印刷品

計算語言學通訊

第十四卷第二期

發 行 人：張俊盛
主 編：曾淑娟
執行編輯：黃琪
地 址：台北市研究院路二段 128 號中研院資訊所
劃撥帳號：19166251
電 話：(02)2788-3799ext.:1502
傳 真：(02)2788-1638
E-mail：rocling@hp.iis.sinica.edu.tw
網 址：http://rocling.iis.sinica.edu.tw/ROCLING/
討 論 區：aclclp-rocling@listserv.nthu.edu.tw

本 期 要 目

徵求 Rocling 研討會論文

學會為促進「計算語言學」相關領域研究資源之溝通，計畫建置歷年來收錄於「計算語言學研討會(ROCLING)」之論文全文資料庫，以供學者上網檢索。因此，誠摯地請歷年來論文曾發表於「計算語言學研討會」之學者，提供全文電子檔，以促進計算語言學學術研究成果交流。論文電子檔請傳至本會秘書處：黃琪小姐 (jessie@hp.iis.sinica.edu.tw)。

全國博碩士論文資訊網

「全國博碩士論文資訊網」為教育部委託國家圖書館執行的專案計畫，此系統已推出全國博碩士論文線上建檔系統，共收集論文全文 28,207 筆、博士全文影像 8,553 筆及摘要 233,526 筆。希望從事「計算語言學」相關領域之博碩士生，本於充實教育研究資源及提供全民共享的學術研究之目標，踴躍參與線上資訊庫建置的行列，論文上網建檔之相關說明請參閱國家圖書館全國博碩士論文資訊網。網址如下：
(http://datas.ncl.edu.tw/theabs/1/)

第三屆碩士論文獎開始申請

申請資格：

1. 國內大專院校碩士班應屆畢業生及其指導教授從事計算語言學相關研究方向者。
2. 參賽限制：每位指導教授以推荐兩篇論文為限(含個人指導與共同指導)。
3. 申請人應注意其所申請之論文絕無抄襲、剽竊情事，若經發覺除追回其所獲得之獎金、獎狀外，一切後果自行負責。

申請期間：六月一日起至七月十五日止
申請辦法請逕自上網查詢：
http://rocling.iis.sinica.edu.tw/ROCLING/
或請電話聯絡 02-27883799*1502 黃琪小姐

中文計算語言學期刊 Vol. 8-1 出刊

「中文計算語言學期刊」2003 二月刊即將於六月底出刊，本期將以特刊方式出刊，共收錄了四篇文章，主題為「Word Formation and Chinese Language Processing」。期刊將於七月上旬郵寄予本會會員，屆時若未收到，敬請通知本會秘書處。

Research on Computational Linguistics Conference XV
第十五屆計算語言學研討會

September 25-26, 2003, National Tsing Hua University, Taiwan, ROC
http://rocling.iis.sinica.edu.tw/ROCLING/conference/conference_cf.htm

CALL FOR PAPERS

Sponsors:
Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

Department of Computer Science and Department of Foreign Languages
National Tsing Hua University

ROCLING

Scope:

Papers are invited on substantial, original, and unpublished researches on all aspects of computational linguistics, including, but not limited to the following topic areas.

(a) cognitive linguistics
(b) discourse modeling
(c) document database/large corpora
(d) electronic dictionaries
(e) information retrieval
(f) language understanding
(g) language processing over Internet
(h) machine translation
(i) NLP and educational applications
(j) morphology
(k) computer assisted language learning
(l) parsing/generation
(m) phonetics/phonology
(n) quantitative/qualitative linguistics
(o) speech analysis/synthesis
(p) speech recognition/understanding
(q) spoken dialog systems
(r) spoken language processing
(s) syntax/semantics
(t) others

Paper Submission:

Softcopy in pdf, rtf, or Microsoft Word format* of a finalized full paper and *one document submission form (sample accessible via the above web page) should be sent to jschang@cs.nthu.edu.tw. The submitted papers should be written in either Chinese or English and in a format of single column, double-spaced with maximum 25 A4-sized pages. The first page of the submitted paper should bear the items of paper title, author name, affiliation and email address. All these items should be properly centered on the top, with a short abstract of the paper following.

The best paper award will be announced at ROCLING XV.

Important Dates:

Preliminary paper submission due:	July 15, 2003
Notification of acceptance:	August 13, 2003
Final paper due:	August 31, 2003

Call for Workshop-Proposals:

The Program committee welcomes submissions of proposals for workshops to associate with Rocling. For details on workshop submission please contact Jason S. Chang (Email: jschang@cs.nthu.edu.tw). Currently, the following satellite workshops are being organized:

Workshop on Information Retrieval
- featuring invited talks and penal discussion on cross language information retrieval, summarization, and question answering

Workshop on Computer Assisted Language Learning
- featuring invited talks and penal discussion on topics related to computer assisted language learning.

Workshop on Speech and Natural Language Processing
- featuring invited talks on speech and natural language processing and some 20 project notes on NSC-sponsored speech and natural language processing projects

ACL02 會後心得

張俊盛
清華大學 資訊工程系

會議名稱：第四十屆計算語言學學會年度會議
（The 40th Annual Meeting of the Association for Computational Linguistics (ACL02)）
會議時間：2002 年 7 月 8-10 日
會議地點：美國賓州費城
發表論文：利用長度與字彙資訊的適應性的句子對應研究
（Adaptive Sentence Alignment based on Length and Lexical Information）

國際計算語言學學會（Association for Computational Linguistics, ACL）為計算語言學學界中，最重要的一個國際性組織，ACL 在歐洲並設有分會。學會的主要活動，一者是出版 Computational Linguistic 期刊，另一個就是每年舉辦的學術會議。ACL 年度會議歷史悠久，此次會議剛好屆滿 40 年，在計算語言學的最重要的研究機構 – University of Pennsylvania 舉行，顯得特別有意義。本屆的參予的學者也達到空前的 674 人。本人和萬能技術學院資訊系的莊暢教授，一起發表台灣唯一入選的一篇研究成果，並做軟體系統的示範。另外致遠科技的蘇克毅博士，也參加大會的主會議及其他的活動。

從七月七至十二日的會期中，共有四個 tutorials、二十一個技術分場、四個軟體示範分場兩個、三個學生論文分場、十三個工作坊。申請人在主會議中，十日上午的 Demo Session 中，以海報及軟體示範的方式，發表研究成果。我們提出中文與英文對譯文件的句子對應的新做法，題為「利用長度與字彙資訊的適應性的句子對應研究」（Adaptive Sentence Alignment based on Length and Lexical Information - NSC 90-2411-H-007-033-MC）。相信該論文已為英文與中文的機器翻譯與跨語言檢索研究注入新的技術。

每年的 ACL 會議中首先登場的 tutorials 大都是由學界的新秀講授，這些 tutorials 常常能相當程度的反應出計算語言學界（Computational Linguistics, CL）的發展趨向。技術論文方面，會議中所宣讀的論文範圍極為廣泛，包括自然語言處理、語言學的許多層面，無法一一詳述。以筆者個人的偏見，覺得以下的論文，以及相關的研究趨勢特別值得注視：

1. **雙語語料庫對應分析、機器翻譯系統之自動化評估**
   1.1 Using Similarity Scoring to Improve the Bilingual Dictionary for sub-Sentential Alignment.
   Katharina Probst, Ralf Brown
   1.2 Adaptive Sentence Alignment based on Length and Lexical Information
   Thomas C. Chuang and Jason S. Chang
   1.3 Discriminative Training and Maximum Entropy Models for Statistical Machine Translation
   (ACL 2002 Best Paper Award Winner)
   Franz Josef Och, Hermann Ney
   1.4 Bleu: a Method for Automatic Evaluation of Machine Translation
   Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu
   1.5 Measuring Text Reuse
   Paul D. Clough, Robert J. Gaizauskas, Scott S.L. Piao, Yorick Wilks
   1.6 Evaluating Translational Correspondence using Annotation Projection
   Rebecca Hwa, Philip Resnik, Amy Weinberg
   1.7 Learning a Translation Lexicon from Monolingual Corpora
   Philipp Koehn and Kevin Knight
2. **機器翻譯機器、輔助人工翻譯**

2.1 Translating Named Entities using Monolingual and Bilingual Resources
Yaser Al-Onaizan, Kevin Knight
2.2 A Decoder for Syntax-based Statistical MT
Kenji Yamada, Kevin Knight
2.3 TransType: Text Prediction for Translators
George Foster, Philippe Langlais and Guy Lapalme

3. **專名實體辨識與自動問答系統**。
3.1 Learning Surface Text Patterns for a Question Answering System
Deepak Ravichandran, Eduard Hovy
3.2 Automated Question Answering in Webclopedia
Ulf Hermjakob, Eduard Hovy and Chin-Yew Lin
3.3 Named Entity Recognition using an HMM-based Chunck Tagger
GuoDong Zhou, Jian Su
3.4 Ranking Algorithms for Named Entity Extraction: Boosting and the Voted Perceptron
Michael Collins
3.5 Performance Issues and Error Analysis in an Open-Domain Question Answering System
Dan Moldovan, Marius Pasca, Sanda Harabagiu, Mihai Surdeanu

4. **非督導式的機器學習、語意歧異解析**。
4.1 Word Translation Disambiguating using Bilingual Bootstrapping
Cong Li, Hang Li
4.2 An Unsupervised Method for Word Sense Tagging using Parallel Corpora
Mona Diab, Philip Resnik
4.3 Acquiring Collocations for Lexical Choice between Near-Synonyms
Diana Zaiu Inkpen and Graeme Hirst

5. **語音辨識、語音對語音翻譯、對話分析**。
5.1 Automatic Interpretation System Integrating Free-Style Sentence Translation and Parallel Text Based Translation
Takahiro Ikeda, Shinichi Ando, Kenji Satoh and Akitoshi Okumura
5.2 AutoTutor: A Conversational Tutoring Environment
Kazama, Jun'ichi and Makino, Takaki and Ohta, Yoshihiro and Tsujii, Jun'ichi
5.3 Speech Translation on a Tight Budget without Enough Data
Robert E. Frederking, Alan W Black, Ralf D. Brown, Alexander Rudnicky, John Moody and Eric Steinbrecher
5.4 What's the Problem: Automatically Identifying Problematic Dialogues in DARPA Communicator Dialogue Systems
Helen Wright Hastie, Rashmi Prasad and Marilyn Walker
5.5 A Phrasebook Style Medical Speech Translator
Manny Rayner and Pierrette Bouillon

大會並設立終生成就獎，今年首次頒發給 U Penn **以** Tree Adjoining Grammar **見知於學界的** Joshi **教授，表彰他理論創新、人才教育、研究資源的建立多方面、長期的貢獻。這也顯示計算語言學界四十而立，邁入成熟的境界，形成一個豐富的研究領域。這一屆的會議吸納了機器學習**（Machine Learning）、**資訊檢索**（Information Retrieval）、**資訊淬取**（Information Extraction）**等領域的研究議題與研究人員，使得計算語言學的研究更蓬勃壯大**。

大會最後的一個節目，特別邀請 Peter Norvig **（主管** Google **公司的品質與研發）的演講。講題是** Better Web Search, with and without Computational Linguistics。Norvig **博士是人工智慧與自然語言理解的專家，執筆寫過一本廣為各大學採用的人工智慧教科書，他指出：資**

訊檢索已經從專家的研究工作，變成全民運動。而搜尋引擎也從過去的簡單的單字的向量模型，轉變成包括網路連結圖的分析。但是目前尚未廣泛的引入完整的自然語言分析的技術。他在演講中，指出有許多的技術，很有應用的潛力，可能會融入下一世代的搜尋引擎。這些高潛力的自然語言處理與資訊檢索技術包括：

1. 高效率的網頁的語言識別（Language Identification）與語言相關的停止字。
2. 簡短查詢關鍵詞的翻譯。
3. 文件與查詢關鍵詞的分類。
4. 自動網頁摘要。
5. 網頁垃圾關鍵詞的防治。
6. 相關網頁的自動分析與呈現。

我在會場提問，請教 Norvig 博士對跨語言檢索研究的看法。他的答覆，指出跨語言檢索，應該包括查詢關鍵詞與文件的翻譯，應用的潛力很大，適當的文件翻譯技術將來在搜尋引擎內，應該會有很大的發展的空間。反過來，他也指出搜尋引擎也可以回過頭來對自然語言處理做出深刻而重大的影響語貢獻。像 Google 這樣的超大型的搜尋引擎，其實就能提供一個簡單的介面，讓自然語言處理系統，查詢有史以來最大最新的語料庫，可以達到人工建立的知識庫所不能及的涵蓋性與精確性。Google 已經在研究途徑讓企業與學界，能夠更直接的利用其資源。現在在進行中的 Google in a Box 就是其中的一種方式。

計算語言學學會（Association for Computational Linguistics, ACL）已經邁出國際化的不可逆的一步，在原有的歐洲分會之外，成立了美洲分會之後，惟亞洲方面受制於日本，目前尚未成立分會。從 1998 年起，改變長年在北美洲舉辦會議的慣例，輪流在美，歐，亞三洲舉辦年會。台灣的計算語言學界經過十多年的耕耘，也卓然有成，向 ACL 之外的國際組織 ICCL 爭取到公元 2002 年 Coling 國際會議的主辦權。希望教育部、國科會等機構能夠大力支持，強化台灣在此一領域的形象，擴大影響力，在計算語言學界的國際組織中扮演積極參與的角色。

由於全球化的腳步愈來愈快，國際情勢的變化，區域政治的勢力消長，國際貿易、軍品的對外採購、基礎建設的對外採購，再再都使得外國文件的處理與外國情報的蒐集，如檢索、翻譯的需求，往往超越現有人力資源的負荷，必須求助於自動化處理。而語言處理、人機介面等研究近年來也逐漸成熟，可以支援可行的應用系統。特別是 Internet 興起，對計算語言學界而言，意味著資訊檢索、機器翻譯等 NLP 應用普及化時代的來臨。廣受注目的 Google 等資訊服務的核心技術與未來的發展都需要自然語言處理與資訊檢索的研究與技術。然而這些資訊技術與服務的技術。我們期待產官學界能夠透視新趨勢、新產品的表象、深入了解其核心，多多支持國內研究機構，加強有關自然語言處理、資訊檢索、機器翻譯等核心層面的建設、研究、開發。

漢語口語語料庫研究

曾淑娟
中央研究院語言學研究所籌備處

前言

在德國八年，我從事的都是德語的研究。雖然偶爾也碰一些西非帶聲調的語言，像是 Togo 的 Ewegbe（調值只有高調和低調），但是重點還是在德語口語語料的分析。在國外，大型的自然口語語料庫很早就已經展開建置工作，因此九０年代已是開始使用大型語料進行研究的時代。反觀臺灣，卻還沒有開始。因此在這幾年當中，我利用不同計畫的機會，嚐試建立不同類型的漢語口語語料庫。在這篇短文中，我會先簡短介紹國外幾個眾所周知的語料庫，再談談建置中的漢語口語語料庫處理。

口語語料庫

通常講到口語語料庫時都免不了要提到情境設計、發音人選取、錄音程序、錄音地點、錄音設備、聲音處理與文字處理。情境設計跟語料類型有直接的關係。語料類型按準備與否可如下分類：

(一) 朗讀語料（read speech）是朗讀文字或是公眾人物完全依講稿內容朗讀演講。
(二) 準備性語料（prepared speech）則是公眾人物依講稿內容背誦演講、記者採訪或談話性節目主持人提出準備好的問題。
(三) 自發性語料（spontaneous speech）是平時自由的交談、無準備的談話。

按參與談話人數，則可如下分：

(一) 獨白（monologues），例如朗讀講稿與事實或故事陳述。
(二) 對話（dialogues）是記者採訪或談話性節目對談、兩人對話。
(三) 多人會話（conversations）是兩人以上交談或兩人以上談話性節目對談。

情境設計可以是完全無限制的自由談話（free conversations），也可以是針對特定主題的會話（topic-oriented conversations），或是執行預先設計好的任務對話（task-oriented conversations）、訪問（interviews）或是請發音人口頭陳述事實（narratives）。

發音人選取影響語料內容極大，絕大多數的語料庫發音人都以大學學生為主。錄音程序包括發音人所獲得的指示、閱讀或簽署的文件以及錄音程序的說明。錄音地點可以是在戶外、錄音間或普通房間。錄音設備則需說明所使用的錄音器械（通常是數位錄音機 DAT）、麥克風、錄音取樣品率、立體聲與否等等資訊。聲音處理相關的有檔案轉錄的過程，聲音檔處理所用的語音軟體與語料是否經過切音處理（labelling）。文字處理則是語料語言內容的轉記與標記。像漢語，通常會有兩層的轉記：漢字與拼音。標記依語料庫建立的目的不同，可以是幾個特定現象，也可以是完整的言談或語法標記。

語料庫研究內容包括標記系統的分析與語料建置的工具開發。標記部分主要是口語對話的言談結構分析與口語語音與語法分析。依語言材料的不同，言談結構的呈現亦會跟著改變。因此言談標記必須依照語料庫的建立目的與形態而設定。口語語音的內容，若非實際操作審視語料，是無法想像其中的多樣性。口語語音的分析與標記，若能由語言學家來執行，可方

便提供語音工程研究具代表性且經專業標注的語音訓練材料。語言內容雖是語料庫的重點，但是語料建置的工具也佔重要的一席之地。開發工具不佳，不僅收集速度慢，所收集與標記的內容，能否前後一致，也會產生問題。口語文字轉寫與資料庫建立的工具，也是口語語料庫研究重要課題之一。

在九０年代初幾個重要的語料庫都陸續完成，在這裡只提幾個較為重要的。

(一). Switchboard Corpus 含有 2430 個特定主題的電話對話(美式英語)，共約 240 小時錄音，3 百萬字。

(二). Map Task Corpus 以含有能誘發語誤的地名設計讓兩人一組的發音人進行地理導引對話（英式英語)，共 128 對話，180 小時錄音。

(三). TRAINS Corpus 更是設計了 20 個不同形式的任務，錄製 98 個對話（美式英語)，共 6.5 小時錄音，55000 字。

(四). BAUFIX Corpus 則是包含 30 小時模擬的人 機對話（Wizard of Oz)，5 小時人 人對話（德語)，對話內容是合作組裝模型飛機。部件的形狀、顏色、大小（名詞片語）以及組裝的動作與順序（動詞片語）是主要要收集的語料。

BAUFIX Corpus 語料收集的目的，是在建立模擬機器人理解與操作組裝的大架構底下。因此設計時考慮到特定的詞彙領域。後續的研究也多放在語法與詞彙分析上。有類似目的的 TRAINS Corpus，語料量雖然不多，但是聲音檔都有音素與字時間點的切音位置（time-aligned phonemic and word transcriptions)，可供後續的語音研究。至於 Map Task Corpus，則因情境設計有語誤的聲韻（phonological analysis）與不流暢語流分析（disfluency)，以及後續的心理語言學實驗。Switchboard Corpus 語料量之大，使得陸續有些大型的標記應用計畫得以應用，例如 60 個言談標記（DAMSL, Dialogue Act Markup in Several Layers)，便標記了 1155 個五分鐘的對話，且用以訓練並建立言談的語言模型。也因其語料為某種程度上自由的對話，而有豐富不流暢語流的語料，故能進行大量的言談分析。

建置中的漢語口語語料庫

從一九九九年到現在，在中央研究院語言學研究所籌備處共錄製了四個不同設計的漢語口語語料庫：現代漢語連續口語對話語音語料庫（Mandarin Conversational Dialogue Corpus)、現代漢語地圖導引口語語音語料庫（Mandarin Map Task Corpus)、現代漢語主題對話語音語料庫(Mandarin Topic-oriented Conversation Corpus)，與現代漢語新聞朗讀語音語料庫(Mandarin Read News Corpus)。

一九九九年錄製現代漢語連續口語對話語音語料庫時，發音人是由中央研究院調查研究工作室依據 16-25 歲、26-35 歲以及 36-45 歲等三大年齡層，男女各半，從台北市市民中隨機抽樣選出。抽樣結果取得 1080 位候選人後，再寄出邀請函詢問是否有意願至中研院參與錄音。而由於計劃初始階段只預計收集 30 個對話，因此我們從回覆同意前來參與錄音的人之中，依他們的回覆順序取前 60 位，共 37 位女性、23 位男性。發音人來自社會不同階層，從高中生到博士班學生，上班族到經理、老師、家庭主婦等都有。所錄製的對話是陌生人之間的對話，會較為生疏。之後的三個語料庫，則是邀請這六十人中的三十人，各自帶一位他們熟識的親友前往錄製，因此是熟識人之間的對話。在現代漢語地圖導引口語語音語料庫中，持詳圖的一律都是曾經參與過第一次錄音的發音人，以便取得同一發音人較多的語料。

在正式錄音之前，我們都會先向發音人說明收集語料的目的以及整個計劃的研究目標，然後說明錄音過程。待他們瞭解同意後，請他們簽署同意書、填寫基本資料和語言使用的問卷，

最後再請他們閱讀過程說明，直到他們清楚瞭解才開始錄音。發音人被告知應儘可能自然地談話，不必特別注意句法或發音，因為計劃的目的是希望能收集到不同的說話腔調及風格。

所有四個語料庫的數位錄音都是採用 SONY TCD-D10 Pro II DAT 的數位錄音機，使用 Audio-Technica ATM 33a 單一指向性手持式麥克風。以取樣本率 48 kHz 將兩位發音人的語料分別錄於左右聲道。錄音地點為普通房間。為方便起見，其他有關語料庫的細節列於表一。

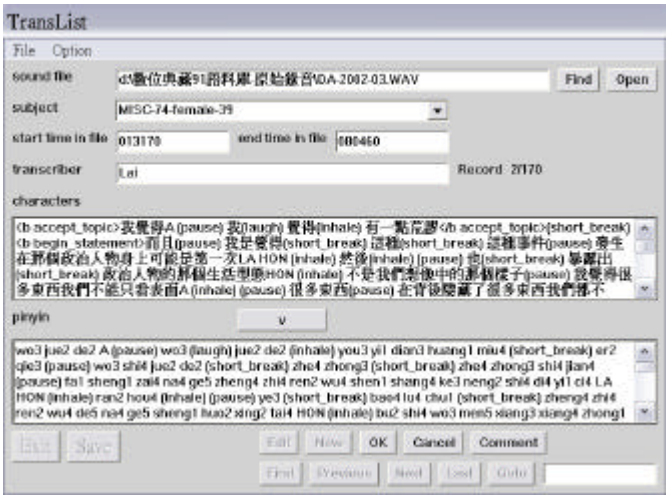
表一：錄製完成的漢語口語語料庫

	現代漢語連續口語對話語音語料庫	現代漢語地圖導引口語語音語料庫	現代漢語主題對話語音語料庫	現代漢語新聞朗讀語音語料庫
情境設計	收錄的是陌生人之間的自然對話。談話內容除了限定的路徑問題及自我介紹之外，並不限定特定主題，隨發音人自由交談。	地理任務導向對話。發音人雙方熟識，一持詳細地圖，一持刪減部分路名與建築物名後地圖，由持詳圖者依序引導持簡圖者至三個指定目的地。	熟識者之間的自然對話。 兩位發音人選定二ｏｏ一年中發生之一特定新聞主題或事件進行對談。	新聞稿的朗讀。新聞內容是從二ｏｏ一年各類十大熱門新聞中選出八篇報導，再統一刪減至適當長度。
收集時間	1999.03 – 1999.07	2002.01 - 2002.03	2002.01 - 2002.03	2002.01 - 2002.03
語料大小	30 個對話，共 25.6 小時，對話平均長度約 50 分鐘	30 個對話，共約 5 小時，對話平均長度約 10 分鐘	30 個對話，共約 11 小時，對話平均長度約 22 分鐘	60 個朗讀，共約 2 小時，每篇新聞朗讀平均長度約 2 分鐘
發音人	共 60 人 (男 23 位，女 37 位) 16-25 歲: 20 位 26-35 歲: 19 位 36-45 歲: 21 位	共 60 人 (男 27 位，女 33 位) 14-25 歲: 14 位 26-35 歲: 14 位 36-63 歲: 32 位	共 60 人 (男 27 位，女 33 位) 14-25 歲: 14 位 26-35 歲: 14 位 36-63 歲: 32 位	共 60 人 (男 27 位，女 33 位) 14-25 歲: 14 位 26-35 歲: 14 位 36-63 歲: 32 位
標記系統	詳細自發性口語現象	特殊語音現象	言談標記	無
文字處理	已轉記十四萬字（完成第一輪標記）	已完成所有對話文字轉記（標記進行中）	已轉記並標記七萬字	朗讀文稿
聲音處理	轉為立體聲 wav 檔與單聲 ptk 檔，並以三分鐘為單位切分	轉為立體聲 wav 檔並以每個任務為單位切分	轉為立體聲 wav 檔	轉為立體聲 wav 檔
音檔大小	18 GB	2.8 GB	6.78 GB	1.3 GB

有關漢語口語語料處理

TransList 是針對以上四個漢語口語語料庫的轉寫與標記所開發的工具，見圖一。欄位有聲音檔案、發音人、標記員、說話輪起訖時間、漢字轉記與拼音轉記和註記。轉寫文字檔格式，分為檔頭與單筆說話輪的轉寫內容。檔頭（header）記載錄音地點、錄音日期、錄音型態、錄音語言，取樣品率與錄音聲道等以及所有轉記欄位的資料，都將自動匯入資料庫中，以便建立日後的後設資料（metadata）。

圖一：轉寫介面 TransList



檔頭

<recordplace> Taipei, Taiwan
<recorddate> June 3, 2001
<speechtypei> spontaneous
<speechtypeii> dialogue
<language> Mandarin
<samplingrate> 48 kHz
<recordtype> stereo

單筆說話輪

****
<voicefile>d:**分割完成的檔**\stereo_01\mcdc-01-01.wav
<speaker>MISC-08-male-25
<start>000000
<end>009514
<translator>Fen
<chinese>
<b particle>EI </b particle><b clear throat>@</b clear throat>**你好我姓賴請問一下貴姓**<b hiccup>@</b hiccup>
<b breathe>@</b breathe>
</**chinese**>
<english>
EI @ ni3 hao3 wo3 xing4 lai4 qing3 wen4 yi2 xia4 gui4 xing4 @ @
</english>
<comment>
</comment>
****

每一筆紀錄代表一個說話輪（turn)。**每個說話輪的內容有聲檔名，發音人資料，說話輪起迄時間點，標記員資料，漢字與拼音轉寫與。若有與當筆說話輪相關的註記，則記於** comment **裡頭。聲音檔以**.wav **檔為主；發音人則附記了年齡與性別資料；起迄時間以毫秒**(msec)**為單位。漢字部分使用** Big5 **編碼，外來語以拉丁字母為主，語氣詞與感嘆詞以大寫的拉丁字母轉記；非語音現象則以@註記。標記格式為**<tag name>**語言內容**</b tag name>**。在漢語拼音部分則有實際發音的標示。以現代漢語地圖導引口語語音語料庫為例，所採用的標記包括自發性口語中特殊語音現象如音節連併，同化現象，拖長音與鼻化音等。處理好的標記語料，會經由** TransList **自動轉為** Access **格式資料庫。預先訂定好的標記集經程式處理會自動編號。語料中有加標記的字詞，欄位就會給定該編號為其對應欄位的值。如此把原先水平的轉記文字化**

為垂直的行列資料庫，可以克服多數口語語料庫無法進行檢索搜尋的問題。例如下列語言內容可以轉成圖二的資料庫。進而可擷取特定標記的字詞，進行後續的分析。

蓋章認可<b inappropriate pronunciation>的</b inappropriate pronunciation><b short break>@</b short break>只有<b assimilation>三分</b assimilation>之一<b inhale>@</b inhale><b marker>NA </b marker>其它的<b clear throat>@</b clear throat><b exhale>@</b exhale><b assimilation>三分</b assimilation>之二是<b inhale>@</b inhale>警察局自己<b pause>@</b pause><b inappropriate pronunciation>就</b inappropriate pronunciation><b inappropriate pronunciation>是</b inappropriate pronunciation>

圖二：匯入後資料庫格式

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
18304	073050	105588	68	蓋	gai4	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18305	073050	105588	69	章	zhang1	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18306	073050	105588	70	認	ren4	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18307	073050	105588	71	可	ke3	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18308	073050	105588	72	的	[le5]	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18309	073050	105588	73	@	@	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18310	073050	105588	74	只	zhi3	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18311	073050	105588	75	有	you3	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18312	073050	105588	76	三	san1	MISC-08-male-25	Fen	mcdc-01-2	0	2	0	0	0	0	0	0	0	0
18313	073050	105588	77	分	[men1]	MISC-08-male-25	Fen	mcdc-01-2	0	2	0	0	0	0	0	0	0	0
18314	073050	105588	78	之	zhi1	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18315	073050	105588	79	一	yi1	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18316	073050	105588	80	@	@	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18317	073050	105588	81	NA	NA	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18318	073050	105588	82	其	qi2	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18319	073050	105588	83	它	ta1	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18320	073050	105588	84	的	de5	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18321	073050	105588	85	@	@	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	4	0	0	0	0	0	0
18322	073050	105588	86	@	@	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	10
18323	073050	105588	87	三	san1	MISC-08-male-25	Fen	mcdc-01-2	0	2	0	0	0	0	0	0	0	0
18324	073050	105588	88	分	[men1]	MISC-08-male-25	Fen	mcdc-01-2	0	2	0	0	0	0	0	0	0	0
18325	073050	105588	89	之	zhi1	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18326	073050	105588	90	二	er4	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18327	073050	105588	91	是	shi4	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18328	073050	105588	92	@	@	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0
18329	073050	105588	93	警	jing3	MISC-08-male-25	Fen	mcdc-01-2	0	0	0	0	0	0	0	0	0	0

藉由口語語料庫的語言素材，不僅可以實證言談結構、語法類型，以及語音變化在標記系統與實際標記結果的相符性，更可對在自發性口語中，語法的使用與語音聲韻的特徵，就其中的交互現象，以量化和實驗語音學的研究方法進行分析。自發性口語對話的語法形態，應可反映出深層語法的骨架在自然語境下的變化，也可以驗證語誤等語言使用現象與漢語語法之間的因果關係。自發性口語中語法元素的省略、句子中斷、插入等現象，都必須經由實際語料來協助理論的分析。漢語語法，目前已有眾多學者從事理論研究或者是功能性分析。但真正從口語語料的語言現象出發，實地探討分析漢語口語語法的類型與本質，以及漢語口語語法與其他語言分支，例如言談結構、聲韻表達和社會環境對語言使用的影響，仍然缺乏。原因之一，在於口語語料資源未能充分分享，且檢索不易。藉由所開發的現代漢語口語語料與轉寫及可擷取的資料庫工具，相信能為漢語口語語法研究注入新血。

結語

學會很高興有清大的巫宜靜同學來幫忙統籌每期的通訊專欄。沒想到巫同學頭一個找上的竟是自己。希望這篇小小的文章能對讀者有所幫助。有關本文或是其他礙於篇幅沒能寫上的，讀者若有指教，誠心歡迎。

參考書目

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G.M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S. and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34, pp. 351-366.

Gibbon, D., Moore, R. and Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter.

Greenberg, S. (1996). The Switchboard Transcription Project. Proc. of the Large Vocabulary Continuous Speech Recognition Summer Research Workshop. Baltimore, Maryland, USA, 1996.

Heeman, P. and Allen, J. (1999). The TRAINS 93 Dialogues. TRAINS Technical Note 94-2. University of Rochester.

Hindle, D. (1983). Deterministic Parsing of Syntactic Non-fluencies. In Proc. of ACL ' 83 (pp. 123-128).

Hirschberg, J. and Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3): 501-530.

Levelt, W. J. (1983). Monitoring and Self-Repair in Speech. *Cognition*. 14. 41-104.

Lickley, R. and Bard, E. (1998). When Can Listeners Detect Disfluency in Spontaneous Speech? *Language and Speech*. 41 (2): 203-226.

Nakatani, C., Hirschberg, J. and Grosz, B. (1995). Discourse Structure in Spoken Language: Studies on Speech Corpora. Presented at the AAA-I-95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.

Sagerer G. and Eikmeyer H. and Rickheit G. (1994). "Wir bauen jetzt ein Flugzeug": Konstruieren im Dialog. Arbeits-materialien, Technical Report. SFB360 "Situierte Künstliche Kommunikation. University of Bielefeld, Germany.

Schegloff, E., Jefferson, G. and Sacks, H. (1977). The Preference of Self-Correction in the Organization of Repair in Conversation. *Language*. 53(2): 361-382.

Tseng, S.-C. and Y.-F. Liu. (2002). Annotation Manual of Mandarin Conversational Dialogue Corpus. Technical Report 02-01. CKIP, Academia Sinica. Taipei. (in Chinese)

Call for Submissions

Computational Linguistics and Chinese Language Processing invites submission of original research papers in the area of computational linguistics in general and Chinese (natural) language processing in particular. Contribution can be written either in English or Chinese. English will be the primary language of this journal for its international readership. A 600-word extended abstract in English is required for all Chinese contributions. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions are welcomed in the following three categories：

Papers：Submissions in this category should report significant new research results in computational linguistics or new system implementation involving significant theoretical and/or technological innovation. The accepted papers are divided into the categories of regular papers and short papers. There is no strict length limitation on the regular papers but it is suggested that manuscripts not exceed 40 double-spaced A4 pages. Short papers are restricted to no more than 20 double-spaced A4 pages.

Survey Papers：Submissions in this category are either invited by the editorial board or voluntary .They should offer a critical overview of either 1) the state of arts of a certain subfield of computational linguistics or Chinese language processing, or 2) of existing systems and/or technologies for a particular natural language application. Page limitation of this category is the same as above.

Project Reports：Submissions in this category should report on 1) new systems, or 2) new resources such as corpora, lexicon, tree-banks etc..

All contributions will be anonymously reviewed by at least two reviewers, with exception of the project reports, which will be reviewed by only one reviewer.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the CLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication.

Style for Manuscripts: The paper should conform to the following instructions.

1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.

2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires to begin with a capital letter. The author's name should appear below the title.

3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.

4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left- hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1-1,1-2...).

5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript

6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.

7. References: In each of the References, please make sure that the following are given

(1) names of all the authors.
(2) title of the paper
(3) full title of the journal
(4) volume number and year of publication
(5) first and last page number.

Example：Dempster, A. P., N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* 39 (B) 1977, pp. 1-38.

No page charges are levied on authors or their institutions.

Final Manuscripts in Electronic Form: If a manuscript is accepted for publication, the author will be asked to supply an electronic form (RTF or ASCII text file) of the manuscript via e-mail or on disk.

All manuscripts must be submitted in triplicate to:

Computational Linguistics and Chinese Language Processing
Professor Keh-Jiann Chen
Institute of Information Science, Academia Sinica, Nangkang, Taipei 115, Taiwan, R. O. C.
Electronic submissions are also welcomed. Please e-mail MS-word, PDF, or Postscript files to:
clp@hp.iis.sinica.edu.tw