「閩客語典藏計畫」—《台中縣東勢鎮客語故事集》之缺字處理說明

中研院資訊所文獻處理實驗室
中央研究院語言所文獻語料小組
2009/10/08　　丁玟伶

壹、《台中縣東勢鎮客語故事集》的缺字狀況說明：

「閩客語典藏計畫」為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《台中縣東勢鎮客語故事集》等辭典文獻建立閩南語、客家語語料庫。在文獻數位化過程中，若遇到電腦無法呈現的缺字，即以構字式[1]來表示。由語言所助理陳巧欣提供了《台中縣東勢鎮客語故事集》一共七集的word電子文獻檔，檔案是以構字式做為缺字表達方式。例：「跲」字用「⿰足⿰厶企」表示。

貳、《台中縣東勢鎮客語故事集》的缺字問題與解決方式

缺字處理是使用程式分析電子文獻檔中的構字式，與漢字構形資料庫比對，整理出漢字構形資料庫未收之字，再與原文進行校對，確認缺字字形是否正確，最後進行造字。處理過程中發現的問題，說明如下：

一、構字式的問題

構字式的表達有一定的準則，而《台中縣東勢鎮客語故事集》缺字的構字式存在以下幾點問題：

1. 構字式使用 Unicode 部件

由於漢字構形資料庫目前無法支援 Unicode，因此無法判讀構字式中的 Unicode 部件，構字式的 Unicode 部件應改為 Big5 部件。例：檔案中「偃」字的構字式「亻⿰厶厓」的「厓」為 Unicode 部件，編碼

[1] 構字式為中研院資訊所文獻處理實驗室開發，有關構字式的表達與使用方式，請參考〈構字式的處理技巧〉一文。

為 5393，修正為 Big5 編碼 88E8 的「厓」。

2. 未正規化的構字式

構字式的部件與組成有一定的標準，因此有以下的問題時需進行正規化：

（1）構字式誤用相似部件

檔案中有些構字式不小心誤用相似部件，與原文校對後，改為正確部件。例：與原文校對後，檔案中的構字式「月⿰赫」不小心誤用相似部件「月」，應改為正確部件「月」，構字式為「月⿰赫」。

（2）手誤之構字符號

有些構字符號因手誤打錯，需修改為正確的構字符號。例：「⿱雷」比對原文字形後，構字式中的橫連符號應改為直連符號「⿱雷」。

二、異體字問題

《台中縣東勢鎮客語故事集》有些缺字是漢字構形資料庫內字形的異體字，因此在處理缺字時，會有該使用原文之字形，或以漢字構形資料庫已收之字取代的考量出現，因此處理異體字時最主要的方式是，如只是細微的差異，皆以漢字構形資料庫之字形取代，不再另外造字。例：檔案缺字字形為「⿰焦」，而漢字構形資料庫已收有「⿰」字，兩者只有「灬」字為右邊部件或下方部件的細微差異，因此以漢字構形資料庫之「⿰」字取代，不再另外造字。

參、《台中縣東勢鎮客語故事集》缺字處理流程

《台中縣東勢鎮客語故事集》缺字處理的詳細流程如下：

一、缺字分析與構字式修改

（一）使用程式分析《台中縣東勢鎮客語故事集》檔案，找出文中所有的構字式，建立《台中縣東勢鎮客語故事集》缺字資料庫。

（二）利用漢字構形資料庫比對《台中縣東勢鎮客語故事集》缺字資料庫，找出漢字構形資料庫內沒有的構字式，並作以下處理：

1. 修改 Unicode 部件。例：「亻⿰厓」的「厓」修正為「亻⿰厓」。
2. 正規化構字式，需修改的部份如下：

（1）正規化誤用的相似部件。例：「月⿰赫」實則為「月⿰赫」，誤用相似部件「月」，應改為正確部件的「月」。

（2）正規化手誤的構字符號。例：「⺮⿰雷」實則為「⺮⿱雷」，構字符號「⿰」應改為「⿱」。

（三）以漢字構形資料庫再次比對修改後的構字式，確認漢字構形資料庫未收錄的缺字將與原文進行校對。

二、缺字校對

《台中縣東勢鎮客語故事集》資料庫中，未收在漢字構形資料庫的缺字構字式與原文一一進行校對，需修改的部份如下：

1. 修改資料庫缺字的異體字字形，使用漢字構形資料庫已收字形。例：「⿰焦」以漢字構形資料庫已收之「⿱」字的「⿱⿰」取代。

三、修改 word 檔

修改 word 文本上的缺字構字式：

（一） 以《台中縣東勢鎮客語故事集》資料庫修正之結果，修改 word 檔案中構字式的 Unicode 部件為 Big5，並把改過之部件製作成「修正 Unicode 部件」清單，列出修改過的 Unicode 部件，修改過之字頻共計 1065 字，完整清單請參考附件二，欄位說明如下：

1. 原構字式：缺字在檔案中的構字式或字形。
2. Unicode 部件：缺字在檔案中使用的 Unicode 部件。
3. 修正構字式：修改後的構字式。

4. 頻次：缺字出現的次數。

5. 檔名：缺字所在的檔案名稱。

（二） 以《台中縣東勢鎮客語故事集》資料庫修正之結果，正規化word 檔案中未正規化之構字式、修改異體字形。將改過的構字式列成「修正構字式」清單，並在備註欄寫上修改原因，以供語言所參考。修改過之字頻共計 70 字，完整清單請參考附件三，欄位說明如下：

1. 原構字式：缺字在檔案中原來的構字式或字形。

2. 修正構字式：修改後的構字式。

3. 頻次：缺字出現的次數。

4. 檔名：缺字所在的檔案名稱。

5. 備註凡例：

（1）異體字，使用漢字構形資料庫之字形：缺字與漢字構形資料庫已收之字形是異體字，只有細微的差異，皆以漢字構形資料庫之字形取代，不再另外造字。修改過之字頻共計 12 字。例：原文缺字字形為「羙⿰焦」，而漢字構形資料庫已收有「鰰」字，兩者只有「灬」字為右邊部件或下方部件的細微差異，因此以漢字構形資料庫之「鰰」字取代，不再另外造字。

（2）正規化構字式：檔案使用未正規化的構字式表達缺字時，需修改為漢字構形資料庫正規化的構字式。修改過之字頻共計 58 字。例：「月⿰赫」正規化成「月⿰赫」,「月」改為正確部件「月」。

四、《台中縣東勢鎮客語故事集》缺字進行造字

確認《台中縣東勢鎮客語故事集》缺字資料庫無誤後，與漢字構形資料庫比對，未收的缺字將進行造字，完成後收至漢字構形資料庫。《台中縣東勢鎮客語故事集》缺字共計 60 字，其中字漢字構形資

料庫已收錄 38 字，待造字共 22 字，因其中 1 字重複，實共計 21 字，將進行造字。為了未來修改的方便性，檔案內缺字仍維持構字式形態，並製作一份包含構字式的造字記錄表供語言所參考。請參考《台中縣東勢鎮客語故事集》缺字造字記錄。

五、確認《台中縣東勢鎮客語故事集》檔案的缺字處理完成

1. 執行word「工具」中的「構字式轉字形」[2]，確認檔案修正後的構字式，可否轉為字形。
2. 無法轉換字形之字，需確認構字式是否正確，再比對漢字構形資料庫，未收字將進行造字。

[2] 如何在 word 文本上看見缺字字形，請參考附錄一

附件一、《台中縣東勢鎮客語故事集》修正Unicode部件表

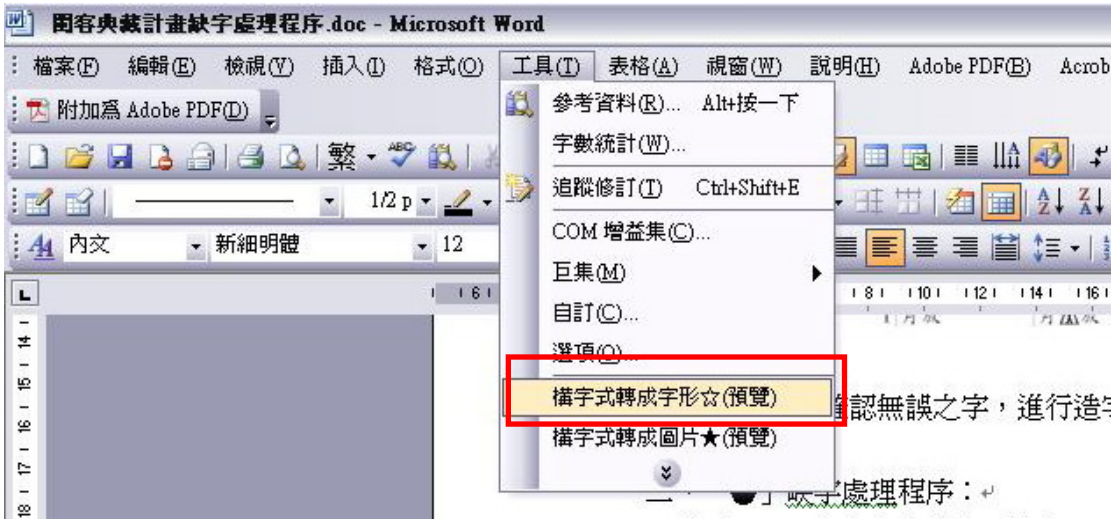
原構字式	Unicode 部件	修正構字式	頻次	檔名
亻⿰厓	厓	亻⿰厓	937	台中縣東勢鎮客語故事集(二)
犀⿰刂	刂	犀⿰刂	113	台中縣東勢鎮客語故事集(二)
發⿰刂	刂	發⿰刂	7	台中縣東勢鎮客語故事集(二)
荅⿰刂	刂	荅⿰刂	5	台中縣東勢鎮客語故事集(四)
目⿰厶	厶	目⿰厶	3	台中縣東勢鎮客語故事集(六)

附件二、《台中縣東勢鎮客語故事集》修正構字式清單

原構字式	修正構字式	頻次	檔名	備註
茉⿰焦	難⿱灬	12	台中縣東勢鎮客語故事集(二)	異體字,使用漢字構形資料庫之字形。
月⿰赫	月⿰赫	15	台中縣東勢鎮客語故事集(五)	正規化構字式，「月」改為「月」。
子⿰牙	子⿰牙	17	台中縣東勢鎮客語故事集(五)	正規化構字式，「子」改為「子」。
竹⿰雷	竹⿱雷	1	台中縣東勢鎮客語故事集(六)	正規化構字式，「⿰」改為「⿱」。
月⿰夌	月⿰夌	25	台中縣東勢鎮客語故事集(四)	正規化構字式，「月」改為「月」。

附錄一：如何在 word 檔案上看見缺字字形

一、安裝漢字構形資料庫後，點選word「工具」表單內的「構字式轉成字形」(如下圖)，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。



二、更新版本或下載新增字檔

如使用 word「工具」表單內的「構字式轉成字形」，卻無法看到新造的字形，有以下兩種解決方式：

1. 使用的漢字構形資料庫為舊版，無法看到新造字形，可以到「缺字處理計畫」網頁更新至最新版的漢字構形資料庫，目前最新的為 2.53 版。
2. 如已是最新的版本，卻仍無法看見新造之字，可至「缺字處理計畫」網頁的「行事曆」下載更新的造字檔。例：下圖為 2.51 版的更新字檔，需下載「檔名」欄內的檔案，更新漢字構形資料庫。

文獻處理實驗室
Home of Chinese Document Processing Lab

下載漢字構形資料庫2.51版更新檔案

請參閱下載說明

名稱	檔名	檔案大小	更新原因	更新日期
楷書構形資料庫	cdphanzi.mdb	20.1M	增加95個新字	2008年2月12日
標楷體外字集一	hzcdp01k.ttf	5.0M	修改1個字形	2008年2月12日
細明體外字集一	hzcdp01m.ttf	3.9M	修改1個字形	2008年2月12日
標楷體外字集二	hzcdp02k.ttf	5.8M	修改1個字形	2008年2月1日
細明體外字集二	hzcdp02m.ttf	4.5M	修改1個字形	2008年2月1日
標楷體外字集六	hzcdp06k.ttf	6.8M	修改1個字形	2008年2月1日
細明體外字集六	hzcdp06m.ttf	5.0M	修改1個字形	2008年2月1日
標楷體外字集七	hzcdp07k.ttf	6.9M	修改1個字形	2008年2月1日
細明體外字集七	hzcdp07m.ttf	5.4M	修改1個字形	2008年2月1日
標楷體外字集八	hzcdp08k.ttf	5.7M	修改2個字形	2008年1月17日
細明體外字集八	hzcdp08m.ttf	4.3M	修改2個字形	2008年1月17日
標楷體外字集九	hzcdp09k.ttf	646K	增加95個新字	2008年2月12日
細明體外字集九	hzcdp09m.ttf	520K	增加95個新字	2008年2月12日
漢字構形資料庫2.51	cdphanzi.exe	904K	修正程式錯誤	2008年2月12日