中華大學生物資訊學系系統開發專題報告

# 現代國語流行歌詞音韻資料庫之建置

# Building of phonological database of modern Mandarin popular songs

專題組員:許舒雯、李昀霈、傅軍雅

專題編號：PROJ2015-BIOINFO-10104

指導老師:董其樺老師

摘要

本專題利用 PHP+MySQL 建置一個流行音樂音韻資料庫，以國語金曲歌手所發表的歌曲為主，利用 PHP 程式去分析計算現代流行歌曲詞句中常用的詞語，並歸類收錄在資料庫中。此外，這些字詞亦紀錄其注音音韻，以韻腳的形式呈現分類。根據此資料庫，使用者可以尋找特定韻腳的單詞，並顯示來源歌曲之範例，協助使用者進行流行樂曲或新詩之創作。

## 1.簡介

一首能深入人心的歌曲，押韻往往是其中主要的關鍵之一，對於押韻的使用，其實早在中國古時就已存在，它是一種常見的創作技巧，在一個韻文作品中，押韻能使其完整、聽起來順耳易記、百聽不厭，是過去至今的詩、詞、曲創作上，不可或缺的重要存在[1]。在一首歌曲中，除了押韻的使用外，詞句的使用也是極為重要的，一首歌之所以能流行，不光只是有個好的押韻技巧，其中還必須有符合當時的慣用詞語，因此流行歌曲是貼近於我們的生活，是各個時代的寫照，也是展現當時代重要的依據。

但就目前可供參考的研究文獻中，我們發現到，對於現代台灣國語流行歌曲的相關的研究數量並不多，其中有針對歌詞文句進行資訊化統計，分析中文歌詞裡常用的字詞[2]。然而，將其資訊化後所歸納整理出的相關資料庫，則是未有相關研究。

在網路上，已有一些資料庫專門蒐集中文語流行歌曲之歌詞，並建置其查詢頁面可供使用者搜尋歌詞，如魔鏡歌詞網[3]。但這些歌詞資料庫僅是傳統倉儲式資料庫，單純記錄每首歌的歌詞，並未針對其文字內容做統計分析。

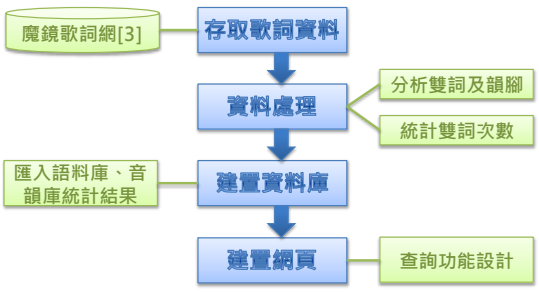
因此本專題對上述議題，進行現代國語流行歌詞數位化的分析與整理，在進行數位化分析與整理中，我們統計歌詞常用的詞彙，並將它與韻腳串聯結合分析，建置一個現代國語流行歌詞音韻資料庫，以供大眾或對於要做此方向研究、創作的人使用。此外，本專題不僅分析近代國語流行歌曲常見的詞語，也提出可供查詢之網頁，分別可進行國字查詢、雙詞查詢、聲韻查詢、歌手歌曲瀏覽等功能，能針對於不同的需求進行使用。

## 2.專題進行方式

本專題從網路上既有的歌詞資料庫進行資料的收集，如魔鏡歌詞網等歌詞資料庫[3]，並依所需內容進行編輯篩選，刪除含有非中文歌詞及過度重複字句等的部分，整理出本專題所需的歌詞資料。我們所收集的內容，是以近幾年第 20 屆至 25 屆的金曲獎

得主作為對象，篩選出所有國語流行歌曲。在歌曲當中，我們會去除部分在同一首歌中重複多次且無意義的字句，或含有非國語文句的句子。待篩選整理後，即為進行資料庫的建置。

我們以 PHP+MySQL 進行歌詞分析程式撰寫與資料庫建置。其資料庫的內容除了含有對於歌詞的統整分析以外，也包含了語料庫、音韻庫等的建置，將歌詞進行音韻標記與分詞整理。關於音韻庫資料的來源是以 CNS11643 中文標準交換碼全字庫[4]所提供的資料進行建置，以其中的文字碼 BIG-5 碼及 CNS 字碼，進行韻腳的標註，找到與此對應的注音和韻腳，再結合所收集的歌詞進行分析，建立成本專題所需的音韻資料庫。關於語料庫的建置，則是參考語言模型的分詞方法(Unigram)的方法進行[5]，將歌詞以一句為單位進行分詞，而分詞的方式，本專題採以兩兩字為一組的方式進行統計分析，並與音韻庫做結合，並運用於後續網頁的建置。



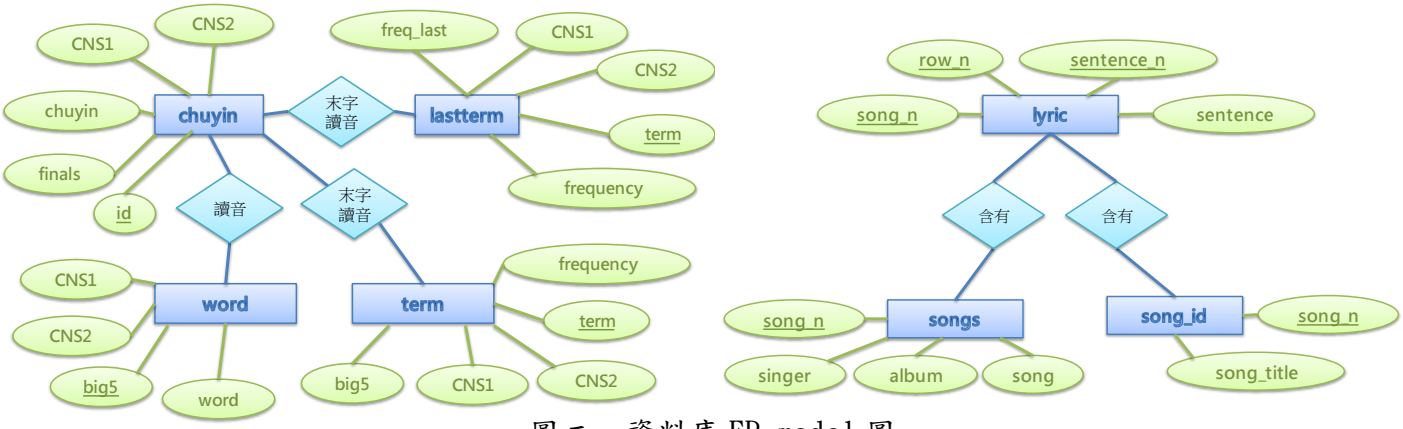

圖一、專題進行流程圖。

如圖一所示，本專題進行的流程分別為收集資料、資料處理、建置資料庫、建置網頁。在收集資料的步驟中，我們先將收集範圍縮小，篩選條件為 20 屆至 25 屆金曲最佳國語男歌手獎、最佳國語女歌手獎、最佳樂團獎、最佳新人獎得獎人，演唱歌曲以國語為主，篩選完後共有 21 個歌手/團體，將這些歌手/團體所收錄專輯裡的所有國語流行歌收集起來。經過蒐集，共得到 276 張專題，1796 首歌，62902 句歌詞。

接著，我們將所收集到的歌的歌詞作初步處理，其中含有特殊符號、標點符號及全形空白等都改為半形空白，以便後續程式辨識。最後在歌詞檔案中，我們將半形空白符號視為一句歌詞的結尾，將換行符號視為一行歌詞的結尾。PHP 程式則分析每一個歌詞檔案，將歌詞以每兩個字做為一字詞單位，記錄此雙詞的出現次數，並且獲得第二個字的聲韻。

分析完歌詞後，我們將其統計資料建置成資料庫。此資料庫，每個資料表的結構，如表一所列。chuyin 資料表主要是紀錄每個字的 CNS 字碼、注音及韻腳。term 資料表主要是紀錄雙詞的出現次數、big5 字碼及 CNS 字碼。lastterm 資料表主要是紀錄雙詞在句尾出現的次數及雙詞第二個字的 CNS 字碼。上述這兩個資料表相當類似，差別在於後者主要統計於句尾之雙詞。

lyric 資料表主要是紀錄句子的來源是第幾首歌的第幾行的第幾句。songs 資料表主要是紀錄歌曲的編號、歌手/樂團名稱、來源專輯及歌名。song_id 資料表主要是紀錄每首歌的編號。上述這兩個資料表的差別，在於後者主要記錄歌詞文字檔之檔名和編號的對應關係。word 資料表之資料則是匯入自 CNS11643 中文標準交換碼全字庫[4]。




圖二、資料庫 ER model 圖

最後，我們使用了網路上公開免費的網頁模板來設計本專題的查詢網頁。在此網頁中，能夠讓使用者進行四種查詢功能：「國字查詢」、「雙詞查詢」、「聲韻查詢」、「歌手歌曲瀏覽」。

表一、音韻資料庫各資料表結構

| chuyin 資料表 | | | | |
|---|---|---|---|---|
| id | CNS1 | CNS2 | chuyin | finals |
| Int (6) | varchar (4) | varchar (4) | varchar (12) | varchar(6) |
| lastterm 資料表 | | | | |
| term | frequency | freq_last | CNS1 | CNS2 |
| varchar (10) | Int (5) | Int (5) | varchar (4) | varchar(4) |
| lyric 資料表 | | | | |
| song_n | row_n | sentence_n | sentence | |
| Int (4) | Int (3) | Int (2) | Varchar (100) | |
| songs 資料表 | | | | |
| song_n | singer | album | song | |
| Int (5) | Varchar (20) | Varchar (100) | Varchar (100) | |
| song 資料表 | | | | |
| song_n | | song_title | | |
| int(4) | | varchar(200) | | |
| term 資料表 | | | | |
| term | frequency | big5 | CNS1 | CNS2 |
| varchar (6) | Int (5) | varchar (4) | varchar (4) | varchar(4) |
| word 資料表 | | | | |
| word | big5 | CNS1 | CNS2 | |
| Varchar (3) | Varchar (4) | Varchar (4) | Varchar (4) | |

3. 主要成果

本專題建置了資料庫及查詢網頁。在網頁中，具有四種查詢功能，且有些可互相連接，供使用者使用更為便利。

第一種功能為國字查詢。如圖三所示，輸入欲查詢國字的注音，可查詢到相同讀音的字，並計算此注音之國字共幾字，列出此國字的 CNS 字碼、big5 字碼，按下國字連結會直接連接到聲韻查詢，記錄歌詞句尾中出現此字的句子共幾句，並列出來源歌曲及句子(圖四)。點選韻腳注音連結會連至聲韻查詢。

圖三、國字查詢

圖四、國字查詢連接聲韻查詢

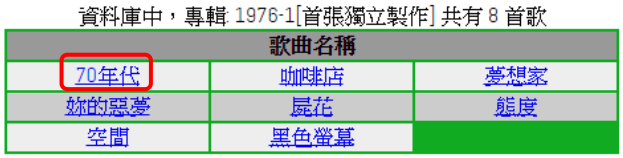
第二種使用者查詢之功能為雙詞查詢。輸入欲查詢的雙詞，可查詢到此雙詞在歌詞中的出現次數，並列出有此雙詞出現的歌詞、此歌詞的來源歌曲、此歌曲之收錄專輯名及歌手/樂團名(圖五)。若按下雙詞連結會直接連接到聲韻查詢，如圖六。此網頁會顯示歌詞句尾中出現此雙詞的句子共幾句，並列出來源專輯、歌曲及句子。

圖五、雙詞查詢

圖六、句尾包含指定雙詞的歌曲列表

第三種功能為聲韻查詢。如圖七，在網頁的查詢欄位中輸入欲查詢的聲韻，可查詢到此聲韻國字共幾字，並列出第二個字符合此韻腳的雙詞、雙詞出現在句中的次數、雙詞出現在句尾的次數及其韻腳等資訊。若按下雙詞連結會直接連接到聲韻查詢，記錄歌詞句尾中出現此雙詞的句子共幾句，並列出來源專輯、歌曲及句子(圖六)。

圖七、聲韻查詢



圖八、專輯歌手及歌曲列表瀏覽

網頁中最後一項查詢功能是歌手歌曲瀏覽(如圖八)，這部分類似網路上常見的歌詞資料庫。不同的是，我們在每句的最後兩個字做了連結，點選連結會連接到雙詞查詢(圖九)。


1976樂團 -- 1976-1[首張獨立製作]

| 70年代 歌詞 |
| --- |
| 是大麻煙還是搖滾樂麻醉了你 |
| 是厭倦還是逃避主宰了生活 |
| 放下來　舉起槍的手　現在擁有的是不是你一直的夢寐以求 |
| 對這不公平的一切的一切　我感到憤怒 |
| 這混亂的世界　這混亂的世界 |
| 放下來　舉起槍的手 |
| 現在擁有的是不是你一直的夢寐以求 |
| 為什麼　愚蠢自私的笨蛋　還是如此麻木 |
| 這混亂的世界　這混亂的世界 |
| 混亂而美麗　混亂而美麗　混亂而美麗的世界 |
| 在最墮落　最骯髒的城市裡　也有我最溫暖的天堂 |

圖九、歌詞查閱

我們在製作本專題上，最容易遇到的狀況是蒐集資料方面我們沒有已建立好的歌詞資料庫可供下載使用，像是我們蒐集歌詞沒有一個歌詞資料庫可以參考，因此需要以人工方式到歌詞網進行手動下載本專題需要的歌手歌曲，依照「歌手>專輯>歌名」方式做存取。

在分析歌詞上，常會發現歌詞中出現半形、全形符號、非中文文字等參差不齊的句子符號參雜其中。由於 PHP 程式辨識全形符號時可能會面臨編碼的問題，所以我們利用人工方式挑出程式可能會辨認錯誤的符號與非中文的句子，以利作業。

有些歌詞作品裡會有 Repeat 字詞表示反覆詠唱，我們會先用人工方式刪除，以利後續程式進行本專題需求的資料存取。或是有的歌詞作品中間會包含我們取材的歌詞網網址，我們在使用 PHP 程式分析存取歌詞時，便利用程式碼編寫跳過將其忽略。

最後我們運用本專題的查詢網頁，進行實例應用。我們嘗試將日常生活中常看見或聽見但較不朗朗上口的廣告詞做些修改，讓原本不押韻的句子有押韻，或是讓原本句尾同字的句子，換成同韻不同字。

例如，在某衛生棉品牌所用的廣告詞：「即使不能隨時換，肌膚也能維持乾爽潔淨」，在第一句的末字「換」與第二句末字「淨」因為音韻不同，因此使得原廣告較不易讓人一聽就熟記。然而，透過本專題的資料庫查詢，我們能快速的找到「換」的音韻為「ㄨㄢ」，並且列出同樣符合此音韻的字詞有哪些。從該列表中，可以挑選適合的字詞來替換原廣告詞，如「雲端」、「柔軟」等。所以最後就能改用如「不能隨時換，肌膚也乾爽柔軟」或「不能隨時換，肌膚也像在雲端」等文案。

又或者，在某相機底片廣告詞提到：「十種表情，百樣心情」，前後句都用了相同的字。為了增加廣告詞中文字的變化，可以查詢「情」字在資料庫中與之相同的音韻的其他字詞。經過簡易的查詢，可以得到諸如「風景」等押同韻的字詞。因此原廣告詞就能試著改寫成「十種風景，百樣心情」。

4. 評估與展望

中文一直都是很優美的語言，一個字會富含許多種意思，所以在字的發音上可能會有一字多音的現象。在本專題中就發現常見的字中，例如「行」就會有「ㄏㄤˊ、ㄏㄤˋ、ㄒㄧㄥˊ、ㄒㄧㄥˋ」等音的不同而造成韻腳的不同，呈現有「ㄧㄥ、ㄤ」的韻腳，但在本專題中還無法辨識這樣一字多音的狀況，所以在查詢的時候會發現不論查哪種韻腳，有關「行」為韻腳的雙詞都會出現。也許未來我

們可以讓每一個雙詞具有意思，去判斷韻腳避免一字多音重複出現的情況。此外，所歸納出來的雙詞是否真的具有中文語意，則可利用中研院平衡語料庫來比對已有意義的字詞[6]。

我們在本專題所挑選的歌曲僅限於近六屆金曲獎得主的專輯歌曲，因此資料庫所統計的雙詞數量不夠完善，而沒有辦法提供最完整的查詢。本專題的發展空間在於繼續擴展我們的歌詞資料庫，抓取更多的歌詞並分析，才能提供越完善的查詢網頁。

目前本專題未利用程式自動化將反覆歌詞刪除，只能以人工的方式篩選，所以我們必須每個檔案都檢查，將手動進行歌詞刪減。但針對於此我們有發現到一些問題，因為在建立資料庫前，一開始就是直接由原檔進行修改，因此在歌手歌曲瀏覽的部分，無法完整呈現原歌詞，對此我們期望將來能在處理歌詞時能進行修正，使我們的歌手歌曲瀏覽的部分能更加的完善。

## 5.結語

本專題著重於歌詞的分析，藉由押韻與分詞的方式來進行歸納整理，統計篩選出常見的詞與韻腳，列出近期台灣流行歌詞常見的詞彙。並利用PHP+MySQL，將這些語料分析統計後建成資料庫，製成國語流行歌音韻查詢網頁，目的在於供欲寫詩、作詞或進行研究參考之人使用。另外，除了現有的國語流行歌音韻查詢網頁外，我們也期待將來能再附加一些其他實用或趣味的新功能。

## 6.銘謝

在此我們感謝董其樺老師耐心的指導與多次提點，在專題的製作討論過程中，讓我們學習到了許多，也感謝在過去所學相關課程中，教導我們的各位老師，讓我們能夠在此次專題中學以致用，使我們獲得了許多寶貴的經驗。最後感謝共同製作的組員，也感謝所有在專題製作期間幫助我們的人，因為有你們才有如今的成果。

## 7.參考文獻

[1] 劉祐銘、陳瑤玲博士，2010，*臺灣國語流行歌曲歌詞用韻研究(1998~2008)*，中國文學研究所碩士學位論文，靜宜大學

[2] Chun-Ju Huang and Joachim Allgaier, *What science are you singing? A study of the science image in the mainstream music of Taiwan.* Public Understanding of Science, 2015. 24(1): p.112-125.

[3] *魔鏡歌詞網*. Retrieved from http://mojim.com/

[4] 國家發展委員會 (2014). *CNS11643中文標準交換碼全字庫*. Retrieved from http://data.gov.tw/node/5961

[5] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *An Introduction to Information Retrieval.* Cambridge University Press, 2009, p.237 - 240.

[6] 中央研究院語言所 (2014). *中央研究院現代漢語標記語料庫 4.0 版.* Retrieved from http://asbc.iis.sinica.edu.tw