中華民國計算語言學學會
通訊 ACLCLP
The Association for Computational Linguistics and Chinese Language Processing
發行人：陳信希
主　編：王新民
執行編輯：黃　琪
第十九卷第二期

E-Mail：aclclp@hp.iis.sinica.edu.tw
Website : http://www.aclclp.org.tw
地 址：台北市研究院路二段128號中研院資訊所
電 話：(02)2788-3799 ext:1502
傳 眞：(02)2788-1638
劃 撥：19166251

本 期 要 目

第八屆博碩士論文獎得獎名單

博士論文獎

優等獎一名：獲獎金二萬元及獎狀

得獎者：潘奕誠（台灣大學資訊工程研究所）
題目：語音文件檢索的進一步研究：基於次詞成分的技術及使用者與系統的互動(Improved Approaches of Spoken Document Retrieval—Subword-based Techniques and User/System Interaction)
指導教授：李琳山 教授

佳作獎一名：獲獎金一萬元及獎狀

得獎者：禹良治（成功大學資訊工程學系）
題目：應用語意相依關係及超空間模擬語言模型於網頁文本探勘及資訊檢索之研究(A Study on Semantic Dependencies and HAL Modeling for Web Text Mining and Information Retrieval)
指導教授：吳宗憲 教授

碩士論文獎

優等獎一名：獲獎金一萬元及獎狀

得獎者：王昱鈞（台灣大學電機工程研究所）
題目：基於網路語料之專有名詞翻譯方法於中日韓跨語言資訊檢索之應用(Web-based Named Entity Translation Method for Korean-Chinese and Japanese-Chinese Cross-language Information Retrieval)
指導教授：許聞廉 教授、顏嗣鈞 教授

佳作獎二名：獲獎金伍仟元及獎狀

1. 得獎者：王昱婷（台灣大學資訊工程學系）
   題目：以機器學習方法處理跨語言檢索合併問題(A Machine Learning Approach for Result Fusion in Multilingual Information Retrieval)
   指導教授：陳信希　教授
2. 得獎者：黃瀚萱（交通大學資訊科學與工程所）
   題目：以序列標記方法解決古漢語斷句問題(Classical Chinese Sentence Division by Sequence Labeling Approaches)
   指導教授：孫春在 教授

ROCLING-2008 圓滿結束

由國立臺灣師範大學及本會共同主辦的「第二十屆自然語言與語音處理研討會」已於 97/9/5 在國立臺灣師範大學教育學院大樓演講廳順利圓滿結束，參與此次盛會的人士分別來自香港及台灣，與會人數多達 200 人次。本次會議共收錄了 17 篇口頭報告論文及 10 篇海報論文，最佳論文是由「A Thesaurus-Based Semantic Classification of English Collocations」獲得，作者分別爲：黃仲淇先生、曾瓊慧小姐、高紅雯小姐及張俊盛博士(清華大學資工系)，最佳論文作者分別於會議閉幕中獲頒獎狀乙紙，並共同獲頒獎金伍千元。會議所有論文已建置於學會網頁(http://www.aclclp.org.tw/pub_proce_c.php)，歡迎有興趣者上網瀏覽。

Call for Papers

Special Issue of The IEEE Transactions on Audio, Speech and Language Processing on Voice Transformation

With the increasing demand for Voice Transformation in areas such as speech synthesis for creating target or virtual voices, modeling various effects (e.g., Lombard effect), synthesizing emotions, making more natural dialog systems which use speech synthesis, as well as entertainment, film and music industry, toys, chat rooms and games, dialog systems, security and speaker individuality for interpreting telephony, high-end hearing aids, vocal pathology and voice restoration, there is a growing need for high-quality Voice Transformation algorithms and systems processing synthetic or natural speech signals.

Voice Transformation aims at the control of non-linguistic information of speech signals such as voice quality and voice individuality. A great deal of interest and research in the area has been devoted to the design and development of mapping functions and modifications for vocal tract configuration and basic prosodic features. However, high quality Voice Transformation systems that create effective mapping functions for vocal tract, excitation signal, and speaking style and whose modifications take into account the interaction of source and filter during voice production, are still lacking. We invite researchers to submit original papers describing new approaches in all areas related to Voice Transformation including, but not limited to, the following topics:

- Preprocessing for Voice Transformation (alignment, speaker selection, etc.)
- Speech models for Voice Transformation (vocal tract, excitation, speaking style)
- Mapping functions
- Evaluation of Transformed Voices
- Detection of Voice Transformation
- Cross-lingual Voice Transformation
- Real-time issues and embedded Voice Transformation Systems
- Applications

Proposed Schedule:
Submission deadline: 1st April 2009
Notification of acceptance: 15 September 2009
Final manuscript due: 30 October 2009
Tentative publication date: January 2010

Lead Guest Editor:
- Yannis Stylianou (yannis@csd.uoc.gr), University of Crete, Crete, Greece
Guest Editors:
- Tomoki Toda (tomoki@is.naist.jp), Nara Inst. of Science and Technology, Nara, Japan,
- Chung-Hsien Wu (chwu@csie.ncku.edu.tw), National Cheng Kung University, Tainan, Taiwan
- Alexander Kain (kaina@ohsu.edu), Oregon Health & Science University, Portland Oregon, USA
- Olivier Rosec (olivier.rosec@orange-ftgroup.com), Orange-France Telecom R&D, Lannion, France

台大台灣南島語多媒體語料庫簡介

沈文琦、宋麗梅
台灣大學語言所
{d97142002、limay}@ntu.edu.tw

一、前言

台灣南島語一直是台灣珍貴資產之一，就南島民族歷史發展、地理位置分佈及其語言多樣性等角度來看，台灣南島語的重要性的確不容小覷，但因目前台灣南島研究多缺乏文獻紀錄，許多語言不是已消失，就是瀕臨滅絕。因此，台大台灣南島語多媒體語料庫建置的主要目的，除了保存珍貴的語料，更致力於有系統性地彙整南島語，以利相關研究之進行。

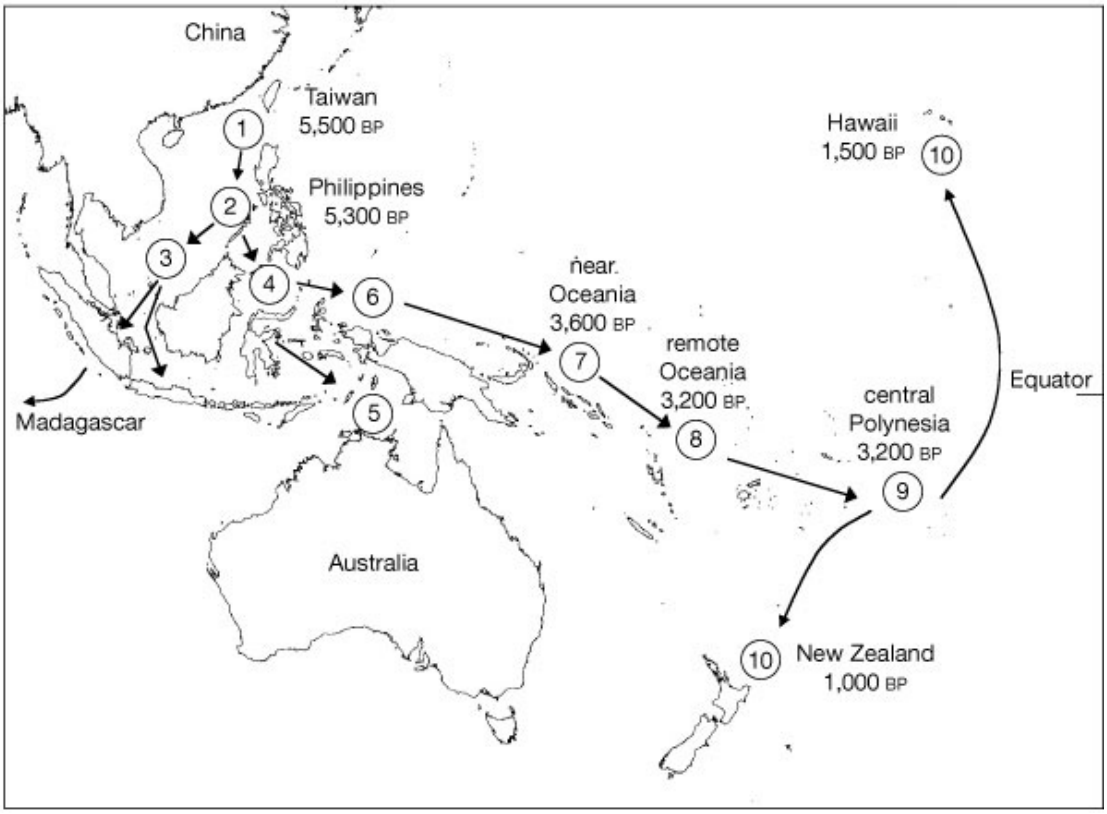
二、台灣南島語簡介

南島民族分佈的範圍廣達 26000 公里，西起馬達加斯加島，東達南美洲東岸的復活島，橫跨太平洋及印度洋（圖一），約佔地球面積三分之一以上的水域，使用南島語的人口約有兩億七千萬，使用的語言達 1200 種以上，是世界上最重要的語系之一。

圖一：南島語系地理上的分佈[1]

UCLA 地理及生理學者 Jared Diamond（2000）曾在 Nature 期刊發表短文 “Linguistics: Taiwan’s gift to the world”（語言：台灣給世界最好的禮物），並清楚點出：研究台灣南島語及其文化、地理、考古、遺傳等等相關議題，對瞭解整個南島族群遷徙擴散及航海造船技術發展，具有相當重要的意義，文中也表達了對台灣南島語極大的讚賞[1]。另外，南島語言學家 Robert Blust 也提出：1200 種南島語言可分爲十語支（linguistic subgrouping），其中九支在台灣都看得見；第十支則

在約 5000-6000 年前，從台灣經菲律賓、婆羅州，往南、東、和西散播，約在 2000-3000 年間擴散至大洋洲各島及澳洲、紐西蘭等地（圖二）。而伴隨語言傳遞而來的，還有航海、農業、畜牧，瓷器等技術，在人類發展史上綻放了光彩的一頁。也因此，當代南島語言學家大都公認：台灣地區的南島語言彼此差異最大，亦最爲紛歧，也最可能是古南島民族向外擴散的原點。



圖二：推測南島語系擴散的過程[2]

Krauss（1922）提到，全世界約有六千種語言，超過半數語言面臨瀕臨絕跡的危機[3]，台灣南島語亦然。由於台灣南島語過去並沒有文字紀錄，加上所有的平埔族語言都已經絕跡，所謂高山族僅佔台灣人口的 2.04%，在社會、經濟等壓力下，部落族人轉學國語、台語等主流語言，在目前已正名的十四族原住民中，如卑南、鄒、賽夏、雅美、撒奇萊雅、邵、噶瑪蘭等過半族語的使用人口均少於一萬人，特別是邵族與噶瑪蘭族，更少於一千人[1]。因此，台灣南島語的語料保存可說是刻不容緩。

三、台大台灣南島語多媒體語料庫歷史沿革

台大台灣南島語多媒體語料庫，原爲國立台灣大學資訊電子科技整合研究中

[1]台灣原住民人口分佈詳見行政院原住民委員會網頁，請見http://www.apc.gov.tw/chinese。

心「多媒體整合實驗室」計劃的子計劃之一（2001.1.1 至 2003.12.31），由語言所黃宣範、蘇以文、宋麗梅等教授共同主持。整個計劃結合了台灣大學數個系所（含語言所、資訊系、圖資系、資管系、電機系、新聞所、戲劇系）的專業人才，以語言爲主軸，藉由資訊科技運用，建置語言資料庫加以典藏，於 2005 年 7 月正式建立起此一線上資料庫之雛型。自 2006 年 3 月起，語料庫由國科會人文中心「台大南島語語料庫之建置」計劃補助，由台灣大學語言所宋麗梅教授主持，以既有的語料庫爲基礎，進行後續的修訂與擴建。

四、台大台灣南島多媒體語料庫的建置

本語料庫的資料皆來自第一手的自然口語語料，主要是希望透過自然語料（naturalistic data）的收集來補引出語料（elicitation data）的不足，畢竟，自然語料是在真實情境（real context）下產生，而非由研究者或發音人依直覺創造[4]；此外，透過詳細紀錄口語中的各種現象，包括停頓（pause）、重複（repetition）、修正（repair）、音調（pitch）等，更能真實呈現語言使用者心理及認知上的表現，對言談篇章分析（discourse analysis）的研究亦有所助益。除了語料本身，語料庫也提供參與成員在語料蒐集過程中的相關田野筆記，這些資訊對研究者而言，亦是十分珍貴的參考資料，不但描述了語言本身的特色，更表現出文化、語言與人類認知系統的交互影響。

語料庫中語料的收集主要是透過田野調查[2]的方式進行，語料內容與發音人的生活對話或部落傳說有關，或是請發音人觀看無聲影片[5]或不含文字的圖畫書[6]之後，以族語口述影片或圖書內容，研究者利用數位錄音機及錄影機紀錄，之後再轉寫成文字紀錄，分割 IU 單位（Intonation Unit）、提供標記及中英翻譯等，在成功上傳至語料庫之後，研究者或南島語相關工作者便可透過網頁瀏覽台灣南島語的相關語料（http://corpus.linguistics.ntu.edu.tw）。

（一）語料轉寫

由於語料庫文本轉寫主要是人力作業，標準化流程更趨重要，研究者除了將聲音檔轉寫成文字檔外，利用軟體 Praat 標記 IU 單位，主要是希望藉由此軟體顯示的波形（wave form）、音調（pitch contour）、聲音強度（intensity）或聲譜圖（spectrogram）等來判斷，將轉寫者切割 IU 的個別差異減到最低，並精確計算每個 IU 的時長，以利日後音檔切割上傳作業之進行。

[2] 基於智慧財產權考量，上傳前我們會先取得發音人的授權同意書，同意將生活智慧、傳說故事、影音檔等上傳至台大台灣南島語語料庫，以供學術研究。

此外，爲能盡可能達到轉寫一致性，根據 Chafe（1987, 1994）以及 Du Bois et al.（1993）的對 IU 的定義[7-9]，列出五項明顯的音韻特徵，作爲處理南島語 IU 切割的準則，此五項特徵分別爲：(1) 停頓（pause）；(2) 聲音長度的改變（change of duration）；(3) 聲音強度的改變（change of intensity）；(4) 音調改變（change of pitch）；(5) 重新集氣（breath reset）[10]。下面依序以三個撒奇萊雅語的例子來說明 IU 切割的原則。

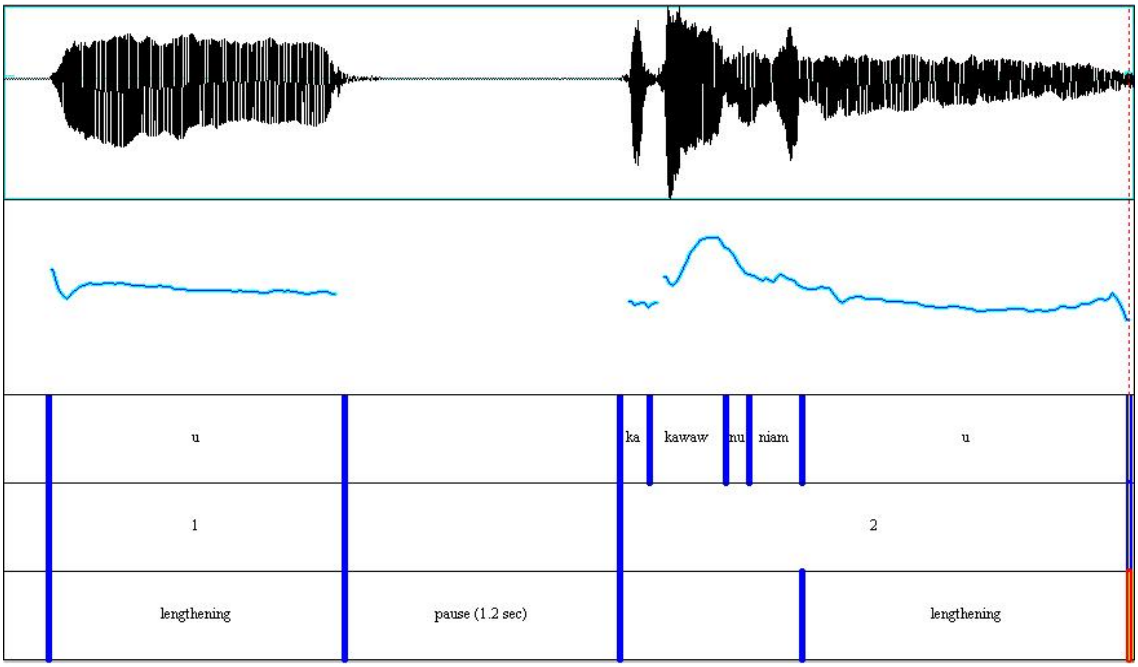
第一個例子出現停頓，以及語尾助詞的拉長，兩個條件都可以作爲切割 IU 的原則（圖三）。

(1) 摘錄自撒奇萊雅語 Skzy_ta'on_story: IU 45-47

45. ...(1.9) u== ,_
CN
CN

46.	...(1.2)	ka-kawaw	nu	niam	u== ,_
		RED-thing	GEN	1EPL.POSS	CN
		重疊-事情	屬格	1PL.排除.所有	CN

47. .. babalaki haw sa .\
elder HAW DM
老人 HAW DM
We old people do everything well.
我們老人做事做的好。

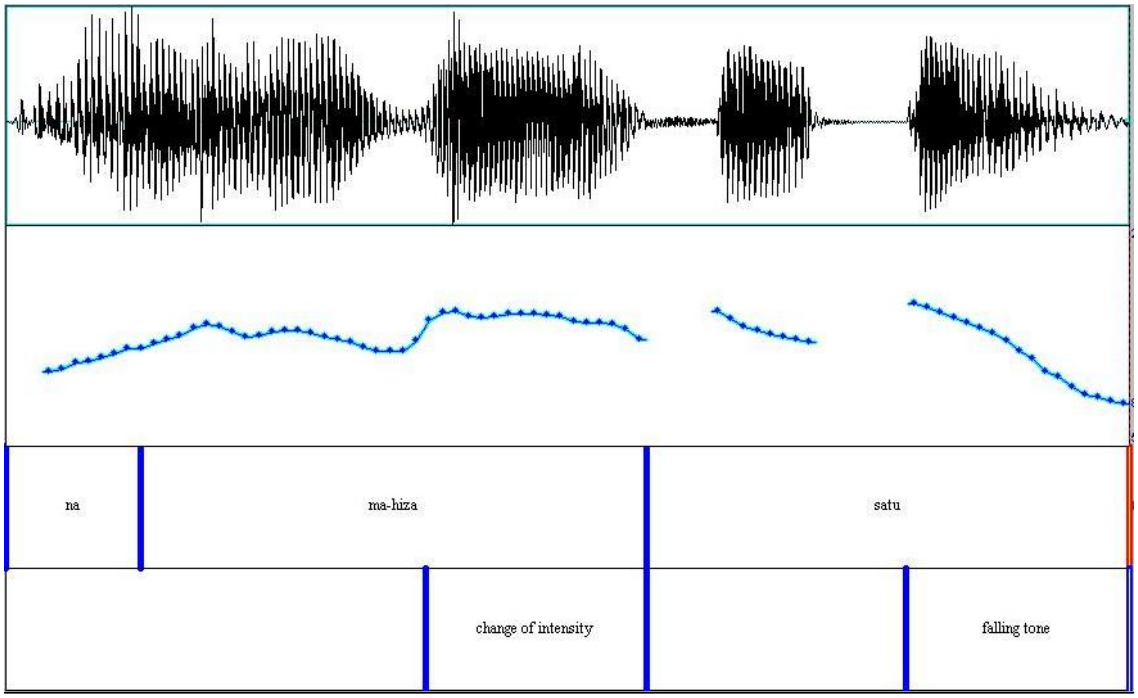


圖三：IU 的停頓及音節拉長

第二個例子主要是因爲語調下降視爲一個 IU 單位的結束，但此 IU 中聲音強度改變僅以符號 “^”表示，並未切割成另一個 IU（圖四）。

(2) 摘錄自撒奇萊雅語 Skzy_amui_putong: IU 48

48. ...(1.9) na=mahi^za satu ,\
past=that.way DM
過去=那樣 DM
That’s how it was.
過去是那樣



圖四：語調下降及改變聲音強度

第三個例子第 33 個 IU 重新集氣，起頭比第 32 個 IU 還高，加上該音節後面出現一個停頓，因此單獨切爲一個 IU（圖五）。

(3) 摘錄自撒奇萊雅語 Skzy_amui_putong: IU 32-34

32. ...(1.4) paluma-han=tu ,\
plant-PF=PFV
種-受焦=完成
33. .. tu ,\
OBL

斜格

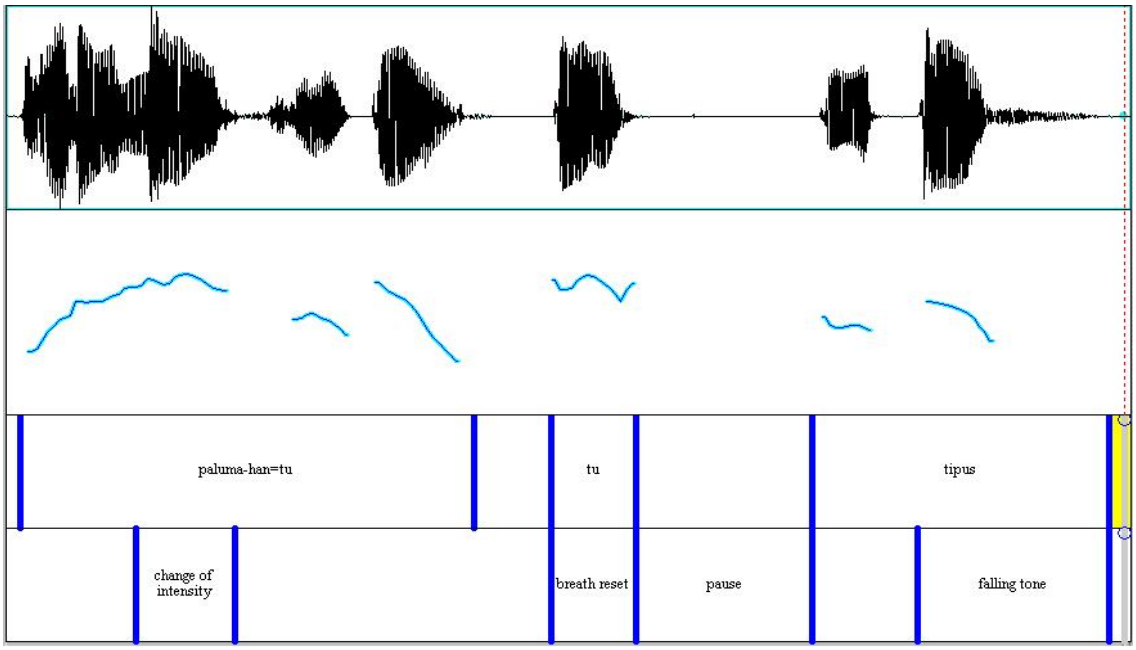
34. ... tipus ,\

rice

稻米

Rice was grown.

稻米種了



圖五：重新集氣

（二）標記

標記又分爲言談篇章標記及語法標記：言談篇章標記是根據 Du Bois *et al.*（1993）的轉寫標記法，語法標記則是參考德國 Max Planck 語言所及 Leipzig 大學語言所共同建置的 Leipzig Glossing Rules 來做標記，並針對台灣南島語各族語言的不同略作修正。詳細修正細節請參照台大台灣南島語料庫網頁上之說明。

（三）資料庫編寫

本語料庫使用 SQLite 作爲資料庫，主要因爲它是獨立平台，又能支援 UTF-8 編碼，此外，此語料庫是以 XML 的格式輸出資料，與大多數相關的語料庫相容，便於資訊之交流。

（四）輸出介面

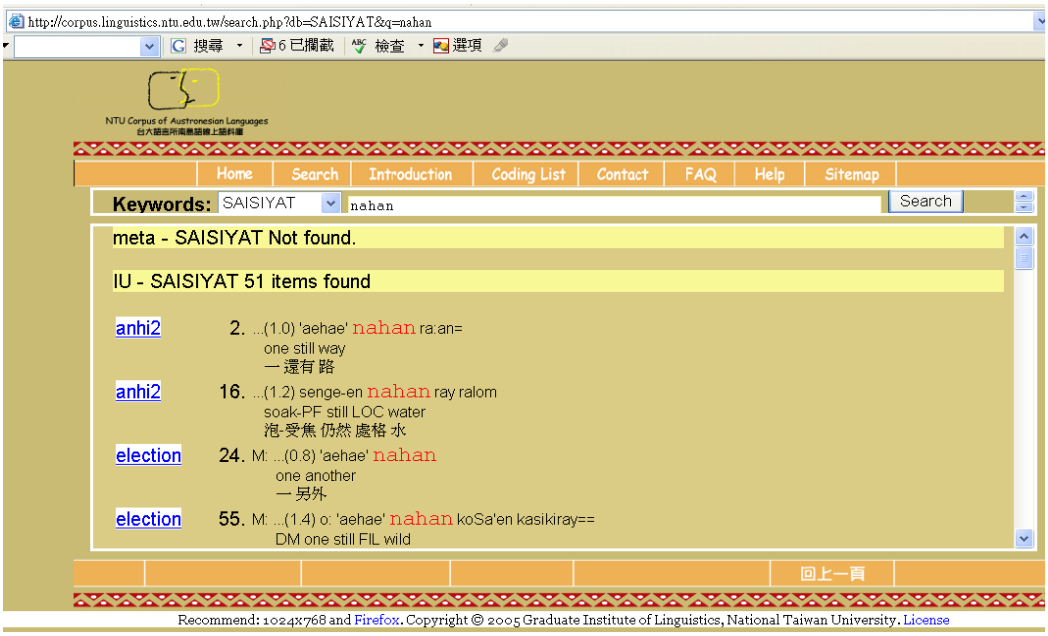
後設資料、文本轉寫、多媒體影音檔等都會出現在輸出頁面上（如圖六）。後設資料特別放在每一份音檔的最上方，包括檔案名稱、主題、形式、語言別、方言別、發音人姓名（族名）、音檔長度、IU 總數、採集時間、校對時間、轉寫者、校對者等十二項資訊。

文本則依照 IU 順序排列，包含中英文的標記及最後句子的中英翻譯，雙擊每個 IU 後面的圖像，即可聽到已切割過的影音檔，頁面左邊放置發音人的照片。照片底下提供瀏覽選項，如不顯示英譯、不顯示中譯等供使用者選擇。

圖六：自本語料庫擷取的頁面之一

（五）語料搜尋

本語料庫提供使用者以英文、中文或南島語等語言搜尋單詞或詞綴。例如，在搜尋介面上鍵入 *nahan*（如圖七），然後選擇賽夏語（亦可搜尋目前語料庫中已建置的任一或所有語言）。



圖七：搜尋賽夏語 *nahan* 的介面

五、目前研究成果

目前語料庫中已建構好的包括賽夏、噶瑪蘭、鄒、阿美及撒奇萊雅語資料庫。噶瑪蘭語有 4 筆口述語料（皆有聲音及影像檔），賽夏語有 22 筆口述語料（只有聲音檔），阿美語則有 2 筆口述語料（皆有聲音及影像檔），鄒語有 2 筆語料（只有聲音檔），撒奇萊雅語則有 2 筆口述語料（只有聲音檔）。其他族語語料的轉寫及校對工作，亦持續進行。

六、語料庫特色

隨著自然語料處理技巧和網路科技的快速精進，建立語料庫成了保存記錄珍貴語料最有效率的方式。台大台灣南島語多媒體語料庫的精神，即在發揮最大使用效益—不論是語言學者或是一般大眾，只要是對南島文化及語言有興趣的人士，皆可藉由本語料庫取得豐富而珍貴的語料。爲此，語料庫的建立乃以下列八大原則爲依歸：

（一）普及性：本語料庫的設計，即不只是提供語言學者研究之用，更期盼能滿足一般大眾的求知需求，推廣台灣南島語言及文化。

（二）口語性：本語料庫的資料來自於第一手的口語資料，根據 Du Bois *et al.* (1993）的轉寫標記法，以 IU 爲單位，紀錄口語中的各種現象，真實反映出語言在生活中的實際使用狀況。

（三）一致性：語料的建立修編，皆採標準化作業程序，以提高使用之便利性。在標記規格方面，語料庫內採用統一規格標示（Leipzig Glossing Rules爲基礎）；篇章標記則大致依照 Du Bois *et al.*（1993）的格式。這套規格除符合國際標準之外，同時亦適用於大多數的台灣南島語，因此，本設計便利使用者進行不同台灣南島語間的交叉查詢、比較與研究。

（四）跨語言性：本語料庫的查詢功能，除了可以針對單一語言查詢，更提供了跨語言的查詢功能。使用者只需輸入英文、中文或任一南島語的詞彙或詞素，即可進行線上查詢。

（五）相容性：爲了使本語料庫發揮最大的效益，本系統所採用的 XML（Extensible-Mark-up Language）格式爲一簡易、靈活的程式語言系統，目前廣泛運用於網路資訊的交換。本語料庫可以 XML 格式輸出所有資料，以滿足不同系統使用者間溝通之需求。

（六）便利性：在新文本上傳的同時，系統即可自動計算詞彙數目，並就所有詞彙之個別資訊編輯成線上字典，減少人力之負擔。

（七）多元性：除文字之外，本系統亦提供影音檔等多媒體的使用，使語料庫的內容更加多元，同時也可提升一般大眾使用的興趣與動機。

七、未來展望

台大台灣南島語多媒體語料庫的目標，除了繼續蒐集台灣各南島語相關資料，更希望能與資訊相關科系合作，爲台灣南島語建置語言資料之典藏。此外，並以半自動化爲目標，包括半自動化轉寫（semi-automated transcription）、聲音與文字同步（sound-to-text alignment）、影像與文字同步（image-to-text alignment），以克服人力轉寫耗時費工之問題。藉由資訊系、電機系等對資訊、語言處理之長才，加上台大語言所研究團隊對台灣南島語之研究成果，期能對台灣南島語語料庫建置之半自動化有所貢獻。

八、文獻回顧

[1] Diamond, Jared M. 2000. Taiwan’s gift to the world. *Nature* 307: 709-710. Available from http://www.nature.com/nature/journal/v403/n6771/full/403709a0.html

[2] Russell D. Gray & Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405: 1052-1055. Available from http://www.nature.com/nature/journal/v405/n6790/full/4051052a0.html

[3] Krauss, Michael E. 1992. The world’s languages in crisis. *Language* 68:6-10.

[4] Su, Lily I-wen, Li-May Sung, Shuping Huang, Fuhui Hsieh and Zhemin Lin. Forthcoming. NTU Corpus of Formosan Languages: A State-of-the-art Report. *Corpus Linguistics and Linguistic Theory.*

[5] Chafe, Wallace L. 1980. The deployment of consciousness in the production of a narrative. In Chafe. Wallace L. (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* 9-50. Norwood, NJ: Ablex,.

[6] Mayer, Mercer. 1980. *Frog, where are you?* NY: Dial Books.

[7] Chafe, Wallace L. 1987. Cognitive constraints on information flow. In Coherence and grounding in discourse, ed. by Russell S. Tomlin, 21-51. Amsterdam: John Benjamins.

[8] Chafe, Wallace L. 1994. *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing.* Chicago: University of

Chicago Press.
[9] Du Bois, J. W, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In J. A. Edwards and M. D. Lampert, (ed.), *In Talking Data: Transcription and Coding for Language Research* 45-90. Hillsdale, N.J.: Lawrence Erlbaum Associates.
[10] Sung, Li-May, Lily I-wen Su, Fuhui Hsieh and Zhemin Lin. 2008. Developing an On-line Corpus of Formosan Languages. Presented at the 7th Annual Wenshan International Symposium, May 19, National Chengchi University.

作者簡介：

宋麗梅爲台灣大學語言所副教授，主要研究包括：語法學、南島語句法、中英比較等。目前著重在台灣南島語句法研究及南島語語料庫建置。
沈文琦爲台灣大學語言所博士生，主要研究爲台灣南島語句法，目前著重在撒奇萊雅族語（台灣第十三族原住民，多分佈於花蓮）。