第五章　相關文字處理技術

本章介紹本系統所提供的技術，在 5.1 節說明在中文克漏詞試題方面，如何提供誘答選項的技術，5.2 節說明在改錯字試題編輯方面，如何提供相似字的技術，5.3 節說明如何判斷句子中是否存在著量詞，5.4 節說明本系統在中文句子重組試題方面所提供的技術。

5.1 中文克漏詞

1. 另方面卻又對名牌所標榜的精緻及___愛不釋手

（1） 品質　→　正確答案
（2） 品嚐　→　誘答選項
（3） 品德　→　誘答選項
（4） 品管　→　誘答選項

圖 5.1 中文克漏詞試題的範例

如圖5.1是一題中文克漏詞試題的範例，挖掉了一個詞的句子稱為題幹（stem），被挖掉的詞即是該試題唯一的答案（key），而其它三個選項稱為誘答選項。

接下來我們要進行編輯誘答選項，本系統提供兩種方式產生誘答選項的詞彙，提供給編輯者做為參考。

第一種方式，在HowNet中，每個詞彙都有八個欄位，尤其最後一個欄位名稱為DEF，描述著這個詞彙的義原關係，HowNet大概使用了一千七百多個義原關係[13]，來定義中英雙語知識辭典中的每個詞彙，並且建立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。

第二種方式是採用中研院“現代漢語語料庫”一詞泛讀的學習工具[18]，本系統會連到一詞泛讀的網頁，假如答案為 A 詞彙，會自動連到中研院一詞泛讀的網頁，並且傳送參數 A 詞彙，網頁會回傳有關 A 詞彙的近義詞，本系統收集網頁所回傳的結果後，就會提供 A 詞彙的近義詞，提供給編輯者。

第三種方式是 HowNet 裡的所有詞彙與指定的詞彙比對，如果有交集相同中文字的詞彙，則提供給編輯者選擇，例如：「明天」與「晴天」二個詞彙，有共同的中文字「天」。

5.2 改錯字

就錯別字試題編輯方面來說明，本系統會針對編輯者所選的字，提供同音字、相似音及相似字三種功能。

首先，先介紹第一部分，在同音字方面，本系統使用詞庫檔為 tsi.src，也是新酷輸入法所採用的詞庫[11]。本系統會根據編輯者所選的字，從詞庫檔找出相同發音的國字，全部列出來，供編輯者做為選擇，例如：別出“心”裁，常會被誤用為別出“新”裁。

第二部分介紹相似音的部分，相似音也就是發音非常的相似，容易讓人產生混淆不清。在相似音的部分，參考網頁的內容（ http://tw.myblog.yahoo.com/linpelu-2006/article?mid=9&prev=169&next=4&l=f&fid=8&sc=1），於是在相似音的部分，我們主要分成三個類別。

✧ 捲舌音與不捲舌音（ㄓㄔㄕ與ㄗㄘㄙ），例如：「知」與「資」的

發音，前者屬於捲舌音，後者是不捲舌音。

✧ 聲母相似音（ㄐㄑㄒ），例如：「雞」與「漆」的發音。

✧ 韻母相似音（ㄣ和ㄥ），例如：「謹」與「景」的發音。

✧ 本系統會根據編輯者所選的字，從詞庫檔找出具有相似發音的國字，全部列出來，供編輯者做為選擇。

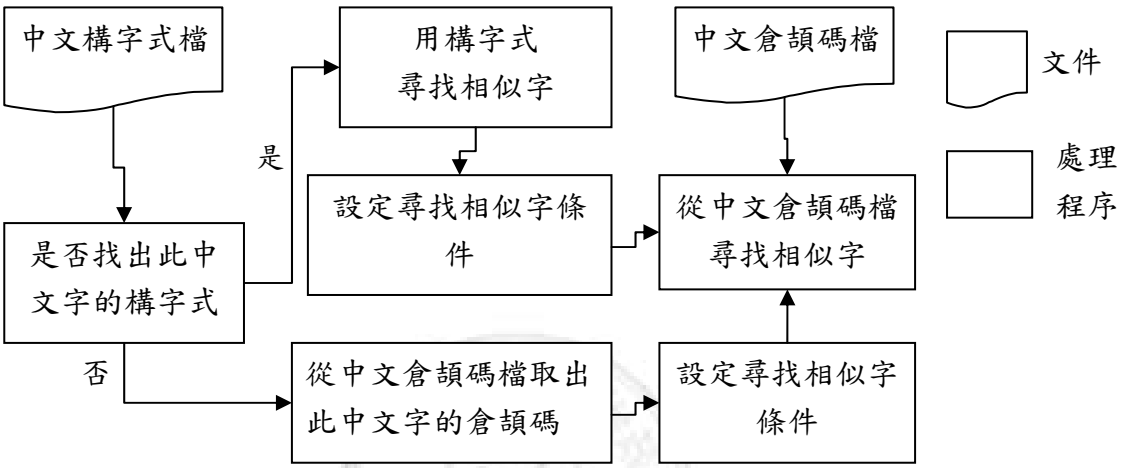

圖 5.2 尋找相似字的流程圖

第三部分介紹尋找相似字的部分，圖 5.2 為尋找相似字的流程圖，圖中最右方說明文件及處理程序各屬於那一類的圖型。首先我們先介紹如何製作中文構字式檔及中文倉頡碼檔，最後再介紹三個處理程序，分別是找出中文字的構字式、設定尋找相似字的條件及尋找相似字。

本系統主要是採用倉頡碼技術來尋找相似字，因為倉頡碼是根據字的形狀來做分類[4]，每個字根有相對應的字根符號。倉頡字母總共可分為四大類，依序為哲理、筆劃、人身及複雜筆劃。以下說明這四類的內容。

✧ 哲學類：與大自然環境有關，例如：日、月、金、木、水、火、土。

✧ 筆劃類：將中文字筆劃分為斜、點、交、叉、縱、橫、鉤，將這七類歸類成筆劃類，分別以竹、戈、十、大、中、一、弓來表示筆劃類的全部七類。

✧ 人身類：與人體器官有關，例如：人、心、手、口。

✧ 複雜筆劃類：將中文字重要且複雜的字型分為側、並、仰、紐、方、卜，把這六類歸類成複雜筆劃類，分別以尸、山、廿、女、田、卜來表示複雜筆劃類的全部六類。

接下來介紹中文字如何取出倉頡碼。首先必須先知道取碼的順序，一開始先判斷中文字是“分體字”或者是“連體字”。一個中文字可以分為兩個或兩個以上的部分，則稱為“分體字”，而“分體字”才有“字首”與“字身”。第一個切割出來的部分，則稱為“字首”，其餘的部分稱為“字身”。“分體字”的分割方式如下。

✧ 字首字身上下重疊：例如：音（上為「立」，下為「曰」）。

✧ 字首字身左右並排：例如：話（左為「言」，右為「舌」）。

✧ 字首由外而內包含字身：例如：國（外為「口」，內為「或」）。

中文字合乎上述三項中其中一項，則為分體字，若不合乎三項中其中一項則為連體字。

再來介紹取倉頡碼的規則，假設此中文字為分體字，則先取字首倉頡碼，再取字身倉頡碼。

首先，取字首倉頡碼的規則分成三點。

✧ 字首倉頡碼為一碼，則取一碼：例如：字身為人，倉頡碼為“人”。

✧ 字首倉頡碼剛好為兩碼，則取兩碼：例如：字身為二，倉頡碼為“一一”。

✧ 字首倉頡碼超過兩碼，仍舊取兩碼（頭、尾碼共兩碼）：例如：字身為剛，倉頡碼為“月山”。

下列四點為取得字身倉頡碼的規則。

✧ 字身倉頡碼剛好為一碼，則取一碼：例如：字身為人，倉頡碼為“人”。

✧ 字身倉頡碼剛好為兩碼，則取兩碼：例如：字身為二，倉頡碼為“一一”。

✧ 字身倉頡碼剛好為三碼，則取三碼：例如：字身為寺，倉頡碼為“土木戈”。

✧ 字身倉頡碼超過三碼，仍舊取三碼（頭、二及尾碼共三碼），取三碼的目的是因為倉頡碼檔裡的每一個國字，字身倉頡碼最多三碼。例如：字身為馬，字身倉頡碼為“尸手火”。

假設中文字為連體字，取連體字倉頡碼的規則分成以下四點。

✧ 倉頡碼剛好為一碼，則取一碼：例如：字身為人，倉頡碼為“人”。

✧ 倉頡碼剛好為兩碼，則取兩碼：例如：中文字為二，倉頡碼為“一一”。

✧ 倉頡碼剛好為三碼，則取三碼：例如：中文字為寺，倉頡碼為“土木戈”。

✧ 倉頡碼剛好為四碼，則取四碼：例如：中文字為馬，倉頡碼為“尸手尸火”。

✧ 倉頡碼超過四碼，仍舊取四碼（頭、二、三及尾碼共四碼）。例如：中文字為愛，字身倉頡碼為“月月心水”。

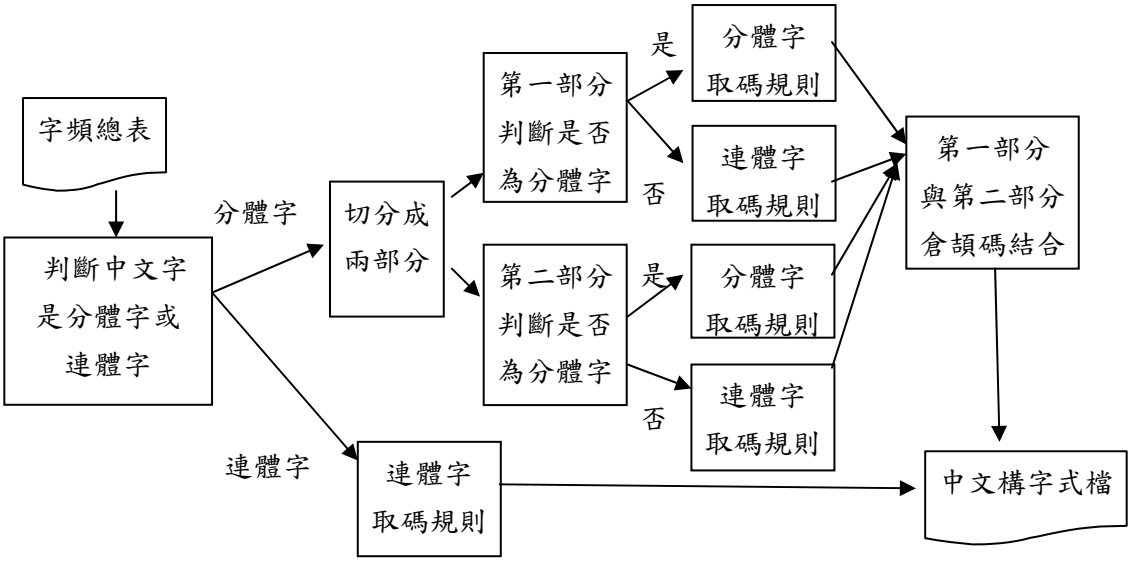

圖 5.3 中文構字式檔製作流程

圖 5.3 為中文構字式檔製作流程，字頻總表共有 5063 個字。我們把字頻總表的每一個中文字，製作成中文構字式。

首先，我們先判斷此中文字是分體字或連體字，如果中文字滿足之前所描述的“分體字”分割方式其中一項，則我們認定此中文字為分體字，於是我們把此中文字分成字首與字身，字首當成第一部分，字身則當成第二部分。我們把第一部分（此中文字的字首）的字，判斷是否為分體字，如果是，就利用分體字取倉頡碼的規則，對第一部分的字取倉頡碼；如果第一部分不是分體字，則利用連體字取倉頡碼的規則來取第一部分的字。同樣地，我們把第二部分（此中文字的字身）的字，去判斷是否為分體字，如果是，就利用分體字取倉頡碼的規則，對第二部分的字取倉頡碼；反之，第二部分的字如果不是分體字，則用連體字取倉頡碼的規則來取第二部分的字，最後再將第一部分的倉頡碼與第二部分的倉頡碼結合（這也是此中文字的構字式），儲存在中文構字式檔裡（在中文構字式檔裡，分體字會有兩組倉頡碼）。如果中文字未滿足之前所描述的“分體字”分割方式其中一項，則我們認

定為此中文字為連體字；我們使用連體字取倉頡碼的規則來取此中文字，儲存在中文構字式檔裡（在中文構字式檔裡，連體字只有一組倉頡碼）。

```
對 廿金廿土 木戈
成 戈竹尸
公 金戈
行 竹人 一一弓
後 竹人 女戈竹水
地 土 心木
而 一月中中
聖 尸十口 竹土
```

圖 5.4 中文構字式檔的內容

上圖 5.4 為中文構字式檔的內容，每一行資料會有一個國字並搭配一組或者是兩組的倉頡碼（每組倉頡碼的個數大於等於一）。中文構字式檔的內容分成兩類，一類為分體字，另一類則為連體字。假設此中文字為分體字時，中文構字式檔裡會有兩組倉頡碼，連體字在中文構字式檔只有一組倉頡碼，將來電腦就可以判斷出中文字是分體字或者是連體字。

下圖 5.5 為中文倉頡碼檔的內容，每一行資料會有一個中文字搭配此中文字的倉頡碼。我們收集了 13685 個中文字的倉頡碼[10]，儲存在中文倉頡碼檔，本系統提供這個檔案，其主要目的是能夠提供更多的相似字。

```
人 人
倜 人月土口
鷓 人月竹日火
俗 人金人口
```

圖 5.5 中文倉頡碼檔

接下來說明本系統如何從中文構字式檔裡來尋找相似字。首先，本系統根據使用者編輯的中文字，從中文構字式檔裡是否能尋找到此中文字的構字式倉頡碼，假設在中文構字式檔找到此中文字，並且能判斷出此中文字是連體字或分體字，如果是連體字，則會有一組構字式倉頡碼，令α等於空字串，β等於這一組構字式倉頡碼；如果此中文字為分體字，則會有兩組構字式倉頡碼，則α等於第一部分的構字式倉頡碼，β等於第二部分的構字式倉頡碼，例如：中文構字式檔找到「地」這個字，這個字為分體字，則會有兩組倉頡碼分別為「土」及「心木」，則α等於「土」，β等於「心木」。

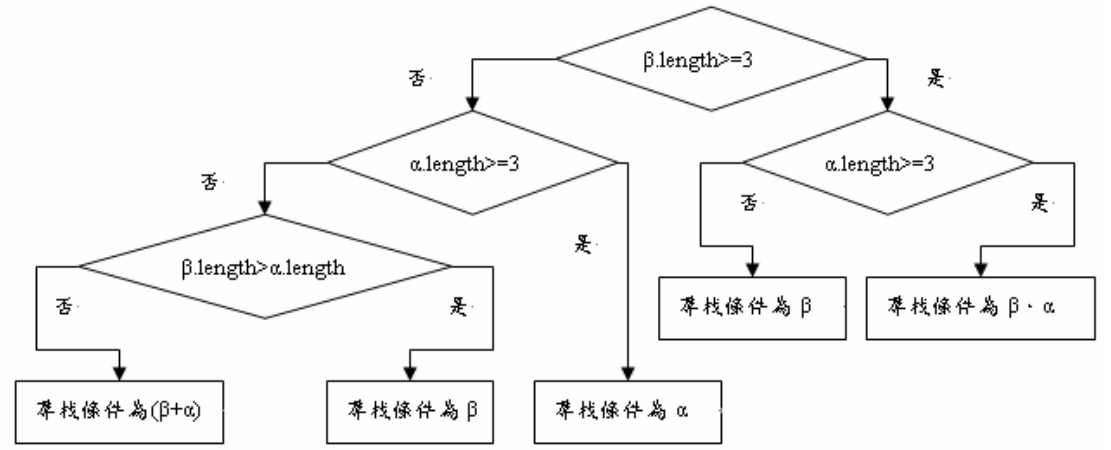

圖 5.6 中文構字式檔設定尋找相似字條件的流程

上圖5.6為中文構字式檔設定尋找相似字條件的流程，假設中文構字式檔找到此中文字，則會找到此中文字的構字式倉頡碼（一組或兩組的構字式倉頡碼），本系統要設計一個尋找相似字的流程，會依據α、β（構字式倉頡碼）的倉頡碼個數來決定尋找相似字（從中文構字式檔裡去尋找相似字）的條件。我們令θ為尋找相似字的條件，本系統會利用尋找相似字的條件，與中文構字式檔每一個字的構字式裡的α（第一部分的倉頡碼）、β（第二部分的倉頡碼）來比對，假如α

或者β的構字式倉頡碼，只要其中一個構字式倉頡碼與θ（尋找相似字條件）相同時，則本系統就認定是相似字。

我們之前收集了中文倉頡碼檔，接下來介紹如何利用θ從中文倉頡碼檔裡尋找相似字。

本系統尋找相似字的方式是利用 θ（尋找相似字條件）去比對中文倉頡碼檔每一個中文字倉頡碼的後幾碼（因為倉頡碼檔裡，每一個中文字的字身倉頡碼最多三碼，而且又位於倉頡碼的後幾碼），所以分成二類來尋找相似字。

- ✧ θ 如果小於等於三碼，則只需比對中文倉頡碼檔裡所有中文字的倉頡碼的後幾碼是否與 θ 相同，如果是，就是相似字，例如：θ為“一弓口”，而“何”的字身為“可”，字身的倉頡碼為“一弓口”，θ與“何”的字身倉頡碼同為“一弓口”，所以就認定“何”這個字為相似字。
- ✧ 如果 θ（尋找相似字條件）總共四碼或四碼以上時，則需要取 θ 的第一碼、第二碼及最後一碼，總共三碼去比對中文字倉頡碼的後三碼，只要比對相同，就認定為相似字。例如：θ 為“愛”這個字的倉頡碼，倉頡碼內容為 “月月心水” ，θ 的倉頡碼個數為四碼（超過三碼），所以就取 θ 的第一碼、第二碼及最後一碼做為尋找相似字的條件，尋找條件為 “月月水” ，而“儇”的倉頡碼為“人月月水”，尋找條件與“儇”的倉頡碼的後三碼同為“月月水”，所以就認定“儇”這個字為相似字。

接下來介紹，如果從中文構字式檔裡是找不到此中文字的構字式倉頡碼時，則會從中文倉頡碼檔去尋找此中文字的倉頡碼，將來做為尋找相似字的條件。假設本系統從中文倉頡碼檔找到中文字的倉頡碼

為λ，本系統判斷λ內容的倉頡碼個數來決定如何去尋找相似字。

- ✧ λ如果小於等於三碼，則只需比對中文倉頡碼檔裡所有中文字的倉頡碼的後幾碼是否與λ相同，如果是，就是相似字；例如：尋找“可”這個字，此字的倉頡碼為“一弓口”，因為未超過三碼，λ為“一弓口”，而“何”的倉頡碼為“人一弓口”，λ的倉頡碼與“何”倉頡碼的後三碼同為“一弓口”，本系統就判斷“何”這個字為相似字。
- ✧ 如果λ（尋找相似字條件）總共四碼或四碼以上時，則需要取λ的後三碼去比對中文倉頡碼檔的所有中文字倉頡碼的後三碼，只要比對相同，就認定為相似字；例如：尋找“河”這個字，此字的倉頡碼為“水一弓口”，因為超過三碼，就取此倉頡碼的後三碼，所以λ為“一弓口”，而“何”的倉頡碼為“人一弓口”，λ的倉頡碼與“何”倉頡碼的後三碼同為“一弓口”，本系統就判斷“何”這個字為相似字。

5.3 量詞

如何判斷那些詞為量詞，其步驟如下。

- ✧ 步驟1：本系統會把句子送到中央研究院中文詞庫小組的中文斷詞系統[2]。
- ✧ 步驟2：本系統會接收中文剖析服務回傳詞性的結果。
- ✧ 步驟3：量詞詞性的標記為“M”，來判定那些詞為量詞。

編輯者如果輸入一句話為“今天天氣很好”，本系統會把句子送到中央研究院中文詞庫小組中文剖析服務，回傳結果為 **今天（N） 天氣（N） 很（ADV） 好（Vi）**，本系統會判斷“今天天氣很好”這

個句子完全沒有任何的量詞，本系統會告知使用者。

與上一個例子比較，假如編輯者輸入一句話為“山上有一座公園”，則回傳結果為 **山（N） 上（N） 有（Vt） 一（DET） 座（M） 公園（N）**，本系統會判斷回傳的結果的詞性，所以“山上有一座公園”這個句子有量詞，本系統會告知使用者。

5.4 中文句子重組

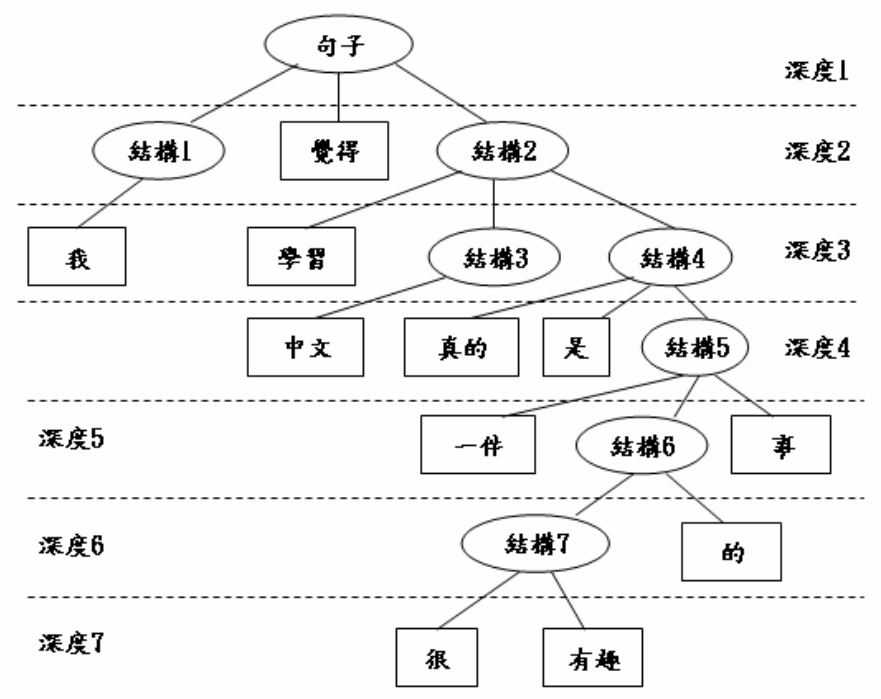

圖 5.7 中文句子的樹狀圖

見圖5.7，我們可以簡單的觀察到一個利用結構樹的深度來調控中文句子重組題目難易度的方式。在深度為3的時候，這一棵結構樹有四個成分，因此連同深度小於3就已經出現的基礎詞彙，我們可以把這一個句子分解為『我』、『覺得』、『學習』、『中文』和『真的是一件很有趣的事』五個成分（constituents）。其中『中文』是『結構3』之下的所有詞彙，而『真的是一件很有趣的事』是『結構4』之下的所有詞彙。如果以深度為4的結構樹做為分解句子的依據，這一例句就可

以被分為『我』、『覺得』、『學習』、『中文』、『真的』、『是』和『一件很有趣的事』共七個成分。

我們一開始會先透過中央研究院－中文句結構樹資料庫檢索系統[1]得到樹狀結構的句子，首先，將樹狀結構的句子儲存於資料庫，學生要測驗句子重組題目時，本系統會提供前後文的提示，讓使用者了解前後文的題意後，進行中文句子的重組。

在編輯題目中文句子重組的題目時，必須先輸入題目前半部內容也就是欲測驗之句子的前文，再來就是輸入題目的內容，並且把要測驗的句子，先以底線取代，把要測驗的句子輸入到欲測驗之句子的空白處，題目後半部內容是欲測驗之句子的後文，按下確定新增的按鈕後，我們會先把句子送到中央研究院－中文句結構樹資料庫檢索系統[1]，該系統產生結構樹後，再儲存到中文句子重組資料庫裡。

學生進入了中文句子重組的考試介面，本系統會亂數從資料庫裡選出一個題目，我們就以上述『我』、『覺得』、『學習』、『中文』和『真的是一件很有趣的事』的例子來說，會把『我』、『覺得』、『學習』、『中文』和『真的是一件很有趣的事』的七個句子元件打亂選項後，使用者必須根據前後文，來決定此題的答案的排列，使用者排完順序之後，等到使用者認為不需要再移動句子元件，接下來按下確定鍵，正確解答會顯示於正確答案的文字方塊裡，然後會產生一個視窗把正確答案與學生作答的結果做比較，讓學生可以即時知道結果。