一套以拼音規則為基礎的台語羅馬字編碼轉換與網頁展示工具

蔡哲民
崑山科技大學資訊管理學系
tjm@mail.ksu.edu.tw

李仁傑
崑山科技大學資訊管理學系
jest.pcst@msa.hinet.net

林俊育
台語信望愛主編
chuniok@fhl.net

摘要

目前一般所使用的台語羅馬字到了 2004 年其特殊字元才被加入 Unicode 標準中，在此之前缺乏統一的編碼標準，以致需要一套編碼轉換工具。又因為輸入法缺乏的問題，使得數字調號的表示法成為取代輸入法的解決方案，因此把數字調號編碼轉換成 Unicode 編碼成為重要的需求。

由於目前台語羅馬字缺乏廣泛使用的電腦字形支援，網頁也無法直接用 Internet Explorer 6.0 以前的瀏覽器閱讀，因此將台語羅馬字換成圖形展示，成為此過渡時期的重要解決方案。

本論文提出一套台語羅馬字編碼轉換與圖形展示工具，利用台語羅馬字的編碼原則，降低台語與英文混合時的轉換錯誤率，並可選擇將輸出轉換成內嵌文字圖形之 HTML 格式。並展示此工具與網頁內容管理系統結合建構之台語網站的範例。

關鍵詞：台語羅馬字、台語網頁、台語編碼轉換

Abstract

Because of their special fonts and browser supporting shortage, Taiwanese-Romanized documents can't be read through general web browser directly. For lack of input method, the numerical tone marks are used to be substituted. We present a tool which translates numerical tone marks documents into Unicode documentation, the error rate is reduced by checking the spelling rule of Taiwanese-Romanization. This tool can also translate the Taiwanese-Romanized documents into graph-embedded HTML documentation to solve the browsing problem caused by special fonts.

This tool is assembled with a content management system to build a Taiwanese-Romanized website.

.

Keywords: Taiwanese-Romanization, Taiwanese-Romanized website, Code translation.

1. 前言

台語是一種弱勢語言，雖然日常生活中使用台語的人數不少，但是真正以台語書寫的文件不多見，更遑論以台語為主要語言的網站了。其主要的原因，除了因早期推行國語運動造成知識份子對台語的不熟悉外，也因為台語一直沒有一套統一的「文字標準」，造成台語文件的書寫不易。現存的台語文件使用的「文字標準」，除了最早期傳教士流傳下來的「台語羅馬字」(或稱「白話字」)之外，還有 TLPA[1]等系統。不過目前存留的台語文件中，還是以「台語羅馬字」為最大宗[5]，因此本研究是以台語羅馬字為標的。

除了文字標準差異之外，台語電腦化也面臨中文字電腦化初期遭遇的「萬碼奔騰」問題。目前，即使是最多人用的「台語羅馬字」，其「調號」與「鼻音」使用的特殊字元，也因為歷史的因素而有 HOTSYS[9]、Taiwanese Package(TP)[7]等不同的編碼系統與相對應的特殊字形[8]。不過這樣的「萬碼奔騰」問題，在 2004 年 ISO 組織通過將白話字元號加入 Unicode 之後會慢慢有所改善。

由於台語羅馬字使用到一些特殊的調號與鼻音字符，即使目前有少數字形支援這些特殊字符[2]，也缺乏廣泛被使用的輸入法來輸入這些特殊字符[3]。為了解決這個困境，許多前輩就用阿拉伯數字來表示調號，用「N」或「nn」來表示鼻音，用「ou」來表示「ɔ」。例如：A-sat e5 si-koa, kau hou7 gak8-koaN 就用來表示 A-sat ê si-koa, kau hō gȧk-koaⁿ。這樣的表示法稱為「數字調號」。

使用數字調號不需要任何特殊輸入法，就可以直接透過標準英文鍵盤輸入台語文章，但是要輸出成文件或網頁時，則還需要轉換成標準的台語羅馬字形始能供使用者閱讀。目前，已有一些離線的工具如：TP[7]等可以進行數字調號與 Unicode 之間的轉換；也有一些網頁工具如：暗光鳥 ê 厝的羅馬字轉換器[4]可以進行類似的轉換。這些編碼轉換工具，除了一般的文件編碼轉換用途之外，在某個程度上也取代了台語輸入法的地位。另外，數字調號表示法使用的字元全部都在 ASCII 編碼的標準可視字元之內，因此也很適合作為資料庫或其他軟體處理台語文件的內碼。

不過現有的這些台語調號轉換系統多半是直接將用來代表台語特殊字符的數字或英文代號轉換成對應的字符，因此如果遇到台語文字中混雜英文時候，就會轉換錯誤，例如：「you」就會被轉換成「yo·」。不過目前國際交流日漸頻繁，文件裡面夾雜英文或其他語言的機會越來越多，因此怎樣進一步提升台語文件編碼轉換的正確率就成為一個重要的課題。

另由於台語字元一直到 2004 年才被接受進入 Unicode，所以現階段大多數的電腦軟硬體環境並無法完整的顯示台語字元，其中「o͘」這類具有葫蘆點的字型，更是目前廣泛使用的商業字形尚未支援的[2]。即使安裝了支援台語字元的 Unicode 字形，如果使用 Internet Explorer 6.0 以前的瀏覽器，網頁不經過特別的處理也無法正確顯示台語字元。因此，台灣網際網路普及率雖然已經將近 50%[6]，但網際網路上還是很難找到以台語為主要語言的網站。即使找到台語網站，使用者也需要安裝特殊字形、使用 FireFox 等瀏覽器才能正確閱讀台語文章。

站在保存台灣語言、文化的立場，為了能夠讓使用者利用一般電腦與瀏覽器，無須下載特殊字型與做特別的設定就能瀏覽台語羅馬字網頁，實在是一個非常重要的課題。

為了解決上述兩個問題，我們利用 Open Source 的開發工具 PHP 配合 FreeType[12]、GD 函式庫[11]等，應用台語羅馬字拼音規則，製作了一套編碼轉換工具，可將數字編碼的台語羅馬字文件轉換成 Unicode 文件，並可選擇將此羅馬拼音文件線上轉換成圖形顯示於網頁上，使得使用者無須做任何設定即可閱讀該網頁。

為了驗證此系統之功能，我們製作了一套獨立的台語羅馬字編碼轉換系統，並將此系統整合進一套內容管理系統中，實際承載一個台語網站之運作。

該內容管理系統使用數字調號當作內部的文件編碼，以降低程式修改的複雜度，網頁發佈時才呼叫本工具將數字調號文件轉換成內容相同的圖形與 Unicode 台語網頁，以提供一般使用者可以在任何電腦上閱讀台語網站文件，擁有台語字形的使用者可以自由複製台語文件。

以下我們將介紹本系統使用的台語字元轉換與網頁展示方式，並分析本系統與英文混雜使用時的正確率，並檢視使用「N」或「nn」作為台語羅馬字鼻音表示法所帶來的影響。

2. 系統結構

本系統用來將數字調號台語文件轉換成 Unicode 編碼之台語文件，並且依照使用者的設定，將輸出之台語文件用 Unicode 輸出或者轉換成圖形網頁輸出。本系統的運作流程如圖 1 所示。

台語文件輸入後，先被切割成一個一個的 token，系統會檢查每一個 token，先利用編碼範圍過濾掉中文，這樣就剩下台語和混雜在台語中的英文字。然後利用台語羅馬字的拼音規則來進行進一步的分析，如果是台語，則針對調號與鼻音進行轉換，否則將不做任何轉換。經過轉換成 Unicode 的台語羅馬字，再依系統設定直接輸出或轉換成圖形，並將此台語羅馬字置換成內嵌此圖形之 HTML 輸出。本系統測試網站在 http://taigi.fhl.net/tai/，運作畫面如圖 2a、2b。

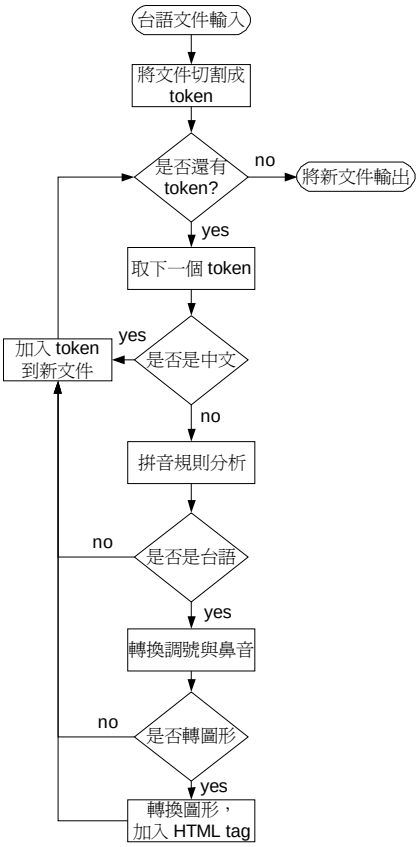

圖 1 編碼轉換系統流程圖

圖 2a 編碼轉換系統輸入畫面

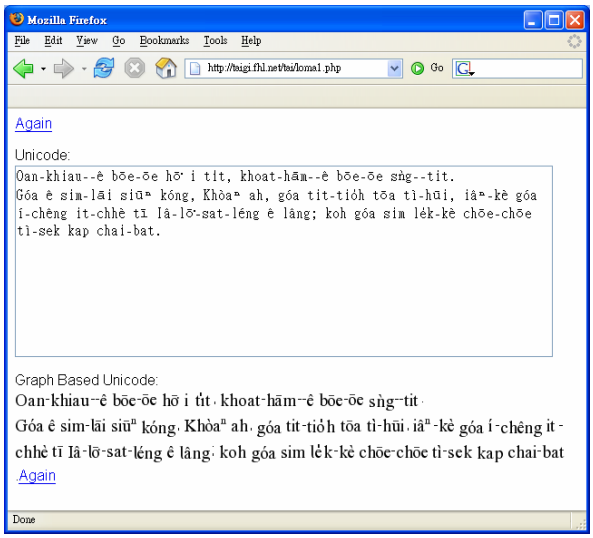

圖 2b 編碼轉換系統輸出畫面

2.1 台語羅馬字拼音規則分析

為了進一步提升本系統正確分辨台語羅馬字之比例，我們使用了台語羅馬字拼音的基本規則：

台語羅馬字＝[子音]母音[入聲尾][調號]　　(1)

亦即一個台語羅馬字可以由子音、母音、入聲尾和調號組成，其中母音是一定要具備的，其他都是可有可無。而代表該四種音的符號如表 1 所示。特別值得注意的是 ng、m、n 這三個符號同時可以代表母音和子音。

真正在進行編碼轉換的時候，由於大部分的字符都是使用標準的英文字母，只有當台語字具有調號（2,3,4,7,8）、鼻音（N 或 nn）或者母音 ou 時才需要進行調符轉換，亦即也只有在這個狀況下才會造成夾雜英文字的誤判。透過比對台語羅馬字的拼音規則，可以更進一步的避免「you」這一類的英文被誤判為台語，提升編碼轉換的準確率。

調號與鼻音轉換為特殊字形時，還需要修正這些字形的位置與母音結合。

表 1 台語羅馬字字符與分類

分類	字符
子音	chh,ch,kh,th,ph,p,t,k,b,g,l,j,h,s
母音	N(或 nn),ou,a,i,u,e,o
子母音共用	ng,m,n
入聲尾	t,p,k,h
調號	2,3,4,7,8

2.2 台語羅馬字網頁圖形轉換

為了讓使用者不必安裝特別的字形就可以閱讀台語網站，我們將一套支援完整台語羅馬字特殊字元的 Unicode 字形放置在伺服器端，當使用者要求的網頁使用到台語調符時，系統透過 PHP 內建的 GD Library，呼叫 FreeType 函式庫，將此字形以每個台語羅馬字為單位轉換成 PNG 圖形，並且改寫文件的 HTML，將此產生的圖形檔內嵌於網頁中展示出來。

為了提升系統效率，我們將已經產生過的台語羅馬字檔案存放在網頁伺服器的檔案系統中。系統於產生圖形前先檢查檔案是否存在，不再重複產生已經產生過的圖形檔案。

使用字型產生文字圖形，雖然耗費系統資源、佔據頻寬，不過實在是一種過渡時期為了避免增加使用者閱讀網頁障礙的解決方案。對於不希望網頁內容被複製的作者，這樣的作法還有保護網頁內容的附帶效果。

圖 3a、3b 是使用本工具的兩個版本網站，兩個網站內容一樣，但是圖形版本的網站不需要安裝特殊字形就可以閱讀。

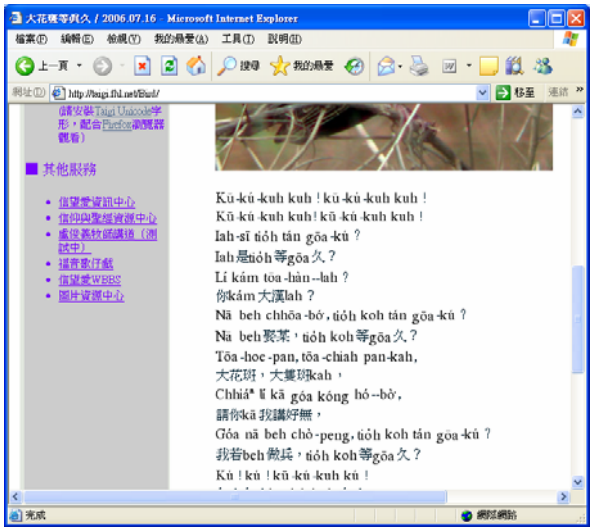

圖 3a 內嵌羅馬字圖形的網頁

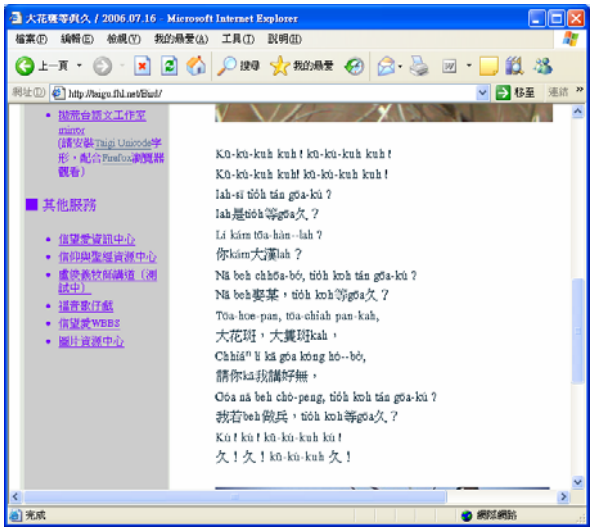

圖 3b Unicode 版本的網頁

3. 與網頁內容管理系統之整合

本工具與一套既有之網頁內容管理系統整合後之架構如圖 4 所示，我們採用了目前網際網路上常用 MySQL 與 PostgreSQL 這類免費的資料庫來製作台語網站的內容管理系統。為了迴避這類資料庫系統繁複的編碼設定問題，我們採用了數字調號來當資料庫的內碼，所有文件的台語羅馬字都以數字調號的形式存入資料庫中。

該網頁管理系統允許網頁管理者經身份認證後對網站的內容與架構進行新增、刪除與修改等功能。當網頁要發佈時，則透過編碼轉換系統將數字調號台語文件轉換成 Unicode 格式，然後逕自輸出成 Unicode 版本靜態網頁，或者啟動台語圖形轉換功能將 Unicode 台語文件轉換成對應的圖形檔與 HTML 文件輸出成圖形版本的靜態網頁。

透過這樣的架構，可以同時產生兩套內容相同的網站，在支援台語羅馬拼音特殊字元的 Unicode 字形檔還沒有普遍被使用前，滿足一般使用者與需要複製文件之使用者的需求。此系統之測試網站 URL 為 http://taigi.fhl.net 與 http://taigu.fhl.net。

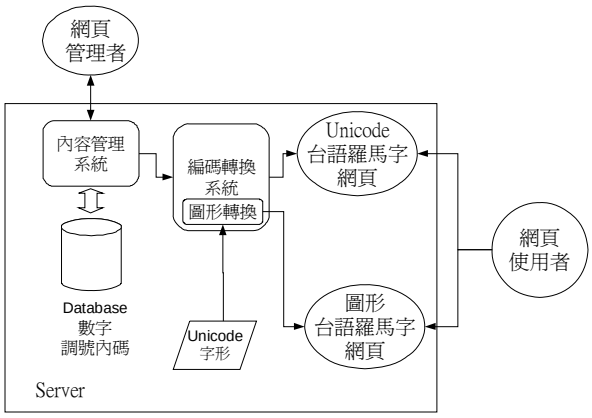

圖 4 與內容管理系統整合架構圖

4. 比較與討論

4.1 與英文混合編碼轉換錯誤率評估

為了驗證本編碼系統的正確率，我們採用由 pyDict 計畫[10]取得之 134183 字英文字典，將每個字當成台語羅馬字輸入系統中，比較 TP 與我們的系統錯把英文字當成是台語羅馬字，且造成轉換錯誤的字數。由於鼻音可能用大寫「N」來代表，所以我們也將字典中所有的字轉換成大寫重新做一次檢驗，檢驗的結果如表 2。

表 2 編碼轉換錯誤字數

方法	錯誤字數	轉大寫錯誤字數
Taiwanese Package	2581	69937
暗光鳥 ê 厝	37	35
本系統使用 N 代表鼻音	15	289
本系統使用 nn 代表鼻音	15	26

由表 2 中可以看出在小寫的情況下，本系統把英文誤認為台語羅馬字的字數是現有的工具中最低的。在英文字典轉大寫的情況下，除了暗光鳥 ê 厝之外，其他的系統的錯誤字數都會提高，不過相較於沒有使用羅馬拼音規則來過濾的 Taiwanese Package 系統，本系統的錯誤率還是有較好的表現。由上述的測試中可以看出引用羅馬拼音規則來過濾英文字，的確可以大量降低轉碼的誤判率。

根據我們的測試，暗光鳥 ê 厝所提供的轉碼工具，並不處理全部大寫的台語羅馬字轉換，亦即 THIN 這樣的字在該系統中並不認為是台語羅馬字。不過這實際上可能是一個大寫的台語羅馬字。在實際的使用上，遇到全部大寫英文字或台語羅馬字的機率不高，因此不接受全部大寫的字母為台語羅馬字，是一個降低誤判率的可能方法。不過我們認為大寫英文與台語羅馬字的混合機會不高，所以正確的轉換全大寫羅馬字可能比一味的希望降低誤判率來得重要。

4.2 效能評估

我們在 3.0G CPU、2G RAM 的 Fedora Core 5 Linux x86 伺服器上，透過 100M 區域網路，用 1.73G CPU、1G RAM 的 Windows XP 電腦透過瀏覽器進行測試。用 32,222 個台語羅馬字的巴克禮版本舊約聖經創世記（約 180,000 個字元）與 9,967 個字的但以理書（約 55,700 個字元）來測試此編碼轉換與圖形網頁展示系統之效率。每個數據都是測試三次以後平均而得，我們並將測試分為自行產生圖形與預先產生圖形兩組，分別測試系統重新產生所有台語羅馬字圖形（自行產生圖形）與檔案系統中已經產生好所有的台語羅馬字圖形（預先產生圖形）的狀況。

表 3 系統效能評估

方法	編碼轉換	編碼轉換與圖形產生
創世記（自行產生圖形）	73 sec	337 sec
創世記（預先產生圖形）	63 sec	325 sec
但以理書（自行產生圖形）	6 sec	35 sec
但以理書（預先產生圖形）	5 sec	32 sec

由表 3 中可以看出系統運作所花費的時間隨著轉換內容的長度而大幅增加，預先產生台語羅馬字圖形能夠增快速度，但無法大幅增快速度，其原因應該是系統仍須檢查檔案系統中是否已經存在產生過的台語羅馬字圖形。

由於轉換成圖形需要消耗大量系統資源，因此建議應該避免在動態網頁之中即時產生，我們的實驗網站是透過「發佈靜態網頁」的方式來將圖形產生的延遲限制在網頁發佈的時候，而避免每個使用者讀取網頁時都去產生台語羅馬字圖形。

5. 結論

我們已經完成一套以台語羅馬字拼音規則為基礎之編碼轉換與網頁展示工具。透過運用台語羅馬字的拼音規則，本系統可以大量減少台語與英文混合的狀況下，進行「數字調號」格式與 Unicode 格式轉換時所造成的誤判。透過將羅馬拼音轉成圖形，並修改網頁成為內嵌此圖形之 HTML，可以使網頁讀者不需要安裝特殊的字型、使用特殊的瀏覽器就能閱讀台語羅馬字網站。

我們並將此工具與一既有之網頁內容管理系統結合，建立一圖形羅馬字與 Unicode 雙入口的實驗網站，本工具的原始碼可以在http://taigi.fhl.net/tai 中取得，使得有志於建立台語羅馬字網站的使用者也可以直接將此工具加入自己的內容管理系統，不必重新打造一套相同的工具。

誌謝

感謝國科會科教處對本研究之贊助，本計畫編號為：NSC 94-2218-E-168-001

參考文獻

[1] 洪維仁等，台灣語言音標，http://zh.wikipedia.org/wiki/TLPA。

[2] 陳鄭弘堯，“ 認 bat 白話字編碼 kap 字型 ê 關係（hō·初用者）”, http://www.lomaji.com/siau-sit/2005/08/890.php ，2005.08。

[3] 陳鄭弘堯，“白話字輸入法（IME）比較表”，http://www.lomaji.com/poj/tools/converters/POJ_IME_chart.html，2006.01。

[4] 陳鄭弘堯，“羅馬字轉換器”，http://www.lomaji.com/poj/tools/converters/convert.php。

[5] 楊允言，“台語符號 ê 競爭—以 TLPA kah 白話字做例”，2002 年 7 月。

[6] 資策會，Focus on Internet News & Data，“2006 年 3 月底止台灣上網人口”，http://www.find.org.tw/0105/howmany/howmany_disp.asp?id=140。

[7] 劉杰岳，http://www.phahng.idv.tw，2001。

[8] 劉杰岳、楊允言，“白話字電腦文書處理 ê 研究”，2005.11。

[9] 蘇芝萌，“ HOTSYS-HAHSYS 台客語文書處理軟體”，1994。

[10] Daniel Gau, http://sourceforge.net/projects/pydict。

[11] “GD Graph Library”, http://www.boutell.com/gd/。

[12] “The FreeType Project”, http://www.freetype.org/。