡是新語料一個很好的來源，其及時搜索並發現匯總語言應用時出現的新狀況和形勢，是監控語言學習語料庫的核心功能。

超大型和監控語言的學習語料庫，為這項研究展現了前沿的發展，開闢了新的研究方向。這意味著，語言學習語料庫將向幾個方向拓展。首先，在規範化方面，語言學習語料庫的數據化文本，其中描述性、標注性的語言和固有格式將實現規範化，這是語言學習語料庫進一步拓展的重要保障。在此前提條件下，可以設計和開發出一整套通用的語言學習語料庫軟件方便客戶使用。此外，可以實現語言學習語料庫更多和更高程度上的共用，實現不同語言學習語料庫的集成，進行不同的比較或互動研究。

其次，自動化原則，替語言學習語料庫建立一個更高的自動化程度，突出表現為語料的選取將不全靠人工進行，而可以由電腦憑藉程序運行自動化完成。再者，網絡語言學習語料庫，可以不全是現在集中管理的模式，它可以散佈在網絡不同的機器上，不但可以把其中的每一部分單獨代表著語言學習語料庫，還可以把這些視為在邏輯上的系列整體。同時，需智能化。語言學習語料庫系統，必然會具備某程度的智慧，完全可以自動更新語料、搜索發現並報告新知識。如果能進一步發展多功能，語言學習語料庫將不僅僅應用在語言學電子數據化研究上，它更代表著龐大的知識庫，幾乎可以用來研究社會生活和政治等更為多樣的問題。

同理，不同領域現已建成的數據化文本數據庫與語音數據庫都可以很輕鬆地轉換為語言學習語料庫，為語言學電子數據化研究服務。語言學習語料庫將和各種數據庫融合在一起，成為數字信息社會的重要信息基礎之一。

在語料庫語言學電子數據化研究的拓展歷程中，各家學派著書立說，共同致力於這門學科的繁榮發展。當前，語言學習語料庫語言學電子數據化研究呈現出一些前沿的拓展趨勢，並向不同的層面延伸。科技與經驗的結合為建立漢語學習語料庫提供了一個明確的指引方針，確是前人功不可沒的例證。

2.2.6 漢語文課程語言學習語料庫和漢語文教學

教育教學是科學研究的自然延伸。早在語言學習語料庫語言學剛剛興起的二十世紀六十至七十年代，海外一批語言學習語料庫語言學的開拓者和語言教育家，就開始把語言學習語料庫，放在語言教學的應用方面，使之成為語言教學的重要部分，因為這兩者是相互滲透的綜合體。Leech 在 *Teaching and Language Corpora* (1997 年) 中提出：「語言學習語料庫和語言教學相互共進的核心，顯現在直接將語言學習語料庫用在教學，也就是『教語言學習語料庫語言學』，『怎樣利用語言學習語料庫』和『怎樣用語言學習語料庫來教』。」⁴⁴ 語料也被間接地用在教學中，例如工具書的刊行和出版、教學書籍的設計和開發，以及語言能力和水準的測試等。此外，尚有些面向教學的語言學習語料庫的建立，例如專門用於語言的、中國漢語母語習得和外語學習者等語料庫，都非常值得去探究。

實際上，語言學習語料庫正在扮演不同角色，它代表了教學的實際內容，普遍面對的是學習語言學中的本科生和研究生群體，語言學習語料庫是他們用來探討和研究語言的工具。語言學習語料庫同時代表了教學的可用資源，服務的是學習各類語言（漢語母語和外語）的各種學生群體，在他們的學習過程中發揮著重要的輔助功能。

44 參考 Leech (1997: 9) 及 Biber, Conrad 和 Reppen (1998: 11)。

語言學習語料庫和語言教學的研究興起於海外，因此，有關語言學習語料庫和英語教學的研究成果非常多。而內地起步較晚，主要是對國外相關研究原理、方法和研究技術的直接引進。經過幾年時間積累，內地的英語語言學習語料庫和外語教學的相關研究也呈現迅速發展的勢頭，因而成為了語料庫研究的又一大前沿課題。然而，真正的漢語教學語料庫研究的設計和開發領域還是較貧乏。雖然有一批用於語言研究的漢語語料庫，例如前文所提到的大學或研究語言學院已經建立的語言學習語料庫等，可是涉及語文教學相關的語料庫依然很少。當前，真正已經設計和開發出和漢語文教學相關的學習語料庫，僅有北京師範大學所建立的，關於中學生語文教科書的語言學習語料庫。當然，尚有些學者正對此研究空白進行彌補，例如北京師範大學魏順平教授就提出小學語文語言學習語料庫的研究與探索⁴⁵。可是，這些語言學習語料庫的設計和開發，都只能填補一角，尚未能達到專業性、一體性的程度。

一般情況下，語言學習語料庫代表了自然語言數字化的文本資源彙編。自然語言的集合幾乎是平面的材料，因此，它對語言學習的理解較狹義；至於廣義的語言學習語料庫，應當是圖形和文本合為一體，是立體化的語言材料集結。

2.3 語言學習語料庫建設現狀（國外－內地－香港）

2.3.1 國外研究現狀分析

目前在國際間，學習者語言學習語料庫研究大致分為幾個階段。首先是規劃與構建，這一過程通常需要語言學家與軟件工程師攜手合

45 魏順平、何克抗（2007）。〈關於小學語文課教學專用語料庫的建構與應用〉。《中國電化教育》，2007年6月。

作，從設計到技術開發共同參與。其次，相關研究涵蓋第二語言習得的各個層面，包括語音、詞彙、語法和篇章等。學習者語料的研究成果廣泛應用於語言教學，例如編寫辭典、語法書、選擇適合的教學文本，以及設計自主學習的電子學習產品等。據不完全統計，目前國際上已經建成或正在建設中的學習者英語書面語語料庫約有八個，口語則有兩個⁴⁶。

相關文獻顯示，當前海外學者在語言學習語料庫、外語教學及語篇理論的研究上，無論深度還是廣度皆較內地更為突出。在國際學界，有關語料庫在英語教學理論與實踐層面的探索與成效均相當突出，相關研究已取得諸多進展，其教學實踐亦深受語料庫理念影響，具備高度的實用和指導價值。這些國際經驗對於本研究亦具有重要的啟示和參考意義。

現在，美國的布朗大學建立了布朗語言學習語料庫。英國蘭卡斯特大學、挪威的奧斯陸大學和貝根大學三校聯合起來，建立了 Lancaster-Oslo/Bergen Corpus of British English (LOB) 語言學習語料庫。美國等海外的許多學者，以這兩個數據庫為基礎，做了許多研究，而最成功的研究，是語法標注性的語言學習語料庫的相關研究。研究人員設計出名為 TAGGIT Tagger/Greene and Rubin's Tagging Program (TAGGIT) 的系統，基本上可以自動化標注。用此系統給 Brown 語言學習資料庫裡一百多萬詞彙量的語言資料，作自動化標注，正確率達到 77%。此外，研究者還設計名為 Constituent Likelihood Automatic Word-tagging System⁴⁷ (CLAWS) 的系統，來給上文中提到的語料庫，

46 王宗炎等 (2003)。〈針對中國英語學習者語料庫進行了系統介紹〉。載於《現代外語》第 4 期，頁 219-221。

47 <https://ucrel.lancs.ac.uk/claws/>

進行自動化標注的實驗研究，依照統計學方法及統計的信息，來建立相應演算法，正確率達到 96%，比 Tagger/Greene and Rubin's Tagging Program (TAGGIT) 系統高 20% 左右。最近，研究組又進行了三個相鄰的標記，計算同時出現的頻率，並將正確率提高到 99.5%。此數據已經超過了人工標注的極限。

目前來看，在國外使用的核心語言學習語料庫還有：

倫敦－隆德口語語言學習語料庫：此庫收錄的篇目約有 87 篇，每一篇的詞量 5000 左右，共計 43.4 萬詞量，且倫敦－隆德口語語言學習語料庫，備有詳細的韻律標注 (Prosodic Marking)。

美國的 American Heritage Intermediate⁴⁸ (AHI) Corpus 語言學習語料庫：由美國傳統刊行和出版社 (Heritage Press) 建立，約有 400 萬的詞量。

英國的 Oxford Text Archive⁴⁹ (OTA) 牛津數據化文本檔案庫：由牛津大學的計算中心籌備建立，約有 10 億位元組的檔案庫。

英國的 British National Corpus⁵⁰ (BNC) 國家語言學習語料庫：此語料庫 1995 年才建立完成，核心運用 Text Encoding Initiative 編碼 (簡稱 TEI 編碼)，並且運用了 1986 年國際標準化組織公佈制定的標記語言標準 SGML (ISO 8879)。

美國 Association for Computational Linguistics /Data Collection Initiative (ACL/DCI)：由美國計算語言學協會 (Association for Computational Linguistics) 推動的數據收集計劃 (Data Collection Initiative)，旨在為非商業學術機構提供無版權阻礙的語言資料，運

48 <https://aclanthology.org/C73-1023/>

49 <https://lids.ling-phil.ox.ac.uk/lids/xmlui/>

50 <http://www.natcorp.ox.ac.uk/>

用 SGML 標準進行統一編碼與數據共享。ACL/DCI 的核心目的是為公益或者不從事商業收益的學術研究機構提供語料，使這些機構，不被高昂的購買版權費用阻礙。ACL/DCI 也是利用 SGML 國際標準統一的置標，因而使數據交換變得便利起來。

美國賓州大學的 Linguistic Data Consortium⁵¹ (LDC) 語言數據聯合會：此聯合會是會員制，成員可以共用 163 個語言學習語料庫，這 163 個庫包含了 text 和 speech。

日本的 RWC Japanese (RWC-JP) Corpus (RWCP) 機構研究設計和開發的 RWC 日語語言學習語料庫：他包括日本的《每日新聞》節目 4 年以來全部的全文語料，詞庫的語素標注大約有 1 億多條。

亞洲不同國家、不同語種的互譯語言學習語料庫的建立，在日本，由國立語言研究所主持研究，由中野洋代表主持，北京語言大學也有參與。

為推動語言學習語料庫領域的發展，海外多地相繼成立了如 Trans-European Language Resources Infrastructure⁵² (TELRI) 和 European Language Resources Association⁵³ (ELRA) 等核心學術組織。TELRI 學會（即跨國語言資源建設學會）由約翰·辛克萊擔任主席，其運作經費主要來自國際學術團體的共同支持。該學會致力於推動涵蓋多語種的語料庫建設，目前已建立如柏拉圖多語語料庫 Plato Multilingual Corpus，以及 TELRI Research Archive of Computational Tools and Resources⁵⁴ (TRACTOR) 等計算資源資料庫，並持續籌備多語種詞

51 <https://www ldc.upenn.edu/>

52 <http://telri.nytd.hu/>

53 <https://www.elra.info/catalogues/free-resources/>

54 <http://telri.nytd.hu/tractor/intro.html>

庫等項目，為多語言學習語料庫的發展奠定基礎。TELRI 學會每年會召開一次研討會，最近的一次研討會在 2000 年 9 月 22 至 26 日於斯洛維尼亞的盧布捷那召開，本次研討會的焦點在於如何從語料庫中自行提取數據。European Language Resources Association (ELRA) 即歐洲語言資源協會，該會由贊伯利擔任主席，學會核心工作是進行蒐集和傳播語言資源，最後將這些資源商品化，並且為其使用者提供法律上的支援。European Language Resources Association (ELRA) 不同的服務處也就是 European Language Resources Distribution Agency (簡稱 ELDA)，ELDA 核心進行 ELRA 的戰略和計劃的研製和推廣，ELRA 也會組織每兩年一次的 Language Resources & Evaluation Congress，也就是語言資源評價會議，簡稱 LREC 會議。LREC 會議在 1998 年於西班牙舉行，第二次在 2000 年 5 月 31 日到 6 月 2 日於希臘雅典召開，2002 年 5 月 27 日到 6 月 2 日在西班牙的拉斯帕爾馬斯召開。

綜觀國外，語言學習語料庫已非常流行，且已具有一定的規範與技術，這個良好的經驗，為國內的漢語語料庫提供了很多借鑒模式，在了解之餘，也可加以相互交流，向他們取經。

2.3.2 內地研究現狀分析

語言學習語料庫語言學逐步崛起，是近幾年來語言學研究領域的一股潮流，在過去的二十多年得到了長足的發展，現已成為現代語言學的重要部分。然而，它不像社會學範疇內的語言學，和心理學範疇內的語言學研究那樣，只研究社會、心理與語言之間的關係。它是以語言學習語料庫為模式來研究語言，以眾多自然產生的語料作為研究對象，用統計學技術、定性分析方法和描述性語言，來表述其具體的構造和用法。這種定量型的研究方法，基本上可以應用在語言研究的

全部領域，例如語法、語言變異、語篇、文體學、詞彙和歷史語言學等。

在語言研究領域，內容和結構主義語言學在研究中很倚重實際使用的語言材料。布龍菲爾德（Bloomfield）和薩丕爾（Sapir）等人都在研究裡使用以啟發形式採集到的話語（Elicited Speech）。二十世紀五十年代末，喬姆斯基提出了內省式的唯理主義研究方法，並逐漸在語言學研究領域裡佔據主導地位。他重視將漢語作為母語的人的語言直覺，認為憑藉內省，可以考察具備完整漢語母語知識的理想講話人與聽話人的語言能力。因此，語言學家不必研究人們實際使用的語言，而應研究將漢語作為母語者的語言直覺，就是研究者自己的語言直覺。

隨著科技和社會進步，特別是電腦技術的進步，用機器大規模處理語料進行語言研究的方法，顯示了極大的優越性。加上語言哲學思潮，向實證方向的回歸，語言學習語料庫語言學方法的優勢日益明顯。語言界越來越多人意識到，語言材料在語言研究中，所能發揮的作用，遠遠超出人類的想像。語言學習語料庫和語言學家的直覺，應該是相互補充，而不是相互排斥的。於此，以實證和統計資料為基礎的研究方法，得到了重新肯定。

針對本課題依據的需要，擬將從以下幾點對內地研究加以述評，因為香港的漢語文教學，許多可以向內地借鑒參考，漢語文語料庫的設計和形式，也可以進行深入探究。

2.3.2.1 內地語言學習語料庫英語教學研究現狀

在 2000 年至 2009 年間，外語教學領域中有關語言學習語料庫的研究呈現出明顯增長的趨勢，顯示這一主題正獲得越來越多關注。這

主要得益於語言學習語料庫的發展，使語言統計分析更為廣泛且貼近實際；相關研究能細緻刻畫語言微觀層面的運用情形，為英語教學提供了穩固的研究基礎。同時，也為英語教師帶來了更新穎的教學方法與研究路徑。

從統計數據看，前五年（2000 年–2004 年）12 種核心期刊共刊載了 34 篇文章，佔總數的 32%，後五年（2005 年–2009 年）卻佔了近三分之二（共 73 篇）。這說明 2004 年以前，因為技術不發達，運用語言學習語料庫進行的英語教學研究還沒有全面盛行。自 2005 年開始，越來越多的學者，開始注重這方面的研究。其中一個重要原因是 2004 年首屆中國語言學習語料庫語言學和英語教育教學研討會的召開。這場會議的舉辦，大大推動語言學習語料庫於英語教學研究的發展，使其成為語言學和英語領域的熱門議題。儘管這熱潮在 2006 年有所緩和（2005 年和 2006 年論文數量均為 12 篇），不過，自 2009 年起，這一領域的研究又出現了明顯的增長勢頭，呈現急劇上升的發展態勢。由此可見，內地在將語言學習語料庫應用於外語教學方面的關注程度，雖一度有所波動，但整體來看仍持續上升。

內地英語語料庫相關研究較早可見楊惠中、黃人傑（1982）發表於《外語教學與研究》第 4 期的〈JDEST 科技英語計算機語料庫〉。此後在二十世紀八十年代，相關研究數量仍相當有限；九十年代後逐步增加，至九十年代後期開始明顯上升，至 2005 年達到最高峰，總數多達 35 篇。

80 篇語言學習語料庫研究論文中，71 篇是基於 100 萬詞次或以下的小型規模的語言學習語料庫；7 篇是基於數百萬至上億詞次，大中型規模的語言學習語料庫；另有兩篇，未曾說明所用語言學習語料庫規模的大小。其中，大中型語言學習語料庫主要包括 British

National Corpus (BNC)、Bank of English 等具有代表性的大型語料庫；至於小型規模的語言學習語料庫，則包括 Lancaster-Oslo/Bergen Corpus (LOB)、Brown University Standard Corpus of Present-Day Edited American English (Brown)、London-Lund Corpus (LLC)，以及內地自主建立的 Chinese Learner English Corpus (CLEC)、College Learners' Spoken English Corpus⁵⁵ (COLSEC)、Spoken English Corpus of Chinese Learners⁵⁶ (SECCL) 和 Written English Corpus of Chinese Learners⁵⁷ (WECCL) 等。

在這 80 篇研究論文中，僅有 14 篇是依據口語語言學習語料庫為研究對象，比例佔 17%，其餘基本上都是關於書面語的研究成果；7 篇是根據雙語語言學習語料庫所做的研究，比例約 9%；還有少數是自建的、以特定語言為研究對象的學習語料庫，例如科研論文和媒體等。在這些研究成果中，利用語言學習語料庫標注屬性的論文，比例相對較高，有 33 項之多，比例超過 40%；用到“CLEC”錯誤標注屬性的，則將近有 20 篇，佔其中的 60% 左右。

在針對小型語言學習語料庫進行的 71 篇研究論文中，有 48 篇引用了學習者語料庫的相關數據作為分析依據。而其中佔據比例最大的，則是 Chinese Learner English Corpus (CLEC)，有 26 篇，佔學習者語言學習語料庫相關研究的 54%。此外，引用 Written English Corpus of Chinese Learners (WECCL)、College Learners' Spoken

55 <https://www.eduhk.hk/fhm/research-and-knowledge-transfer/digital-humanities/achievements-in-digital-humanities/the-english-speech-corpus-with-different-proficiency-levels-more-effective-teaching-and-learning>

56 https://corpus.eduhk.hk/english_pronunciation/index.php/our-spoken-english/

57 <https://www.uu.se/en/departament/english/research/english-linguistics/electronic-resource-projects/corpus-of-chinese-learner-english-ccle>

English Corpus (COLSEC) 和 Spoken English Corpus of Chinese Learners (SECCL) 有各 3 篇，還有少數以 English as a Foreign Language (EFLCT) 和 Learner International Spoken English Interlanguage (LINSEI-China) 為對象的研究。

為進一步了解內地語言數據資源的現況與未來走向，學界持續追蹤其於英語學習資源上的運用。有學者對過去五年（2007 年至 2011 年）發表於 10 種外語類核心期刊的 172 篇相關論文進行了統計分析（數據來自 China National Knowledge Infrastructure⁵⁸（中國知識基礎設施工程，簡稱 CNKI））。這 10 本主要期刊包括：《外國語》、《現代外語》、《中國翻譯》、《中國外語》、《外語教學與研究》、《外語界》、《外語學刊》、《外語教學》、《外語電化教學》及《外語教學原理與實踐》等。

根據統計資料，這五年間，聚焦於語言學習語料庫與英語教學相關的研究論文數量明顯增長，說明此領域日益受到學界重視。語言學習語料庫具備多方面優勢，例如提升語言統計分析的真實性和廣泛性，以及加快數據檢索的效率，這些特點正是推動該領域發展的關鍵因素。數據亦顯示，2007 年至 2009 年期間，10 本核心期刊共發表了 76 篇相關論文，佔總量的 44.1%，而 2010 年至 2011 年則有 55.9%。這反映自 2007 年起，語料庫與外語教學結合的相關研究熱度逐漸升高，越來越多學者投入此領域。早期主要以原理探討為主，因此在 2008 年論文數量略有回落，隨後學者們逐步在語料庫於英語教學各層面的應用取得進展，紛紛發表相關成果。特別是 2009 年之後，相關研究快速繁榮，2010 至 2011 年，學術成果在範圍和深度上都顯著擴展，應用價值亦日益明顯。總結來說，儘管期間曾有波動，內地在語

58 <https://oversea.cnki.net/index/>

料庫應用於英語教育研究方面的關注度總體持續上升。

中國英語語言學習語料庫研究成果顯著，在口語研究方面，呈現在教師教學話語的分析研究層面，他們對大學生英語口語的話語分析研究、語音研究、語言學習語料庫口語教學研究、語法、詞彙，甚至交際策略的研究等內容，都很感興趣，對書面語的話語研究則更是熱衷。從英語教育教學，到中介語學習研究，詞彙和語法的有效搭配，到用語的正確性和語篇的合理性，都取得了客觀的研究成果。

中國英語語言學習語料庫語言學電子數據化研究，也還有很多缺陷，要真正達到和國際語言研究水準的平等接軌，除了英語教師需要持續提升自身專業素養外，廣大的研究者，還應當具備相應的電腦操作知識。在這兩種影響原因得以改進的前提下，中國語言學習語料庫的研究事業，才可以拓展並持續成熟。基於漢語數據的很多設計是向英語取經，故在發展歷程上，掌握英語語料庫的優勢與不足，將為有效地設計漢語語料庫，建立起良好基礎。

2.3.2.2 內地語言學習語料庫漢語教學研究現狀

內地自上世紀二十年代起，已有學者嘗試構建數字化文本的語言學習語料庫，透過數據整理與分析，探索漢字在實際語料中的出現情況，並以此篩選常用字，擬訂適合教學應用的核心字表。雖然當時的語料庫規模有限，只能作為現代語言學習語料庫的雛形，且僅能以機讀方式處理，但這些早期嘗試為後來的漢語語料庫發展奠定了基礎，在漢語語料庫拓展歷程中具有開創性意義。

教育學者陳鶴琴基於語料統計分析，為教學設計編纂了《語體文應用字彙》一書，此書於 1925 年完成編撰，並在 1928 年由上海商務印書館發行。當中緒論對中文常用字彙的複雜性進行了詳盡說明，並

香港語文教學語篇難易度測定：
基於漢字測試語料庫的研究

作者：彭志全

編輯：張珏熒

設計：Dawn

出版：紅出版（青森文化）

地址：香港灣仔道 133 號卓凌中心 11 樓

出版計劃查詢電話：(852) 2540 7517

電郵：editor@red-publish.com

網址：<http://www.red-publish.com>

香港總經銷：聯合新零售（香港）有限公司

出版日期：2026 年 10 月

圖書分類：社會科學／教育

ISBN：978-988-8972-01-2

定價：港幣 130 元正

本書立足香港語文教學環境，結合語料庫語言學理論與近年 AI 技術成果，綜合參考國內外語料庫建設經驗，自主研發漢字語料庫，並制定從初小到初中的《九級字表》，建立量化文本篇章難易度的標準。

作者更設計並開發「九級字表分析」網上平台，利用 AI 自動分級功能，為每篇教材快速評定難度，生成科學、客觀的難易度數據，供教師實現校本分級課程、精準評估與個性化教學。

從語料庫建設到智能平台應用，本書為漢語教學提供一套完整而具操作性的科學框架，既有助減輕教師備課負擔，亦可提升學生的語文素養。



專業出版 國際銷售



紅出版文化平台

加入我們：www.red-publish.com

Mod^{pro}E.

上架建議：社會科學 / 教育

定價：港幣130元正