中華民國計算語言學學會
通訊 ACLCLP
The Association for Computational Linguistics and Chinese Language Processing
第十八卷第五期
發行人：陳信希
主　編：王新民
執行編輯：黃　琪

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

本 期 要 目

2008 International Conferences List

ACL-2008: HLT

The 45th Annual Meeting of the Association for Computational Linguistics ： Human Language Technologies
- Conference Date：June 15-20, 2008
- Location：Ohio, USA
- http://www.ling.ohio-state.edu/acl08/

ACM-SIGIR-2008

The 31st Annual International ACM SIGIR Conference
- Conference Date：July 20-24, 2008
- Location：Singapore
- http://www.sigir2008.org/

AIRS-2008

Asia Information Retrieval Symposium
- Conference Date：January 16-18, 2008
- Location：Harbin, China
- http://ir.hit.edu.cn/airs2008/

CLSW-2008

Chinese Lexical Semantics Workshop
- Conference Date：July 14-17, 2008
- Location：Singapore
- http://www.chineselex.org/

Coling-2008

The 22nd International Conference on Computational Linguistics
- Conference Date：August 18-22, 2008
- Location：Manchester, UK
- http://www.coling2008.org.uk/

GWC-2008

The 4th Global WordNet Conference
- Conference Date：January 22-25, 2008
- Location：Szeged, Hungary
- http://www.inf.u-szeged.hu/projectdirs/gwc2008/

ICASSP-2008

The 33rd International Conference on Acoustics, Speech, and Signal Processing
- Conference Date：March 30 - April 4, 2008
- Location：Las Vegas, USA
- http://www.icassp2008.org/

IJCNLP-2008

The 3rd International Joint Conference on Natural Language Processing
- Conference Date：January 7-12, 2008
- Location：Hyderabad, India
- Website：http://www.ijcnlp2008.org/

Interspeech-2008-ICSLP

The 10th International Conference on Spoken Language Processing
- Conference Date：September 22-26, 2008
- Location：Brisbane, Australia
- http://www.interspeech2008.org/

LREC-2008

The 6th Language Resources and Evaluation Conference
- Conference Date：May 26 - June 1, 2008
- Location：Marrakech , Morocco
- http://www.lrec-conf.org/lrec2008/

MATBN 中文廣播新聞語料庫 簡介

民國90年8月至93年7月間，國內從事語音處理研究之相關學校及研究單位聯合執行國科會語料蒐集計畫－『中文自發性語音語料庫之建立』(Spontaneous Mandarin Speech: Corpus and Processing；計畫編號：NSC-90-2213-E-009-109, NSC-91-2219-E-009-039, NSC-92-2213-E-009-021)，參與的單位共有國立交通大學電信工程學系、國立台灣大學電機工程學系、國立清華大學電機工程學系、國立成功大學電機工程學系、中央研究院資訊科學研究所、工研院前瞻研究中心及中華電信研究所。該計畫完成一個MATBN中文廣播新聞語料庫，語料來源是198個小時之公共電視晚間新聞，內容包括音檔、人工標記及文字轉寫（transcription）。為讓該項成果與國內外從事中文語音處理研究之單位分享，計畫執行團隊將MATBN語料庫技轉予本學會，授權本學會發行予各界使用。

MATBN語料庫的錄製時間自民國90年11月17日起，至民國92年4月3日止。每個一小時的錄音檔案都是先以建置在公共電視播音室的數位錄音座(DAT)，以取樣頻率44.1kHz，16位元，立體聲的設定錄製。之後，每一個數位錄音檔再取其左聲道，轉成取樣頻率16kHz，16位元的微軟音檔(Microsoft Windows wave)儲存及進行人工標記。

MATBN語料採用DGA&LDC所開發的Transcriber [1][Barras et al. 2001]進行人工分段、標記和文字轉寫(如圖一所示)。結果以XML檔案儲存(如圖二所示)，內容包括：新聞邊界(story boundary)、語者轉換邊界(speaker turn boundary)、背景雜訊、文字轉寫等，文字編碼則採用大五碼(Big 5)。上述標記均包含時間註記，可與音檔對應。198個小時的語料中，經過人工標記，共包含4,100則新聞報導(story)、581則節目重點內容介紹(headline)、197則氣象預報、197則結尾。所有新聞、節目重點內容介紹、結尾及10則氣象預報共約143小時的語音有完整的標記及文字轉寫，其餘氣象預報及廣告、純音樂等段落則僅有時間註記。這些語音來自7位新聞主播(anchor reporter)、386位現場記者(field reporter)及5,900位被採訪者(interviewee)。部份現場記者及被採訪者無法確認身分，所以實際人數應低於上述數字。文字轉寫部份共包含約兩百三十萬字。

MATBN 語料已設定發展集和測試集作為評比基準(benchmark)。發展集為 2003/01/24、2003/01/27、2003/02/07、2003/03/05 及 2003/03/06 五日的新聞，測試集則為 2003/01/28、 2003/01/29、2003/02/11、2003/03/07 及 2003/04/03 五日的新聞。發展集和測試集的選取原則如下：(1)盡量包含最多的新聞主播；(2)必須是 2003/1/1 之後的新聞，方便使用者利用 2003 年之前的文字新聞訓練語言模型，避免弄錯。

授權語音資料庫共五片 DVD 光碟：

MATBN_1: TRAIN_1 (36個音檔與對應的標記檔)
MATBN_2: TRAIN_2 (36個音檔與對應的標記檔)
MATBN_3: TRAIN_3 (36個音檔與對應的標記檔)
MATBN_4: TRAIN_4 (41個音檔與對應的標記檔)

[1] 可以到 http://www.ldc.upenn.edu/mirror/Transcriber.old/en/menu_web.html 下載

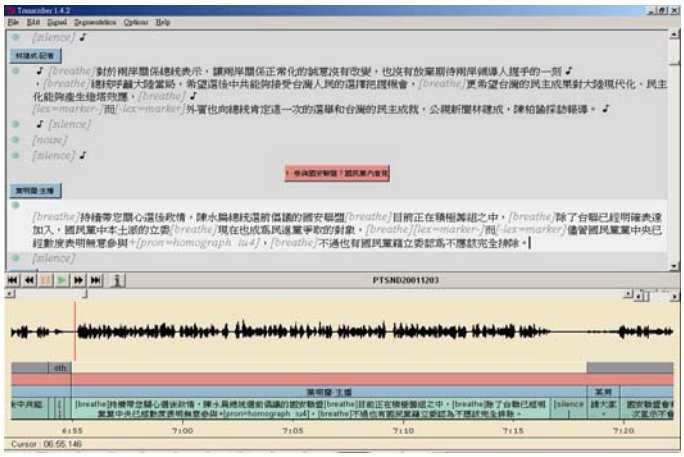
MATBN_5: TRAIN_5 (39個音檔與對應的標記檔),
DEVELOPMENT (5個音檔與對應的標記檔),
EVALUATION (5個音檔與對應的標記檔)

關於 MATBN 語料的詳細介紹，請參見[Wang et al. 2005] or http://sovideo.iis.sinica.edu.tw/SLG/corpus/MATBN-corpus.htm.

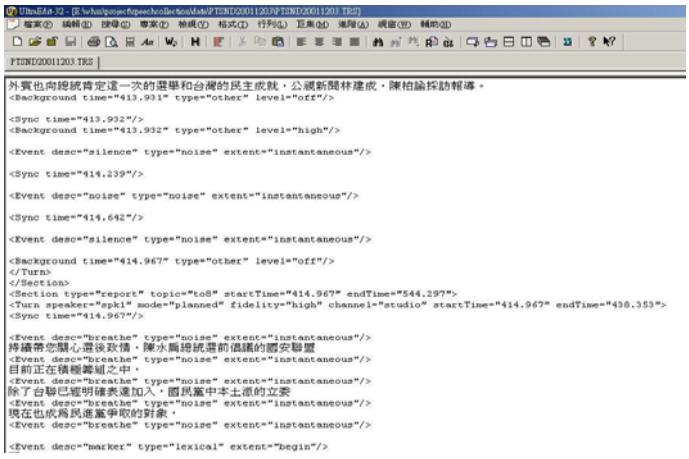
參考資料

[Barras et al. 2001] Barras, C., E. Geoffrois, Z. B. Wu and M. Liberman, "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production," *Speech Communication*, 33, 2001, pp. 5-22.

[Wang et al. 2005] Wang, Hsin-Min, Berlin Chen, Jen-Wei Kuo, and Shih-Sian Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2), June 2005, pp. 219-236.



圖一：以 Transcriber 標記新聞語音的實例



圖二：Transcriber 輸出的 XML 標記檔案

語料庫申請：

- 申請辦法：請參閱學會網站 http://www.aclclp.org.tw/corp_c.php。
- 費用：學術研究單位　　NT$ 40,000 元
  一般企業　　NT$ 200,000 元

A Brief of the MATBN Corpus

The MATBN Mandarin Chinese broadcast news corpus is a product of a joint project sponsored by the National Science Council, Taiwan. It contains a total of 198 one-hour news shows from the Public Television Service Foundation, Taiwan with corresponding transcripts. The primary purpose of this collection is to provide training and testing data for continuous speech recognition evaluation in the broadcast news domain.

The MATBN corpus spanned the period November 17, 2001 through April 3, 2003. Each one-hour broadcast news episode recording was first made in stereo with a 44.1kHz sampling rate and 16 bit resolution by a DAT recorder set up in the TV broadcasting studio. Then, each DAT recording was converted into a single Microsoft Windows wave file. Finally, the signal was down-sampled to 16 kHz with a resolution of 16 bits. During this operation, only the left channel was selected.

The MATBN corpus has been segmented, labeled, and transcribed manually using the DGA&LDC Transcriber[2] [Barras *et al.* 2001]. The transcripts are in Big5-encoded form, with SGML tagging to annotate acoustic conditions, background conditions, story boundaries, speaker turn boundaries, and audible acoustic events, such as hesitations, repetitions, vocal non-speech events, and external noises. These tags include time stamps that are used to align the text with the speech data. In the 198-hour broadcast news corpus, based on hand-segmentation results, there are 4,100 stories, 581 headlines, 197 weather forecasts, and 197 ending sections. Around 143 hours of speech from 10 weather forecasts and all the stories, headlines, and ending sections were carefully transcribed, while the remaining weather forecasts and segments containing advertising or pure music were just annotated with time stamps without orthographic transcripts. There are 7 anchor reporters, 386 field reporters, and 5,900 interviewees. The identities of some field reporters and interviewees could not be determined. Since the unidentified field reporters and interviewees could correspond to the same person, the true numbers of field reporters and interviewees could be lower than the above numbers. The transcripts contain around 2.3 million Chinese characters in total.

A development set and an evaluation set have been defined for the benchmark test. The development set consisted of five shows recorded on 2003/01/24, 2003/01/27, 2003/02/07, 2003/03/05, and 2003/03/06, while the evaluation set consisted of five shows recorded on 2003/01/28, 2003/01/29, 2003/02/11, 2003/03/07, and 2003/04/03. The basic guidelines for making selections are as follows: First, we wanted to include as many studio anchors as possible. Second, the test shows had to be broadcast after January 1st, 2003 so that we could use the newswire text before January 1st, 2003 to train the language models.

The corpus is distributed on 5 DVDs:

MATBN_1: TRAIN_1 (36 speech files and the corresponding transcription files)
MATBN_2: TRAIN_2 (36 speech files and the corresponding transcription files)
MATBN_3: TRAIN_3 (36 speech files and the corresponding transcription files)
MATBN_4: TRAIN_4 (41 speech files and the corresponding transcription files)
MATBN_5: TRAIN_5 (39 speech files and the corresponding transcription files),
DEVELOPMENT (5 speech files and the corresponding transcription files),
EVALUATION (5 speech files and the corresponding transcription files)

For details of the MATBN corpus, please refer to [Wang *et al*. 2005] or http://sovideo.iis.sinica.edu.tw/SLG/corpus/MATBN-corpus.htm.

References

Barras, C., E. Geoffrois, Z. B. Wu and M. Liberman, "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production," *Speech Communication*, 33, 2001, pp. 5-22.

Wang, Hsin-min, Berlin Chen, Jen-Wei Kuo, and Shih-Sian Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2), June 2005, pp. 219-236.

Application：Please see the ACLCLP website at http://www.aclclp.org.tw/corp.php.

Application fee：

- Non-profit organization : US$ 1,350.-
- Commercial organization : US$ 6,750.-

[2] Transcriber can be downloaded at http://www.ldc.upenn.edu/mirror/Transcriber.old/en/menu_web.html.

General Chair
Lin-shan, Lee
National Taiwan University

General Vice-Chair
Jung-Ho Lee
Chunghwa Telecom Co.,Ltd.

Secretaries General
Tsungnan Lin
National Taiwan University
Fu-Hao Hsing
Chunghwa Telecom Co.,Ltd

Technical Program Chairs
Liang-Gee Chen
National Taiwan University
James R. Glass
Massachusetts Institute of Technology

Technical Program Members
Petar Djuric
Stony Brook University
Joern Ostermann
Leibniz University Hannover
Yoshinori Sagisaka
Waseda University

Plenary Sessions
Soo-Chang Pei (Chair)
National Taiwan University
Hermann Ney (Co-chair)
RWTH Aachen

Special Sessions
Shih-Fu Chang (Chair)
Columbia University
Lee Swindlehurst (Co-chair)
Brigham Young University

Tutorial Chair
Tsuhan Chen
Carnegie Mellon University

Publications Chair
Homer Chen
National Taiwan University

Publicity Chair
Chin-Teng Lin
National Chiao Tung University

Finance Chair
Hsuan-Jung Su
National Taiwan University

Local Arrangements Chairs
Tzu-Han Huang
Chunghwa Telecom Co.,Ltd.
Chong-Yung Chi
National Tsing Hwa University
Jen-Tzung Chien
National Cheng Kung University

Conference Management
Conference Management Services

IEEE Signal Processing Society

Taipei, Taiwan 2009

ICASSP

2009 TAIPEI Signals over the Horizon

IEEE International Conference on Acoustics, Speech, and Signal Processing
April 19 - 24, 2009
Taipei International Convention Center
Taipei, Taiwan, R.O.C.

The 34th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) will be held at the Taipei International Convention Center in Taipei, Taiwan, April 19 - 24, 2009. The ICASSP meeting is the world's largest and most comprehensive technical conference focused on signal processing and its applications. The conference will feature world-class speakers, tutorials, exhibits, and over 50 lecture and poster sessions on:

- Audio and electroacoustics
- Bio imaging and signal processing
- Design and implementation of signal processing systems
- Image and multidimensional signal processing
- Industry technology tracks
- Information forensics and security
- Machine learning for signal processing
- Multimedia signal processing
- Sensor array and multichannel systems
- Signal processing education
- Signal processing for communications
- Signal processing theory and methods
- Speech and language processing

Taiwan: The Ideal Travel Destination. Taiwan, also referred to as Formosa – the Portuguese word for "graceful" – is situated on the western edge of the Pacific Ocean off the southeastern coast of mainland Asia, across the Taiwan Strait from Mainland China. To the north lie Japan and Okinawa, and to the south is the Philippines. ICASSP 2009 will be held in Taipei, a city that blends traditional culture and cosmopolitan life. As the political, economic, educational, and recreational center of the country, Taipei offers a dazzling array of cultural sights not seen elsewhere, including exquisite food from every corner of China and the world. You and your entire family will be able to fully experience and enjoy this unique city and island. Prepare yourself for the trip of your dreams, as Taiwan has it all: fantastic food, a beautiful ocean, stupendous mountains and lots of sunshine!

Submission of Papers: Prospective authors are invited to submit full-length, four-page papers, including figures and references, to the ICASSP Technical Committee. All ICASSP papers will be handled and reviewed electronically. The ICASSP 2009 website www.icassp09.com will provide you with further details. Please note that the submission dates for papers are strict deadlines.

Tutorial and Special Session Proposals: Tutorials will be held on April 19 and 20, 2009. Brief proposals should be submitted by August 4, 2008, to Tsuhan Chen at tutorials@icassp09.com and must include title, outline, contact information, biography and selected publications for the presenter, a description of the tutorial, and material to be distributed to participants. Special sessions proposals should be submitted by August 4, 2008, to Shih-Fu Chang at specialsessions@icassp09.com and must include a topical title, rationale, session outline, contact information, and a list of invited speakers. Tutorial and special session authors are referred to the ICASSP website for additional information regarding submissions.

Important Deadlines

Tutorial Proposals Due	August 4, 2008
Special Session Proposals Due	August 4, 2008
Notification of Special Session & Tutorial Acceptance	September 8, 2008
Submission of Regular Papers	September 29, 2008
Notification of Acceptance (by email)	December 15, 2008
Author's Registration Deadline	February 2, 2009

電腦語言學的研究發展概論：過去、現在與未來

楊孝慈
國立雲林科技大學 應用外語學系
yanght@yuntech.edu.tw

一、源起

電腦語言學的研究源自於 1950 年代，當時一些美國語言學者試著想把俄文的學術論文由機器自動翻譯成英文，以方便能快速蒐集並閱讀資料。爲此，語言學家必須了解這兩種語言的文法，包括句構學、語型學、語意學、以及語用學等語言結構。語言學家把文法知識交由電腦語言學家來設計語言轉換的程式，希望能使電腦了解俄文和英文的文法，然後把俄文翻譯成英文。剛開始由機器翻譯出來的文章，只能讓美國學者從俄文寫的論文裡，得到很粗淺的概念，它並不是一個精緻的翻譯。

到了 1960 年代，電腦語言學家意識到機器翻譯比想像的還要困難，因爲人類語言的文法層次非常複雜，所以機器翻譯的研究發展就遇到了困境，無法突破。之後，電腦語言學家開始從其他方面發展，例如運用電腦來做語言資料的整理，或是開發一些軟體來辨識語言的意義 [Cole *et al.* 1998]。

二、定義

什麼是電腦語言學呢？從字義上來說，電腦語言學就是語言學家和電腦程式設計者共同研究語言的文法，以電腦來進行分析語料。目前，電腦語言學家常研究的問題就是人工智慧，如何使得電腦能夠了解人類的語言，然後進行一些跟語言相關的工作，包括它能夠翻譯、能夠處理一些語言的資訊，甚至於能夠了解人類的話，跟人類互動。

有些學者給電腦語言學的定義比較嚴謹，他們認爲，如果只是純粹地讓電腦進行對文本的統計分析，那這並不是電腦語言學主要的研究方向。電腦語言學主要的研究是如何讓電腦能夠了解人類的語言。因此，他們設計了程式，灌輸電腦有關人類語言的文法。但是，另外有一些學者認爲，應該從更廣的層面來看電腦語言學的研究發展，電腦語言學並不是只有讓電腦能夠辨識人類的語言；他們認爲，電腦語言學應該包括能讓電腦來分析語言的資料、測試語言學的理論，或者是了解人類和電腦的相同和相異性。

整體來說，電腦語言學的應用範圍很廣，包括了機器翻譯、人工智慧、資料檢索、文法檢測、人機問答、電腦協助的語言教學，或是自動寫作評分等等，都是電腦語言學的研究發展，當然還包括其他的應用 [Bosch *et al.* 1999]。接下來我將進一步地談論電腦語言學的過去、現在和未來發展。

三、過去發展

電腦語言學從 1950 年代開始就進行機器翻譯，但到目前爲止，機器翻譯技術並沒有突破多少。以前只是希望能把俄文或是其他外國語言翻譯成英文，讓美國學者或是科學家能大略地了解國外的學術和科學研究發展，做爲情報上的分析或交流。但是，機器翻譯的技術遇到瓶頸，原因是人類的語言文法非常複雜，並不能簡化成只有一些公式或程式而已，這些複雜的文法觀念必須考慮語音、音調等等。

所幸，機器翻譯已發展出二種重要的技術，一種是借由人工先行把想要翻譯的文本簡化句型，讓機器能夠便於分析語料，然後翻譯，這種技術是由先行的人工編輯，簡化文本，使得機器的翻譯能夠順暢，也就是把一些文法句型複雜的句子簡化成爲簡單的句子，讓翻譯機器能夠辨讀，進行翻譯。

另外一種技術就是由機器來協助翻譯，也就是先建立一個已經翻譯好的語料庫，讓翻譯者來運用，當要翻譯一個句子的時候，語料庫就會檢測它內部的翻譯文本，搜尋比較接近的句子，然後提供譯文出來給翻譯者參考。翻譯者可以完全採納資料庫所提供的譯文，或者是將資料庫的譯文進行修改，成爲更適當的翻譯。這種技術是借由資料庫搜尋的方式，讓翻譯者參考相似的譯文，再進行編輯，以使翻譯的工作變得簡單。因此，對於內容相似性很高的文本，在做翻譯的時候，使用機器來協助翻譯，會比較有效。例如，資料庫裡如果已經有從英文翻成中文的 A 牌冷氣機操作的翻譯文本，那麼，當 B 牌的英文操作文本要翻成中文時，就可以用 A 牌的翻譯文本資料庫來進行輔助，因爲冷氣機的操作方式有很多都很類似，所以當遇到相似的操作說明和句子時，資料庫就會搜尋 A 牌冷氣機操作資料庫的譯文，提供出來給翻譯者做參考，進行編輯，這樣就能使得翻譯的工作輕鬆簡單 [Shih 2006]。

大體來說，機器翻譯的技術已經從過去的機器翻譯文本，變成整合人工編輯的方式，進行機器翻譯。翻譯者可以先行將文本簡化，讓機器翻譯順暢無礙。或是先建立譯文資料庫，然後藉由搜尋相似句型的譯文，讓翻譯者參考和編輯。這兩種技術各有缺點，如果是由翻譯者先行簡化句型，那麼，機器翻譯出來的句子，就會比較粗糙而沒有文采，若欲要求高品質的翻譯，就需先建立譯文資料庫，讓翻譯者參考，進行人工編輯；但是，要如何能夠收集精確而良好的譯文做爲資料庫，將是日後發展機器翻譯技術的其中一項重要課題 [Daelemans 2005]。

四、現在發展

最近這幾年，電腦語言學者都集中研究語音辨識系統，他們的目標是希望電腦能夠辨識人類說話的內容。但是，這樣的目標在現階段是不太容易達成，因爲有很多的因素都必須考慮。目前，許多的語音辨識系統的研究和開發，都是在無噪音

或安靜的環境下測試；同時希望使用者能夠使用麥克風，正確發音；另外，所測試的字彙範圍也非常有限，大約二千字。所以，當各種不同的使用者，在各種不同的環境下使用語音辨識系統時，電腦的辨識能力就大幅下降 [Rayner 2007]。

那要如何提升電腦語音辨識系統的能力呢？首先，我們可以先建立一個良好的字彙資料庫，可從兩個方向來進行；第一，先建立一個日常用語的字彙資料庫。第二，根據使用者的工作需要來建立資料庫，例如，如果我們要爲一個專爲從事國際貿易的商人設計一套語音辨識系統，我們除了建立一個日常用語的資料庫外，也要建立一個國際貿易常用字的資料庫，整合這兩種資料庫，做爲電腦語音辨識系統的訓練，那麼就可提升電腦辨識的能力。

接下來的問題是，這個字彙資料庫應該要多大呢？大部份的實驗都是建立一個二千字左右的字彙資料庫，來進行測試語音辨識系統，但在實際的使用層面上，二千字是否適當、足夠，就因人而異，最好的方式是根據特別一群的使用者或者個別的使用者來設計他們的字彙資料庫；例如，如果我們想要專門爲從事國際貿易的人來設計語音辨識系統，那麼就要考慮到他們工作場合常用單字的字彙大小來決定字彙資料庫裡面的容量。根據有些學者的研究，一般中文的常用字使用，大約在五千到六千單字左右 [Cheng 1996]；但是，如果我們想要專門爲某一特定的使用者或族群來設計語音辨識系統，這樣的字彙資料庫建製，將是非常的耗時又昂貴。

儘管如此，字彙資料庫未必包括所有使用者的用字遣詞。當說話者的用字並不包括在字彙資料庫內，那麼語音辨識系統就會呈現錯誤的訊息，在這種情況，要如何解決呢？或許我們可以設計語音辨識系統在遇到無法辨識說話者的用字時，出現空白的欄位，讓說話者輸入單字，使電腦能夠學習，並納入這個單字到資料庫內。或者能進一步地設計，讓說話者能進行錄音，使語音辨識系統在碰到類似的發音時，能針對該使用者的發音，辨識出所使用的單字。

除此之外，我們也可以設計如何讓電腦在遇到無法辨識用字時，能針對上下文來進行修復，使得文意通順。另外，未來語音辨識系統的研究，也可以考慮納入自然情境之下的說話方式，而不止是在安靜的實驗室裡，用麥克風說話。同時，也可以考慮如何將說話的語調和韻律納入辨識系統的訓練 [Bates 2006]。

在多元文化的現代社會裡，也有必要開發多語言的語音辨識系統。例如，在機場時，如果工作人員聽不懂一位乘客的語言，那麼就可以使用這個多語言的語音辨識系統來進行偵測，辨識這個說話者的母語是什麼；當能夠正確地辨識這個說話者的母語，那麼機場的工作人員就能即時派遣會說這名乘客母語的服務人員來溝通並予以協助。這種能辨識多語言的語音系統，是由兩種資料庫來進行設計和開發。第一，彙整各種語言的主要語音資料，包括獨特的子音或母音。第二，根據每一種語言獨特的語音結構和音節的設限來進行比對，電腦將會根據說話者

的語音和音節結構來進行分析，然後篩選出最有可能是這名說話者的母語 [Lyu 2006; Reiter 2006]。

其實，台灣人說中文的語音特徵和大陸中文腔調有很多不同處，那要建立中文語音辨識系統的語音資料庫時，該如何取捨呢？針對這個問題，我們可以先將比較頻繁的台灣中文語音特徵納入在語音辨識系統的資料庫內，例如，在台灣有很多的人，尤其是四十歲以下的台灣人，無法區分ㄣ和ㄥ這兩種鼻音，當母音是ㄧ或ㄜ的時候，大多數的台灣人都把ㄥ唸成ㄣ，如：把「平」唸成「頻」，把「耕」唸成「跟」，這種台灣人的中文語音特徵非常頻繁，所以可以納入在特別爲台灣人設計的中文語音辨識系統內，以提升辨識的效果 [Yang 2007]。

綜觀而言，語音辨識系統的未來發展要考慮許多層面，包括如何辨識不清楚的語音訊息，例如接收電話線的訊息；還有如何在有噪音的自然環境裡辨識語音訊息；以及如何將音調和韻律的語言訊息加入辨識模式的訓練，以提升辨識效果。另外，也要考慮多元文化社會下，交換各種語言的溝通方式；也要考慮說同一種語言有不同腔調的現象，並考慮外國人的腔調等等。這些因素，就現階段來說，有的比較難達成，有的比較上手，可立即獲得改善。比較難克服的問題，包括如何將音調和韻律的文法訊息加入辨識模組的訓練；還有如何有效地辨識在噪音環境下的自然說話訊息；例如，如何提升辨識從電話線發送的訊息，是非常困難的突破技術，因爲就目前爲止，語音辨識系統還是必須依賴說話者在安靜的環境下使用麥克風，清楚地發音。

儘管如此，比較容易改善的現象，包括開發辨識多語言的語音辨識系統，因爲這種偵測多語言的辨識系統只需考慮各別語言之間的特別音素，或是獨特的音節結構，這樣電腦就可以整合所有的語音訊息，判斷說話者的最可能母語。

另外一個比較容易改善的現象，就是開發一種可以包容各種說話者腔調的語音軟體，因爲，只要將不同腔調的獨特語音特徵加入在語音訓練模組內，就可以有效地提升語音辨識效果。

五、未來發展

除了改善語音辨識系統的現況之外，電腦語言學的研究發展也可以從其他許多層面來進行。從本人研究實驗語音學和語言習得的層面來說，至少有三個層面可以在未來努力研發。第一，電腦語言學者可以開發軟體，藉由機器的輔助來整理並分析語言資料，像這類的語言分析軟體對語言學的研究非常有幫助，可以有效地節省分析語料的時間，借由電腦程式的補助，自動運算語言資料，分析結果，將可大幅改善費時又費工的語言分析工作。例如，語音社會學者常要分析語音變異的資料，從量化的資料當中去找出語音變異的現象，並檢視這種語音變化與社會族群之間的關係。這種語料的分析，非常繁瑣，因此電腦語言學者可以開發一個

語音分析軟體，協助語音社會學者在龐大的語料庫當中，分析檢索語音變化的測驗單字以及出現率。進一步，這種語音分析軟體也可以統計語料，輔助語言學者比較兩個說話族群之間的語音差異程度是否明顯重要。

未來的研究也可以運用電腦來運算並檢視兩種語言的差異是如何影響到彼此的互相溝通度。這種語言分析軟體的開發，將有助於我們分辨兩種語言的差異，以及這差異是否會影響到溝通度，藉由運算這兩種語言溝通度的結果來分類這兩種語言之間的關係，探討他們是否爲同一種語言的變體，還是已經演化成爲兩種不同的語言體系 [Asher 2005]。

坦白來說，爲語言學者而設計的語料分析軟體，似乎沒有實際的商業市場，但是，對學術的研究，卻有相當大的貢獻。本人懇切期望台灣的電腦語言學者或從事資工的研究者能針對語言學家的需要，來開發分析語料的軟體，如果這種語料分析的軟體可以提供在網路上，讓其他國家的語言學者下載使用，如此台灣在國際上對語言學研究的學術貢獻，將功不可沒。

最後，本人認爲電腦語言學的未來研究發展也可以開發一些口語學習訓練的軟體；雖然目前有許多的網路資源提供英語學習，也可以見到許多英語學習的電腦軟體，但是針對英語口語學習的軟體，寥寥可數，而且有些問題有待改進。本人認爲一個針對英語口語訓練的電腦軟體必須可以診斷出英語學習者的口語問題，進而提供專業的建議，並提供聽講練習，以有效幫助英語學習者增強口語的能力。要達到這個目的就要先了解台灣人在英語學習時常會碰到的發音困難，因此，建立一個台灣英文的語音資料庫，可以提供語言學者分析台灣人說英語的常見語音特徵，將這些語音特徵整合在語音辨識軟體內，將可以預測台灣人說英語的發音困難，進而提供改善發音的方法，以及配合聽講的練習，期許英語學習者能有效改善口語的能力。

最後，本人也認爲，一個英語聽講練習的電腦軟體，有必要建置一個世界英語腔調的資料庫，提供給英語學習者做聽力練習，甚至是口語的訓練。世界英語的概念在英語廣泛做爲國際語言的趨勢下，十分重要。過去二十多年來，世界英語的研究已經受到愈來愈多英語語言學者和從事英語教學工作者的重視，這股研究世界英語的趨勢，以及熱烈的論壇，已經得到許多國際英語協會或測驗組織的肯定。例如，多益考試已經在聽力測驗上融入四種英語腔調，包括英式英語，美式英語，加拿大英語以及澳洲英語。新的托福考試在聽力測驗上，也包含有美式英語和英式英語，因此考生必須熟悉不同的英語腔調，以增強聽力能力。所以，一個英語聽講練習的電腦軟體，有必要提供不同英語腔調的語音檔，並解釋說明語音腔調的異同性，讓英語學習者瞭解及熟悉不同的英語腔調，如此不僅可以增強英語的聽力，同時也能有效地使用英語與不同國籍的人進行國際溝通。

總而言之，未來的電腦語言學研究可以從許多層面來發展，就如本文所提到的，電腦語言學者可以開發更方便使用的電腦程式或軟體，包括機器翻譯機、語音辨識軟體、多國語言偵測器、語料分析工具、語言統計分析比較軟體，以及世界英語聽講練習的軟體等等，以上這些研究發展，本文已略述其過去、現在和未來的研究發展情況，希望對電腦語言學有興趣的人，有所助益。

註：本文作者感謝陳怡君同學的資料打字與整理。

作者簡介：

本文作者楊孝慈是美國伊利諾大學香檳校區的語言學博士，現任國立雲林科技大學應用外語學系的專任助理教授，主要研究包括：實驗語音學、社會語言學、資料庫語言學、語言習得，以及世界英語。目前研究台灣人說英語的語音特徵，以及分析英語的語音差異對溝通度的影響。

參考文獻

呂仁園，<台灣三種漢語（華語、台語、客語）之多語語音辨識與語音合成引擎之研製及其於嵌入式系統之應用> NSC, Taiwan, 2006.

Asher, N. and A. Lascarides (Ed.), *Logics of Conversation.* Cambridge, 2005.

Bates, M. and R. M. Weischedel (Ed.), *Challenges in Natural Language Processing*. Cambridge, 2006.

Bosch, P. and R. van der Sandt (Ed.), *Focus: Linguistic, Cognitive, and Computational Perspectives.* Cambridge, 1999.

Cole R. *et al.* (Ed.), *Survey of the State of the Art in Human Language Technology.* Cambridge, 1998.

Cheng, C. C., Quantifying dialect mutual intelligibility. In C. J. Huang & Y. Li (Eds.), *New horizons in Chinese linguistics.* Dordrecht, Netherlands: Kluwer, 1996, pp. 269-292.

Daelemans, W. and A. van den Bosch (Ed.), *Memory-Based Language Processing. Cambridge, 2005.*

Rayner, M. *et al.* (Ed.), *The Spoken Language Translator.* Cambridge, 2007.

Reiter, E. and R. Dale (Ed.), *Building Natural Language Generation Systems.* Cambridge, 2006.

Shih, C-L, *Helpful Assistance to Translators: MT & TM.* Bookman Books, 2006.

Yang, J.H. The Role of Sound Change in Speech Recognition Research. 19th Conference on Computational Linguistics and Speech Processing, Taiwan, 2006.