國客語翻譯及讀稿系統之設計與實作

作者：張凱揮、范潤萃、郭斯彥

學歷：國立臺灣大學電機工程研究所碩士、美國密西根州立大學流行病學研究所碩士、美國伊利諾大學香檳分校博士

所屬單位：
張凱揮、范潤萃：國立臺灣大學客家研究社
郭斯彥：國立臺灣大學電機工程研究所教授

簡介

本文將提出一套「國客語翻譯及讀稿系統」。利用「客語字典」及「客語詞典」爲其有限的資料庫，配合客語語音檔，它可以將國語依文讀或白讀方式翻譯成客家話並附註音標，再將翻譯結果以客語語音讀出。其使用的翻譯方法主要是長詞優先的詞代換，並對於語音輸出時的客語變調問題加以處理。文中提出翻譯時所遭遇的困難，也說明了克服的方法，並以其爲下一步改進的目標。另外，本文還提出一新方法，探討如何利用聲韻學原理，讓客語亦能使用國語的資料庫，此乃本論文的一大創見。本系統尚有一些可改進的地方，但基於客語和國語相似性甚高，多數基本句子皆能被正確地翻譯，因此本系統在客語學習及其他應用上，將可以提供極大的幫助和效益。

1. 系統介紹

1-1 開發環境及介面

本套系統爲於 UNIX 環境下利用 C++撰寫的 CGI，以網頁形式做爲與使用者溝通的介面。語音輸出的部份則利用 Java Applet 撰寫，利用 Java 的音效功能播放客語語音。

使用者先在網頁上輸入欲翻譯的國語[1]語句，接著交由 CGI 處理，處理之後會動態產生一網頁。該網頁中內含一固定的 Java Applet，其參數爲語音檔檔名，按下 start 鍵後該 Java 程式會於每秒鐘放一個聲音檔而產生出語音。

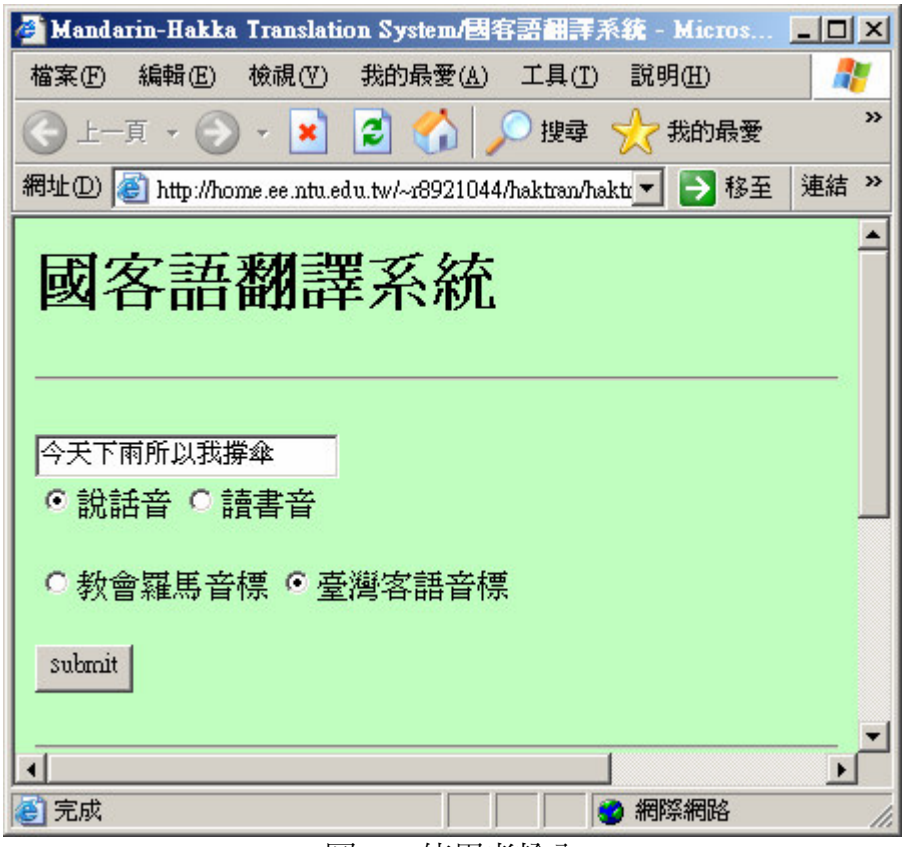
本套系統以一國語文句爲輸入，待翻譯完成之後，系統會輸出客語文句及其音標。客語以四縣腔爲準[1]，支援教會羅馬音標及臺灣客語音標[2]，聲調利用趙元任五度制調值法標記。本文中的國語拼音使用漢語拼音，客語拼音除資料庫使用教會羅馬音標將特別註明外，其餘皆使用和漢語拼音較接近的臺灣客語音標。文中所使用之音標聲母與國際音標對照表如附錄。

1-2 使用方法

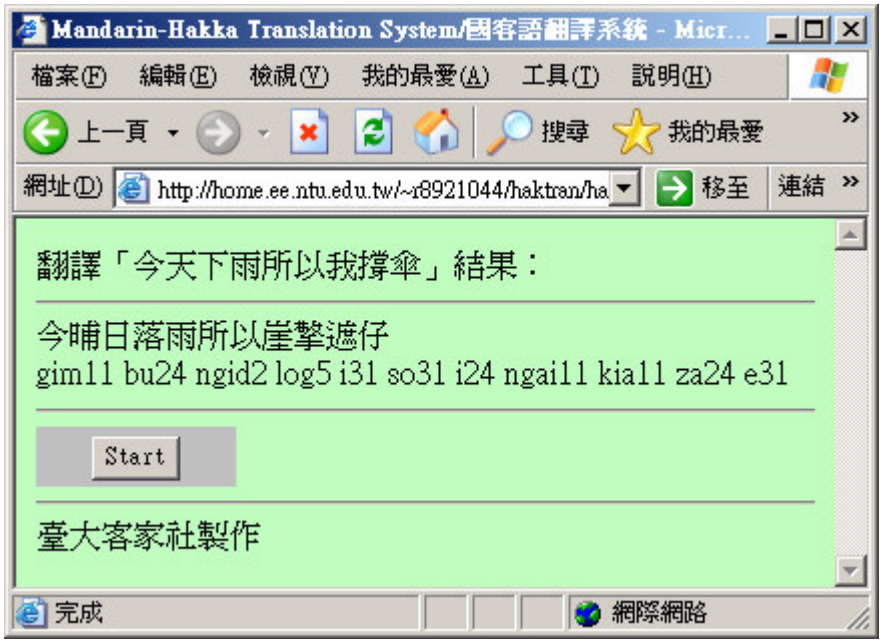
本系統的網址爲 http://club.ntu.edu.tw/~hakka/中點選「華客語翻譯系統」。

首先在網頁中打入想要翻譯的國語文字，例如「今天下雨所以我撐傘」。然後選擇要翻譯爲說話音（白讀）或者讀書音（文讀），接著按網頁上的「submit」鍵將資料送出（如圖一），之後畫面上就出現翻譯之後的客語文字和音標。再按下網頁上的「start」鍵，即可聽到該文句以合成的客語語音讀出（如圖二）。

[1] 以北方官話中的北京話爲基礎



圖一、使用者輸入



圖二、翻譯結果

翻譯分成文讀和白讀兩種。文讀爲逐字翻譯，將每一個字以客語讀音讀出。白讀則會將國語句型和詞彙代換成客語。例如在上例中，文讀只是將「今天下雨所以我撐傘」的讀音查出，輸出結果爲「gim11(gin24) tien11 ha24(ha55) i31 so31

i11 ngo11 cang55 san31」。而白讀會翻譯成客語，其文字爲「今晡日落雨所以崖擎遮仔」，而其讀音爲「gim11 bu24 ngid2 log5 i31 so31 i24 ngai11 kia11 za24 e31」。

有時一個字會有兩種以上的唸法（一字多音），或是一個國語詞會有兩種以上的客語詞對應（一詞多譯）。在這種情況下，多音或多譯的部份會用小括弧括起來，表示有多種對應。在語音合成時括弧中的部份會跳過不唸。例如在「gim11(gin24) tien11 ha24(ha55) i31 so31 i11 ngo11 cang55 san31」這一句話中，實際唸出的語音爲「gim11 tien11 ha24 i31 so31 i11 ngo11 cang55 san31」 。

2. 翻譯原理

2-1 資料庫

本詞典的資料庫使用的是教會羅馬音標，因此本節的範例使用的是教會羅馬音標，在輸出時若選擇臺灣客語音標再由程式自動轉換。國客語翻譯系統的資料來源主要有兩個：其一爲客語字典[3]，其二爲客語詞典[4]。在客語字典中紀錄了每一個中文字的客語讀音，而詞典則是紀錄了國語詞和客語詞的互相轉換關係。兩者的格式相同，爲“國語=客語=客語音標”，欄位中間用等號隔開。若一個國語字有兩個以上的客語翻譯，則利用逗號隔開。如“下=下,下=ha24,ha55”。字典檔的範例如表一，詞典檔的範例如表二。在字典檔中本來只需記載一個字的客語讀音，若遇到客語破音字，則將讀音一一列出即可。但爲求和詞典檔格式相同，所以每一個國語字在客語文字的部份仍會重複，每多一種讀音也會增加相對應的客語字。

又=又=iu55
三=三=sam24
下=下,下=ha24,ha55
丈=丈,丈=tshong24,tshong55
上=上,上,上=song24,song31,song55

表一、客語字典檔範例

一大早=打早=ta31 tso31
一樣=共樣=khiung55 iong55
下雨=落雨=lok5 i31
土地公=伯公=pak2 kung24
生病=破病,得病=pot2 phiang55,tet2 phiang55

表二、客語詞典檔範例

詞典中的資料包括有以下三類：

1. 國語和客語說法不同的詞，如國語「下雨」客語爲「落雨」。

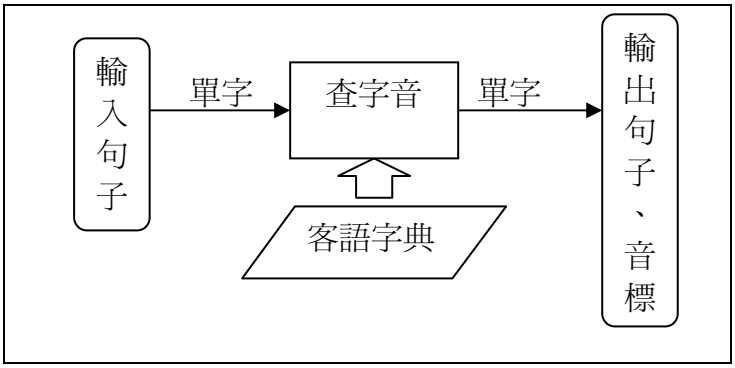
2. 國語和客語說法相同的詞，如國語「苦瓜」客語亦爲「苦瓜」。

3. 國語和客語皆有該詞，但意義不同的詞。如國語「客人」一詞客語中亦有，但指的是「客家人」之意，故國語的「客人」在客語中應翻譯爲「人客」。

詞典中的資料以第一類及第三類爲主，第二類收錄的目的主要是爲了協助斷詞。

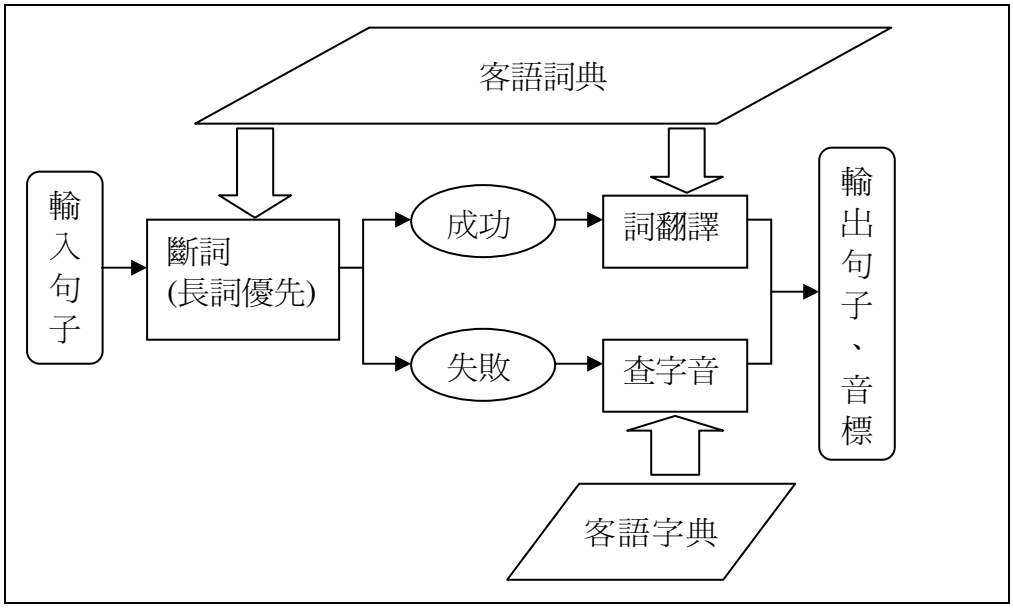
2-2 翻譯演算法

翻譯分爲文讀和白讀兩種。其中，文讀翻譯較爲簡單。做法是直接查詢該中文字的客語讀音，再加以輸出即可。其過程可用圖三表示。



圖三、文讀翻譯

白讀翻譯則較爲困難。第一步需要斷詞，也就是把一句中的每一個詞都斷出來。例如本文的例子應該斷爲「今天 下雨 所以 我 撐 傘」。目前的斷詞方法是查詢客語詞典。如果在詞典中有找到該國語詞相對應的客語詞，就利用該客語詞代換。在代換時是以長詞優先爲原則。例如「我們」在客語中對應爲「恩」，而「我」對應爲「崖」。若是使用者輸入的詞句爲「我們」，則要用「恩」代換而不是拆成「我」「們」再分別代換。若是有查不到的詞，則假設該詞的客語用法和國語用法相同，然後直接查詢客語字典，再將字音輸出即可。其過程可用圖四表示。



圖四、白讀翻譯

2-3 變調處理

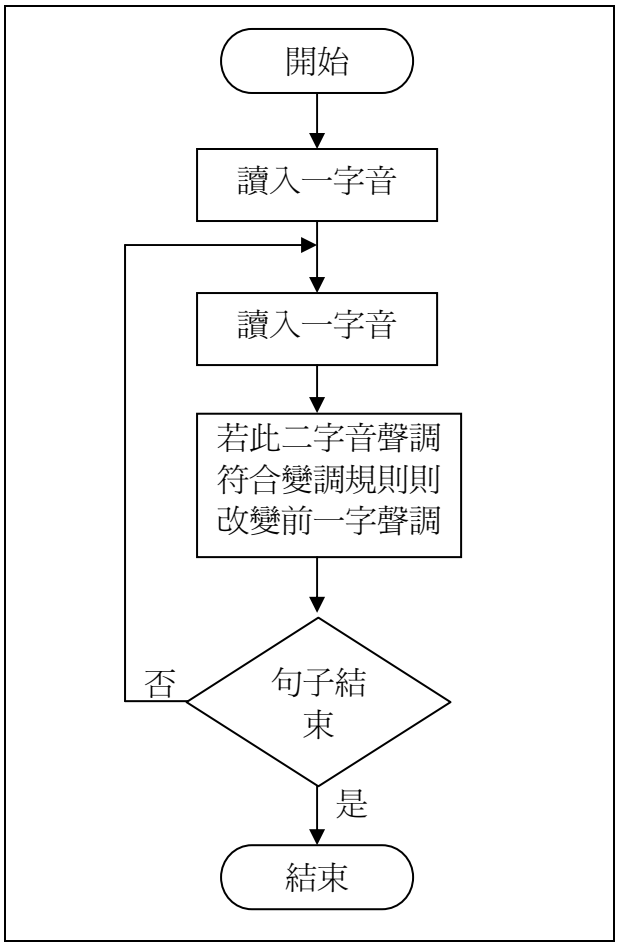
不管是文讀或是白讀，在語音合成前皆有變調的問題要處理。所謂的變調，是指某些調值的音在連續發音時，聲調會改變的現象。例如在四縣腔中，陰平接陰平會變成陽平接陰平，寫成調值即爲 24+24→11+24。例如「東西」一詞照字音唸本來是「dung24 si24」，但實際上要唸成「dung11 si24」才對。在程式處理上是從左而右處理，碰到需要變調的情形時就加以變調。

客語四縣腔的變調規則如表三。

變調規則	調值變化	舉例
陰平+陰平→陽平+陰平	24+24→11+24	東西 dung24 si24→dung11 si24
陰平+去→陽平+去	24+55→11+55	思念 sii24 ngiam55→sii11 ngiam55
陰平+陽入→陽平+陽入	24+5→11+5	音樂 im24 ngog5→im11 ngog5

表三、四縣腔變調規則

變調之演算法如圖五所示：



圖五、變調演算法

2-4 語音合成

經過上面幾個步驟之後，就可以得到一句完整的音標。此時系統將利用這句音標來進行語音合成。方法是預先將所有可能出現的音全部錄下，存成一個一個的聲音檔，以其音標爲檔名。此時將這些音標檔一個一個依序放出來即可。在網

頁上是利用 Java 程式達到這個功能，將每個聲音檔每隔一秒放出。如此即完成所有的翻譯和語音合成步驟。

3. 系統實作問題討論

目前「國客語翻譯系統」還在初步發展的階段，仍然有一些需要改進的地方，並在此節一一討論如下。

3-1 斷詞

目前是利用客語詞典中的詞庫做斷詞的動作。由於客語詞典中的詞大部份爲客語和國語不相同的詞，所以客語和國語相同的詞有可能無法正確地斷詞，於是在詞彙的翻譯上會產生問題。例如國語的「美」一詞，在客語中通常會翻譯成「靚」，但是當「美」和其他的字合成一個名詞時，就不須做此翻譯了。例如「美國」是專有名詞，應該不要翻譯而直接用文讀輸出。在目前的做法中，因爲「美國」一詞不在詞典中，也就不能正確斷詞，而會將「美」和「國」斷成兩個詞。因此會將「美國」翻譯成「靚國」，這樣就不正確了。如果有國語的詞庫，知道「美國」是一個名詞，就可以避免上述情形的發生。

此外，斷詞上還可能有混淆的情形。另如在「她的美國際公認」一句中，雖然「美國」是專有名詞，但把這句斷成「她的 美國 際 公認」是不合理的。正確的斷詞法應該是「她的 美 國際 公認」。對於這樣的問題，需要從句法和語意層次來解決。美國賓州大學的中文句結構樹資料庫(The Penn Chinese TreeBank)中對這個問題有詳細的探討[5]。

賓州大學的中文句結構樹資料庫可以用來對中文句子做斷詞(word segmentation)、詞性標記(part-of-speech tagging)以及文法分析(syntactic bracketing)。在斷詞上，他們使用統計法並利用最大熵策略(maximum entropy

approach)[6]來訓練斷詞器。根據一個字在詞中的位置，他們將該字分成四種型式，分別爲 LL(left，該字在詞的最左邊)、RR(right，該字在詞的最右邊)、MM(middle，該字在詞的中間)及 LR(single-character word，單字詞)。以「產」字做例子，四種標記法如表四所示：

位置	標記	舉例
Left	LL	產生
Single character word	LR	產 小麥
Middle	MM	生產線
Right	RR	生產

表四、四種斷詞標記

因此「她的美國際公認」的兩種斷詞法可以分別標記成以下兩種形式：

她/LL 的/RR 美/LL 國/RR 際/LR 公/LL 認/RR（她的 美國 際 公認）

以及

她/LL 的/RR 美/LR 國/LL 際/RR 公/LL 認/RR（她的 美 國際 公認）

在目前的斷詞器中，他們使用了字之前的兩個字、之後的兩個字和之前的斷詞標記來做斷詞。利用這項技術，以 238000 字做訓練及 13000 字做測試，他們達到了 94.89%的斷詞正確率。

3-2 語法翻譯

在國語句法和客語句法不相同之時，語法的翻譯會出現問題。例如國語中表示「只有、只是」的概念，在客語中則是於句尾加「定定」。如「我只有一隻表」，在客語的講法爲「崖有一隻表定定」。目前，本系統對於此用法就沒有加以處理。現在還很難自動地從國語句子中找出「只有」這樣的概念，這是語法翻譯中最主要的問題。如果光是在發現「只」這個字時將它刪除，而在句尾加「定定」，這種做法也不一定正確。例如「這是一只戒指」就不能翻成「這係一戒指定定」。這部份應該可以利用諸如 Gale 的字義辨異的演算法[7]等自然語言處理常用的演

算法[8][9]加以解決。然而，要使用這類演算法，首先必須建立相當大的語言資料庫，並利用這個資料庫來訓練這些演算法，之後才能正確地判別所要解決的字義問題。

3-3 一字多音及一詞多譯

一字多音就是一個中文字有多種不同的客語讀法，而一詞多譯是一個國語詞有多種客語譯法。目前的做法是利用亂數隨機選一個。在一詞多譯的情形下，如果對應的客語詞並沒有意義上的差別，那麼這是正確的。例如「下雨」可以有「落雨」、「落水」兩種講法。但有時是不正確的。例如「很」一詞在客語中可對應爲「一」、「當」、「恁」等，但意義不是完全一樣，所以實際上並不是用亂數選一個就可以。一字多音也有相同的問題，例如「正經」的「正」一定要唸成 ziin55，而不能唸成「正月」的 zang24，否則就不正確了。

以上仍是字義辨異的問題，可從以下兩方面著手解決。在一字多音方面，有時破音字代表著不同的詞性。例如「行」一字，做動詞用時唸爲「hang11」，如「行路」。而做爲名詞用時唸爲「hong11」，如「銀行」。若能知道原本字的詞性，就能找出正確的翻譯。除了詞性不同而有不同的音之外，一字多音還出現在文讀和白讀上。例如「平」這個字有文讀 pin11 和白讀 piang11 兩種唸法。遇到這種情形，最好的方法還是增加資料庫的量，將不同的用法紀錄下來。至於一詞多譯的情形就複雜的多。比方說上例「很」的客語翻譯，或是語氣詞如「誒、了」等就不容易翻譯的恰到好處。對於這種情形，可以參考 Bender[10]對國語中「把」字的分析及處理。利用不同的語法標記將該字不同的使用情況視爲不同的字，然後利用語料庫做訓練來加以學習。

3-4 變調處理

客語的四縣腔變調十分簡單，除了少數特別的用法之外，依照本文的演算法去處理，通常不會有什麼問題。而有些固定的特殊變調如三疊詞「頭剃到光光光」中的「光光光」，只需將該詞的唸法加入資料庫即可處理。

但其他客語次方言並非如此，變調處理也就相對地較爲困難。例如海陸腔上聲變調在「強調」時不變調，一般語氣時要變調。如何判別是否爲「強調」語氣，是非常不容易的。而大埔腔中的超陰平和去聲變調，這兩種特殊的變調也無法用規則推衍出來。關於這些問題，目前還沒有更好的解決方法，尚待進一步的研究。

3-5 語音合成

第五個問題爲語音合成。目前的做法是預先將聲音錄好再放出。正確的音一個個被播放出來，卻常常讓人有間斷而不自然的感覺。未來可以將聲音檔重新整理，刪除前後空白的部份，使得兩個音之間的間隔不會太長。此外還可以導入語音合成的演算法，將聲調和語音用數學模式合成，這樣將可以得到更真實的聲音，也可以節省較多的記憶空間。

4. 語言資料庫問題探討

4-1 前言

建立一個語言的語料庫是相當費時且困難的工作，但對自然語言處理而言，一個足夠大的語料庫絕對是必需的。因爲不同語言的語料庫之間無法相通，所以想要對一種語言做處理，就必需針對該語言建立語料庫。比方說英文的自然語言處理不能使用中文語料庫，或日文的自然語言處理不能使用法文語料庫等等。

然而中文是相當特別的語言，其中包含了七大方言。方言之間雖不相同，但彼此之間並不是毫無關聯的。目前七大方言中以國語（北方官話）[11]的語料庫

最爲齊全。若要針對其他方言做自然語言處理，則可利用方言之間互有關係的特點來使用國語語料庫，如此將會爲方言的自然語言處理帶來極大的便利性。在本章中將討論國語和客語之間的聲韻關係，以及如何利用國語語言資料庫來解決客語中自然語言處理的問題，如一字多音的問題。

4-2 廣韻音系

要了解爲什麼共用語料庫的方法可行，就必須先了解廣韻音系[12][13][14]。廣韻音系爲漢語聲韻學上「中古音」的代表，代表了唐朝時中國的語音系統。因爲廣韻一書編輯的原則是「若兩方言有分則從其分」，所以除了閩南語之外，現代方言中的各種語音特徵幾乎都可以在廣韻音系中找到。閩南語因爲年代比較久遠，說話音中有一部份的特徵是廣韻音系之前的上古音，就沒有辦法完全適用[15]。

廣韻是利用反切法記音，反切上字代表聲母，反切下字代表韻母和聲調。例如「廣」爲「古晃」切，也就是說這個音的聲母和「古」相同，而韻母和聲調和「晃」相同。拿「古」的聲母 g 和「晃」的韻母及聲調 uang214 組合起來的音爲 guang214，即廣的音。經過語言學家的分析，中古所有的聲母、韻母和聲調都已經差不多區別和歸類出來。根據這些資料，就可以推出一個字在中古時大概是唸什麼樣的音。

中古音在演變爲各種現代方音時是有規律的，依據不同的規律即演變成不同的方言。因此不同方言之中看起來很不相同的特性，卻極有可能來自於同一個語源。在有國語語料庫的情形下，若能確定一個想要解決的問題和國語有著相同的語源，就能利用國語語料庫來尋求解決。

例如「教」一字，在廣韻中有兩種唸法。第一種爲古孝切[16]，見母去聲效韻，韻鏡[17]外轉第二十五開。國語中唸法爲 jiao51，去聲調。在客語中唸法爲

gau55，亦爲去聲調，如教室。第二種唸法爲古肴切，見母下平肴韻，韻鏡外轉第二十五開。在國語中唸法爲 jiao55，陰平調。在客語中唸法爲 gau24，亦爲陰平調，如教書。在這個例子中，國語和客語的讀法皆符合中古到現代的語音演變原則，故知其爲同源音。因此可以直接借用國語語言資料庫到客語之中，以解決客語一字多音的讀音判別問題。

這個方法最大的困難在於，如何決定一個字在不同方言中的發音，是否同源自中古音。每一個方言在分化的過程中，或多或少都會有該方言獨特的字音或詞彙出現，這些字音和詞彙就不能利用另一個語言的資料庫來處理。若不能找出這些字音和詞彙，就可能得出錯誤的結果。例如「林」字，在客語中有 lim11 和 na11 兩個音。這個字在廣韻中記載有「力尋切」一個音，也就是說只有 lim11 這個音是源自中古音的。在國語中也只有 lin35 一個音。因此並不能利用國語的資料庫來判別客語中的「林」字應發何音。

若有廣韻音系的資料庫，就能直接確定一個字音的來源，甚至可以直接由廣韻記載的資料來推測該字的發音。由於廣韻音系語料庫的建立是相當不容易的，所以後面提出的方法將著重在：如何利用聲韻學上的規則，推衍出兩不同方言的字彙是否同源於中古音。若是，則可直接利用另一方言的語料庫。

中古音到現代方音演變的學問相當複雜，因此以下只提出一個例字加以說明。其他的字可依照語言學的規則，依循下例的方法來加以判別。

4-3 推衍規則

以「教」字爲例。首先敘述其從中古音演變爲現代方音的規則，之後敘述現代方音倒推中古音的方法，最後說明如何判別是否爲同源的規則。

4-3-1 中古到現代之規則

「教」爲「古孝切」及「古肴切」。其聲母皆爲牙音全清見母字，韻母皆爲效攝二等，僅聲調不同，分別爲去聲及平聲。

在國語中，牙音全清見母字讀爲 g 聲母。而效攝二等韻，除影母字外皆唸爲 iao 韻母，所以應爲 giao 的音。但國語中牙音聲母的細音字會鄂化爲舌面音，故應唸爲 jiao。在聲調上，中古去聲在國語唸爲去聲，而中古平聲清音聲母在國語唸爲陰平，所以古孝切的音在國語應唸爲 jiao51，古肴切的音在國語應唸爲 jiao55，和國語中的「教」字音韻完全相合，故知此二音皆源自中古音。

在客語中，牙音全清見母字讀爲 g 聲母，而效攝二等韻唸爲 au，所以應爲 gau 的音。在聲調上，中古去聲在客語唸爲去聲，而中古平聲清音在客語唸爲陰平，所以古孝切的音在客語應唸爲 gau55，古肴切的音在客語應唸爲 gau24，和客語中「教」字的音韻完全相合，故知此二音皆源自中古音。

4-3-2 現代到中古的反推

整理所有類似 4-3-1 節中所列出的規則，即可得到中古音在現代方音中應當發什麼音。相反的，倒著推衍這些規則，也可推知一個音在中古的來源有哪些。仍以「教」的兩個音爲例子。

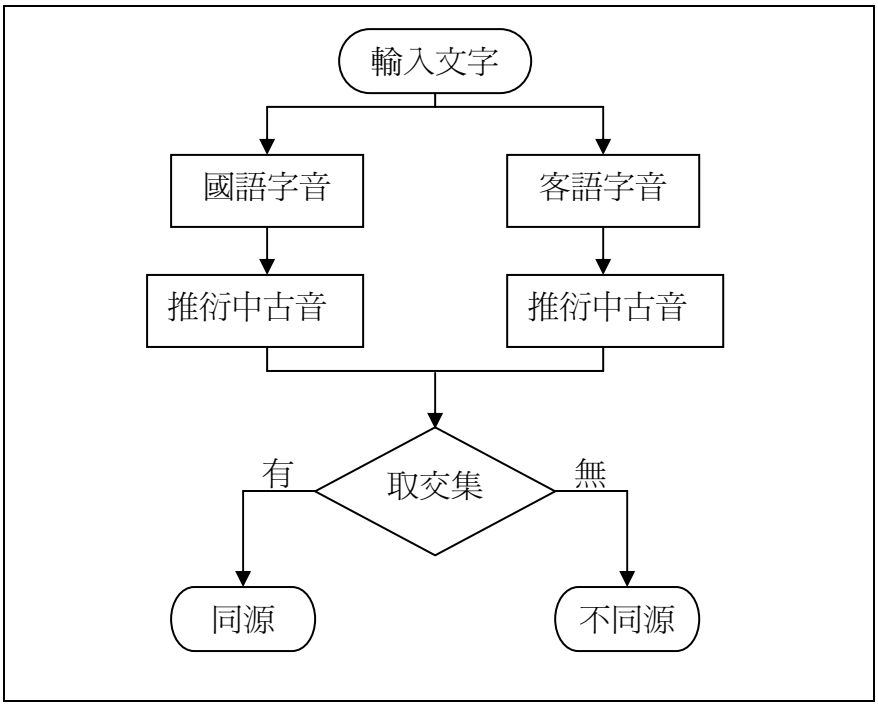
在國語中，「教」唸爲 jiao55 和 jiao51 兩音，分別爲陰平及去聲調。在聲母上，因爲 jiao 爲細音而非洪音，所以這裡的 j 聲母可能的來源有中古精母、從母、見母及群母。在韻母上，iao 的來源有效攝二三四等及流攝三等。國語的陰平調在中古音有平聲和入聲兩個來源，而去聲調在中古有全濁上、去聲和入聲三個來源。

在客語中，「教」字唸爲 gau24 和 gau55，分別爲陰平及去聲調。在聲母上，g 只有中古見母一個來源。在韻母上，au 的來源只有效攝二等。客語的陰平調來

源有中古的清音平聲、部份次濁上聲及部份去聲，而客語去聲在中古有全濁上聲和去聲兩個來源。

4-3-3 同源判別的規則

同源判別的規則，首先要從該方言的讀音反推出中古可能的來源，也就是中古可能唸哪些音。若一個字在兩不同方言中倒推中古音，發現兩者有交集，即可判別其爲同源字。例如在「教」這個例子中，在音韻上國語和客語有「中古見母效攝二等」這個交集，而聲調上陰平聲的有「中古清音平聲」這個交集，唸去聲的有「中古全濁上或去聲」這個交集，故可知其爲同源。其演算法以圖六表示。



圖六、同源判別演算法

判別出一個字在客語和國語中的讀音爲同源之後，就可以利用國語語料庫來解決客語中一字多音的問題。例如在「教室」一詞中，「教」字國語唸去聲，客語亦可判別爲去聲。在「教書」中，「教」字國語唸陰平調，客語亦可判別爲陰平調。

4-3-4 例外情形處理

雖然漢語中的各方言都可以從中古音系找到規則，但各方言長時間個別演化的結果，也會各自有例外的情形發生。舉「畜」字爲例，在國語中當名詞用時讀爲 chu，而作動詞用時唸 xu。客語中的「畜」字有 hiug 及 cug 兩讀，而廣韻中有許竹、丑六及許救三切。國語和客語的讀法皆源自中古音。但國語和客語間「六畜興旺」一詞中的「畜」字唸法並不相同。國語唸爲 chu，但客語唸爲 hiug。如果利用上面的原則判別，在客語中就會得到不正確的音。此時可以將其讀音加入客語資料庫，以避免這樣的例外情形發生。

但更一般性的方法是建立中古漢語的讀音資料庫。若是能從中古漢語直接判斷讀音，就能避免所使用的另一方言語料庫特殊變化帶來的錯誤。例如「六畜」一詞在康熙字典中爲許六切、許救切，在客語中應讀爲 hiug 音。這樣就能得到正確的讀音。

因而在實際應用上，若能先建立中古漢語語料庫，再建立中古漢語到各方言間的個別差異語料庫應是最有效的做法。

5. 應用

目前「國客語翻譯及讀稿系統」還在起步和發展的階段。相信在可預見的未來，待此系統發展更加完善成熟之時，它將在許多方面被廣泛應用，爲人們的日常生活帶來便利，也將爲客家話的傳承與發揚做出極大的貢獻。

5-1 幫助學習客家話

本系統應用於客語教學和學習，相信將產生極大的效益與成果。對於一客語初學者而言，這套系統好比一部詳盡的字典，輸入不會唸的詞句，立刻可以知道

其客語發音。不但能利用聲音記音，還能利用音標來學習發音。對於一客語教學者而言，這套系統如同學生的隨身小老師，可以協助老師教學，糾正學生發音，並可當作複習客家話的有聲書。對於有心傳承母語的家長們而言，這套系統就如同一本教學指導手冊，一方面可透過語音和字句來傳授客語，二方面也可藉由查閱單詞和音標來印證所學。

5-2 輔助中文識字

考慮到以下的情況，本系統將可被進一步應用到輔助中文識字方面：海內外有些華人彼此只用客家話來溝通；在台灣也有少數人只懂得客家話。此外，一些嫁到臺灣的外籍新娘雖然會說客家話，但很可能不會讀寫中文字，這些情形多少局限了他們的日常生活和人際關係。若能應用本系統，將可以幫助他們學習中文字，以便和他人進行更多的互動與交流。將文句輸入網頁中，以客語的語音讀出，再利用語言認知的原理，就可以幫助他們初步認識中文的字形與字義。

5-3 電視的雙語配音

面對現代多元化的社會，電視節目也應該有雙語甚至多語配音，以滿足各階層觀眾之需求。應用國客語翻譯及讀稿系統於電視配音，一方面可以滿足中壯老年客語人士的需要，還能使客家年長者藉由收看客語配音節目，更容易接受新觀念、新資訊。另一方面，將使客家子弟以及有心學習客語的人，能夠經由客語配音的電視節目來學會客家話。

5-4 電台廣播的配音

這套系統能逐字逐句地將國語詞句翻譯成客語發音，所以可應用在廣播電臺的一些例行播放中，例如電台台呼以及中間串場的廣告。若廣播電臺找不到合適的客語播音員，也可以應用國客語翻譯及讀稿系統來做客語配音。另一方面，利用此系統將預錄的節目翻譯成客語，可解決電台在深夜時缺乏客語播音人員的困

擾。此國客語翻譯及讀稿系統，還可以幫助較無經驗的播音人員在讀稿之前，自我檢閱及修正客語發音，使正確的客家語音能經由廣播傳送出來。

5-5 公共場所客語廣播

政府爲了照顧到社會上各個族群，開始在公共場合中推廣多語廣播。然而，有經驗的專業客語播音員目前仍嫌不足。若能應用此系統，則可協助政府渡過這段青黃不接的時期。在公共場合，例如捷運、火車站、客運站、以及公車上，都可以運用此系統來進行客語的廣播。在某些需要現場廣播的公共場合，例如各大火車站，有時沒辦法找到會說多語的播音人才，此時可運用國客語翻譯及讀稿系統。播音員在輸入中文之後，就可以直接將標準的客語播放出來。

5-6 電話語音轉爲客語

在現代化的社會中，資訊傳播迅速，凡事都講求效率，所以公私立機構和許多公共場合，皆廣泛運用了電話語音系統。例如醫院掛號看病、郵局的郵件查詢、信用卡的諮詢服務等等。由於電話語音的使用十分普遍，若能應用此系統將電話語音轉化爲客語，不但可以爲各公司機構開創一片商機，還可以進一步便利廣大的客家族群。

6. 結論

本文介紹了「國客語翻譯及讀稿系統」的設計方法及運作原理，並對於目前所遭遇的問題、未來改進方向和應用層面做了詳盡的描述與討論。其中利用中古音和國語資料庫協助其他漢語方言翻譯及讀音判別的方法，將有助於減輕方言資料庫不足的問題，這是本文所提出的一大新創見。

在此特別感謝台大資工所陳信希教授和林川傑同學的技術協助，台大客家社葉時霖、劉欣怡同學的資料協助，張景富先生的記音協助以及張其穎同學的語音輸入。

附錄

國語聲母之國際音標和漢語拼音對照如下：

發音方法 發音部位	塞音及塞擦音		鼻音	清擦音	濁擦音 及邊音
	不送氣	送氣			
唇音	p/b	p’/p	m/m	f/f	
舌尖音	t/d	t’/t	n/n		l/l
	ts/z	ts’/c		s/s	
捲舌音	tʂ/zh	tʂ’/ch		ʂ/sh	ʐ/r
舌面音	tɕ/j	tɕ’/q		ɕ/x	
舌根音	k/g	k’/k		h/h	

註：國際音標/漢語拼音

客語聲母之國際音標和臺灣客語音標、教會羅馬拼音對照如下：

發音方法 發音部位	塞音及塞擦音		鼻音	清擦音	濁擦音 及邊音
	不送氣	送氣			
唇音	p/b, p	p’/p, ph	m/m, m	f/f, f	v/v, v
舌尖音	t/d, t	t’/t, th	n/n, n		l/l, l
	ts/z, ts	ts’/c, tsh		s/s, s	
舌根音	k/g, k	k’/k, kh	ŋ/ng, ng	h/h, h	

註：國際音標/臺灣客語音標, 教會羅馬音標

參考書目

[1] 羅肇錦，台灣的客家話，臺原出版社，1990

[2] 古國順、何石松、劉醇鑫，客語發音學，五南出版社，2002

[3] 楊政男、徐清明、龔萬灶、宋聰正，客語字音詞典，臺灣書店，1998

[4] 涂春景，苗栗卓蘭客家方言詞彙對照，國家文化藝術基金會，1998

[5] Nianwen Xue, Fei Xia, Fu-dong Chiou, and Martha Palmer, “The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus”, Natural Language Engineering, 10(4): 1-30, June 2004

[6] Adwait Retnaparkhi, “Maximum Entropy Models for Natural Language Ambiguity Resolution”, Ph. D. Thesis, University of Pennsyvania, 1998

[7] Gale, William A., Kenneth W. Church, and David Yarowsky, “A Method for Disambiguating Word Senses in a Large Corpus”, Computers and the Humanities, 26 pp 415-439, 1992

[8] Gerald Gazdar and Chris Mellish, Natural Language Processing in Lisp, Addison Wesley, 1989

[9] Christopher D. Manning and Hinrich Schutze, Foundations of Statistical Natural

Language Processing, The MIT Press, 1999
[10] Emily Bender, “The Syntax of Mandarin – basb”, Journal of East Asian Linguistics, 9(2), 2000
[11] 詹伯慧著，現代漢語方言，新學識文教出版中心，1991
[12] 陸法言撰，陳彭年等修校，廣韻校本，世界書局印行，隋朝
[13] 董同龢著，漢語音韻學，文史哲出版社，1998
[14] 林燾、耿振生著，聲韻學，三民書局，1997
[15] 周長楫著，閩南語的形成發展及在臺灣的傳播，台笠出版社，1996
[16] 王引之等撰，康熙字典，世界書局印行，清康熙年間
[17] 張麟之刊行，龍宇純校注，韻鏡校注，藝文印書館印行，南宋