中文到閩南語之線上翻譯及閩南語之語音合成

林川傑・陳信希

國立台灣大學資訊工程學研究所
cjlin@nlg2.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

一、摘要

臺灣是一個多語的社會，在許多地方提供了多語的服務。如何花最少人力資源來提供這樣的服務，便是目前閩南語語言處理研究的重點。

本篇論文提出國語—閩南語翻譯系統的設計方法，並說明翻譯所需處理的問題，包括選詞、未知詞的翻譯、變調問題、閩南語呈現方式等。

目前實驗成果有網際網路展示系統，並已將此架構擴展應用在國語—客家語的翻譯系統上，亦做線上展示，請參考網址：
http://nlg3.csie.ntu.edu.tw/TWmain.html

二、動機

臺灣是一個多語並存的社會，國語、閩南語、客家語和原住民族語言是許多人日常生活的溝通工具。在目前，保存不同地域的語言文化，增進不同族群間的了解，漸漸引起社會大眾的注意。

近來閩南語的研究日趨熱絡，但在資訊科學領域，過去這方面的研究較少，其主要的原因在於大量機讀資料的嚴重缺乏。機器翻譯系統在臺灣大學自然語言處理實驗室已研之有年，已有相當的基礎，是以我們決定跨出第一步，設計第一套屬於自己母語的機器翻譯系統。詳細資料請參考碩士論文(林川傑, 1997)。

三、簡介

3.1 閩南語

閩南語有七個聲調，十七個聲母和七十五個韻母。在我們所使用的詞典中，共有 2,617 個不同的音節。若考慮各種變調後，則有 3,241 個不同的音節。

閩南語最主要的問題在於如何用文字表示。目前有許多種不同的表示法，包括用漢字表示及各種拼音方式。用漢字是最易被大多數人讀懂的，但其最大的問題在於許多用詞的對應漢字仍有爭議。比方說：

我 beh 去學校

現在已有學者進行台語本字的研究(楊秀芳, 1996)，如果能在未來達成一致的意見，這將是較易明瞭的表示方式。

各拼音方式則沒有上述的爭議，有音一定可拼出來。但是拼音法各家分歧，且要看懂拼音需經過學習，不是每個人都能看懂。

本系統提供三種結果輸出方式：漢羅並用、教會羅馬字(鄭良偉, 1993; 中國大百科全書, 1988)、變調後結果及其語音合成。

3.2 系統架構

機器翻譯系統已有多種不同的解決方式，不論是哪一種模型，機器翻譯系統都必須去抓到對譯語言間用詞與語法等不同處。因此典型的機器翻譯系統包含了下列各模組：剖析器將輸入文句做語法分析，並產生其語法剖析樹；詞轉換器選擇最佳的對應詞；結構轉換器則做對譯語言間語法

結構的對應轉換，而生成器則將翻譯結果產生出來。

國語和閩南語間有極相似的語法結構。1968 年時，趙元任教授提出

> 「在文法方面，中國各地方言在文法上最有統一性。除去一些小分歧，…咱們可以說，中國話其實只有一文法。即使把方言也算在，…實質上，其文法結構不僅跟北平話一致，跟任何方言都一致。因此把北平話的文法稱中國話的文法，比把北平話叫中國話更爲有理。」(Chao, 1968; 丁邦新譯, 1980)

根據這個理論，我們便假設國語和閩南語的語法構是相同的，因為國語即自北平話而來。這假設在大部份的句子中皆成立，如：

國語　他 今天　心情 很 好
閩南語 伊 今仔日 心情 真 好

上例兩句話的語序完全相同。在我們的模型中，便假設語序是不變動的。圖 1 為整個國語閩南語機器翻譯系統之架構。

首先對輸入文句做斷詞與詞性標註，之後從一國語閩南語對譯詞典選出適當的對譯詞，若該詞並未收錄在對譯詞典中，則查單字音典將它逐字唸出。為了做語音的輸出，接著再處理變調問題，以得流暢的語音結果。以下分述各階段所會碰到的問題。

四、問題

4.1 斷詞及詞性標註

國語文句先做斷詞，其原因在於對譯的基本單為是詞，而不是字。如：

今天　今仔日 (kin-a2-jit8) ○
　　今天　(kin-thian) ×

本系統所使用的斷詞及詞性標註系統乃由中研院平衡語料庫(CKIP, 1995)訓練而得。斷詞正確率及未知詞 (如人名、地名等專有名詞) 的辨識率將影響翻譯結果。

4.2 選詞歧義性

系統中所用國語閩南語對譯詞典為「台華對譯詞庫」，為鄭良偉教授所主持發展的。在詞典中，國語對應至閩南語仍有選詞歧義性存在，例如：

會　e7, e7-tang3, e7-hiau2, e7-tit-thang, hoe

其中 e7-tang3 和 e7-tit-thang 是同義詞，其餘則不是。我們對歧義度做了統計，前一千個最常用的國語詞，其對應候選詞數平均有 2.49 個；若再以詞頻加權，則候選詞數更升至了 3.51！這表示選詞仍是一項嚴重的問題。

要降低候選詞數，上下文的資訊是必須的。因而閩南語詞間同時出現的機率便很重要，此即為馬可夫模型 (Markov Modek)。然而這方法而要大量斷過詞的閩南語語料庫來訓練出可靠的機率值，這是目前最缺乏的資源。在這種情形下，我們只能以詞性降低候選詞數，再以範例做為選詞的知識。詞典中收有詞性的資訊，考慮之後，候選詞數降為 2.02 和 2.28，是一項很有用的資訊。選詞範例則取自「國語常用虛詞及其台語對應詞釋例」(鄭良偉, 1989)，這本書取材自「現代漢語八百詞」(呂淑湘, 1980)，書中有國語各常用詞，例句和其閩南語對譯例句，做為我們選詞的知識。

4.3 未收錄詞的翻譯

在前一千名最常用的國語詞中，竟有 109 個詞未被對譯詞庫收錄，更別提人名地名等專有名詞，因此需要處理未收錄詞的問題。

當碰到未收錄詞時，我們便查詢

單字音典，將之逐字讀出。單字音典收錄了 5,680 個常用中國字的閩南語讀音，如：

資 chu

讀音參考「國台雙語辭典」(楊青矗, 1992)。單字音典中並有文白讀、漳泉腔等讀音資訊，增加翻譯結果的多元性。

4.4 變調

中文屬音調語言的一種，變調是一個很常見的問題。在國語中，當兩個三聲字緊接在一起時，前面的三聲會變調為二聲，如：

導演 dao3-ian3→dao2 ian3

一個字單獨被唸出時所用的調即稱為本調。在閩南語文句中，大部份的字都會變調，只在下列情形時會讀本調：

(1) 單獨唸一個字
(2) 單獨唸一詞時詞尾的字
(3) 語法段落的最後一個字(楊秀芳, 1991; 鄭良偉, 1993)，包括句子的最後一字

舉個例句：

他	今天	心情	很	好
伊	今仔日	心情	真	好
i	kin-a2-jit8	sim-cheng5	chin	ho2
i7	kin7 a jit8	sim7	cheng5	
	chin7	ho2		

其中“日”和“情”因為在名詞片語尾、“好”在句尾，所以不變調，其餘所有字均變調。

閩南語中的變調規則已經有其定論，歸納有一般變調規則、仔前變調規則、輕聲調變調規則及三疊形容詞變調規則等，又因腔調而不同(楊秀芳, 1991; 鄭良偉, 1993)。問題在於必須決定何處該不該變調、應用何變調規則。目前我們的策略是：

(1) 若詞中有“仔”字，則前一字依仔前變調規則變調。
(2) 三疊形容詞第一字依三疊形容詞變調規則，第二字依一般變調規則。
(3) 名詞詞尾讀本調(暫行)。
(4) 其他字依一般變調規則變調。

其中第(3)條為暫時規則，未來將加入淺層的文法剖析，可更清楚地找到名詞片語等語法段落的邊界，段落尾字讀本調，餘者變調。

另外我們並沒有處理輕聲調變調的情形，原因在於同樣的文句會因語意的不同而有不同的唸法，如：

無去 bo7 khi3 (沒去)
bo5=khi0 (不見了)

此時需考慮到語法結構和上下文，限於閩南語資源的缺乏，目前僅將“的”的前一字維持本調。

五、成果

本實驗的終極目標乃在建立「臺灣本土語言互譯及語音合成系統」，目前已經整合國語斷詞與語音合成系統、國語—閩南語翻譯系統，以及最近和台大客家社合作發展的國客語翻譯系統。這些系統都在網際網路作線上展示，網址為：

http://nlg3.csie.ntu.edu.tw/TWmain.html

六、誌謝

非常感謝鄭良偉教授，提供我們台華對譯詞庫。

七、參考文獻

Chao, Y.R. (1968), *A Grammar of Spoken Chinese*, University of California Press, 1968, pp.13-14.

林川傑(1997), 國語-閩南語機器翻譯系統之研究,碩士論文,國立台灣大學資訊工程學研究所,1997.

CKIP (詞庫小組) (1995), “研究院語料庫的內

容及說明,” 中文詞知識庫小組技術報告#95-02, 中央研究院, 1995.
丁邦新(譯) (1980), 中國話的文法, 中文大學出版社, 香港, 1980.
中國大百科全書, 語言文字編, 中國大百科全書出版社, 北京, 1988, pp.227-228.
呂淑湘 (1980), 現代漢語八百詞, 商務印書館, 北京, 1980.
楊秀芳(1991), 臺灣閩南語語法稿, 大安出版社, 台北, 1991.
楊秀芳(1996), 閩南語研究本字專案, 教育部, 1996.
楊青矗(1992), 國台雙語辭典, 敦理出版社, 台北, 1992

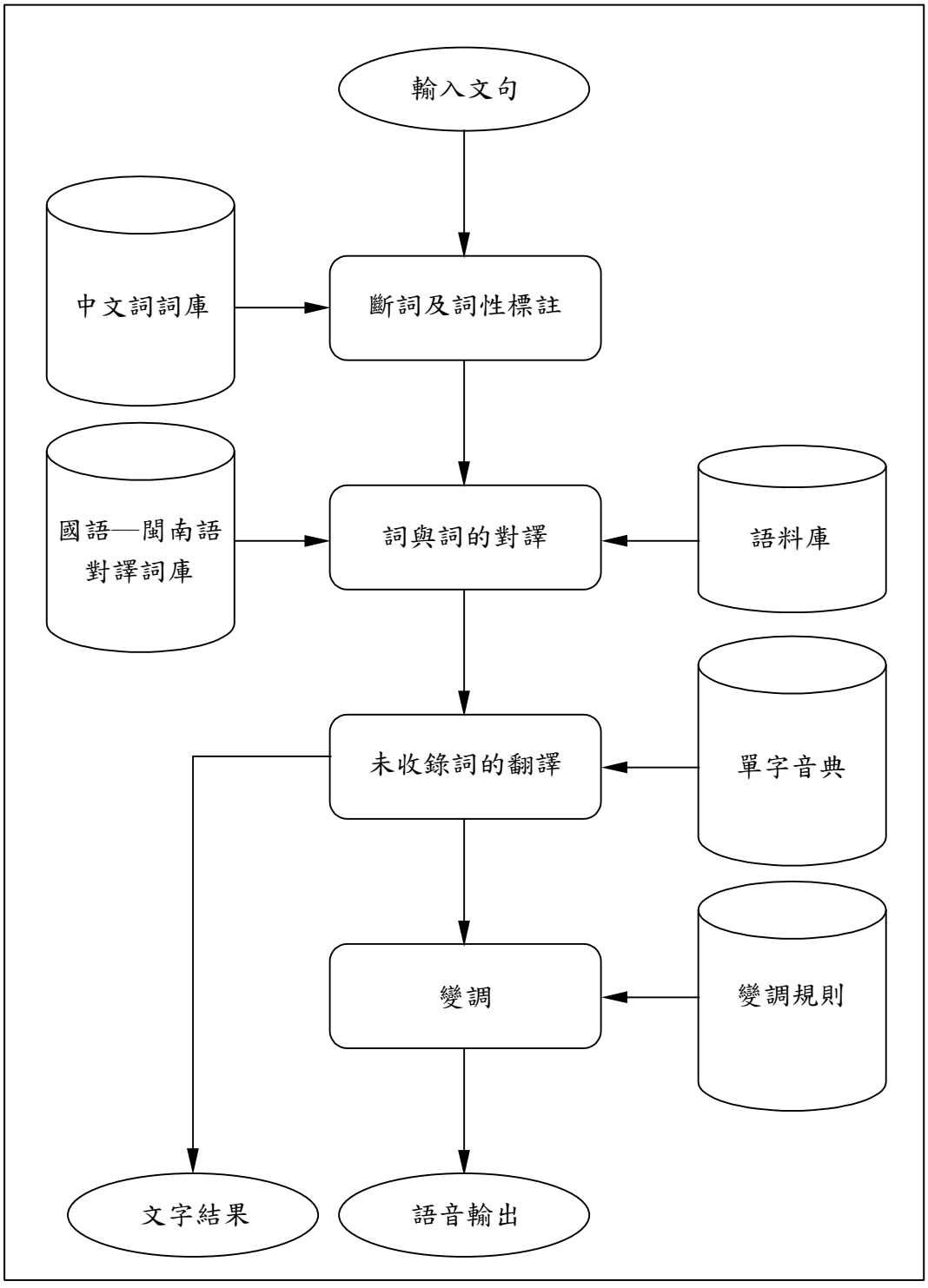

圖 1　國語閩南語機器翻譯系統之架構