# 「閩客語典藏計畫」—《英廈辭典》缺字處理程序

中研院資訊所文獻處理實驗室
2007/11/5 丁玟伶製作

「閩客語典藏計畫」爲中央研究院語言學研究所「漢語典藏與典藏架構」的五個子計畫之一，擬收錄《英廈辭典》等辭典文獻建立閩南語、客家語語料庫。文獻首先以外包打字，之後經過助理們二次電腦螢幕逐字校對，未來將放上網路供檢索之用。在文獻數位化過程中，不可免遇到電腦無法呈現的缺字，文中先以構字式或符號來表示，因此這次主要工作即是處理《英廈辭典》的缺字問題，使缺字可以完整呈現在二校電子檔中。

語言所提供了一校完成的《英廈辭典》word 電子文件檔，文中的缺字問題採取兩種方式表示，第一種是可用構字式之字皆以構字式表示，第二種是構字式無法呈現之字形則以「●」表示。因此當我們處理缺字時會分成三個階段，第一階段爲處理構字式之字，第二階段則處理「●」缺字，最後造字完成後，處理 word 文本上之缺字，使之可呈現於電腦。

第一階段、構字式處理程序：

一、缺字分析：

使用程式分析 word 文本，找出文中所有的構字式並加到《英廈辭典》缺字資料庫。

二、利用漢字構形資料庫比對《英廈辭典》缺字資料庫的構字式，找出漢字構形資料庫內未有之缺字。

三、比對原書，確認未有的缺字之字形，並修改資料庫內錯誤構字式：

由 word 文本轉出的缺字資料庫，存在許多錯誤和無法確定之構字式，因此需比對《英廈辭典》原文，一一確定缺字字形並檢查修正。需修改之構字式大致分爲以下三種：

1. 由於漢字構形資料庫目前仍爲 Big5 版本，因此《英廈辭典》資料庫中採用 Unicode 的部件，需以 Big5 部件取代，否則無法判讀。

例如：文本中「毛⿲⺍⺍」的「⺍」，需以 Big5 的「⺍」取代 unicode 部件的「⺮」。

2. 構字式表達錯誤，手動以正確構字式修改。

例如：文本中的「口口⿲正」，正確爲「⿰口⿲正」，而非兩個口字。

3. 無法確定的構字式，在資料庫中的「備註」欄做上記號。

例如：文本中的月⿲放，無法決定偏旁是月或月，因此將保留原構字式，並在資料庫中的「備註」欄寫上可能的組合和問號，交還原單位決議。

四、比對原書，檢查資料庫內已有缺字字形是否正確：（可省略）
以《英廈辭典》原文比對該資料庫裡，漢字構形資料庫已有之字，此步驟只爲確認辭典上之字是否真爲漢字構形資料庫內的字形。

五、把修改過和無法確定的構字式製作問題字清單：（如下圖）
找出修改過與無法決定之問題構字式，以 excel 製作一個方便的檢表，另附上英文與頁數。可使原單位清楚瞭解修改過的部份，和那些構字式需再度確認。


Microsoft Excel - 問題字整理.xls

F70 不確定字形

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | 英文 | 連接符號 | 部件序 | 構字式 | 檔名 | 備註 | 字典頁 |
| 2 | Airy | 2 | ⿰正 | ⿰⿰正 | 英廈辭典A_1校.doc | ⿰⿰改為⿰ | 11 |
| 3 | Asperity | 3 | 屈毛 | 屈⿰毛 | 英廈辭典A_1校.doc | 屈內包 | 23 |
| 4 | Barnacle | 1 | 巾⿰ | 巾⿰⿰ | 英廈辭典B_1校.doc | 虫⿰⿰? | 32 |
| 5 | Bilge | 1 | 月放 | 月⿰放 | 英廈辭典B_1校.doc | 月⿰放? | 38 |

六、比對過原書確認無誤之字，進行造字，放入漢字構形資料庫。

第二階段、「●」缺字處理程序：
一、找出 word 文本中的「●」缺字，放進《英廈辭典》缺字資料庫。

二、比對原書確認「●」缺字的類型：
1.《英廈辭典》原文缺漏不清之字皆以「●」代表。
2. 「●」字無法以構字式表示。

例如： 

3. 「●」可能爲某字，但不確定，放入問題字清單，交還原單位決定。
例如：扌⿰●，比對原書後覺得可能爲「扌⿰舟⿰犬」一字，因此放入問題字清單。

三、經原單位確認過的字形，進行造字，放入漢字構形資料庫。

第三階段、《英廈辭典》word 文本的構字式處理程序：

一、由缺字資料庫整理出問題字修正清單，經過原單位確認後，依清單修改有問題的構字式或缺字。清單內容包含以下三種：

1. 修改 word 文本構字式中的 Unicode 部件以 Big5 部件取代。
2. 修改 word 文本中錯誤的構字式。

例如：「巾⿰昌⿱日」對照原文後可能爲「虫⿰昌⿱日」，經確認後即爲「虫⿰昌⿱日」。

3. 修改 word 文本中確定字形的「●」缺字。

例如：「●●吠」中的●字，對照原文可能爲「⿰犭六右犬⊙」，經確定後●字即爲該字形。

二、造字完成後，製作造字字形與缺字構字式之對照表：

爲了未來修改的方便性，word 文本的缺字仍維持構字式形態，因此造字完成後，製作一份構字式與完整字形的對照表以供原單位參考。

※ 如何在 word 文本上看見完整缺字：

使用word「工具」表單內的「構字式轉成字形」，即可呈現出漢字構形資料庫內已有之完整字形。其他使用word的相關問題，請參考「文獻處理實驗室」的「缺字處理計畫」網頁。