

# 古漢語轉碼程序說明—以《左傳》為例

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  
2008/03/27 丁致伶

「中央研究院古漢語語料庫」建構始於一九九〇年，是應漢語史研究需求而建構的語料庫。這個語料庫又依是否經過斷詞及加標詞類而分成兩類，即未加標的素語料庫以及有標注的標記語料庫。目前素語料庫所蒐集的語料已涵蓋上古漢語（先秦至西漢）、中古漢語（東漢魏晉南北朝）、近代漢語（唐五代以後）大部分的重要語料，約 403 本書；在標記語料庫方面，已完成部份上古漢語及近代漢語的語料標注工作，約 40 本書，總計 7 千多萬字，將視結果逐步提供線上檢索。

古漢語語料庫的中文字碼是採用Big5（收錄 13,051 字），並使用瀚典全文檢索系統造字檔<sup>1</sup>（以下簡稱為舊版造字）以解決缺字問題，處理缺字的方式是在Big5 的使用者造字區（FA40~C8F1）增補新字，由於Big5 造字區只能容納 6,217 字，不敷使用，且在電腦中文編碼環境整個轉換為Unicode之下，舊版造字於線上檢索不易，在參考史語所新版漢籍全文資料庫轉為Unicode之經驗後，古漢語語料庫決定也改用Unicode（收錄 20,902 字的版本），將原為Big5 編碼的舊版造字轉為Unicode，但Unicode仍然無法解決大量低頻缺字的問題，因此另外採用構字式<sup>2</sup>來表達Unicode缺字。

這次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 70,194 個字，而

---

<sup>1</sup> 瀚典全文檢索系統造字檔由中研院計算中心開發，民國八十四年正式上線，主要運用於中央研究院舊版的[漢籍電子文獻資料庫](#)，故又稱為舊版漢籍造字，或簡稱為舊版造字。

<sup>2</sup> 構字式由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。

(<http://www.sinica.edu.tw/~cdp/service/documents/T960419.pdf>)

Windows XP 只能支援其中的 20,902 個字（詳如表一），不支援之字將以構字式表達。例：舊版造字編號 978 的「戴」字，Unicode 編碼是 4CD2，由於 XP 並不支援，仍需使用構字式「戔△鳥」表達。

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

語言所需要轉碼的電子文獻檔案共有兩種：分別為素語料檔和標記語料檔，皆以舊版造字來處理缺字，雖然檔案中有構字式，但是與漢字構形資料庫的構字式編碼不同，仍屬於舊版造字，需要轉碼。

古漢語語料庫在轉碼的程序上分為兩個階段：第一階段處理舊版造字的轉碼，將舊版造字轉為 Unicode，若無對應之 Unicode，則轉為構字式；第二階段處理構字式部分，確認構字式的正確性，並將未收錄的缺字補入漢字構形資料庫。

## 一、古漢語缺字轉碼問題與解決方式

缺字轉碼的原理應是比對新舊版本造字的碼區對應，整理出一張字碼對應表，然後由程式依照此對應表進行轉碼即可；此次古漢語缺字轉碼是將舊版造字對應到 Unicode 或構字式。然而古漢語缺字轉碼並非如此單純的電腦作業，在轉碼的過程中仍有些需人工處理的問題，說明如下：

### （一）異體字問題

古籍文獻數位化的過程中，異體字的採用標準一直為眾人所討論，要如何在遵照文本原形和方便使用者查詢理解之間取得平衡，是異體字採用標準的主要爭議點。由於古漢語文獻中有不少缺字是異體

字，因此在處理缺字時，會有該採用原書異體字，或以標準字（或稱「正體字」）取代的考量出現，在與語言所討論後，為了使用者查詢和使用的方便，在處理異體字時最主要的方式是以標準字取代，除非是專有名詞，如：人名、地名等或特殊情形，則依原書異體字表示。以《左傳》為例：校對原書使用中華書局的《春秋左傳注（修訂本）》，雖然原書與檔案皆使用舊版造字的「衛」字，但「衛」是「衛」的異體字，因此以標準字「衛」取代。

## （二）校對問題

轉碼過程中，每個舊版造字至少需與原書比對確認一次，檢查使用的舊版造字是否正確，若發現錯字，則處理方式如下：

1. **錯字，以程式全部取代**：以原書為準，與書上字形不同，並查詢教育部異體字字典確認非異體字，皆歸類為錯字，以程式取代為正確字形。以《左傳》為例：檔案中使用舊版造字的「菟」字，而原書（《春秋左傳注（修訂本）》）用字為「菟」，查詢教育部異體字字典確認「菟」並非「菟」之異體字，因此歸為錯字，以程式取代為正確字形「菟」。
2. **部份使用不同字形，手動修改**：檔案中為同一舊版造字，校對時卻發現原書使用兩個以上不同的字形，可能部份舊版造字為錯字或原書使用兩種異體字形，導致同一舊版造字轉碼時對應到兩個以上的字形；將一部份手動修改，剩餘的頻次則用程式取代成對應的 Unicode 或構字式。以《左傳》為例：檔案使用舊版造字的「髡」字 9 次，原書（《春秋左傳注（修訂本）》）中卻分別寫成「髡」和「髡」兩種異體字形，依書將代表人名的 2 字手動修改為「髡」，其他 7 字皆轉為標準字「髡」。

## 二、轉碼資料庫簡介

缺字轉碼是比對新舊版本造字的碼區對應，整理出一張字碼對應表，然後由程式依照此對應表進行轉碼，因此製作對應表時需使用到

Microsoft Access 資料庫，為方便理解，將先介紹支援轉碼並記錄轉碼資料的資料庫。由於語言所的古漢語與史語所的舊版漢籍電子文獻轉碼皆使用舊版造字檔，為了作業方便，沿用同一轉碼資料庫，轉碼資料庫中有三個重要的表格：造字頻次表、造字表、漢籍用字表。造字頻次表主要記錄舊版造字在檔案中出現的頻次；造字表則記錄了舊版造字的基本資料，以及所對應的 Unicode 和構字式；漢籍用字表則是用以記錄轉碼時所遇到的問題和解決方式。針對上述三個表格，進一步說明如下：

(一)「造字表」：Big5 造字空間為 6217 個，編號由 1 到 6217，造字都存在這 6217 個碼區裏。造字表記錄舊版造字之字形、字碼、相對應的構字式、Unicode，是舊版造字檔的基本資料表，也是程式執行轉碼時的依據；在一般狀況下，轉碼程式會根據造字表的資料進行轉碼，將舊版造字轉成對應的 Unicode 字形；若無對應之 Unicode 字形，才會轉成構字式。另外，造字表中有一個「確認」欄，是因每個舊版造字至少比對原書一次，確認造字字形無誤，因此當比對原書字形無誤時，便要在確認欄填上 1，表示已確認字形無誤，如有誤則填上 2。

(二)「造字頻次表」：當程式分析電子文獻時，會自動計算每個舊版造字在檔案中出現的頻次，分別記錄在這個資料表中。

(三)「漢籍用字表」：校對過程中發現的問題和處理方式記錄在此。

此表以一個轉碼檔案為單位，每個檔案各有兩個欄位：

1. 字形記錄欄：當電子文獻檔案所使用的舊版造字字形與原書不同，或有其他特殊情形（如：異體字、錯字……）時，在此記錄轉碼後希望呈現的字形，轉碼程式會優先將舊版造字轉為此處所記錄的字形。
2. 修訂代號欄：處理人員將所遇到的特殊情形，以代號記錄於此，以便日後檢視。修訂情形如下：

- (1) 錯字，以程式全部取代：以正確的 Unicode 字形或構字式把錯誤的字形用程式全部取代，修訂代號為 1。
- (2) 異體字問題：確認在檔案中並非作專詞或特殊情形使用，即可將依標準字取代檔案中的異體字，修訂代號為 2。
- (3) 部份使用不同字形，手動修改：同一個舊版造字，但書上卻有兩個以上不同的字形，可能部份舊版造字為錯字或原書使用兩種異體字形，導致同一舊版造字轉碼時對應到兩個以上的字形；將一部份手動修改，剩餘的頻次則用程式取代成對應的 Unicode 或構字式，修訂代號為 3。

### 三、舊版造字轉碼流程

此階段的轉碼目的是將舊版造字轉為 Unicode，若無對應之 Unicode，則轉為構字式。以《左傳》為例說明轉碼程序，詳細處理方式請參考附錄一〈《左傳》舊版造字轉碼說明〉。

- (一) 以 NotePad 開啟《左傳》電子檔，檢查該檔編碼是否為 ANSI，若否，則另存檔案為 ANSI 編碼，再以程式計算每個舊版造字出現的字頻，結果會自動記錄於造字頻次表中。《左傳》電子檔共有 105 個舊版造字，頻次 1,995 次。
- (二) 分析完成後把《左傳》電子檔另存為 Unicode 編碼檔，以免轉碼後無法判讀 Unicode 字形或構字式。
- (三) 將 105 個舊版造字一一比對原書，依核對字形之結果處理如下：
  1. 造字非異體字且符合原書字形，則在「造字表」的「確認欄」填上 1。例：舊版造字「𧈧」字分析後頻次為 3，表示該字在檔案中出現 3 次，比對原書，發現 3 次皆符合原書字形，因此在「造字表」的「確認欄」填上 1，表示此字正確。
  2. 舊版造字字形錯誤或是為需轉換成標準字的異體字，在「造字表」的「確認欄」填上 2，依以下幾種情況有不同的處理方式：

- (1) 造字不符合原書字形，且查詢教育部異體字字典確認非異體字，即為錯字，則在「漢籍用字表」的「字形記錄欄」寫上正確的 Unicode 字形或構字式，並在「修訂代號欄」填寫修訂代號 1。  
例：舊版造字的「菟」字分析後頻次為 1，表示該字在檔案中出現 1 次，比對原書字形為「菟」，查詢教育部異體字字典確認「菟」並非「菟」之異體字，因此歸為錯字，將「菟」填至「漢籍用字表」的「字形記錄欄」，並在「修訂代號欄」填上 1。
- (2) 造字為異體字，且非作專詞或特殊情形使用，以標準字取代，在漢籍用字表填上標準字形，並在「修訂代號欄」填寫修訂代號 2。  
例：舊版造字的「衛」字分析後頻次為 777，表示該字在檔案中出現 777 次，而「衛」是「衛」的異體字，因此在「漢籍用字表」填上標準字形「衛」，並在「修訂代號欄」填上 2。
- (3) 同一個舊版造字，但書上卻使用兩個以上不同的字形，可能部份舊版造字為錯字或原書使用兩種異體字形，導致檔案也需改為兩個以上字形時，則在「漢籍用字表」的「修訂代號欄」填寫修訂代號 3，一部份用程式取代為對應的 Unicode 或構字式，一部分在校對時手動修改電子檔，手動修改皆會記錄於「手動取代表」，以便日後檢視。例：舊版造字「髡」分析後頻次為 9，表示該字在檔案中出現 9 次，比對原書後卻有「髡」和「髡」兩種異體字形，「髡」為現今通用的標準字，因此在「漢籍用字表」填上標準字形「髡」，檔案中 7 字用程式轉為「髡」，而人名的 2 字則在校對時依原書手動修改檔案字形為「髡」，並在「修訂代號欄」填上 3。手動修改部份會另存在「手動取代表」word 檔上，以便日後檢視。
- (四) 以程式執行轉碼，將舊版造字轉為對應的 Unicode 或構字式；轉碼時，程式將優先採用「漢籍用字表」的「字形記錄欄」之字形，若未記錄其他字形，則採用「造字表」所記錄之 Unicode

或構字式。轉碼完成檔案為「左傳.xml」。

(五) 檢查構字式，將轉碼完成的「左傳.xml」檔另存為「左傳.doc」檔，在 Word 裏執行構字式轉字形的功能，得到轉為字形後的「左傳☆.doc」檔。確認構字式的正確性，並檢查 xml 檔內的構字式字形，是否皆已收錄於漢字構形資料庫中，亦即令 xml 檔內無缺字。未成功轉換為字形的構字式可能原因如下：

1. 缺字，將進行造字。《左傳》中並沒有未收錄於漢字構形資料庫的缺字，因此不需造字。
2. 構字式表達錯誤，必須人工一一檢視，再回到轉碼完成的 xml 檔案修改。例：《左傳》檔案中使用舊版造字編碼的構字式「𠄎 Δ 殳」表示缺字，但原書字形為「殷」，因此將「𠄎」字轉為「士𠄎月」，而「殳」轉為「殳」，因此該構字式轉碼後變成「士𠄎月 Δ 殳」，需回到轉碼完成的「左傳.xml」修改構字式為「𠄎士月殳□」。

(六) 除了舊版造字之外，轉碼完成的 xml 檔中還可能有錯字或以「●」、「？」……等符號代表的缺字，由於校對錯字本不屬於轉碼工作範圍，除非偶然發現才會修改。而「●」、「？」等符號將會在完成缺字轉碼後檢查 xml 檔，如 xml 檔中有這些符號且與原書不符時，則將在 xml 檔手動修改為符合原書字形的 Unicode 或構字式，以上手動修改的字形皆會填入手動取代表，做為日後檢視或修正的依據。

#### 四、整理轉碼文件

完成轉碼後，交回語言所的檔案共有三個：

- (一) 完成轉碼的 xml 檔。其中包含 WinXP 支援的 Unicode 字形，以及用構字式表達的造字。
- (二) 轉碼程序說明文件。包括以下幾個部份：
  1. 轉碼程序說明：說明轉碼的原因和轉碼處理流程。
  2. 轉碼分析表與說明：記錄每個舊版造字轉碼的結果，並於備

註欄記錄人工處理的情形，請參考附錄一〈左傳轉碼說明〉。

3. 手動取代表與說明：記錄手動取代的詳細內容，做為日後檢視或修正的依據，請參考附錄一〈左傳轉碼說明〉。

(三) 統計舊版造字轉碼表。統計每個檔案的舊版造字轉碼字數，請參考附錄二〈古漢語轉碼成果統計〉。

## 附錄一、《左傳》舊版造字轉碼說明

中研院資訊所文獻處理實驗室  
中央研究院語言所語言典藏計畫  
2008/2/26 丁致伶

《左傳》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 70,194 個字，而 XP 只能支援 20,902 個字（詳如表一），不支援之字將以構字式表達。例：造字編號 978 的「戡」字，Unicode 編碼是 4CD2，由於 XP 並不支援，仍需使用構字式「戡△鳥」。

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：

《左傳》使用舊版造字 105 個，字頻 1995 次，這 105 個造字中，94 個可轉成 Windows XP 能顯示的字，字頻 1974 次；另外 11 個字必須轉成構字式，字頻 21 次。

轉碼完成製作轉碼分析表，請參考附件一《左傳》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. 異體字問題：為了使用者查詢和使用的方便，在處理異體字時最主要的方式是以標準字取代，除非是專有名詞，如：人名、地名等特殊情形。例：編號 4134 的「衛」字，是「衛」的異體字，以標準字「衛」取代。
  2. 部份使用不同字形，手動修改：檔案中為同一舊版造字，校對時卻發現原書使用兩個以上不同的字形，可能部份舊版造字為錯字或原書使用兩種異體字形，導致同一舊版造字轉碼時對應到兩個以上的字形；將一部份手動修改，剩餘的頻次則用程式取代成對應的 Unicode 或構字式。例：舊版造字編號 4237 的「髡」字頻次為 9，原書卻分別寫成「髡」和「髡」兩種異體字字形，依書將代表人名的 2 字手動修改為「髡」，其他 7 字皆轉為標準字「髡」。手動修改部份請參考附件二：《左傳》手動取代表。
  3. 錯字，以程式全部取代：與原書字形不符，並查詢教育部異體字字典確認非異體字後，皆歸類為錯字，以程式取代為正確字形。例：檔案中使用舊版造字編號 4353 的「菟」字，而原書使用之字為「菟」，查詢教育部異體字字典確認「菟」並非「菟」之異體字，因此歸為錯字，以程式取代為正確字形「菟」。

附件一、《左傳》轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
297	𠄎	2	FBED			𠄎	
299	𠄐	1	FBEF			𠄐	
300	𠄑	1	FBF0			𠄑	

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
436	𠂔	1	FCDB		𠂔		士𠂔月，錯字，以程式全部取代。
626	𠂔	1	FDFC		𠂔		𠂔，錯字，以程式全部取代。
763	𠂔	1	FEE8		𠂔		
793	𠂔	1	8E47	97E4	𠂔		
978	𠂔	1	8F63	4CD2		𠂔△鳥	
980	𠂔	12	8F65	72AB	𠂔		
982	𠂔	1	8F67	80F7	𠂔		胸，異體字問題。
988	𠂔	8	8F6D	7F8B	𠂔		
1062		1	8FD9			馬△𠂔	駟，異體字問題。
1077	𠂔	1	8FE8	7A3E	𠂔		稟，錯字，以程式全部取代。
1565	𠂔	9	92F9	7DDC	𠂔		
1756	𠂔	1	945C	84D8	𠂔		蓑，異體字問題。
1857	𠂔	3	94E3	86A0	𠂔		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
1885	逢	4	9540	882D	逢		
1893	袞	1	9548	886E	袞		袞，異體字問題。
1937	詢	2	9574	8A3D	詢		
2043	踳	2	9641	8E01	踳		
2449	雒	3	98BF			名 隹	雒，錯字，以程式全部取代。
2478	鞏	6	98DC	978C	鞏		
2479	鞞	1	98DD	979B	鞞		
2651	鸛	9	99EC	9E1C	鸛		
2836	三	5	9B49		三		
2837	三	14	9B4A		三		
2878	叟	3	9B73	5081	叟		
2906	竽	1	9BB1	659A	竽		竽，異體字問題。
3068	三	8	9CB6		三		
3069	三	10	9CB7		三		
3071	三	10	9CB9		三		
3072	三	9	9CBA		三		
3073	三	7	9CBB		三		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3074	彡	17	9CBC		彡		
3770	个	4	8141	4E2A	个		
3780	儻	5	814B	5101	儻		
3787	冢	6	8152	51A2	冢		
3791	况	5	8156	51B5	况		况，異體字問題。
3834	嗥	1	81A3	5637	嗥		
3846	壻	5	81AF	58FB	壻		
3883	徧	8	81D4	5FA7	徧		
3891	慙	3	81DC	6159	慙		
3921	皙	23	81FA	6673	皙		
3929	黎	2	8243	68C3	黎		
3930	憵	2	8244	6901	憵		
3935	槩	7	8249	69E9	槩		
3949	泣	29	8257	6D96	泣		
3974	牀	5	8270	7240	牀		
3976	犁	12	8272	7282	犁		部份使用不同字形「犁」，手動修改。
3987	穀	3	827D	7474	穀		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4010	皐	3	82B6	7690	皐		皐，異體字問題。 部份使用不同字形「皐」，手動修改。
4011	皦	8	82B7	76A1	皦		皦，異體字問題。
4030	秆	1	82CA	79C6	秆		
4038	竈	17	82D2	7AC8	竈		
4080	羣	96	82FC	7FA3	羣		
4085	耻	1	8342	803B	耻		
4091	皐	18	8348	81EF	皐		部份使用不同字形「皐」，手動修改。
4107	葱	2	8358	8471	葱		
4132		159	8371			血𠂇𠂇	眾，異體字問題。
4134	衛	777	8373	885E	衛		衛，異體字問題。
4142	褻	1	837B	8943	褻		褻，異體字問題。
4145	羈	5	837E	898A	羈		羈，異體字

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
							問題。
4156	諂	1	83AB	8B1F	諂		
4172	屨	2	83BB	8E77	屨		
4184	迹	3	83C7	8FF9	迹		
4185	邊	1	83C8	9037	邊		
4224	頤	2	83EF	983F	頤		
4237	髡	9	83FC	9AE0	髡		髡，異體字問題。 部份使用不同字形「髡」，手動修改。
4238	髡	1	83FD	9AE2	髡		
4242	鬪	58	8442	9B2D	鬪		
4250	鮓	6	844A	9C0C	鮓		
4252	鰓	3	844C			魚𩺰眾	鰓，異體字問題。
4259	麇	1	8453	9E95	麇		
4265	鼃	1	8459	9F03	鼃		
4307	罵	1	84A5	99E1	罵		罵，異體字問題。
4308	劫	5	84A6	5227	劫		劫，異體字

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
							問題。
4312	皞	1	84AA	66AD	皞		皞，異體字問題。
4353	菟	1	84D3	83AC	菟		菟，錯字，以程式全部取代。
4469	勻	41	85AA	5304	勻		
4507	吳	419	85D0	5434	吳		吳，異體字問題。
4598	鞞	2	866C	292FE		革△韋	鞞，異體字問題。
4753		3	876A			尗△比△大	尗△大，錯字，以程式全部取代。
4762	旃	1	8773	3ACB		舛△冉	旃，異體字問題。
4768	妬	1	8779	59AC	妬		
4822	尗	1	87D1	5C29	尗		尗，異體字問題。
4894		1	885A			宀△心△冉	甯，異體字問題。
5034	弃	9	8949	5F03	弃		棄，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
5035		3	894A			具 	郟，錯字，以程式全部取代。
5085	芊	4	897C	26B12		艹  艸	芊，錯字。部份使用不同字形「芊」，手動修改。
5156	季	14	89E5	353B	季		
5248		1	8AA4			扌  間	擱，異體字問題。
5290	携	8	8ACE	643A	携		攜，異體字問題。
5299	戟	2	8AD7	39B8		卓  戈	戟，異體字問題。
5302	昏	1	8ADA	662C	昏		昏，異體字問題。
5310	烏	1	8AE2	8204	烏		
5321		1	8AED			圭  隹	隼，錯字，以程式全部取代。
5404	鴟	1	8BA3	4CED		即  鳥	
5456		1	8BD7			韋  未	韎，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
5505	耆	1	8C49	264BE		老厶占	考厶口，錯字，以程式全部取代。
5622	邛	5	8CE0	48B5		云厶卩	
5652	蝨	2	8CFE	27312		或厶虫	
5655	躑	3	8D42	56BB	躑		躑，異體字問題。
5705	廐	5	8D74	5ED0	廐		廐，異體字問題。
5713	躑	2	8D7C	481E		足厶戚	
5752	揔	1	8DC5	63D4	揔		

## 二、手動取代表說明：

《左傳》檔案中之舊版造字，因字形部份錯誤而無法全部以程式取代，皆手動修改成Windows XP能顯示的字或構字式，並製作手動取代表，請參考附件二。欄位說明如下：

- (一) 編號：Big5 造字空間為 6217 個，編號由 1 到 6217。
- (二) 檔案原文摘錄：摘錄檔案《左傳》中的文句段落，其中紅字底線的部分，為需要手動取代的原文。文句後方標示(全部取代)者，表示執行全部取代的動作，不一一列舉所有原文。
- (三) 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。文句後方標示(全部取代)者，表示執行全部取代的動作，不一一列舉所有原文。
- (四) 備註：記錄修改相同辭句的次數或其他事項。

附件二、《左傳》手動取代表

編號	檔案原文摘錄	手動取代結果	備註
3976	秋邠 <u>島</u> 來來朝(全部取代)	秋邠 <u>犁</u> 來來朝(全部取代)	有 2 處
	伯州 <u>島</u> 奔楚(全部取代)	伯州 <u>犁</u> 奔楚(全部取代)	有 5 處
	使沈尹壽與師祁 <u>島</u> 讓之	使沈尹壽與師祁 <u>犁</u> 讓之	
	晉人執宋行人樂祁 <u>島</u>	晉人執宋行人樂祁 <u>犁</u>	
4010	公曰行也·及 <u>当</u> 鮑	公曰行也·及 <u>皐</u> 鮑	
4091	以肥子 <u>肩</u> 歸	以肥子 <u>皋</u> 歸	
4237	之師于雞父·胡子 <u>差</u> (全部取代)	之師于雞父·胡子 <u>髡</u> (全部取代)	有 2 處
	宋樂 <u>差</u> 帥師伐曹	宋樂 <u>髡</u> 帥師伐曹	
5085	國民信之· <u>丕</u> 姓有亂	國民信之· <u>半</u> 姓有亂	

## 附錄二、古漢語轉碼成果統計

2008 年 3 月

中研院語言所文獻語料小組製表

欄 位 說 明	檔案：轉碼檔案名稱
	總字數：需轉碼檔案的總字數（單位為萬字）
	漢籍造字數：該檔案使用舊版造字的缺字數量
	造字字頻：該檔案使用舊版造字的缺字總頻次
	轉 Unicode 字數：舊版造字可直接轉成 WinXP 可顯示的 Unicode 字數
	轉 Unicode 字頻：舊版造字轉成 WinXP 可顯示的 Unicode 字之頻次
	轉構字式字數：無法轉成 Unicode 的舊版造字字數，改以構字式表達
	轉構字式字頻：無法轉成 Unicode，改以構字式表達的舊版造字頻次。
	新增造字數：漢字構形資料庫尚未收錄的缺字，需要另外造字的數量。

檔名	總字數 (萬字)	舊版造 字數	造字字 頻	轉Unicode 字數	轉Unicode 字頻	轉構字 式字數	轉構字 式字頻	新增造 字數	轉碼者	完成時間
左傳	26	105	1.995	94	1.974	11	21	0	丁玫伶	2008/2/26
左傳(標 記)	50	106	1.624	97	1.607	8	17	0	丁玫伶	2008/2/26