

《歲時習俗資料彙編》舊版造字轉碼說明

中研院資訊所文獻處理實驗室
中研院史語所漢籍電子文獻工作小組
2007/12/10 黃榮順 製作

一、《歲時習俗資料彙編》一書(歲時習俗資料彙編.xml)使用舊版造字 44 個，字頻 2461 次，詳如附件一。這 44 個造字中，27 個可轉成 Windows XP 可顯示的字，字頻 2194 次；另外 17 個字必須轉成構字式，字頻 267 次。

二、附件一的造字分析表說明如下：

甲、編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

乙、造字：舊版造字

丙、字頻(txt)：造字在「.txt」文件的出現次數

丁、字頻(xml)：造字在「.xml」文件的出現次數

戊、Big5：造字的 Big5 碼

己、Unicode：造字所對應的 Unicode 碼

庚、WinXP：造字在 Windows XP 的對應字形

辛、構字式：Windows XP 若無對應字形，則改採用構字式

壬、備註凡例：

1、錯字：與原書字形不符，屬於意義不同的字，則用正確的字取代原造字，如編號 5849 的「|  口」字，原字應為「𠂔」。

2、構字部件，不轉碼：txt 檔使用的造字到 xml 檔中已被修改完畢，但 xml 檔中其他的構字式碰巧使用了該字碼所對應的構字部件，故雖有字頻差異，但不需手動處理。本書 44 個造字中，有 42 個都是這個情況。如編號 4176 的「輒」，對應的構字部件為「灾」。比對原書後發現，txt 檔的 152 例都被修改為「輒」，xml 檔的 1 例是另出的，與 txt 檔的 152 個詞例都不相同，採用的則是構字部件「灾」。由於 txt 檔的 152 例已都正確修改，所以即使頻次不同，也無須進行手動取代，只要在資料庫中註記即可。除非 xml 檔中所使用的構字部件需要以其他字形取代時，才以手動方式處理，並紀錄在「手動取代表」（詳參第四點說明）中。如編號 4933 的造字「崙」，該字字碼所對應的構字部件為「瓜」，實即「瓜」字，經查原書，並無專取其形的用例，為廣查詢利用，本書皆

以「瓜」字取代。

三、 Unicode 目前收錄的漢字總數為 70194，分屬於三個不同區段，詳如表一。目前 Windows XP 只支援 CJK 認同表意文字區的 20902 個字，內碼為 4E00-9FFF。所以造字編號 5335 的「𧈧」字，Unicode 編碼為 459F，由於 Windows XP 並不支援，仍須改用構字式「亡𧈧虫」。

表一、Unicode 的字數及編碼區段

Unicode	字集子集合	新增字數	新增編碼區段	總字數	WinXP
1.1 版	CJK 認同表意文字區	20902	4E00-9FFF	20902	支援
3.0 版	CJK 認同表意文字擴充 A 區	6582	3400-4DFF	27484	不支援
3.1 版	CJK 認同表意文字擴充 B 區	42710	20000-2A6D6	70194	不支援

四、 造字分析表「備註」欄標示「手動取代」者，詳細取代內容可參考附件二的手動取代表。附件二的手動取代表說明如下：

- 甲、 編號：Big5 造字空間為 6217 個，編號由 1 到 6217。
- 乙、 部件：漢字構形資料庫收錄之構字部件。
- 丙、 取代：實際使用字形與部件欄不同時，於此欄列出。
- 丁、 檔案原文摘錄：摘錄檔案(歲時習俗資料彙編.xml)中包含該造字的文句段落，其中紅字底線的部分，為需要手動取代的原文。
- 戊、 手動取代結果：變更過後的檔案內容，其中藍字底線的部分，為修改後的結果。

附件一、「歲時習俗資料彙編.xml」造字分析表

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
4176	輒	152	1	83BF	8F19	輒		灾，構字部件，不轉碼。
4223	顯	1	45	83EE	984B	顯		葡，構字部件，不轉碼。

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
4262		94	2	8456			麥△𠂇	昏，構字部件，不轉碼。
4278	亘	43	1	8466	4E98	亘		𠂇，構字部件，不轉碼。
4350	侶	2	14	84D0	5058	侶		𧈧，構字部件，不轉碼。
4512		1	2	85D5			口△𠂇	火，構字部件，不轉碼。
4552	趣	12	118	85FD	27F08		走△𧈧	𠂇，構字部件，不轉碼。
4888	荔	5	2	8854	8318	荔		𠂇，構字部件，不轉碼。
4933	崙	10	178	88A3	5CDD	崙		瓜，構字部件，不轉碼， 以「瓜」字手動取代。
4964	媛	1	9	88C2	59AD	媛		尔，構字部件，不轉碼。
5289	慤	9	1	8ACD	6164	慤		𠂇，構字部件，不轉碼。
5335	𧈧	5	22	8AFB	459F		亡△𧈧	𧈧，構字部件，不轉碼。
5336	睹	2	2	8AFC	668F	睹		𧈧，構字部件，不轉碼。
5338	暎	68	89	8AFE	668E	暎		𠂇，構字部件，不轉碼。
5339	𧈧	5	16	8B40	59E7	𧈧		𠂇，構字部件，不轉碼。
5341		1	3	8B42			𠂇△嬖	𠂇，構字部件，不轉碼。
5347	憾	2	26	8B48	617D	憾		𠂇，構字部件，不轉碼。

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
5353	役	3	1	8B4E	4F07	役		丩，構字部件，不轉碼。
5355	𡗗	1	18	8B50	20AE4		𡗗	广，構字部件，不轉碼。
5378	巛	3	2	8B67	5DD9	巛		亼，構字部件，不轉碼。
5395	規	3	5	8B78	4093		矢△見	卜，構字部件，不轉碼。
5418	邨	3	4	8BB1	28697		冉△卩	衤，構字部件，不轉碼。
5427		4	1	8BBA			𠂇△𠂇△𠂇△𠂇△伐	匕，構字部件，不轉碼。
5464	煥	19	3	8BDF	7157	煥		戔，構字部件，不轉碼。
5470	棟	2	1	8BE5	68C5	棟		畝，構字部件，不轉碼。
5471	契	16	1	8BE6	22346		契	巳，構字部件，不轉碼。
5479	蠲	1	11	8BEE	4D77		圭△黽	冫，構字部件，不轉碼。
5515		11	5	8C53			車△丸	亠，構字部件，不轉碼。
5541	蠹	30	1	8C6D	8827	蠹		果，構字部件，不轉碼。
5558	龕	5	4	8C7E	9E84	龕		冗，構字部件，不轉碼。
5568	开	12	4	8CAA	5F00	开		冫，構字部件，不轉碼。
5574	愔	2	1	8CB0	612D	愔		萑，構字部件，不轉碼。
5583	殮	24	1	8CB9	98F1	殮		殮，構字部件，不轉碼。

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
5585	寓	7	68	8CBB	3762		宀△禹	鳥，構字部件，不轉碼。
5588	恠	2	3	8CBE	6061	恠		彡，構字部件，不轉碼。
5589	廚	2	1	8CBF	2228D		形广豆寸口	畀，構字部件，不轉碼。
5631	浪	1	16	8CE9	6CFF	浪		勹，構字部件，不轉碼。
5633		3	1	8CEB			日△台	左，構字部件，不轉碼。
5721	頓	3	2	8DA6	9814	頓		疋，構字部件，不轉碼。
5752	摠	14	2	8DC5	63D4	摠		疋，構字部件，不轉碼。
5797	灑	36	1768	8DF2	7054	灑		△，構字部件，不轉碼。
5799		4	4	8DF4			形灑炎口	口，構字部件，不轉碼。
5849		1	1	C6C8			丨△コ	尸，錯字。
5859	支	1	1	C6D2	6534	支		

附件二、「歲時習俗資料彙編.xml」手動取代表

編號	部件	取代	檔案原文摘錄	手動取代結果
4933	瓜	瓜	瓜 (全部取代)	瓜 (全部取代)