

閩南語兒童語料數位化工作流程簡介

更新日期：2005/11/16

計畫單位：國立中正大學語言學研究所

計畫名稱：台灣兒童語料庫計畫

計畫簡介：

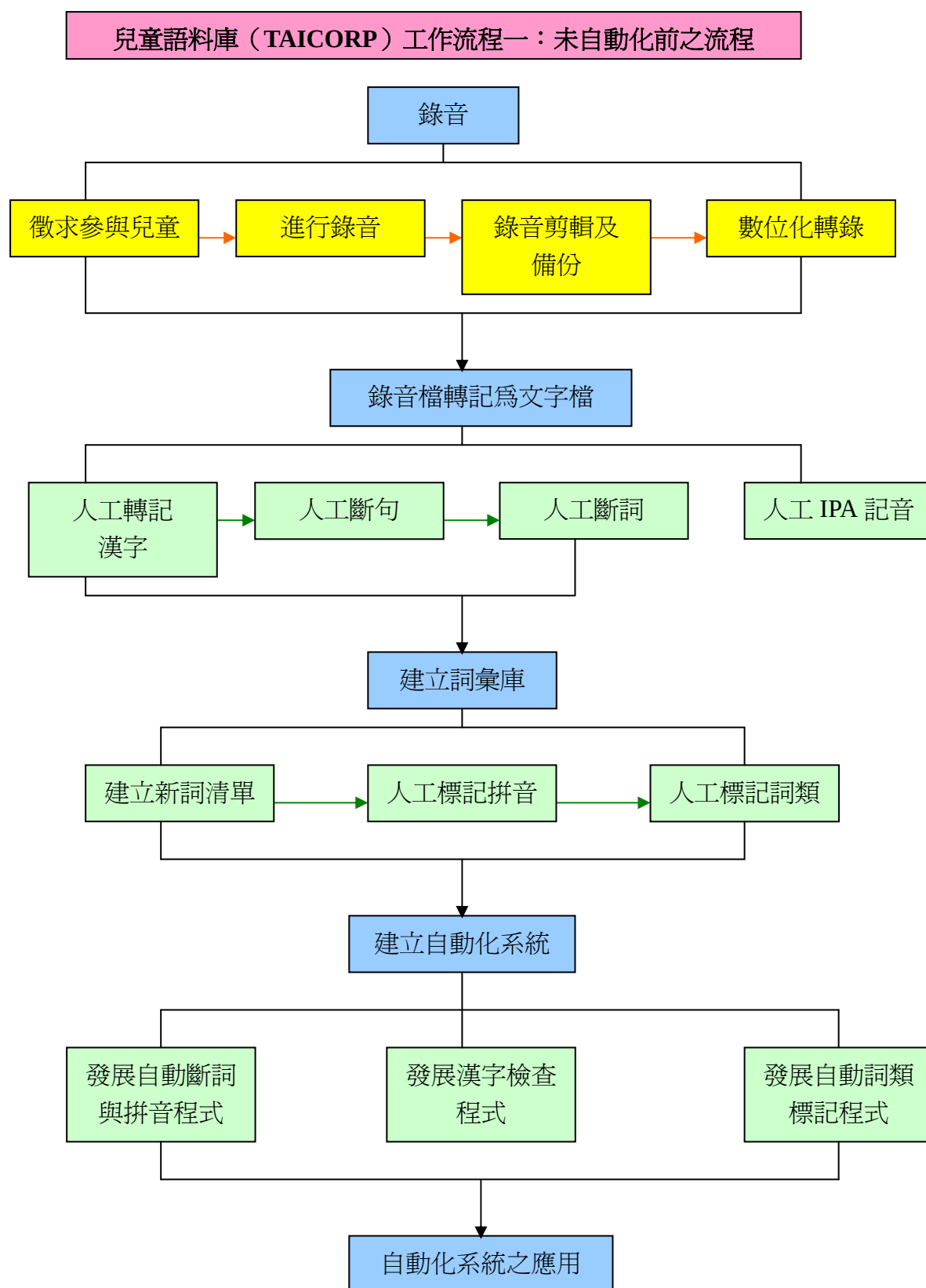
「台灣兒童語料庫」Taiwan Child Language Corpus（簡稱 TAICORP）是將所收集之台灣兒童口語錄音語料，依照世界標準的兒童語料交換系統 Child Language Data Exchange System（簡稱 CHILDES; MacWhinney and Snow 1985, MacWhinney 1995）格式，建構成語料庫。其主要目的在（1）提供國內外學者語料共享的便利性與語料分析工具；（2）藉由標準規格的設定，使台灣兒童語料的收集能更有系統、更有效率，並且快速地涵蓋台灣地區所有語言。語料庫最終將設立網站，開放國內外學者使用。

在新生一代普遍使用國語的時代背景之下，台灣閩南語兒童語言習得的語料彌足珍貴。本語料庫可提供語音學、音韻學、構詞學、句法學、語意學、語用學等不同層面的語言學與兒童語言習得研究，也可提供語音工程方面的研發與應用。本計畫由國立中正大學語言學研究所蔡素娟教授主持，從 1997 年 10 月開始錄音，經轉記、標記、格式化等過程，歷時將近九年。共收錄 431 人次錄音檔案，錄音總長共約 330 小時。文字檔共約五十萬句，一百六十多萬詞。

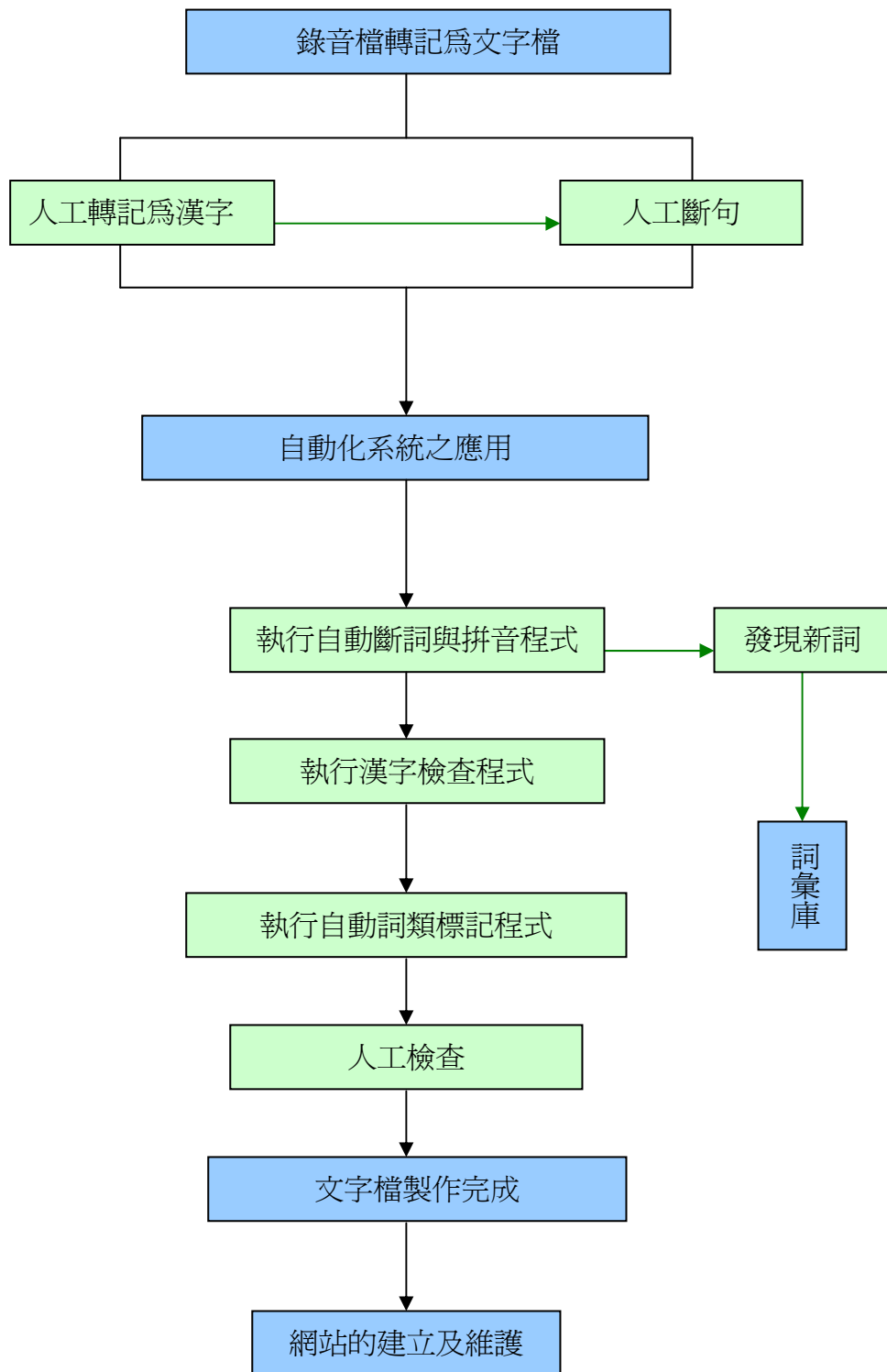
數位化工作流程說明

該計畫的數位化作業，大致依照下列五項步驟進行，依序分別為：一、錄音；二、錄音檔案轉記為文字檔；三、建立詞彙庫；四、建立自動化系統；五、自動化系統之應用等五個方面，共細分二十三項步驟進行。茲分別介紹如次。

閩南語兒童語料數位化工作流程圖如下：



兒童語料庫 (TAICORP) 工作流程二：自動化後之流程



製作日期：2005/11/17

流程圖提供者：

國立中正大學 語言學研究所 蔡素娟教授

一、 錄音

「錄音」部分分為五個步驟進行，分別為「訓練研究助理」、「徵求參與兒童」、「進行錄音」、「錄音剪輯及備份」、「數位化轉錄」。

(一) 訓練研究助理：由計畫主持人訓練研究助理。最核心的研究助理有三名。簡述如下：具語言學碩士級背景知識，並以閩南語為母語。透過每星期三到六小時的討論會，訓練助理，瞭解閩南語音韻及書寫系統、閩南語詞彙、句法、語意及詞類標記系統、CHILDES 系統及兒童語言習得相關文獻；並熟悉 IPA 國際音標記音。

(二) 徵求說閩南語家庭之兒童：目標選定中正大學附設托兒所、幼稚園及鄰近鄉鎮，徵求來自說閩南語家庭，年齡在一歲至三歲之間的幼兒。陸續共選出 14 名兒童。

(1) 以海報及網路發布廣告；利用幼稚園家長日到場對家長說明，徵求說閩南語家庭的兒童。

(2) 排定錄音時間：聯絡家長；並排定錄音時間表

(三) 進行錄音

(1) 準備錄音器材：錄音器材選擇方便攜帶、機動性強、容量較大、易長期保存語料之錄音器材。下圖左起為迷你光碟片、專業用耳機、專業用麥克風、迷你光碟隨身錄音機。



圖一：錄音器材

(2) 進行錄音訪談：至兒童家中進行訪談錄音。錄音為週期性，寒暑假

亦不間斷。二歲以下者，每週訪談一次；二至三歲者，每兩週訪談一次；三至四歲者，每二至三週訪談一次。每次訪談約 1 至 2 小時不等，實際錄音時間 40 至 60 分鐘。

錄音期間：1997 年 10 月至 2000 年 5 月。共錄音 431 人次，約 330 小時。訪談方式為：錄下兒童在家長或保姆陪同下，在自己家中的日常對話。錄音的內容除了自然言說，還藉助圖畫簿、故事書、玩具、布偶、剪紙、摺紙或其他遊戲，引發兒童主動說話。

（四） 錄音剪輯及備份

- (1) 錄音剪輯：由助理將錄音光碟中不相關的錄音或太長的空白錄音刪除，將錄音切割為較小段落，在光碟中標記段落編號；於光碟中輸入錄音日期、檔名。每 1 小時的錄音約需耗時 1.5 小時剪輯。總工作時間： $1.5 \times 330 \text{ 小時} = 495 \text{ 小時}$ 。



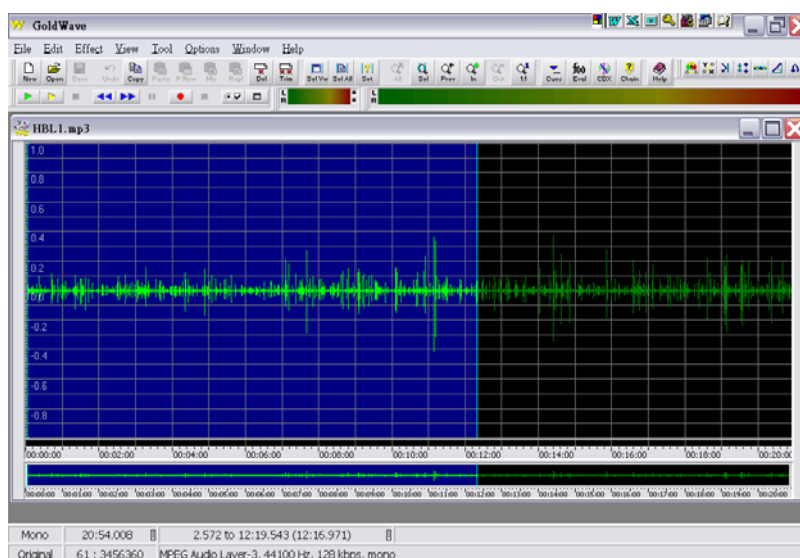
圖二：進行錄音剪輯（示範者：謝沛諭）

- (2) 錄音備份：使用迷你光碟錄音座及迷你光碟隨身錄音機進行迷你光碟備份製作。



圖三：進行錄音備份 (示範者：謝沛諭)

- (五) 數位化轉錄：將迷你光碟錄音檔轉為較不佔空間之 MP3 格式，以方便儲存。於日後可隨時轉為語音分析所需之格式（如 WAV 格式）。所使用之轉錄軟體為 GoldWave Digital Audio Editor（GoldWave Inc. 研發）。



圖四：進行數位化轉錄

二、 錄音檔轉記為文字檔

「轉記」分為四個步驟，依序為：「人工轉記漢字」、「人工斷句」、「人工斷詞」、「人工 IPA 記音」。

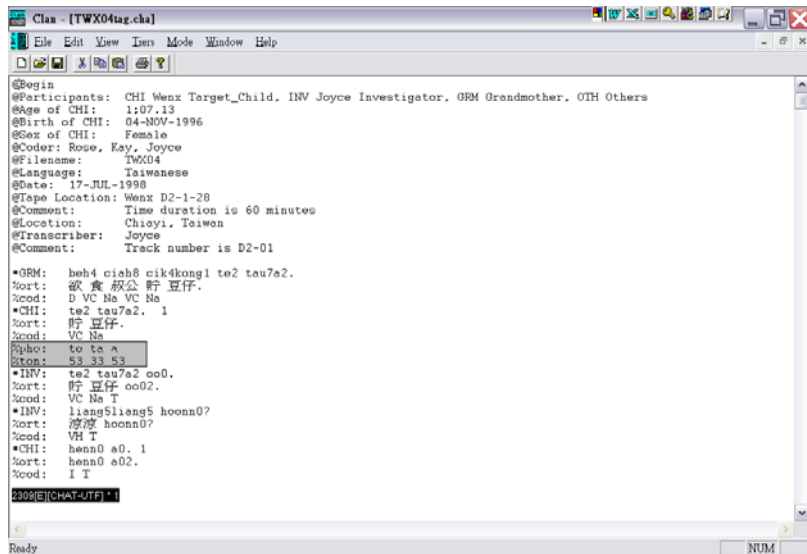
- (一) 人工轉記漢字：由於閩南語的漢字書寫系統目前並沒有定案，再加上有

許多本字無法確定，或者有音無字的情形，因此有必要訂定文字轉記的原則。故在進行文字轉記前，首先需確立閩南語書寫系統，本計畫所參考的辭典主要有四本，依優先順序排列為：《臺灣閩南語辭典》《台灣話大辭典》《廈門方言詞典》《閩南語詞彙》如下圖由左至右。轉記平台為 CHILDES 兒童語料交換系統。每 1 小時錄音需要花約 10 小時不等的時間轉記成文字檔。總工作時間：330*10=3,300 小時。



圖五：閩南語辭典

- (二) 人工斷句：由於本語料庫之語料為口語語料。助理需參考言談分析之斷句原則，將自然言談切分成獨立意義句子。
- (三) 人工斷詞：由於目前無閩南語斷詞標準，故本計畫根據中華民國計算語言學學會所訂定之「資訊處理用中文分詞規範調查研究及草案研擬」，將語句切分為獨立意義、且扮演特定語法功能的字串。
- (四) 人工 IPA 記音：採語音轉記 (phonetic transcription) 的方式詳細轉記兒童實際發音。在音段方面，以 Unicode IPA 符號記音，參考書目為 Handbook of the International Phonetic Association (1999)；聲調採用五度標音法。每小時的錄音約需花 4.5 小時記音。共 $4.5 \times 330 = 1485$ 小時。



圖六：完成 Unicode IPA 記音之文字檔

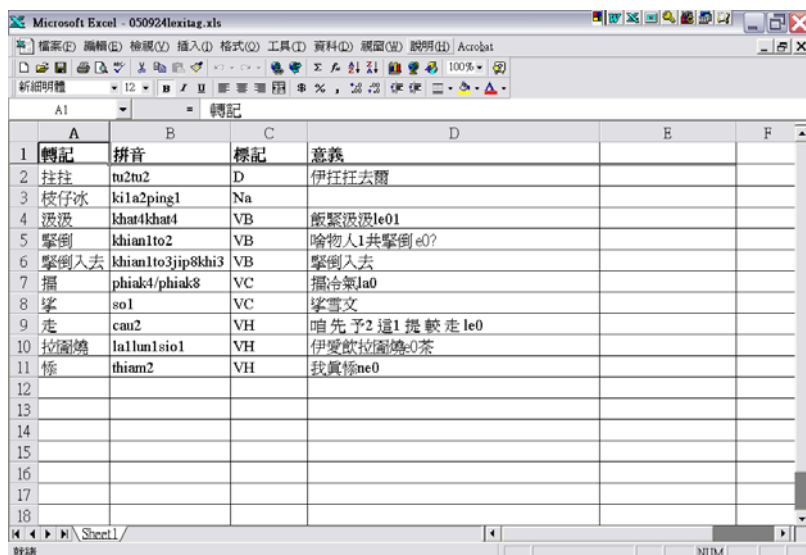
三、建立詞彙庫

錄音以人工轉記為文字很費人力，因此最終目標還是要建立自動化系統。而自動化系統的建立需要詞彙庫作基礎。「建立詞彙庫」依序分為三個步驟進行：「建立新詞清單」、「人工標記拼音」、「人工標記詞類」。

- (一) 建立新詞清單：以轉記好之文字檔中之所有詞彙建立清單，經由人工確認詞彙清單中的漢字與詞典是否一致。
- (二) 人工標記拼音：根據教育部於民國八十七年所公佈之「閩南語羅馬拼音第二式」人工標記詞彙清單中的漢字之拼音。

圖七：人工標記拼音

- (三) 人工標記詞類：參考中央研究院詞庫小組「詞類標記原則」以及 CANCORP: The Hong Kong Cantonese Child Language Corpus, Lee and Wong (1998)、台灣閩南語動詞分類研究 曹逢甫 (1996) 等相關文獻。採用中研院詞庫小組的詞類標記，但是僅限於 46 個簡化標記，以避免詞類劃分過細時產生主觀強制性的歸類。



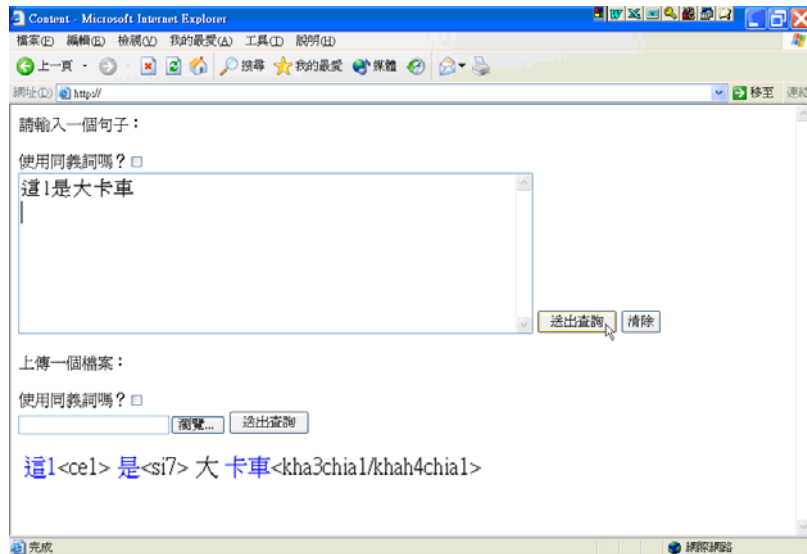
	A	B	C	D	E	F
1	標記	拼音	標記	意義		
2	往往	tu2tu2	D	伊往往去爾		
3	校仔冰	ki1a2ping1	Na			
4	汲汲	khat4khat4	VB	飯緊汲汲le01		
5	擊倒	khian1to2	VB	嗰物人1共擊倒e0?		
6	擊倒入去	khian1to3jip8khi3	VB	擊倒入去		
7	搵	phiak4/phiak8	VC	搵冷氣la0		
8	穿	so1	VC	穿雪文		
9	走	cau2	VH	咱先予2這1提較走le0		
10	拉爾德	la1lu1sio1	VH	伊愛飲拉爾德e0茶		
11	係	thiam2	VH	我真係ne0		
12						
13						
14						
15						
16						
17						
18						

圖八：人工標記詞類

四、 建立自動化系統

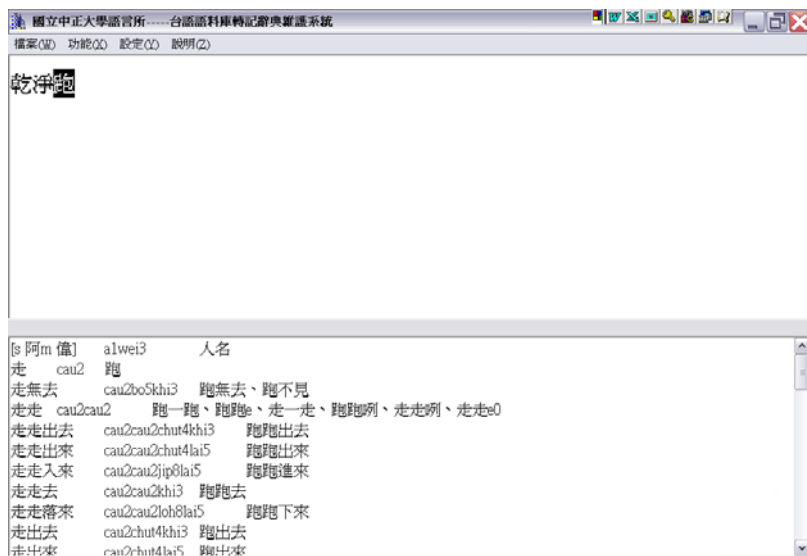
「建立自動化系統」以上述詞彙庫為基礎。分為三個部分：「發展自動斷詞與拼音程式」、「發展漢字檢查程式」、「發展自動詞類標記程式」。

- (一) 發展自動斷詞與拼音程式：將輸入之句子或整個文字檔案，根據本計畫修訂「資訊處理用中文分詞規範調查研究及草案研擬」所撰寫之「閩南語斷詞原則」及詞彙庫之詞項，根據長詞優先之準則，與詞彙庫比較。若所輸入之漢字與詞彙庫一致，則以黑色呈現，並在其後標注拼音；若所輸入之漢字尚未建立於在詞彙庫，則以藍色呈現。此程式除了斷詞及標注拼音之外，還可以將新詞納入詞彙庫。



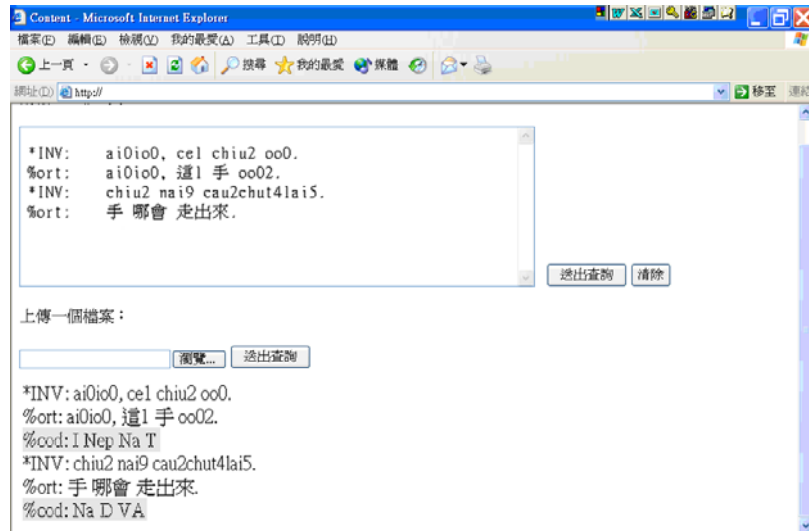
圖九：自動斷詞與拼音程式

- (二) 發展漢字檢查程式：目的為求漢字與詞彙庫所列之標準之一致。搜尋之方式有三：一為輸入閩南語羅馬拼音、二為輸入可能之漢字、三為輸入國語之相對詞；透過此三種任一，皆能擷取出詞彙庫中含有該詞之詞條。但若該詞未建立於詞彙庫中，查詢後則不顯示。



圖十：漢字檢查程式

- (三) 發展自動詞類標記程式：以人工標記詞類之文字檔作為基礎，發展自動詞類標記程式。將輸入之句子（已完成斷詞工作），自詞彙庫中擷取出其詞類標記；當該詞有多個詞類標記時，程式則以頻率最高之標記為優先考量並標記之。若該詞在詞彙庫中未標記詞類，則以三個問號（???)呈現。



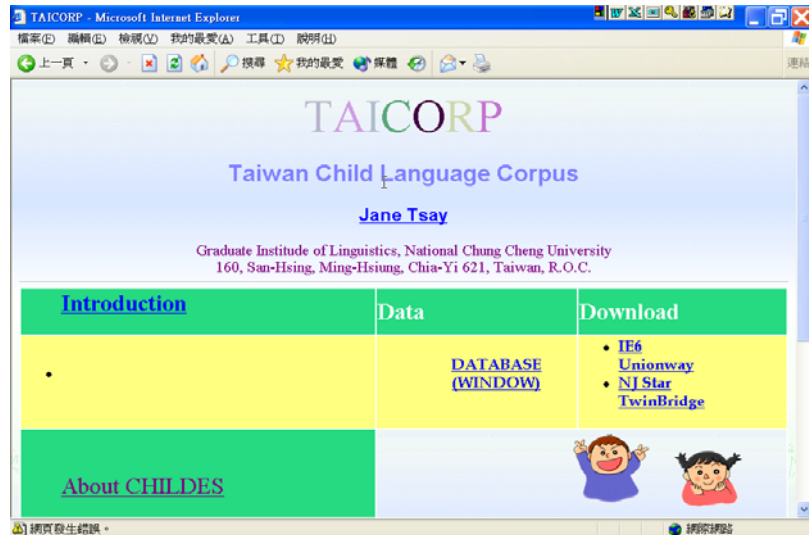
圖十一：自動詞類標記程式

五、 自動化系統之應用

- (一) 執行自動斷詞與拼音程式：將語句切割成詞，並標注拼音。
- (二) 執行漢字檢查程式：檢查漢字與詞彙庫所列之標準是否一致
- (三) 執行自動詞類標記程式：標記詞類。
- (四) 人工檢查：檢查程式輸出檔，如詞有不只一個詞類，則檢查其自動標記是否正確。

六、 網站的建立及維護

- (一) 網站架構及內容之編纂：計畫主持人與研究助理討論網站內容及所呈現之介面。網站內容包含語料庫簡介、資料庫、使用手冊、相關程式以及相關網站之連結。
- (二) 網站之建立及維護：為語料庫建立專門網站，以供世界各地學者研究之用。完成最後檢測之後，網站將開放給外界瀏覽。



圖十二：網站首頁

-
- ※ 製作單位：數位典藏國家型科技計畫---內容發展分項計畫
國立中正大學語言學研究所---台灣兒童語料庫計畫
 - ※ 文字撰寫：數位典藏國家型科技計畫---內容發展分項計畫
語言主題小組助理 賴佳旻
國立中正大學語言學研究所台灣兒童語料庫計畫 助理 謝沛諭
 - ※ 圖片拍攝：數位典藏國家型科技計畫---內容發展分項計畫
語言主題小組助理 賴佳旻、林淑惠、陳秀華
 - ※ 圖文編輯：數位典藏國家型科技計畫---內容發展分項計畫
語言主題小組助理 賴佳旻、陳美智、陳秀華
 - ※ 感謝國立中正大學語言學研究所---【台灣兒童語料庫】之計畫主持人 蔡素娟教授、前任助理黃婷鈺小姐、劉慧娟小姐及現任助理謝沛諭小姐撥冗指教及協助拍攝與提供資料，特別致謝。
-