

# 漢字難度分析暨回饋系統之建置與發展 The Design and Development of a System for Chinese Character Difficulty and Features

Jung-En Haung<sup>1</sup>, Hou-Chiang Tseng<sup>1</sup>, Li-Yun Chang<sup>2</sup>,  
Hsueh-Chih Chen<sup>3</sup>, Yao-Ting Sung<sup>3</sup>

<sup>1</sup>Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology

<sup>2</sup>Department of Chinese as a Second Language/Chinese Language and Technology Center, National Taiwan Normal University

<sup>3</sup>Department of Educational Psychology and Counseling, National Taiwan Normal University

samuel8781@gmail.com, tsenghc@mail.ntust.edu.tw, { liyunchang, chcjyh, sungtc }@ntnu.edu.tw

## 摘要

漢字特徵分析（如：部件頻次、功能及結構組合等）在字本位教學中具有關鍵作用，但目前缺少部件結合難度的文本分析系統，本研究根據中文部件組字與形構資料庫以及學習者難度分級，建置漢字特徵與難度分析系統，功能包含：分析文本的漢字難度、提供文本漢字部件解構以及文本部件衍生字資訊，給予使用者深入且直觀的文本分析結果。本系統之預期效益能在教學面裨益教學者編排識字教材與出題，在研究面應用漢字特徵於中文自然語言處理之任務。

## Abstract

Feature analysis of Chinese characters plays a prominent role in "character-based" education. However, there is an urgent need for a text analysis system for processing the difficulty of composing components for characters, primarily based on Chinese learners' performance. To meet this need, the purpose of this research was to provide such a system by adapting a data-driven approach. Based on Chen et al.'s (2011) Chinese Orthography Database, this research has designed and developed an system: Character Difficulty – Research on Multi-features (CD-ROM). This system provides three functions: (1) analyzing a text and providing its difficulty regarding Chinese characters; (2) decomposing characters into components and calculating the frequency of components based on the analyzed text; and (3) affording component-deriving characters based on the analyzed text and downloadable images

as teaching materials. With these functions highlighting multi-level features of characters, this system has the potential to benefit the fields of Chinese character instruction, Chinese orthographic learning, and Chinese natural language processing.

關鍵字：漢字難度、漢字特徵、字本位教學、漢字教學系統

**Keywords:** character difficulty, character features, character-based education, instructional system for Chinese character education

## 1 緒論

### 1.1 研究背景

隨著全球各地的華語使用者人數不斷上升，華語教育成為近年來重要的研究方向。特別在近期中國孔子學院式微後，臺灣致力擴展華語師資、教材與資源之輸出，如國家教育研究院自 2013 年執行華語文教育八年計畫，以推動華語組織，並整合華語文資源之系統資源為目標（教育部，2013）。其中「建置應用語料庫及標準體系」從漢字、詞語、語法點各面向提供教學、教材設計與測驗評量的參考標準（國家教育研究院，2020），凸顯了難度建置與分級之於因材施教的重要性。

漢字作為組成文本的基礎單位，從形音義來看皆具有多層次的特徵。如：一字多音、一字多義，特別是漢字的數量與複雜的字形組成，常是華語學習者的難點（葉德明，2000）；為將漢字特徵進行整理能應用於華語教育，陳學志等人（2011）針對 6097 個正體中文字從筆畫、部件、偏旁、結構位置關係到整字，進行竭盡式的探究。該研究發展之中文部件組字與形構資料庫包含成字部件 246 個、非成字部件 193 個，並找出 11 種結構關係：

單獨存在、垂直組合、水平組合、封閉包圍、上方三面包圍、下方三面包圍、左方三面包圍、左上包圍、右上包圍、左下包圍、左右夾擊等結構與部件組合的關係；上述漢字字形特徵之分析，搭配字音、字義面的語言學特徵，有助於對漢字教學提出以證據為基礎之建議，並且此類漢字特徵亦有別於目前自然語言處理中的語言模型，如 BERT (Devlin, 2019)。未來可望進一步探討是否有助於中文自然語言處理的任務。例如，在文本可讀性模型加入漢字部件來作為訓練模型的特徵，期望可以有效提升模型之效能。

在漢字教學中，「字本位」的教學導向在語境和詞本位的教學常態下經常被探討，原因在於詞本位教學容易使漢字配合課文語境或詞語的功能出現，而忽略了漢字本身的特性和學習規律，以致在漢字的教學上成效不彰 (呂必松, 2005; 張金蘭, 2016)。為提升漢字學習成效，「字本位」教學導向的核心觀念是以漢字為教學的基本單位，藉由形音義的特性設計學習進度。其中，字形特性的教學，又以分析漢字部件對識字教學助益最大(黃沛榮, 2001)。學習者透過掌握部件與結構位置的組合規律，更高效地建立識字的辨認系統，例如：當學生在遇到「馬」這個部件，教師可運用部件衍生字的概念，引導學生留意當「馬」部件出現在左右結構的左側位置，衍生的「騎」、「駐」、「馭」、「駛」、「駝」等字，多半表示對使用馬所做的衍生動作；出現在右側位置，例如：「媽」、「嗎」、「瑪」、「碼」等，則多與「馬」部件的讀音相似。透過部件衍生字的教學，有助於學生發展組字覺識 (orthographic awareness) 且有助於促進識字量提升 (王瓊珠, 2005)。

若要具備基本讀寫能力，學生識字量至少須達 3000 字以上，且數量隨教育程度提高 (王瓊珠等人, 2008) 教師若想要運用部件衍生字等教法輔助閱讀學習，勢必需要拆解漢字至部件或更小的單位；其過程費時費力，即便運用字典或知曉漢字資料庫，若是一筆一筆資料逐筆搜查，恐怕勞神費時。因此多半尋找漢字語料網站作為工具，讓教材製作或試題編寫更省時省力。

以下本研究列舉以華語作為第一語言或第二語言的教學領域中，常見的能提供檢索功能的漢字語料網站，選取依據為漢字資料、

難度分析和部件衍生字分析等做整理 (如表 1 所示)。

## 1.2 漢字部件與難度分析相關系統整理

平台名稱	漢字資料、難度、部件相關功能	教學效益
<a href="#">漢語多功能字庫</a> (香港中文大學人文電算研究中心)	<ul style="list-style-type: none"> <li>整理漢字演變，以及常見字的背景介紹</li> <li>以古字部件樹的方式呈現字的組成結構</li> <li>字的衍生成語</li> </ul> 尚缺： <ul style="list-style-type: none"> <li>文本分析功能</li> <li>部件衍生字的整理系統</li> <li>漢字的難度分析等級</li> </ul>	<ul style="list-style-type: none"> <li>古字部件輔助學習理解漢字的形成</li> <li>部件古字和結構有益拆解漢字學習</li> </ul> 可精進： <ul style="list-style-type: none"> <li>文本整體分析和單一部件資料整理</li> </ul>
<a href="#">國際電腦漢字及異體字知識庫</a> (中央研究院)	<ul style="list-style-type: none"> <li>漢字的讀音、部首和字卡</li> <li>部件查詢可找尋包含部件的衍生字</li> </ul> 尚缺： <ul style="list-style-type: none"> <li>非成字部件之衍生字查詢</li> <li>詞語和文本部件查詢</li> <li>漢字難度分析</li> </ul>	<ul style="list-style-type: none"> <li>漢字資訊和教學字卡</li> <li>部件查詢益於同部件衍生難字教學</li> </ul> 可精進： <ul style="list-style-type: none"> <li>文本分析和非成字部件衍生字</li> </ul>
<a href="#">ACCESS 全漢字檢索系統</a> (國語中心)	<ul style="list-style-type: none"> <li>漢字讀音、結構、部件、筆畫分解等資訊</li> <li>將部件衍生分成四級的難度</li> <li>文本分析並分解成部件，但排列方式只依照出現順序</li> </ul> 尚缺： <ul style="list-style-type: none"> <li>同部件在文本者中出現的整理</li> </ul>	<ul style="list-style-type: none"> <li>使用者文本部件分析</li> <li>漢字資料的字卡作為教學用途</li> <li>等級難度作為教學先後的參考</li> </ul> 可精進 <ul style="list-style-type: none"> <li>部件衍生資訊為固</li> </ul>

	<ul style="list-style-type: none"> <li>更細分的學習者漢字難度分級</li> </ul>	定內容，應該隨文本做變化
華語教學標準體系應用查詢系統-漢字分級標準檢索系統(國家教育研究院)	<ul style="list-style-type: none"> <li>將漢字分為基礎、進階和精熟三個等別以及十一個級別</li> <li>提供漢字在書面和口語每百萬出現字頻</li> </ul> 尚缺: <ul style="list-style-type: none"> <li>沒有部件相關漢字資料</li> <li>無法進行文本的整體分析</li> </ul>	<ul style="list-style-type: none"> <li>掌握漢字的難度以安排教學的順序</li> <li>透過字頻高低和類型安排學習情境可精進</li> <li>查詢難度以文本會更便於使用者，搭配部件更有助於教學</li> </ul>

表 1. 相關系統平台整理

漢字相關系統平台整理如表 1，可以發現許多平台側重於讓學習者或教學者去認識漢字的演變或背景故事，進而對部件為何而組成有印象，雖在資料的完整上有充足的呈現，對於許多漢字的特徵也有很詳盡的解釋，但也容易讓學生將每個字以獨立的個體來記憶和識別，無法像部件教學那樣以部件衍生出新的漢字；「全漢字檢索系統」落實了文本分析和按照部件分類，若教學者欲查詢文本及其中的漢字部件是十分方便且清楚呈現的，可惜的是系統目前只能按照文本中出現的部件順序(如圖 1 所示)，而且延伸學習的部件是按照學習者難易度做多寡呈現，作為教學補充尚且足夠，但教學者若是欲分析授課文本，會需知悉其他相同部件的生字，在此教學者需求此功能就較為不足。

在漢字難度面向，具有分析功能的系統為「華語教學標準體系應用查詢系統」和「全漢字檢索系統」，教學者可透過專家訂定之詞組和漢字難度，判別學習者掌握的難易，然而兩者在查詢上皆以單獨漢字對應查詢為主，教學者若要進行文本分析，要重複的鍵入生字進行查詢，操作上較為耗時費力。

Component	Frequency	
一	1	檢
子	1	字
丿	1	系
美	1	漢

圖 1. 全漢字檢索系統部件順序

綜上所述，過往系統在漢字教學相當需要的難度分級以及特徵分析資訊兩面向，然而上述系統能提供的教學功能仍有突破的空間。再者，國家教育研究院的系統雖有難度分析功能，但由於其難度分級來自專家共識，未考量漢字各特徵的複雜度，亦未蒐集學習者於上千個漢字的學習表現，所以可能在客觀的漢字特徵指標計算以及難度的心理實質性上有待提升；為填補此研究缺口，Sung 等人進行實徵研究，以 675 位學習者為對象，蒐集每位學習者在 3,190 個漢字的作答反應，並以試題反應理論(item response theory, IRT; Drasgow & Hulin, 1990)分析學習者作答反應，估算每一個漢字的難度，改進了主觀認定的限制，為漢字難度提供更高的外部效度。本研究奠基於該 IRT 分析成果，彙整學習者表現之實徵資料以及中文部件組字與形構資料庫(陳學志等人, 2011)，發展漢字難度分析暨回饋系統，整體目標在提供使用者信而有徵的漢字難度與特徵資訊，並且透過系統之建置，提供更便捷的使用管道，以下分述系統設計、方法與各項功能說明。

## 2 漢字特徵分析平台系統設計

### 2.1 系統設計理念與架構

本研究設計的漢字特徵分析系統，包含三項主要功能:文本分析、難度分級與部件分布。第一、以文本分析功能對輸入的文字進行難度分析，並提供所組成的部件、結構、生詞等資訊，將文本整體和段落難度進行顏色的視覺化，並提供點選詳細漢字資訊以下載字卡之功能。第二、難度分級則期望讓使用者用數據化的方式，掌握高字頻以及在文本中高頻率出現之漢字的難度。本研究所有的難度分級，皆依據學習者的作答反應分析研究，因此適合教學者判別學習者可能遇到瓶頸的漢字。教學者除可安排讓學習者從高頻漢字進行認讀，亦可依照文本漢字難度作為選擇



教學文本的依據。第三、部件分布功能結合部件和結構，讓教學者能以部件切入，參照文本中相同部件的衍生字，按照難度進行教學，例如：以「魚」部件搭配字卡，教授文本中其他含有「魚」部件且較難的漢字如：「鯉」、「鯛」等字。

## 2.2 開發環境

本系統以微軟的網頁開發框架 Visual Studio ASP.NET 進行系統開發。此框架具有相容性高和編程穩定等特性，使系統架設更符合普遍使用者的規格需求。考量網頁系統美觀和使用者體驗的完善，系統前端程式引入 CSS 及 jQuery 等技術來進行前端程式之開發，讓使用者操作流暢兼具美觀。此外，本系統亦大量採用 Visual Studio.NET 內建的程式工具來進行開發。例如：將具有大量數據的漢字資料以 GridView 功能來進行表格的分類呈現，讓使用者可以一目瞭然系統分析的結果，並有條理的查閱漢字特徵資料。目前系統可以處理兩千字以內的文本分析，並過濾漢字以外字元。系統需對應使用者欲分析之文本難度、部件等資訊呈現，由於語料庫已經將 6097 個漢字以特徵和部件等相關資料整理成 Excel 表格，因此系統使用「C# Dictionary」功能進行資料儲存和排序，再依據功能做查表和資訊呈現。針對語料庫外的漢字和其他字符，如：英文字母、標點符號等，系統設計有過濾和不計入分析結果的功能，使用者輸入文本時，不會因為語料庫外的字元影響分析結果。依據不同功能，系統設計有相應的漢字分析的演算法，以「渣」的部件拆解為例，語料庫中的部件資料為「|(氵, -(木, -(日, -)))」，將資料依結構順序進行左右、上下拆解成「氵、木、日、一」四個部件，進而呈現字卡、文本高頻部件排序等功能。系統所使用的語料庫詳盡地歸納漢字資料，因此在各功能的呈現都能以簡潔快速的演算法做運算。目前此網站支援主流的瀏覽器，如 Google Chrome、Microsoft Edge 及 FireFox 等，以下針對本系統功能詳細描述與舉例。

## 3 漢字特徵分析系統各功能設計呈現

### 3.1 功能一：文字分析

#### ● 使用步驟

- (1) 在左框輸入文本
- (2) 點選「分析」按鈕
- (3) 在右框，點選想查看的字；可依需求，下載字卡

#### ● 分析結果

首先，使用者輸入欲分析的文本，系統會顯示每個漢字的難度(見圖 2-1、圖 2-2)，再者點選漢字連結會呈現漢字字卡(見圖 3)，內容包含漢字、簡體寫法、筆劃、注音、六書、難度、漢語拼音、文字結構、部首和部件，最後使用者可下載成字卡圖片。特別值得注意的是本系統給予每個難度對應顏色，在文本分析時，能在視覺上做直觀辨別。

#### ● 教學效益

教學者可透過此功能直觀檢視文本中的漢字難度分布，點選所需教學漢字獲取資訊，下載字卡圖片，在教學上彈性使用，例如：字卡上有【選擇呈現詞首、詞中、詞尾】的功能，教學者可選擇不呈現字的詞尾，引導學生自行發想衍生詞彙，或者進行形成性評量(以字造詞、造句；寫出課本中包含該字的詞…；亦可在完成難度分析或部件分析後，將欲教學的衍生字以字卡方式呈現給學習者。

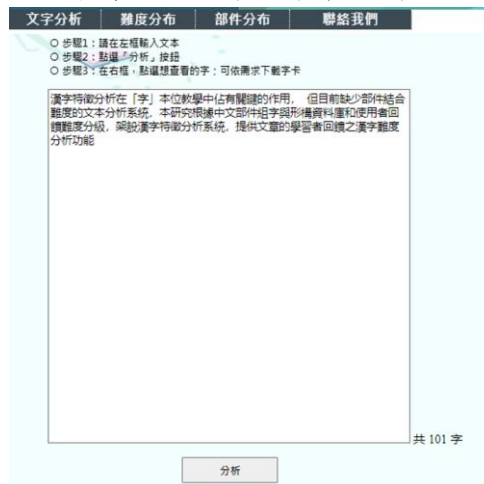


圖 2-1.漢字特徵分析輸入方格



圖 2-2.文本漢字難度呈現



圖 3. 漢字字卡呈現

### 3.2 功能二：難度分布分析

#### ● 使用步驟

- (1) 在左框輸入文本
- (2) 點選「分析」按鈕，系統自動呈現文本的漢字次數統計在右框
- (3) 在右框，請依需求，選擇結果呈現方式：依照出現次數由高到低排列；依照難度排列分析結果

#### ● 分析結果

難度分布功能的呈現方式有二：首先，依據漢字在文本的出現次數做排序(見圖 4)，讓使用者檢視文本中高頻率出現的漢字及其難度；再者，使用者亦可以按照難度做分級呈現(見圖 5)，檢視文本的平均漢字難度以及哪些難度出現較頻繁，透過上方難度按鈕，選取顯示特定難度。重要的是，當使用者以兩者交叉比較，能以數據看見最高頻率出現的字為何種難度，以及最常出現的是哪些難度的漢字。

#### ● 教學效益

教學者可運用此功能，評估特定文本難度對學習者可能造成的學習負荷，例如學習者可能會在哪些較難的漢字遇到問題，進而在教學上做先備識字的輔助或閱讀文本後的加強，例如：文本中(見圖 5)「的」出現比例最高，該字之難度因為「難度低」(在 9 個等級中，排第 2 級)，因此可判斷為基礎且重要的漢字，在教學中值得優先介紹；或者選擇整篇文本中，難度較高的第 7 級漢字：「徵」、「析」、「率」等字，以部件拆分搭配字形結構的拆解，讓學生更細部了解較難的漢字，減輕識字學習上的負荷。

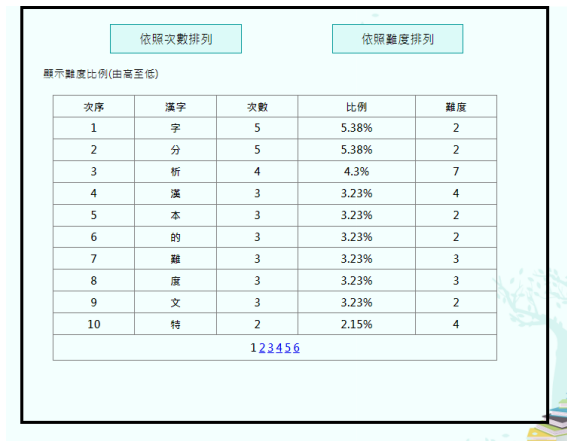


圖 4. 漢字按出現頻率排序功能

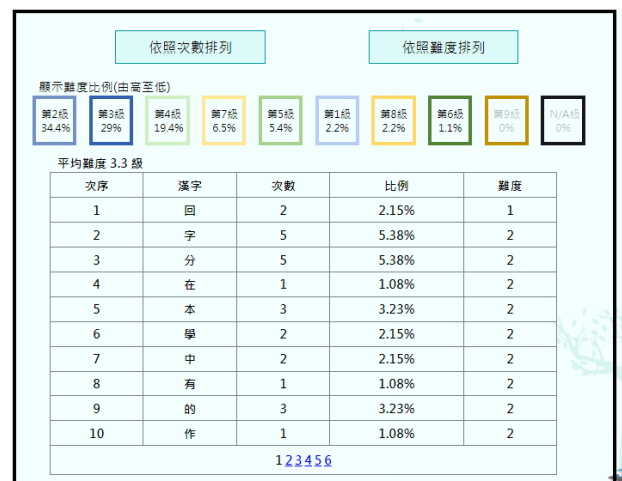


圖 5. 漢字按難度排序功能

### 3.3 功能三：部件分布分析

#### ● 使用步驟

- (1) 請在左框輸入文本
- (2) 點選「分析」按鈕，系統自動拆解文本中漢字的部件組成
- (3) 在右框，點選想查看的部件；可依需求，下載字卡

#### ● 分析結果

首先，部件分布功能能讓使用者透過分析結果，檢視部件在文本中出現的頻率(見圖 6)，接著，使用者點入特定部件的連結後，會呈現該部件是否成字，以及文本中還有哪些漢字有此部件(見圖 7)，最後可下載字卡。值得注意的是，衍生字會顯示難度分級，使用者可以依據難度，進行部件的衍生字教學。

#### ● 教學效益

教學者可搭配漢字難易度，針對同部件的衍生字，由難度低至高進行教學，讓學生學習較高難度級別的漢字前，能有循序漸進的

鷹架 (scaffolding)，例如：「口」部件在文本中的出現頻次高，教學者可安排讓學生學習「口」這個基礎部件字後，再學習較進階的衍生字：在左右結構中，兩個部件的「和」，三個部件的「結」，一直到包含三個口部件的「讓」，學生透過慢慢疊加的部件數來學習，相對於直接學習生難字，預計更能漸進掌握生字。



圖 6. 漢字部件分布功能



圖 7. 部件資料和衍生字字卡呈現

#### 4 討論與未來發展

本研究所發展之「漢字難度分析暨回饋系統」提供使用者多面向地了解文本中漢字特徵的功能，包含：每個字的難度與特徵資訊（功能一：文字分析）、文本中漢字的難度分布與頻率分布（功能二、難度分布）、文本中的部件頻率與相關特徵（功能三：部件分布）。對使用者而言，如果能靈活搭配各功能，將快速掌握文本，進而編寫教材、發展教具或設計識字之評量題目，裨益教師、家長、出版社、教材開發者或任何對漢字難度與特徵感興趣之使用者。

為了有效地運用蒐集自學習者的難度資料，本系統未來將開發「建議學習內容」之功能，透過輸入文本的漢字難度，提供更豐富的訊

息，讓使用者透過選單等方式，除了理解文本中的資料外，能查詢所需衍生學習的漢字資料，例如：搭配漢字難度（比現階段難度更低的漢字選項，或難度更高的漢字選項）與字形特徵（都是上下結構或左右結構的漢字、都具有特定部首的漢字）落實語言教學中的 i+1 教學觀點 (Krashen, 1981)，即透由此項「建議學習內容」，讓使用者更快速掌握鷹架，選擇下一階段所適合學習的漢字。

除此之外，本研究未來也將進行驗證並蒐集使用者意見，進而優化本系統。例如：邀請教師使用系統、蒐集教學者和網站設計專業人員之回饋，期能透過問卷分析和質性使用者體驗晤談等方式，對系統進行功能和介面上的修正以符應教學需求。本研究設計之漢字難度特徵分析系統目標能成為漢字教學的輔助系統，讓教學者透過分析文本後平台提供之漢字的難度分級、部件分布等資訊，更有規劃的進行教學設計，裨益於識字學習，未來也可能透過數據蒐集與分析，將本研究的漢字特徵納入自然語言處理的判斷依據，對中文文本的可讀性相關研究作出貢獻。

#### 參考文獻

- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.
- Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp.577–636). Consulting Psychologists Press.
- Krashen, S. D. (1981). *Second Language Acquisition and Second Language Learning*. Pergamon Press Inc., Oxford
- 王瓊珠、洪儷瑜、張郁雯、陳秀芬 (2008)。一到九年級學生國字識字量發展。教育心理學報，394(4)，555-568。
- 王瓊珠 (2005)。高頻部首／部件識字教學對國小閱讀障礙學生讀寫能力之影響。臺北市立師範學院學報。36(1)，95-124。
- 教育部 (2013)。邁向華語文教育產業輸出大國八年計畫。中華民國教育部。

- 張金蘭 (2016)。中國傳統識字教育給對外漢字教學帶來的啟示。《第六屆「開創華語文教育與僑民教育之新視野」國際學術研討會》，未出版。
- 國家教育研究院 (2020)。遣詞用「據」－臺灣華語文能力第一套標準。國家教育研究院。
- 陳學志、張璣勻、邱郁秀、宋曜廷、張國恩 (2011)。中文部件組字與形構資料庫之建立及其在識字教學的應用。《教育心理學報》，43(閱讀專刊)，269-290。
- 黃沛榮 (2001)。漢字教學的理論與實踐。樂學書局。
- 葉德明 (2000)。外籍生漢字書寫策略探討。第六屆世界華語文教學研討會論文集(三)，311-320。世界華語文教育學會。