第三章　系統架構

本章介紹本系統的系統架構，以及本系統所需的語料和詞典。本章分成四個小節，3.1 節說明系統需求，3.2 節介紹系統的架構，3.3 節介紹語料來源，3.4 節介紹本系統所需要的詞典。

3.1 系統需求說明

人與人之間溝通是透過語言，故人們從幼兒時期即開始學習語言能力，然而真正有系統化教學是在國民小學階段，迅速建立起中文基礎能力的關鍵期。在現今國小教育當中，國語科目在課程上的份量占大多數，其目的是希望能夠奠定良好的語文基礎，所以語文基礎的好壞，則會影響到日後的寫作和表達能力，倘若此基礎能力有所不足，則在中文知識的學習上將會有所限制。

故本研究選擇國小國語科目做為教學標的，提出了電腦輔助中文試題出題系統，希望藉由電腦程式來提供教師試題編輯服務，讓試題編撰者不需要花費大量功夫在句子的篩選上，並且能夠設計出不同類型的題目，達到試題多樣化的目的。早期製作試卷時必須完全用手工的方式來完成。電腦程式雖然很難完全取代試題編輯者，使系統能夠自動化出題，不過可以協助編輯者收集相關資料以及編輯試卷，減少不必要的人工編輯試卷時間。

本系統在設計上，事先從建立一套出題的流程開始，與試題相關的資料庫切分成兩個部分，一個是題庫資料庫，一個是試卷資料庫。編輯好的試題並不是成為試卷上的試題，而是儲存於題庫資料庫，這流程的作法能幫助教師管理試題資料，下次教師出試卷時，本系統就

可以從題庫資料庫裡擷取試題（預先編輯完成的試題），來輔助教師出題，編輯好的試卷，儲存在試卷資料庫，將來學生接受測驗時，本系統就會從試卷資料庫擷取試卷提供給學生做測驗。

3.2 系統架構介紹

圖 3.1 為輔助出題的系統架構圖，圖中矩型為資料庫，而圓角的部分則為處理程序。基本語料庫的資料加工成所需要的格式後，提供給教師做為編輯題目的來源，編輯好的試題儲存於題庫資料庫，且可以依題型的不同，細分成個別的題庫資料庫。

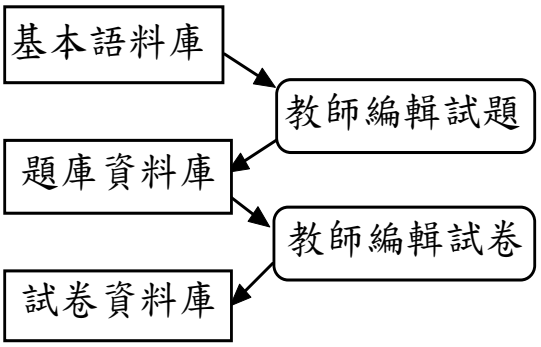

圖 3.1 系統架構圖

我們建立四聲辨識、中文克漏詞、國字注音、改錯字、量詞及中文句子重組等題庫。教師編輯試卷的時候，本系統會提供各類題型的題庫供教師選擇，編輯好的試卷檔，則儲存於試卷資料庫。將來學生接受測驗時，選擇所要的試卷，本系統就會從試卷資料庫擷取試卷檔供學生做測驗。學生測驗後會記錄在學生成績檔裡。教師可以根據學生資料檔，查詢特定學生的答題資料，或者是個別班級的答題資料，了解學生整體表現，做為以後教學的評量及依據。

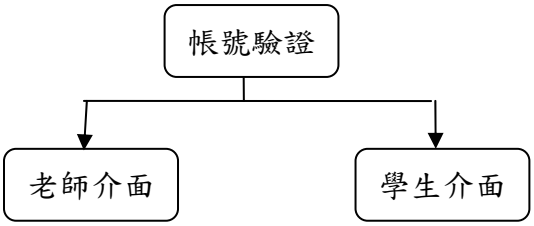

圖 3.2 本系統的功能架構圖

本系統的功能架構圖如圖 3.2 所示，一開始使用本系統必須先經過帳號驗證（也就是輸入使用者帳號及密碼），來判斷使用者的身份是老師還是學生，如果使用者身份是老師，則進入老師介面；身份是學生，則進入學生介面。

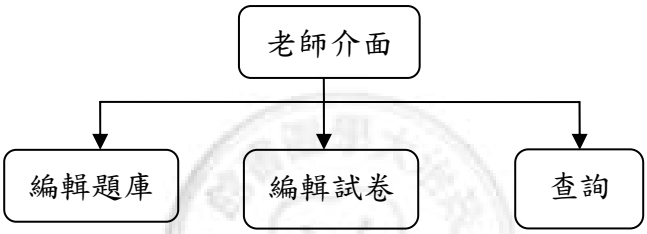

圖 3.3 老師介面功能架構圖

老師介面功能架構圖，如圖 3.3 所示，老師具有編輯題庫、編輯試卷及查詢的功能。編輯題庫的功能裡提供編輯四聲辨識、中文克漏詞、改錯字、量詞及中文句子重組等題目。

編輯試卷的功能提供編輯四聲辨識、中文克漏詞、改錯字、量詞、國字注音及中文句子重組等試卷編輯，使用者進行編輯試卷中，從題庫裡選取題目後，編成一張試卷儲存在試卷資料庫。

查詢的功能包含查詢學生作答情形及題庫內容。在查詢學生作答情形的介面上，顯示學生所有作答的內容，並且在作答的內容旁提供正確答案，供老師做查詢比對。在查詢題庫內容的介面上，提供所有老師編輯四聲辨識、中文克漏詞、改錯字、量詞及中文句子重組的題目。

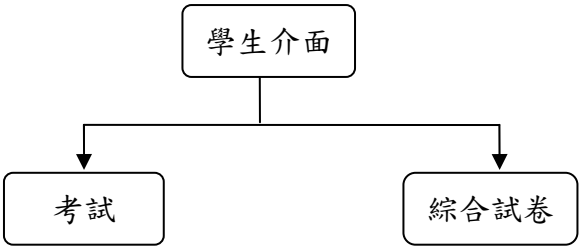

圖 3.4 學生介面功能架構圖

學生介面功能架構圖（圖 3.4），本系統提供學生考試及綜合試卷的功能。本系統在考試的功能上提供四聲辨識、中文克漏詞、改錯字、量詞、國字注音及中文句子重組單一類型考題測驗。在綜合試卷的功能上，則採用中文克漏詞、改錯字、量詞及國字注音混合測驗四種考題類型來測驗學生。

3.3 語料來源

基本語料庫的資料是經過處理後的文字檔（例如：國小國語課本的內容，把一些標題符號刪除等），我們製作本系統時，預先收集坊間的國小國語科的參考書裡的生字表，並且利用網路爬梳器（web crawler）去下載有關國小課本內容的資料[1]，下載的檔案都為 html 的檔案，檔案裡包含了許多標記式語言，所以我們必須篩選出主體內容的資訊，加工後成為我們所需要的格式，提供給教師為指定的語料來源，這些語料提供給教師來編輯試題。

我們利用網路爬梳器主要收集的資料如下，並且加以說明。

1. 課文的句子：本系統由網路爬梳器所收集的中文句子，總共有 57800 個句子（共有 625424 個中文字），以及南一書局出版的國小國語參考書內容[8]，將來可以使用在編輯中文克漏詞題目。
2. 課文句子的斷詞部分：我們把所收集課文的句子，把每個句子傳送到中研院斷詞系統[2]處理後，再傳回斷詞結果中所附帶的詞類

訊息，本系統會根據詞類訊息使用在編輯量詞的題目。

3. 國字的倉頡碼：我們收集了許多國字的倉頡碼，把一些無法顯示在電腦螢幕上的中文字刪除後，一共收集了 13685 個中文字的倉頡碼[10]，製作成一個中文倉頡碼檔。我們從教育部的網站下載字頻總表（以中文字常出現的頻率，排列順序是由高到低排序），而字頻總表所採取樣本的資料來源，總共有五類，分別為：八十七年度出版之雜誌、八十七年度暢銷之書籍、八十七年度印行之報紙、八十七年奇摩站分類索引中之各類網站及八十七年口語調查之資料，以上這五項之細目，則在此網頁內容（ http://www.edu.tw/EDU_WEB/EDU_MGT/MANDR/EDU6300001/result/87news/page1-3.htm?open）。
4. 從字頻總表中的每一個文字，利用倉頡碼的規則，編輯中文構字式（這部分在第五章會詳細敘述），製作成中文構字式檔，本系統會採用中文倉頡碼檔及中文構字式檔，使用在編輯改錯字的題目。

3.4 詞典

本系統執行輔助編輯者出題，尋找字的發音及相似詞，採用了兩個辭典，一個為國字注音辭典（http://chewing.csie.net/chewing_dict_edit.html），一個為英漢雙語的辭典 HowNet（http://www.keenage.com）。

國語字典，本系統採用詞庫檔為 tsi.src，也是新酷輸入法[12]所採用的詞庫檔，其格式如下圖 3.5 所示。

看 28254 ㄎㄢ
看 28254 ㄎㄢˋ
看一下 206 ㄎㄢˋ ㄧˊ ㄒㄧㄚˋ
看一看 154 ㄎㄢˋ ㄧˊ ㄎㄢˋ
看了 2071 ㄎㄢˋ ㄌㄜ•
看了又看 9 ㄎㄢˋ ㄌㄜ• ㄧㄡˋ ㄎㄢˋ
看人 57 ㄎㄢ ㄖㄣˊ
看人 57 ㄎㄢˋ ㄖㄣˊ
看人而定 0 ㄎㄢˋ ㄖㄣˊ ㄦˊ ㄉㄧㄥˋ
看人嘴臉 0 ㄎㄢˋ ㄖㄣˊ ㄗㄨㄟˇ ㄌㄧㄢˇ

圖 3.5 辭典資料格式（Ｉ）

tsi.src 辭典的每一行格式由 「詞」+「一格空白」+「數字」+「一格空白」+「詞的注音」所組合而成的。「詞」為各個字的相關詞，「數字」為詞的優先順序，數字越大則優先順序越高，而「數字」是統計語料庫所得到。「詞的注音」則為 「詞」的發音，例如：「看一下」這個詞彙的發音為 「ㄎㄢˋ ㄧˊ ㄒㄧㄚˋ」。

HowNet 是一個以中文和英文的詞所代表的概念為描述對象，以揭示概念與概念之間以及概念所具有的屬性之間的關係為基本內容的常識資料庫[23]。由董振東與董強兩位學者所編撰完成收錄約十一萬條詞條。本系統所採用 1999 年的 HowNet 辭典，其辭典的格式如圖 3.6：

NO.=001502
W_C=爸爸
G_C=N
E_C=
W_E=dad
G_E=N
E_E=
DEF=human|人,family|家,male|男

圖 3.6 辭典資料格式（Ⅱ）

如圖 3.6 所示，我們分別介紹 HowNet 辭典資料格式裡的各個欄位，在 HowNet 中，每個詞彙都有八個欄位，“NO.”代表該詞彙的號碼，“W_C”代表中文詞條，“G_C=N”代表中文詞性，“E_C”為保留欄位，“W_E”代表該詞彙的英文詞條，“G_E”代表英文詞條，“E_E”為保留欄位，DEF 為義原關係，也就是描述此詞彙的概念關係。HowNet 大概使用了一千七百個多個義原關係[13]，來定義中英雙語知識辭典中的每個詞彙，並且建有描述各個義原之間的關係的分類樹，例如：“讀書”一詞由“從事”、“學”與“教育”三個義原定義而成，所以我們定義 s 為讀書義原的集合，s={從事，學，教育}，我們去尋找 HowNet 所有的詞彙，並且把詞彙的義原與集合 s 做比對，與 s 有交集的詞彙都找出來後，接下來做排序（由多到少），把排序後的詞彙提供給編輯者做為誘選答案。