《大唐三藏取經詩話》(標記語料)舊版造字轉碼說明

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
2012/08/22　丁玟伶

轉碼主要工作是把檔案中的舊版造字轉換成 Windows XP 能支援的 Unicode 字形，Unicode 目前共收錄漢字 70194 個字，而 XP 只能支援 20902 個字（詳如表一），不支援之字將以構字式表達。例：造字編號 2322 的「鏻」字，Unicode 編碼是 4948，由於 XP 並不支援，仍需使用構字式「金⿰乏斧」。

表一、Unicode 的字數及編碼區段

Unicode	新增字數	新增編碼區段	總字數	WinXP
1.1 版	20902	4E00-9FFF	20902	支援
3.0 版	6582	3400-4DFF	27484	不支援
3.1 版	42710	20000-2A6D6	70194	不支援

一、舊版造字轉碼分析：

《大唐三藏取經詩話》(標記語料)使用舊版造字 16 個，字頻 87 次，這 16 個造字中，13 個可轉成 Windows XP 能顯示的字，字頻 82 次；另外 3 個字必須轉成構字式，字頻 5 次。

轉碼完成製作轉碼分析表，請參考附件一《大唐三藏取經詩話》(標記語料)轉碼分析表，欄位說明如下：

（一）編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

（二）造字：舊版造字。

（三）頻次：舊版造字在文件的出現次數。

（四）Big5：造字的 Big5 碼。

（五）Unicode：造字所對應的 Unicode 碼。

（六）WinXP：造字在 Windows XP 的對應字形。

（七）構字式：Windows XP 無法對應字形改用構字式。

（八）備註凡例：備註欄中記錄轉碼後字形及修改原因，凡例如下：

1. Unicode 字型呈現差異：Unicode 字型與舊版造字有些微差異，但只是字體風格差異，實際上仍為同一個字，因此仍取 Unicode 字型。如編號 5932 的「㤅」字，Unicode 字型呈現為「㤅」，實際上仍為同一字。

附件一、《大唐三藏取經詩話》(標記語料)轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
297	𠙵	2	FBED			𠙵	
770	扌	1	FEEF	624C		扌	
1024	虵	16	8FB3	8675	虵		
1491	擷	2	92AF	6527	擷		
2322	鏊	2	97DD	4948		金𠙵荐	
2325	瞷	1	97E0	77A4	瞷		
3801	剗	1	8160	5257	剗		
3891	慙	1	81DC	6159	慙		
3996	踈	3	82A8	758E	踈		
4193	鈎	2	83D0	920E	鈎		
4240	鬅	1	8440	9AF4	鬅		
4242	鬬	1	8442	9B2D	鬬		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4460	勑	4	85A1	52D1	勑		
4957	嵒	2	88BB	5D53	嵒		
5392	悤	4	8B75	60A4	悤		Unicode 字型呈現差異。
5860	无	44	C6D3	65E0	无		