

本 期 要 目

- | | |
|--|-----------|
| 壹. 「CLCLP 國際研究生」招生訊息 | 第二頁 |
| 貳. 學術活動預告-PACLIC-18 | 第三頁 |
| 參. 中文資訊檢索標竿測試集之建置-陳光華 | 第四~十二頁 |
| 肆. 電子郵件申請書及服務條款 | 第十三~十六頁 |
| 伍. 期刊徵求論文-Human Computer Speech Processing | 第十七頁 |
| 陸. 第二屆語言學卓越營心得報告-胡佳音 | 第十八~二十頁 |
| 柒. 第二屆學生計算語言學研討會紀要-巫宜靜 | 第二十一~二十八頁 |

賀！黃居仁博士應邀擔任「計算語言學國際委員會」終身委員

中研院語言所副所長黃居仁博士應邀擔任「計算語言學國際委員會」(International Committee on Computational Linguistics) 委員乙職。「計算語言學國際委員會」每兩年開會一次，委員為終身職，目前共有 20 位委員。亞洲地區除了日本以外，黃博士是第一位亞洲地區受邀之委員。此項殊榮不但是黃博士個人之榮譽，亦是本會之榮耀。

致謝捐款

本會終身會員中研院語言所副所長黃居仁博士惠贈現金一萬元，特此感謝！

學會網頁正式更新

本學會已更新原有網頁，新網址為 <http://www.aclclp.org.tw>。網頁含中英文內容，並加入學會出版品(期刊與會議論文集)檢索功能。請會員多多利用，若有指教請洽秘書處。

中文資訊檢索標竿測試集

本計畫已完成一組包含文件集、問題集以及答案集的測試集；同時也發展了一套建構測試集的程序。在各界急於研發中文資訊檢索系統的今日，中文資訊檢索標竿測試集第三版(CIRB030)應能解決國內無從取得中文測試資料的現狀，使資訊檢索系統的研究與發展能有更高的可行性。本測試集之建構是依據資訊檢索評估的相關理論，期望成為中文資訊檢索領域中一項重要的測試資源，從事中文資訊檢索或是跨語資訊檢索研究的學者專家均可使用該測試集以評量所研發之檢索系統的績效。本測試集得國立台灣大學圖書資訊學系陳光華教授授權學會發行；詳細說明請參閱第 4 12 頁。

電子郵件帳號開始申請

會員委員會為擴大對本會會員之服務，將免費提供電子郵件帳號，凡為本會之會員皆可提出申請，申請使用者需遵守本會所訂定使用「服務條款」，申請書及服務條款請參閱第 13 16 頁。

「CLCLP 國際研究生學程」已開放申請

中研院語言所將於 2005 年加入 TIGP(Taiwan International Graduate Program)國際研究生學程計畫，與清華大學資訊工程學系合作開設計算語言學與中文語言處理(CLCLP, Computational Linguistics and Chinese Language Processing)之分項學程。

此學程將針對中文語言處理及計算語言學提供具國際競爭的課程，讓優秀學子接受進階的專業訓練、且提供研究的機會。

CLCLP 將於 2005 年正式開始，申請學程的截止日期為 2005 年 3 月 31 日，學程提供名額有限，請及早申請。

申請學程所需備齊的文件如下：

1. Undergraduate and graduate (if applicable) academic records, or transcripts.
2. Graduate Record Examination (GRE) scores. (recommended)
3. TOEFL: All applicants whose first language is not English must submit a TOEFL score of 550 (213 on computer-based test) or higher. Applicants residing in Taiwan can also show a GEPT certificate for the High-Intermediate level instead. Applicants who have recently completed a degree program in an English speaking country can be exempted with a letter of proof from an advisor.
4. Three letters of recommendation commenting on the applicant's personal character and qualifications for independent study, including intellectual ability, research potential, and motivation.
5. A statement of purpose that includes a research plan.
6. (Strongly preferred) Working knowledge of Chinese: equivalent of 6 credit hours of college level Mandarin Chinese, or other proof of near-native fluency.
7. (Desirable) Proven ability in computational linguistics, such as published paper, completed project etc.

學程相關網站：

Taiwan International Graduate Program, Academia Sinica <http://www.tigp.sinica.edu.tw>

Institute of Linguistics, Academia Sinica <http://www.ling.sinica.edu.tw/>

Institute of Information Science, Academia Sinica <http://www.iis.sinica.edu.tw/>

Department of Computer Science, National Tsing Hua University: <http://www.cs.nthu.edu.tw/english/>

相關訊息請洽：

115 台北市研究院路二段 128 號 中研院語言所 胡雅蘋 小姐

Tel. 886-2-2786-3300 ext. 325

Fax. 886-2-2785-6622

e-mail. phdclclp@gate.sinica.edu.tw

PACLIC 18

The 18th Pacific Asia Conference on Language, Information and Computation
December 8 (Wednesday) - 10 (Friday), 2004
International Conference Center, Waseda University, Tokyo

Schedule and Important Dates (all dates in JST)

Pre-registration until: October 25

Registration Fees

Pre-registration online (by October 25 (JST)): 7,000 JPY

General walk-in: 10,000 JPY

Student (must show relevant ID) walk-in: 7,000 JPY

Student (without reception/proceedings) walk-in: 1,000JPY

NB: All fee categories above except the last
inclusive of informal reception on the second day
and a proceedings volume with CD-ROM.

Tentative Conference Program

<http://www.decode.waseda.ac.jp/PACLIC18/program.html>

Invited Speakers (in alphabetic order, all confirmed)

<http://www.decode.waseda.ac.jp/PACLIC18/invited-speakers.html>

Chu-Ren HUANG, Academia Sinica

Kiyong LEE, Korea University

Yuji MATSUMOTO, Nara Institute of Science and Technology

Satellite Workshops

Satellite Workshop 1: December 8 (Wednesday)

<http://www.decode.waseda.ac.jp/PACLIC18/satellite-workshop-1.html>

Satellite Workshop 2: December 10 (Friday)

<http://www.decode.waseda.ac.jp/PACLIC18/satellite-workshop-2.html>

Acknowledgements

Preparations for PACLIC18 was financially supported in
part by Waseda University Grant for Special Research
Projects: International Joint Research No. 2003C-201.

PACLIC18 is financially supported by Waseda University
Grant for International Conference Operation.

Inquiries should be sent to: pacllic18-sec@decode.waseda.ac.jp

PACLIC-STEERING mailing list

PACLIC-STEERING@ns.decode.waseda.ac.jp.

<http://ns.decode.waseda.ac.jp/mailman/listinfo/paclic-steering> ACLIC 18

跨語言資訊檢索測試集之建置

陳光華

國立台灣大學圖書資訊學系

khchen@ntu.edu.tw

一、計畫緣起

資訊檢索系統不論在設計、研發、運作等各階段，評估均是其中不可或缺的重要環節。透過此程序，研究者能藉以驗證系統效益、比較各種檢索技術的優劣，以作為改進之參考，使資訊檢索系統的運作及效能更臻完善。資訊檢索系統評估的研究發展，自 1950 年代至今，已有四十年以上的歷史。(註 1) 早期此方面相關的實證研究，大多是在規範化的環境 (Laboratory Environment) 中進行測試 (Test)，透過一些量化或質化的準則，衡量不同技術或不同系統間檢索效益之優劣。最早採用此評估模式的是 1966 年 Cleverdon 所進行的 Cranfield II 計畫，它以文件集 (Document Set)、查詢問題 (Question) 及相關判斷 (Relevance Judgment) 構成一組測試集 (Test Collection) 作為測試的基礎資料，並訂定一套效益測量準則 (Effectiveness Measurement)，以評估多種索引方式之優劣。(註 2) Cranfield 研究採用的實驗模型與測試方法，在系統評估的領域中一直廣受仿效與援用，直至今日仍佔有舉足輕重的開創性地位。然而，早期的測試集規模通常不大，與真實檢索環境間存在頗大的差距，因此植基於其上所發展的檢索系統，在實際運作時往往無法達到良好的效益。(註 3)

1992 年，美國國防部高等研究計畫署 (Defense Advanced Research Projects Agency, 簡稱 DARPA) 與美國國家標準暨技術局 (National Institute of Standards and Technology, 簡稱 NIST) 共同舉辦了文件檢索會議 (Text REtrieval Conference, 簡稱 TREC)，透過大型測試集的建構，以及測試項目、測試程序、評估準則的訂定，提供不同檢索系統與檢索技術之間的標準評比環境，並舉辦論壇提供參與者討論及分享結果。(註 4) 它首創了前所未有的大型測試集，使測試環境得以更接近真實的情況，對檢索技術的發展與系統效益的提昇具有相當重要的貢獻。

影響資訊檢索系統效益的因素十分廣泛而複雜，系統評估工作亦應考量到各個層面，並不能僅依據單純的量化準則。無可否認的，如同 Cranfield II 及 TREC 這般的測試機制，的確在許多方面都有其侷限與爭議性，但是至目前為止，它們確實是少數能得知系統可能效益的具體可行方案，對資訊檢索系統的研究與發展來說，還是具有十分重大的意義。

在今日資訊檢索研究蓬勃發展之際，各界紛紛意識到建立一致性評比環境的必要性。目前除了 TREC 之外，已有一些針對不同語言設計的類似機制嘗試開始運作，如 NTCIR (NACSIS Test Collection for IR Systems) 計畫 (註 5) 與 IREX (Information Retrieval and Extraction Exercise) 計畫 (註 6) 分別建立了日文測試集，AMARYLLIS 計畫則建立了以法文為主的測試集 (註 7)，台灣則建置了 CIRB (Chinese Information Retrieval Benchmark) 1.0 版與 1.1 版，以及 CIRB 2.0 版 (註 8)，韓國亦建置了 HANTEC 測試集。(註 9)

然而在今日跨語言資訊檢索的需求益形重要且迫切，國際性的合作與交流，以建置跨語言資訊檢索測試集，更是各國資訊檢索研究者關注的課題。目前，國際間形成三大跨語言資訊檢索評估機制，其一為美國的 TREC，其二為歐洲的 CLEF (Cross-Language Evaluation Forum) (註 10)，其三為東亞的 NTCIR (註 11)，而這三個機制彼此亦有合作的模式。東亞

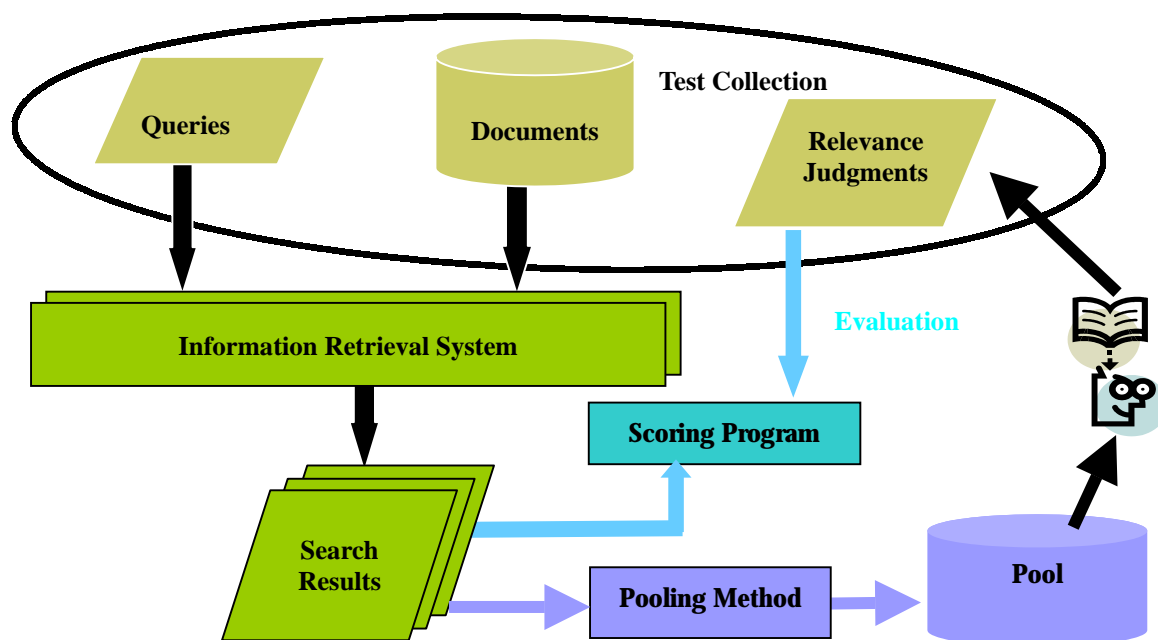
的 NTCIR 由日本、韓國、台灣合作，提供許多的資訊檢索評估項目，如跨語言資訊檢索，專利檢索，網頁檢索等。其中跨語言資訊檢索項目由國立台灣大學陳信希教授與本人共同擔任主席，負責規劃與執行。

二、計畫目的

有鑑於資訊檢索評估是長久的研究議題，資訊檢索測試集不僅對資訊檢索研究者是重要的研究資源，同時對從事計算語言學研究的學者專家而言，也是語言學研究的重要資源。中華民國計算語言學學會自成立以來，即積極蒐集建立語料庫資源，然多屬純粹的語料庫，如平衡語料庫 (Balanced Corpus)，詞性標記語料庫 (Tagged Corpus)，樹狀語言庫 (Tree Bank)，資訊檢索測試集則尚付之闕如。本人自 1999 年起開始進行資訊檢索測試集的建置，並推動使之實際地應用，NTCIR 資訊檢索評估會議，已經使用了本人建置的 CIRB010 (CIRB1.0 版)、CIRB011 (1.1 版)、CIRB020 (2.0 版)。測試集的廣泛使用，不僅協助資訊檢索系統的研究與發展，同時也讓本測試集在中文檢索方面具有一定的影響力。本計畫是持續建置 CIRB 測試集 (第三版, CIRB030)，測試集由三個部分組成：文件集 (Document Set)，問題集 (Topic Set)，答案集或稱為相關判斷 (Answer Set or Relevance Judgment)。

三、研究方法

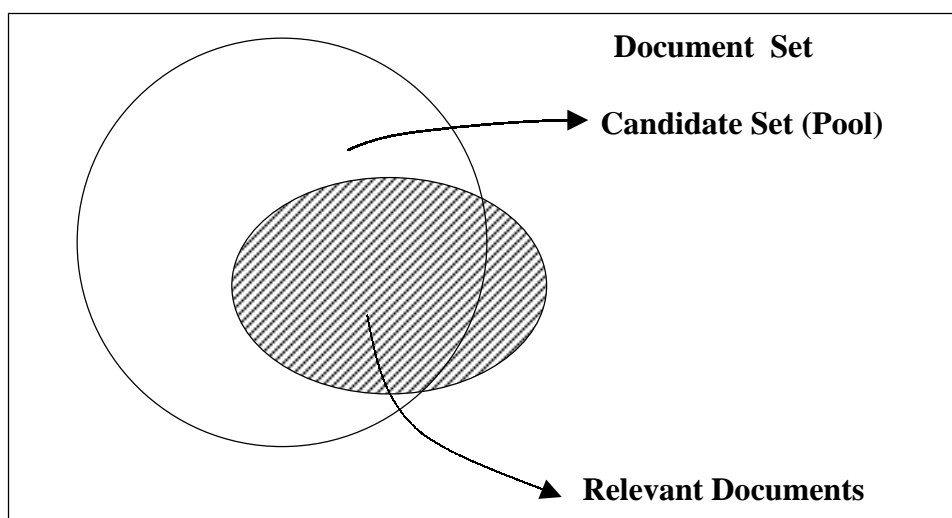
完整的資訊檢索系統評估模式如圖一所示，有測試集、待評估之系統，以及評估者之參與。如前述，測試集分為三部分，其中的文件集與問題集是事先完成的，而答案集是透過眾多資訊檢索系統的參與，以所謂的 Pooling Method 建構而得。



圖一：資訊檢索評估的運作模式

早期規模較小的測試集 (如 Cranfield II 與 CACM)，我們可以一篇一篇閱讀文件，以判斷文件的相關性；然而，目前測試集的文件都極為龐大，不可能閱讀每一篇文章，因此發展 Pooling Method。這個方法的假設是真正相關的文件，應該會被多數的資訊檢索系統找出來，所以將

眾多資訊檢索系統檢索而得的文件，建構為一個 Pool(也就是候選相關文件集，請參見圖二)，評估者僅需判斷這個集合的文件，可以降低製作測試集的時間與人力成本，亦有研究指出 Pooling Method 並不影響資訊檢索系統績效的相對排序使得 Pooling Method 廣為學界接受。



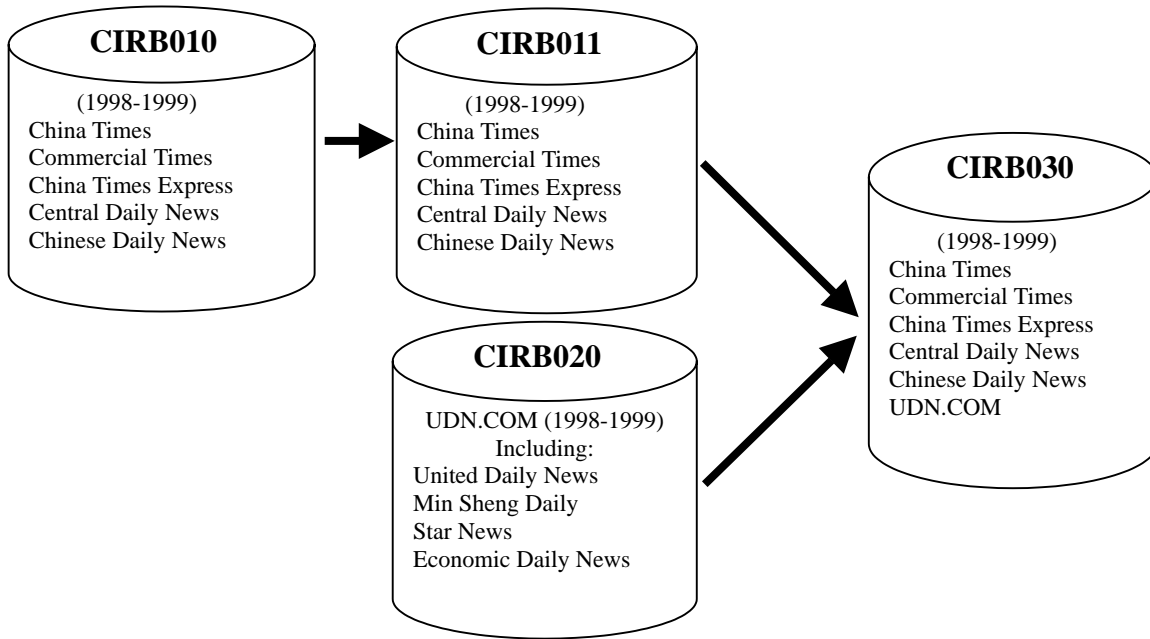
圖二：Pooling Method

問題集的建構亦有一定的檢測方法，而非恣意地蒐羅題目。CIRB030 的題目，經過兩次的過濾程序。第一次是必須通過中文新聞文件的前測；第二次是必須通過英文、日文、韓文新聞文件的前測。前測指的是每一個題目都必須符合“3-in-S+A”原則，S 代表非常相關，A 代表相關，3 代表文件數，因此，“3-in-S+A”的意思是適用的題目必須有 3 篇以上非常相關或相關的文件。

文件集則必須與新聞機構洽談，這部分是相當困難的，必須讓新聞機構理解，長久而言，測試集的建構對於新聞機構是有益的，因此，必須與機構的主事者討論，必要時，也要機構內部的成員演講。接著便是使用授權書的議定，通常需費時 6 個月，才能完成所有的程序。

四、計畫成果

基本上，CIRB030 的文件集是由 CIRB011 與 CIRB020 的文件集合組而成，但修改了部分文件亂碼的問題與部分文件 HEADLINE 與內文不符的問題，同時刪除無內文的文件，因此文件的數量與 CIRB011 及 CIRB020 的文件總數有些許出入。而 CIRB011 和先前發行的 CIRB010 僅有標記上的不同；CIRB011 與 CIRB020 的標記則完全相同。相關標記會於下文說明。這一次我們決定直接發行 CIRB030 而跳過 CIRB020，原因就是整合 CIRB011 的文件集與 CIRB020 的文件集。因此，使用者無須擔心如何取得 CIRB020 的問題，它不會在台灣單獨發行，除非您曾參與 NTCIR 資訊檢索評估會議。為了更清楚地說明 CIRB 版本的演變情形，請參考圖三。



圖三：CIRB 版本演變

另外必須要注意的是, CIRB030 的新聞文件已經整合為 7 個文件檔案, 它們是 cdn1998-1999 (中央日報), chd1998-1999 (中華日報), ctc1998-1999 (工商時報), cte1998-1999 (中時晚報), cts1998-1999 (中國時報), udn1998 (聯合報系 1998), 與 udn1999 (聯合報系 1999); 而這些文件在 CIRB011 與 CIRB020 時是各自獨立的, 因此有 381,681 個文件檔案。下文將分別說明文件集、問題集與答案集三部分。

(一) 文件集

CIRB030 的文件集是由不同的新聞機構合法取得的, 表一說明這些新聞文件的來源與數量。新聞文件的內碼為 BIG5, 且經過後續處理, 加上適當的 XML 標記, 這些標記是經過 NTCIR 執行委員會的討論而制訂的, 表二羅列這些標記, 並說明其意義。圖四則是一個標記後新聞文件的例子。

表一：文件集的組成

中央日報 (cdn1998-1999)	27,770
中華日報 (chd1998-1999)	34,728
中國時報 (cts1998-1999)	38,116
中時晚報 (cte1998-1999)	5,747
工商時報 (ctc1998-1999)	25,811
聯合報系 (udn1998 and udn1999)	249,203
總計	381,375

表二：CIRB030 文件的標記

必要標記		
<DOC>	</DOC>	The tag for each document
<DOCNO>	</DOCNO>	Document identifier
<LANG>	</LANG>	Language code: CH, EN, JA, KR
<HEADLINE>	</HEADLINE>	Title of this news article
<DATE>	</DATE>	Issue date
<TEXT>	</TEXT>	Text of news article
選擇標記		
<P>	</P>	Paragraph marker
<SECTION>	</SECTION>	Section identifier in original newspapers
<AE>	</AE>	Contain figures or not
<WORDS>	</WORDS>	Number of words in 2 bytes

```

<DOC>
<DOCNO>udn_xxx_19980101_0001</DOCNO>
<LANG>CH</LANG>
<HEADLINE> 南華早報報導中共內部兩件與台灣相關新發展： </HEADLINE>
<DATE>1998-01-01</DATE>
<TEXT>
<P>香港英文南華早報今天報導，一九九八年的台灣新聞，可能會和一九九七年的亞洲金融風暴一樣成為最大的新聞事件。</P>
<P>該報說，中共在過去一周傳出兩件與台灣有關的新發展。一是以中共國家主席江澤民為組長的中央對台灣工作領導小組，將增加幾名重量級的文職和軍職成員 而軍方代表包括中央軍委兩位副主席張萬年和遲浩田。</P>
<P>無論從任何角度分析，中共計畫擴大中央對台領導小組，顯然都是旨在加快打破兩岸目前的僵局。</P>
<P>南華早報說，第二項發展是中共通過內部文件向各級幹部，特別是台灣事務幹部，傳達海峽兩岸僵局一兩年內一定會達成突破的信息。</P>
</TEXT>
</DOC>

```

圖四：新聞文件範例

(二) 答案集

CIRB030 的問題集是由日本、韓國、台灣、以及 TREC 共同製作的，換言之，問題集具有國際化的特色，而且，每一個問題皆有四個語言的版本，亦即中文、英文、日文、以及韓文。我們使用<SLANG>標記表明該問題的製作國家或機構，如<SLANG>CH</SLANG> 表示該問題是由台灣製作的；<SLANG>EN</SLANG> 是由 TREC 製作的；<SLANG>JA</SLANG> 是由日本製作的；<SLANG>KR</SLANG> 則是由韓國製作的。<TLANG>標記則用於表明該問題目前的呈現語言。圖五展示一個 CIRB030 製作的問題的例子，而表三說明問題集使用的標記。

```

<TOPIC>
<NUM>013</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>NBA labor dispute</TITLE>
<DESC>
To retrieve the labor dispute between the two parties of the US National Basketball Association at the end of 1998 and the agreement that they reached.
</DESC>
<NARR>
The content of the related documents should include the causes of NBA labor dispute, the relations between the players and the management, main controversial issues of both sides, compromises after negotiation and content of the new agreement, etc. The document will be regarded as irrelevant if it only touched upon the influences of closing the court on each game of the season.
</NARR>
<CONC>
NBA (National Basketball Association), union, team, league, labor dispute, league and union, negotiation, to sign an agreement, salary, lockout, Stern, Bird Regulation.
</CONC>
</TOPIC>

```

圖五：問題範例

最初，我們共製作 50 個問題，然而我們在 NTCIR 資訊檢索評估會議後，刪除了 8 個較不適用的問題，留下的合格的問題的編號為 1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、17、18、19、20、21、22、23、24、25、27、32、33、34、35、36、37、38、39、40、42、43、45、46、47、48、49、以及 50。因此，您可能疑惑為何在 CD-ROM 的 TopicSet 文件夾下的問題似乎有遺漏的狀況，但是這是正確的，因為前述不適用的 8 個問題並沒有收錄於此次的 CIRB030 問題集。表四說明問題集的來源與數量。

表三：CIRB030 問題集的標記

<TOPIC>	</TOPIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language code: CH, EN, JA, KR
<TLANG>	</TLANG>	Target language code: CH, EN, JA, KR
<TITLE>	</TITLE>	The concise representation of information request, which is composed of noun or noun phrase.
<DESC>	</DESC>	A short description of the topic. The brief description of information need, which is composed of one or two sentences.
<NARR>	</NARR>	A much longer description of topic. The <NARR> has to be detailed, like the further interpretation to the request and proper nouns, the list of relevant or irrelevant items, the specific requirements or limitations of relevant documents, and so on.
<CONC>	</CONC>	The keywords relevant to whole topic.

表四：CIRB030 問題集的組成

Created By	Original	After filtering (This release)
Japan	15	12
Korea	12	10
Taiwan	13	13
TREC	10	7
TOTAL	50	42

(三) 答案集

相關判斷是由台灣、日本、韓國的工作同仁一起合作完成的，而台灣負責整合最後的相關判斷。CIRB030 的相關判斷分為四個層級：非常相關、相關、部分相關、與不相關，每一個層級都會設定代表的符號與數值，表五說明這些符號與數值。

然而，TREC_EVAL 程式採用的是二元層級的相關判斷，而 TREC_EVAL 被視為是資訊檢索評估的標準作法，因此我們決定製作兩組答案，一組為嚴謹相關 (Rigid Relevance)，也就是非常相關與相關視為相關；一組為鬆散相關 (Relaxed Relevant)，也就是非常相關、相關、部分相關皆視為相關。您可以在 CD-ROM 的 AnswerSet 文件夾找到二個檔案，CIRB030RJCH-Rigid 為嚴謹相關；CIRB030RJCH-Relax 為鬆散相關。

表五：相關判斷的層級

相關層級	符號	分數
非常相關 (Highly Relevant)	S	3
相關 (Relevant)	A	2
部分相關 (Partially Relevant)	B	1
不相關 (Irrelevant)	C	0

五、測試集的使用

從事資訊檢索研究的學者專家可使用本計畫建構的測試集，以評估自己研發的資訊檢索系統的績效。系統評估基本上是使用資訊檢索領域中廣泛使用的量化指標，而非質性指標。

(一) 評估標準

評估系統將參與測試者送回之檢索結果與測試集之標準答案比對，以求全率 (Recall) 與求準率 (Precision) 為主要之評估公式：

$$\text{求全率} = \frac{\text{檢索到的相關文件數}}{\text{所有的相關文件數}}$$

$$\text{求準率} = \frac{\text{檢索到的相關文件數}}{\text{檢索到的文件數}}$$

參與測試的系統依據此二準則產生的測試結果，再加以計算分析，產生下列幾種評估報告：

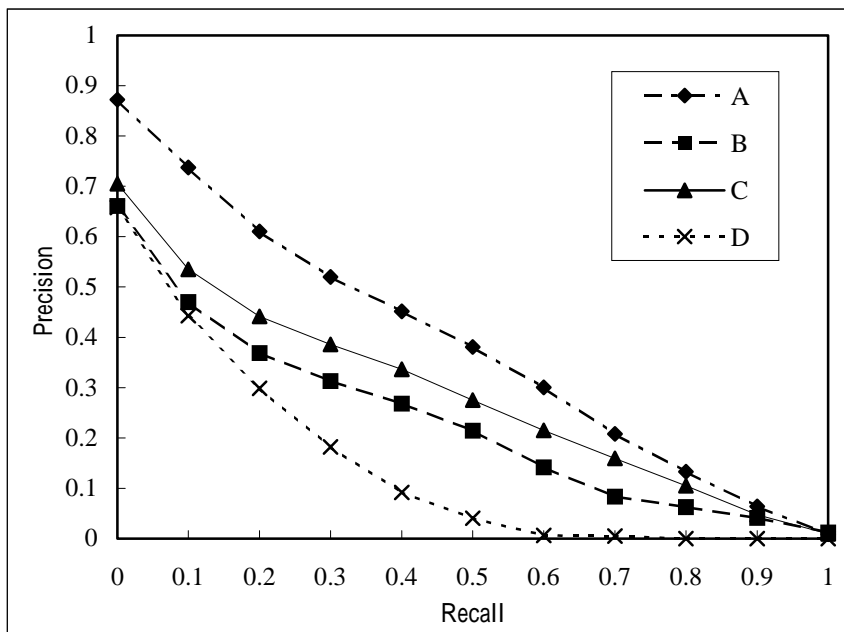
1. 摘要統計表：為系統檢索所得文件與相關文件之各項統計，包括下列項目：

- (1) 系統檢索所得之文件總數。
- (2) 真正與查詢主題相關之文件總數。
- (3) 系統檢索所得文件中，真正與查詢主題相關之文件總數。

2. 求全率與求準率對應表(Recall Level Precision Averages Table)：

(1) 11 點求全率內插值對應之求準率平均 (Interpolated Recall-Precision Averages)：

當求全率為 0.00、0.10、0.20...至 1.00 等 11 個值時，計算其所分別對應到的 11 個求準率平均值，並可據此繪製求全率/求準率圖(Recall/Precision Graph)，比較各系統之檢索效益，如下圖所示。



圖六：求全率/求準率圖(Recall/Precision Graph)

(2) 計算每篇相關文件被檢索出來時，所對應之求準率，並取其平均值。

3. 平均求準率 (Document Level Averages Table)

- (1) 計算在 X 個特定的文件數量被檢索出來時 (5、10...1000)，所對應之求準率，並取其平均值。
- (2) 計算系統檢索出 R 篇文章時，所對應的求準率，並取其平均值。R 值等於查詢主題中所有相關文件的數量 (R-Precision)。

(二) 檢索結果輸出格式

您可以使用 TREC_EVAL 程式評估您的資訊檢索系統產生的檢索結果，這個程式可以在 CD-ROM 的 TREC_EVAL 文件夾找到 Windows 版與 unix 版的程式。要注意的是，TREC_EVAL

程式要求固定的格式，檢索結果的格式如下：

qid iter docid rank sim runid

qid 代表問題編號；*docid* 代表文件標號；*rank* 為檢索出文件的排序；*sim* 為文件與問題的相似性；*runid* 為該次檢索結果的編號（您可以自行賦予）；*iter* 沒有特別的功用，設定為 1 或 0 即可。各個欄位以‘TAB’ (\x0A, \t) 字元分隔。

六、結論

本計畫已完成一組包含文件集、問題集以及答案集的測試集，亦發展一套建構測試集的程序。在各界急於研發中文資訊檢索系統的今日，CIRB030 測試集應能解除國內無從取得中文測試資料的現狀，使資訊檢索系統的研究與發展能有更高的可行性。

資訊檢索評估涉及的層面相當廣泛且多元，而建立一個合適有效之測試集的困難點，除了在具體實施時必須耗費大量的時間與人力之外，測試集實際應用的效能及其是否能反映真實檢索情境與系統評估的客觀性，有待進一步的探討。因此，未來測試集在建構、研究與應用仍有很大的發展空間。

中華民國計算語言學會自成立以來，及努力推動各項研究資源的蒐集與製作，尤其是中央研究院詞庫小組多年來建構各式語料庫，提供學界進行自然語言處理與計算語言學的相關研究，貢獻極大。本人此次在學會的協助下，建構資訊檢索測試集，不僅更加充實學會的研究資源，亦期盼有助於國內外從事中文資訊檢索研究的學者與專家。

致謝

感謝中華民國計算語言學學會提供經費，協助建構 CIRB030 資訊檢索測試集。謝謝學會秘書處黃琪小姐行政事務的協助。

附註

- 註 1： 陳光華，江玉婷。「中文資訊檢索測試集之設計與製作」。 資訊傳播與圖書館學第六卷第三期（民國 89 年 3 月），頁 61-80。
- 註 2： Cyril W. Cleverdon, “The Cranfield Tests on Index Language Devices,” Aslib Proceedings 19, no. 6 (1967): 173-194.
- 註 3： Donna K. Harman, “Evaluation Issues in Information Retrieval,” Information Processing and Management 28, no. 4 (1992): 439.
- 註 4： Donna K. Harman, “The First Text REtrieval Conference (TREC-1),” Information Processing and Management 29, no. 4 (1993): 411-414.
- 註 5： K. Kageura et al., eds., “NACSIS Corpus Project for IR and Terminological Research,” In Natural Language Processing Pacific Rim Symposium '97, Phuket, Thailand, December 2-5, 1997, 493.
- 註 6： “IREX (Information Retrieval and Extraction Exercise) Homepage,”
<<http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html>> (Oct. 31, 1998)
- 註 7： Alan F. Smeaton and Donna K. Harman, “The TREC Experiments and Their Impact on Europe,” Journal of Information Science 23, no. 2 (1997): 173.
- 註 8： CIRB (Chinese Information Retrieval Benchmark) Homepage,
<<http://lips.lis.ntu.edu.tw/cirb/index.htm>> (30 Aug. 2001).
- 註 9： HANTEC Homepage, <<http://hantec.kordic.re.kr/>> (30 Aug. 2001).
- 註 10： CLEF (Cross-Language Evaluation Forum) Homepage,
<<http://www.iei.pi.cnr.it/DELOS/CLEF/>> (30 Aug. 2001).
- 註 11： NTCIR Project (NACSIS Test Collection for IR Systems) Homepage,
<<http://research.nii.ac.jp/ntcir/>> (30 Aug. 2001).

中華民國計算語言學學會

Web Mail 帳號申請表

申請表填妥後請傳真至學會

傳真：(02)2788-1638 電話：(02)2788-3799 ext:1502

申請日期： 年 月 日

申請人資料(請親自填寫，打星號者務必填寫，請先看下列說明後再填寫)：

*姓名：_____

*單位：_____

*會員編號：_____

*Username： 1. _____

2. _____

3. _____

其他信箱：_____

聯絡電話：_____

聯絡住址：_____

說明：

- (1) 申請資格必須為學會成員，若您忘記您的會員編號，請向本會秘書處黃琪小姐查詢。
- (2) Username 即為登入帳號，需為英文字母小寫或數字但首字必須為英文字母，為了避免帳號已經有人使用，請填寫你的帳號順位。
例: Chia-Sheng Wu → cswu
- (3) 密碼會在申請通過後會給予預設密碼，請您重設此密碼，更改密碼時，第一字元請用英文字母，為避免盜用請勿與 Username 一致，且勿用一般英文單字，最好是英文字母和數目字混合使用。
- (4) Username 請以大寫印刷體填寫以利辨識，但在使用時則須用小寫英文字母輸入。
- (5) 本表可從學會網站<http://www.aclclp.org.tw/>下載。
- (6) 已申請過本帳號者，若無法使用請先查詢，請勿重複申請。
- (7) 使用時請遵循網路使用公約及禮節並請閱讀使用條款，請勿將帳號借予他人使用，一經查覺即停止使用。
- (8) 本申請表經審核通過後，本學會將再另行通知，所以務必留下聯絡方法。

申請人簽名：_____

承辦人簽名：_____

中華民國計算語言學學會

Web Mail 服務條款

一、 認知與接受條款

中華民國計算語言學學會(以下簡稱本學會)係依據本服務條款提供本 web mail 服務 (<http://www.aclclp.org.tw/webmail/> , 以下簡稱「本服務」)。當您使用本服務時,即表示您已閱讀、瞭解並同意接受本約定書之所有內容。本學會有權於任何時間修改或變更本約定書之內容,建議您隨時注意該等修改或變更。您於任何修改或變更後繼續使用本服務,視為您已閱讀、瞭解並同意接受該等修改或變更。如果您不同意本約定書的內容,您應立即停止使用本服務。

二、 您的註冊義務

為了能使用本服務,您同意以下事項:

- 依本服務申請表之提示提供您本人正確、最新及完整的資料。
- 維持並更新您個人資料,確保其為正確、最新及完整。若您提供任何錯誤或不實的資料,本服務有權暫停或終止您的帳號,並拒絕您使用本服務。

三、 隱私權政策

關於您的會員註冊以及其他特定資料,本學會僅用作帳號建置用途,絕不會提供給第三方。

四、 會員帳號、密碼及安全

完成本服務的登記程序之後,您將收到一個密碼及帳號。維持密碼及帳號的機密安全,是您的責任。利用該密碼及帳號所進行的一切行動,您將負完全的責任。您並同意以下事項:

- 您的密碼或帳號遭到盜用或有其他任何安全問題發生時,您將立即通知本學會。
- 每次連線完畢,均結束您的帳號使用。

五、 使用者的守法義務及承諾

您承諾絕不為任何非法目的或以任何非法方式使用本服務,並承諾遵守中華民國相關法規及一切使用網際網路之國際慣例。您若係中華民國以外之使用者,並同意遵守所屬國家或地域之法令。您同意並保證不得利用本服務從事侵害他人權益或違法之行為,包括但不限於:

- 公布或傳送任何誹謗、侮辱、具威脅性、攻擊性、不雅、猥褻、不實、違反公共秩序或善良風俗或其他不法之文字、圖片或任何形式的檔案於本服務之上。
- 侵害他人名譽、隱私權、營業秘密、商標權、著作權、專利權、其他智慧財產權及其他權利。
- 違反依法律或契約所應負之保密義務。
- 冒用他人名義使用本服務。
- 濫發廣告郵件。

- 其他本學會有正當理由認為不適當之行為。

六、系統中斷或故障

本服務有時可能會出現中斷或故障等現象，或許將造成您使用上的不便、資料喪失、錯誤、遭人篡改或其他經濟上損失等情形。您於使用本服務時宜自行採取防護措施。本學會對於您因使用（或無法使用）本服務而造成的損害，不負任何賠償責任。

七、下載軟體或資料

本學會對於您使用本服務或經由其他網站而下載的軟體或資料，不負任何擔保責任。您應於下載前自行斟酌與判斷前述軟體或資料之合適性、有效性、正確性、完整性及是否侵害他人權利，以免遭受損失（例如：造成您電腦系統受損、或儲存資料流失等）。本學會對於該等損失不負任何賠償責任。

八、關於使用及儲存之一般措施

您同意本學會得就本服務訂定一般措施及限制，包含但不限於本服務將保留電子郵件訊息、佈告欄內容或其他上載內容之最長期間、本服務一個帳號當中可收發電子郵件訊息的數量限制、本服務一個帳號當中可收發電子郵件訊息的最大檔案、本伺服器為您分配的最大磁碟空間等措施。若本學會將本服務維持或傳送之任何訊息、通訊和內容刪除或未予儲存，您同意本學會毋須承擔任何責任。您亦同意，長時間未使用的帳號，本學會有權關閉。您也同意，本學會有權依其自行之考量，不論通知與否，隨時變更這些一般措施及限制。

九、免責聲明

您明確了解並同意：

本學會對本服務不提供任何明示或默示的擔保，包含但不限於特定目的之適用性及未侵害他人權利。本學會不保證以下事項：

- 本服務將符合您的需求。
- 本服務不受干擾、及時提供、安全可靠或免於出錯。
- 由本服務之使用而取得之結果為正確或可靠。
- 是否經由本服務之使用下載或取得任何資料應由您自行考量且自負風險，因任何資料之下載而導致您電腦系統之任何損壞或資料流失，您應負完全責任。
- 您自本學會或經由本服務取得之建議和資訊，無論其為書面或口頭，均不構成本服務之保證。

十、會員行為

由會員公開張貼或私下傳送的資訊、資料、文字、軟體、音樂、音訊、照片、圖形、視訊、信息或其他資料（以下簡稱「會員內容」），均由「會員內容」提供者自負責任。本學會無法控制經由本服務而張貼之「會員內容」，因此不保證其正確性、完整性或品質。您了解使用本服務時，可能會接觸到令人不快、不適當、令人厭惡之「會員內容」。在任何情況下，本學會均不為任何「會員內容」負責，包含但不限於任何錯誤或遺漏，以及經由本服務張貼、發送電子郵件或傳送而衍生之任何損失或損害。

您了解本學會並未針對「會員內容」事先加以審查，但本學會有權（但無義務）依其自行之考量，拒絕或移除經由本服務提供之任何「會員內容」。在不限制前開規定之前提下，本學會及其指定人有權將違反本服務條款和令人厭惡之任何「會員內容」加以移除。您使用任何「會員內容」時，就前開「會員內容」之正確性、完整性或實用性之情形，同意必須自行加以評估並承擔所有風險。

您了解並同意，本學會依據法律的要求，或基於以下目的之合理必要範圍內，認定必須將「會員內容」加以保存或揭露予政府機關、司法警察或未成年人之監護人時，得加以保存及揭露：

- 遵守法令或政府機關之要求律程序。
- 執行本服務條款。
- 回應任何侵害第三人權利之主張。
- 保護本學會及其使用者及公眾之權利、財產或個人安全。

十一、 智慧財產權的保護

本學會所使用之軟體或程式、網站上所有內容，包括但不限於著作、圖片、檔案、資訊、資料、網站架構、網站畫面的安排、網頁設計，均由本學會或其他權利人依法擁有其智慧財產權，包括但不限於商標權、專利權、著作權、營業秘密與專有技術等。任何人不得逕自使用、修改、重製、公開播送、改作、散布、發行、公開發表、進行還原工程、解編或反向組譯。若您欲引用或轉載前述軟體、程式或網站內容，必須依法取得本學會或其他權利人的事前書面同意。尊重智慧財產權是您應盡的義務，如有違反，您應對本學會負損害賠償責任（包括但不限於訴訟費用及律師費用等）。

十二、 拒絕或終止您的使用

您同意本學會得基於其自行之考量，因任何理由，包含但不限於缺乏使用，或本學會認為您已經違反本服務條款的明文規定及精神，終止您的密碼、帳號（或其任何部分）或本服務之使用，並將本服務內任何「會員內容」加以移除並刪除。本學會亦得依其自行之考量，於通知或未通知之情形下，隨時終止本服務或其任何部分。您同意依本服務條款任何規定提供之本服務，無需進行事先通知即得終止，您承認並同意，本學會得立即關閉或刪除您的帳號及您帳號中所有相關資料及檔案，及停止本服務之使用。此外，您同意若本服務之使用被終止，本學會對您或任何第三人均不承擔責任。

十三、 準據法與管轄法院

本約定書之解釋與適用，以及與本約定書有關的爭議，均應依照中華民國法律予以處理，並以台灣台北地方法院為第一審管轄法院。

Call for Papers
International Journal of
Computational Linguistics & Chinese Language Processing
Special Issue on
Human Computer Speech Processing

Human computer interaction via speech has been increasingly attractive in the area of speech and language processing. Many practical technologies have been explored to demonstrate the potential real-world applications such as mobile speech communication, multimedia information retrieval, adaptive dialogue systems, speaker authentication, speech translation and human computer interfaces. The goal of this special issue is to report the state-of-the-art speech technologies and, hopefully, bridge them with up-to-date advances in international societies. Prospective authors are invited to submit their innovative works to this special issue. We are soliciting paper submissions in topical areas including,

but not limited to:

- Human computer speech interaction
- Speech user interface design
- Speech recognition in real-world environments
- Language modeling and understanding
- Information retrieval of spoken documents
- Expressive speech synthesis
- Spoken language systems and applications
- Speaker recognition, multimodal person authentication
- Multilingual speech processing and speech translation
- Speech summarization
- Speech and language tools for the disabled
- Speech and linguistic corpora

Schedule

Submission deadline: March 31, 2005
Notification of acceptance: October 31, 2005
Final manuscript due: December 15, 2005
Tentative publication date: March 2006

Instructions for Authors

All manuscripts are subject to anonymous peer review. The style for manuscripts is available at the homepage of the *International Journal of Computational Linguistics & Chinese Language Processing* (<http://www.aclclp.org.tw/journal/index.php>). The authors should submit their papers in PDF format to Prof. J.-T. Chien.

Guest Editors

Prof. Jen-Tzung Chien
Dept. of Computer Science
and Information Engineering,
National Cheng Kung University
Tainan, Taiwan
Email: jtchien@mail.ncku.edu.tw

Prof. Tan Lee
Dept. of Electronic Engineering
The Chinese University of Hong
Kong Hong Kong, China
Email: tanlee@ee.cuhk.edu.hk

Prof. Bo Xu
Institute of Automation Chinese
Academy of Sciences Beijing,
China
Email: xubo@hitic.ia.ac.cn

2004 年語言學卓越營小感

胡佳音

國立交通大學 語言與文化研究所 語言學組

I. 緣起

2004 年第二屆語言學卓越營的語料庫與計算語言學課程為國內學生提供了一個絕佳的機會，我個人尤其感到獲益良多。對所謂「語料庫與計算語言學」我一直不得其門而入。曾經我想去資工系旁聽自然語言處理課程，但我太過膽怯不敢跨進全是資工系四年級學生的 NLP 課堂。我心裡猜想：老師大概會講一大堆程式語言吧？遲遲不敢提起勇氣，讓我一直對這個語料庫與計算語言學的朦朧世界感到卻步。

今年五月我因緣際會看到第二屆語言學卓越營的簡介。其中提到「台灣境內的語言學研究所提供給研究生語料庫與計算語言學課程不足，導致年輕一代鮮少有機會接觸語料庫與計算語言學相關研究內容」，真是說中我的痛處。因此我排開了原定出遊瑞士的計畫，趕緊印了報名表請系主任簽名讓我報名卓越營。等待錄取名單的日子和當初等待大學上榜名單一樣難熬。我心想，簡介中說了這個卓越營是為國內研究生所設計，但我還只是個大學生，而報名又只看歷年成績單，我真不曉得我的強烈學習動機能不能在那些林林總總修過的語言學相關課程讓人看出來。我又從工作人員那邊聽說大約錄取六十個學員。我的心裡真是七上八下，一點兒也沒有把握我能得到這個機會。

「放榜」的日子來臨，我忐忑不安地上網察看，一開始大約是太過緊張，怎麼看就是沒有自己名字。我的心一沉，果然，早知道我就去瑞士啦。不死心，我運用現代科技，利用 IE「尋找」功能，鍵入自己的名字。就在那此時此刻，我由衷感謝不管誰讓我錄取的人，謝謝他給我這個機會，而且還把我分配在學分組！我何德何能有這個資格被分在學分組！

「開學」第一天，我像剛要註冊上小學的孩童，七早八早就到了中研院學術活動中心。想不到我去得太早，第一個到場的工作人員似乎用異樣的眼光偷看我，大概在想，怎麼學術活動中心也闖入了流浪之人。

II. 活動期間

這兩個星期裡，我無時無刻不感受到工作人員的用心。見到整個課程安排，師資的安排，活動的規劃，我都深深地想向工作人員九十度大鞠躬，但是膽怯，怕被遣走，直接往杜鵑窩送去，因此我偷偷利用兩行的字句，表達我的感謝。

課程安排部份，課程總共是五十九小時專業課程及八小時的實習課程，包含三部份，分別為「中文計算語言學概論」、「語料庫」及「中文語音處理」。這三部份課程所邀請的師資全部都是該領域的專家，為學員從最基礎的概論講起，再談論他們本身的研究主題。

開幕式邀請國內學者為學員講話，他們勉勵學員在未來兩個星期要好好表現，也提供他們的經驗談。第一場專題演講中，謝清俊老師提到人們生活方式隨著科技進步改變，語言也不例外；研究語言的人，應該要意識到這點。這或許是為什麼近來陸續有研究網路文字的文章出現。這場演講也讓我明白數位典藏的重要性。物體能保存的質與量會隨著時間遭到破壞及影響。數位化能讓珍貴的文化資產得到較完整的保護，也縮減保存所需要的空間，使各地的人都能透過電腦及網路欣賞文物。

鄭錦全老師給我們上計算機概論的第一堂課，他特別講解計算語言學的發展歷程，更慷慨的將他寫的程式，包含 CCLang, CCEsense 等分享給學員，詳細說明並示範其使用方法及應用。他也報告他透過研究各代文學作品，發現人的基本用字數量大約在八千字左右。他另一個研究是關於計量的方法研究大陸方言的親疏度，是一個科學方法研究語言非常好的示範。

講中文計算語言學的還有從北京大學來的俞士汶教授。他風趣的上課方式我想一定讓所有學員印象深刻。他說他很高興來到卓越營，看到這麼多卓越的人，但是只有他不卓越。俞教授真是謙虛。不管是什麼人，看到他帶來的【現代漢語語法資訊詞典詳解】，一定會認為這本詞典是一本卓越的參考書。我對俞教授感到印象深刻的還有他耐心的對文組的學員不厭其煩地解說「n 元語法」、「隱馬爾柯夫」等名詞。他也非常注意每位學員提出的問題，晚上的 coffee time 中他仔細地將提問的學員名字及學校記在本子上，看得出他的用心。

黃居仁老師也講計算語言學，但因為他的課排在卓越營的最後一天，他的講題則著重於「從語言形式到語意及知識內容」。其中提到通用詞類集的必要性，詞意和意面的區分，知識本體等概念。可惜的是黃老師的課時間太少，而黃老師又體諒學員十二天累積下來的倦容，大略提示重點。雖然時間有限，黃老師還是竟然將重要部分詳述並舉例，使得原本對這些概念並不非常熟悉的我也粗粗地獲得一個整體的理論架構。

第二部份的課程是「語料庫」。首先講課的是陳克健老師。每個部份課程負責第一堂的老師似乎總是發覺到了自己背負的重責大任，因此除了後面中研院的「平衡標記語料庫」和「句結構樹資料庫」外，陳老師在第一堂課更簡介了語料庫語言學及詞類標記的原則。由於老師投身語料庫的研究多年，他給學員提了好幾個非常寶貴的意見：其中一個寶貴意見是，我們千萬不能盲目地相信電腦給的數據，我們要能判斷出奇怪不合理的數據。也就是說，頭腦要能和語料庫相輔相成。

清大的張俊盛老師則是講機器翻譯。實驗課時他和隨行的助教們示範 CANDLE、TotalRecall、還有 TANGO 這幾個語料庫應用在英語教學的程式。此外還有模擬人類如何了解問題的 Question Answering，幫助資料取回的研究。我個人認為以上的研究最重要的貢獻是能幫助人類更了解人類自己的認知能力發展。

從賓州大學過來的薛念文老師給我們講賓州大學發展出來的 Penn Chinese Treebank 還有 Proposition Bank 以及電腦要如何從語料庫中自動取得語言訊息。薛老師並分析各種語料庫的優、缺點，以及除了 Treebank 外我們還需要發展 Proposition Bank 的原因。

第三類課程是語音處理，由石基琳老師講用語料庫的方法來看漢語的音韻，還有交通大學陳信宏老師和台北科技大學廖元甫老師講中文的語音辨識和合成。

石基琳老師介紹了中文語音、時長、音調及語調的特性及各別處理方法。對於幾個術語，如 Zipf's law 和 greedy algorithm，老師也仔細的講解。我想每個學員對老師的一句話一定印象深刻：她常強調「電腦可以做的就別浪費人力」，畢業人的一生也不怎麼長，重要、該做的事很多。我們可以學習程式語言，寫一些小小的程式幫助我們節省時間。實習課是老師則是讓我們操作 praat，讓我們練習該軟體的幾個重要功能。

陳信宏老師介紹中文語音的辨認與處理的理論概論，而廖元甫老師則是介紹工研院及台北科技大學的電話總機經由語音辨認後的自動轉接功能。這一點非常有趣。現在手機已經發展出用語音操作電話簿再撥出電話的功能，但個人電話簿裡儲存的名稱是固定的，但是前述的兩個電話總機則是設計能夠辨識某個人的名字、職謂。實習課時，廖老師讓我們實地用軟體找出子音母音的頻率。這是非常實用的課程。

除了上述的專業課程及實習課程外，還有安排專題演講，包括已經描述過的謝清俊老師的演講還有從英國來的 Adam Kilgarriff 講 Web as Corpus 的概念，黃居仁老師講 Domain Lexico-Taxonomy，鄒嘉彥教授講法律語言學，鄭秋豫博士講中文的音調，石基琳老師講用語料庫的方法測試音韻理論，還有最後一天下午俞士汶老師再講北大語言知識庫的概況。

課程和專題演講都使得我受益良多，除外我更喜愛卓越營的安排是每天晚上的 coffee time，每個到場的學員都因此有機會能直接向老師們請益。另外是週六的「語言學前瞻研討會」，邀請語言學領域外的學者或專家，講他們心中的語言學。此外研討會下午還安排擁有語言學或計算語言學學位的前輩們講他們目前的工作和語言學訓練的關聯。一般語言學的學生總是會想，或被問，「念了語言學能做什麼呢？」這些不論是進入貿易界，或機器翻譯等不同領域的前輩的經驗談，不啻是一劑強心針。原來，念語言學並不是只能當老師，甚至只能當英文老師。念語言學最大的收穫將是邏輯分析以及將雜亂資料統整的能力。

III. 緣落

這兩個星期是我人生中最有收穫的其中一段時間。兩個星期雖短，但我因此有機會結識同樣對語言學議題有興趣的同學們，以及能向語料庫、語音處理等計算語言學的專家學者們學習、請益。他們每個不但學有專精，更是耐心的好老師，將困難的理論解釋讓初學者能明白其中意義。這段時間我真是如魚得水，除了有中研院提供的舒適住宿及學習的場所，因而我能心無旁騖地吸收新知享受學習的樂趣，並且，隨時想提出語言學有趣之處向人討論，也不怕找不到同好，或被別人認為「走火入魔」。這樣的感覺真的很好。最重要的是，因為參加了這個卓越營，終於我有機會能走進這個陌生但有趣的世界一探其中。這是一個美夢的實現。

謝謝所有給我這個機會學習的人。

第二屆學生計算語言學研討會紀要

巫宜靜、吳鑑城、吳典松、鍾曉芳

◆ 研討會概況

學生計算語言學研討會，今年是第二屆，已於九月三日，隨同本屆 ROCLING 大會圓滿結束。本屆學生研討會，總共收取八篇文章。海報三篇、報告五篇。參見表(一)、(二)。在領域分佈上，偏屬計算語言學和偏屬語言學者，各四篇。報告者由南而北，分別來自中正大學(1篇)、暨南大學(1篇)、清華大學(2篇)、交通大學(3篇)與台灣大學(1篇)。不論是論文的數量，或是與會人員的數量上，都比第一屆增長許多。較之第一屆，在籌辦人員方面，除了學生委員人數增加，也邀集了十三位資訊、語言所的老師，共同組成指導委員會。在學生研討會的定位上，也比第一屆更清楚，明訂希望學生研究中或研究後的發現與成果，能有公開發表、向老師請教、以及與同好相互切磋的機會。在報告形式上，也增加了海報展示項目，並且除了靜態海報展示，也讓研究者簡介海報內容，接受老師講評，並且能有和與會者討論切磋的機會。整體而言，比起上一屆，有長足的發展，但仍有很多進步的空間。期待下一屆能在時間規劃安排上、論文質量上，以及各項籌辦事務方面，都能有更好發展，能提供學生們更貼心、有效的服務。

表一：海報短文：

序號	作者及論文
1	江亭儀 (Chiang, Ting-Yi) 交通大學 Affective Chunk of Mandarin Wo JueDe (我覺得) and its Discourse-Pragmatic Functions
2	蕭佩怡 (Pei-Yi Hsiao) 暨南大學 Syllable Reduplication in A-not-A Questions
3	林鴻銘 中正大學 上下義詞作為語篇回指詞的認知語言學探討

表二：論文報告：

序號	作者及論文
1	林哲民 (Lin, Zhemin) 台灣大學 POS-Tagger for SaiSiyat: Using Fieldwork Notations and Transformation-Based Error-Driven Learning
2	張群 (Chang, Chun Edison) 交通大學 Cognition and Conceptual Manipulation: A Corpus Study of Collocational Asymmetry in Chinese Gradable Predicative Adjectives
3	余鍵亨 (Yu, Jian-Heng) 清華大學 Alignment of Bilingual Web pages based on the MT evaluation method of BLEU
4	張俊欽 (Chang, Chun-Chin) 清華大學 Word Translation Disambiguation Using Two Monolingual Corpora
5	陳以理、林蘭綺、吳典松 交通大學 自然語言處理技術於專利文件分析之應用

◆ 幕後推手

學生計算語言學研討會能舉辦，首先要感謝前任計算語言學會理事長，清大張俊盛教授，在任期內，大力推動學生研討會的籌備事項。張老師是「始作俑者」，也是學生委員會最佳精神支柱之一。在張老師的鼓勵、協助與指引下，我們才比較能忘卻自身的無知與缺乏，有勇氣、歡歡喜喜、摸索著一路走來。

我們還要感謝本屆計算語言學會理事長，成大王駿發教授，以及學會理事的老師們，讓學生計算語言學會能再度與 ROCLING 大會一起舉辦，提供全額的經費。更要感謝本屆 ROCLING 大會主席簡立峰、王新民老師，計算語言學會秘書黃琪小姐，以及其他所有的大會工作人員，在各項籌備事務上的鼎力協助。從網頁建置、時間進度與議程規劃、會議論文集的印製、以及各項前置作業與交通、場地庶務、會場佈置與管理等等，都極盡細心、體貼地考慮到學生研討會，省卻我們相當多的工作負荷，大大地減少我們出差錯的可能，我們才能順利舉辦本屆研討會。

還要感謝所有的指導委員會的老師，中研院的鄭錦全老師、陳克健老師、黃居仁老師、曾淑娟老師，台灣大學的安可思老師，東吳大學的柯淑津老師，交通大學的劉美君老師、梁婷老師，清華大學的王旭老師、連金發老師、張俊盛老師，以及中正大學的蔡素娟老師、成功大學的吳宗憲老師。十分感謝老師們接受我們的邀請，毫無理由地相信我們，願意當我們的指導老師，協助我們、引導我們。讓我們能在茫然、不知所措的焦慮窘境中，彷彿吃了定心丸般，能勇往直前。

也要感謝本屆的講評老師們，於百忙中撥冗參加研討會。從研究題目、研究方法、研究內容、論證邏輯、書目格式、報告內容的呈現、報告時應注意的語言、身體姿勢、以及接受講評的態度等等，都給予學生報告者，相當多有益的指引，連帶也使我們獲益良多。參見表(三)。

表三：講評老師給予報告者與學生委員會的建議

<p>一、 作研究時應注意的事項：</p> <p>(一) 應注意研究題目、方法與內容的科學性、應用性、創新性，與解釋性。</p> <p>應思考以下問題：</p> <ol style="list-style-type: none">1. 「科學在哪裡？」 研究的基礎、理論是什麼？ 何以見得你的看法是對的，你提出來的模式是好的、是有用的？2. 「可如何應用？」 為什麼要作這篇研究？這篇研究對其他研究有什麼助益？3. 「有什麼創新？」 這篇研究和前人研究有何不同？有什麼貢獻？4. 「應如何解釋？」 數據分佈、語言現象告訴我們什麼？ 問題的成因是什麼？該如何解決？為什麼要那樣處理？ 雖然也許不可能面面俱到，但應儘量提醒自己注意。

- (二) 應注意題目、問題和術語的界定。
- (三) 應注意論證時的邏輯推理。不能以論證的結果與目的來證明論證方法和手段的適當性。

二、 論文撰寫時應注意的事項：

- (一) 文獻探討是否充分？
- (二) 書目格式是否妥當、一致？
- 學生委員會可編定指導手冊，參考其他期刊（如 ACL），擬定標準格式，以供投稿學生參考。

三、 準備報告內容時應注意的事項：

- (一) 應注意報告的訊息焦點，應避免重點過多而失去主要焦點。
- (二) 應考慮聽眾的需求。

四、 報告時應注意的事項：

- (一) 應注意語言用詞，減少不必要的口語詞，如「然後」、「那樣子」，以及「對」等詞語。勿自言自語地說「嗯...對」。應對報告語言，有更高的自我要求，應避免如「做一個排序的動作」的冗贅說法，可以用簡潔的「排序」來表達。
- (二) 報告前，應有充分的準備。可錄下預講練習的錄音帶，從自己的錄音中，改進自我的小毛病。也可寫下報告講稿，注意報告時的轉折用語。充分的準備，有助於避免緊張、增進信心。
- (三) 應注意報告時的身體姿勢，應避免背對聽眾，以及手叉腰、雙腿斜張等姿勢。

五、 接受講評、提問時，應注意的事項：

- (一) 別人的批評是對我有利的。
- (二) 批評別人總是比較容易的。
- (三) 虛心受教可避免重蹈覆轍。
- (四) 審慎判斷可用的批評建議。

本屆的報告者，以碩士班學生居多，他們虛心求教、高度配合的態度，包容我們許許多多未盡美善之處，並且主動於會議開始前詢問我們是否有需要幫忙之處、提供我們必要的協助，也都令我們倍感溫馨。他們的報告，或生澀、或沈穩，也讓人頗有長江後浪推前浪之感，有著後有來者的欣喜，與歲月如梭的感慨。會後與幾位報告者共餐、閒談，聆聽他們與會的心得、未來的研究方向、學術懷想，以及對學生研討會的建議，也讓我們深深自我期許，期盼將來能改進現有的不足，能讓學生研討會成為更多學子的學術舞台，能為學生提供更多有益的服務。

◆ 心路與心得

本屆學生研討會籌備的工作，雖然不輕鬆、感受到許多時間與責任的壓力，但是因為與其他三位學生委員共事，歡樂足以忘憂。本屆的學生委員，除了三 Wu（吳鑑城、吳典松、巫宜靜）之外，也加入了台大語言所鍾曉芳同學。我們很高興曉芳能加入我們，並且主動、積極地幫助我們草擬摘要徵文稿、設想徵求其他學生委員的管道，幫助我們邀請指導老師，發佈會議訊息。曉芳熱誠的參與，令人差點忘卻她是個要準備期末考、碩士班升博士班口試，以及國內外許多會議文章的碩士班學生。實力堅強的曉芳毫無問題地通過口試，直升博士班，還順利有喜，也使我們學生委員會沾光，同感喜氣洋溢。典松和鑑城今年也得準備資格考、準備多項會議的文章，但仍積極、盡責地參與學生研討會籌備的事務。與他們共事，真是令人愉快的事情。

曉芳和宜靜是學語言學的，典松和鑑城是學資訊的。我們分屬不同學校、地區，平常以電子郵件聯絡。聚會討論時，總是能迅速達成共識、分配任務，並且聊得開懷、分享彼此的喜悅。許多「不可能的任務」，就在大夥兒談笑間，順利愉快地達成了。例如，上學期末、暑假開始期間，邀請指導老師本是件不容意的事情。因為暑假是國際會議的旺季，許多老師都已經或者正準備周遊列國，不便收閱電子郵件、與我們討論。學生委員也因為要出國參加會議，所以彼此間的聯絡，也不容易。不過幸好，我們也輪流藉著參加國際會議，在機場、國外會場，順利逮著機會邀請到指導老師，將好消息傳回給在台灣的伙伴們。我們彼此都有共識，出國的肩負使命，回傳訊息，而在台灣的就負責統籌、決議事項。因此才能在短時間之內，確認指導老師委員名單，發佈徵稿訊息。這屆研討會的籌辦，難以預期、掌控的因素很多，幸賴彼此合作無間的幫補，才往往能在千鈞一髮之際，達成任務。

籌備過程中，最大的困難，可能來自於溝通協調的問題。正如研討會的名稱，是由「計算」和「語言學」組成，此二領域，不論在研究方法、論文寫作、會議形式方面，都各自有傳統。要達成某種程度的共識，讓雙方領域的學生都能適應、接受，有時不容易。舉例來說，在徵稿形式上，計算機的會議大多徵全文或者半全文，徵求摘要者較少。語言學的會議，大多先徵求摘要，比較少徵求全文。在文章長度方面，計算機的會議，大多長度在 6-10 頁。也常見以 6 頁為上限。語言學的會議，上限大多在 15 頁左右。以台灣學生語言學會議，上限是 2 萬字，約 20 頁。研究方法與內容方面，計算機大多著重於系統模組的提出、修正，因此重統計公式、與數據評估。語言學多著重語言現象的觀察、理論辯證，與對問題的解釋。若從科學性、應用性、創新性與解釋性等四大指標來看，計算機的研究似乎比較重視應用性，而語言學似可謂較著重解釋性。就學生參加會議的經費而言，在台灣，語言學所舉辦的相關會議或研討會，絕大多數都是免費、也可獲得免費的論文集，報名時只需填寫所屬學校系所即可，沒有太多的關卡設限。計算機可能常與業界有合作關係，因此，學生與會，可能也大多需收取費用。因為二領域各有各的傳統習慣，想要取得一個雙方都可以接受的平衡點，也在在都考驗著我們。

我們採用的是土法煉鋼法，也就是試誤法。從錯誤中學習。第一屆，我們徵求全文，長度上限為 6 頁，包含摘要、附錄。結果是收到許多語言學領域過長的投稿文章。作者表示實在很難濃縮。有的因此而忍痛放棄。第二屆，我們徵摘要，文章全文長度增到 12 頁，海報全文 4 頁。不過由於本屆時間規劃得較為匆促，所以語言學的學生選擇投海報的，佔多數。他

們也反應 4 頁稍嫌不足，希望下屆能增加一些頁數。

此外，由於學生研討會是大會議程的一個場次，因此，和大會承辦單位之間的合作，也考驗著我們的配合度。有時，會出現即使和大會已經達成主要共識，但在實際執行的細節上，雙方仍有不同的理解與處理方式。希望能汲取本屆籌辦的經驗與教訓，於下屆改進本屆不足之處，使得學生計算語言學研討會，能夠成為更完善的學術交流禾場，讓更多學子有更豐富的收穫。

回顧整個研討會籌備與舉辦的路程，路況並非總是平坦的，視線並非總是清楚的。不過，我們很幸運，在崎嶇坎坷顛頗中，有同伴的笑聲相隨；在昏暗不明驚懼中，有師長的指引相依。

◆ 未來展望

第二屆學生計算語言學研討會，結束了。未來呢？是否能從此踏上坦途，一帆風順？還是還有更大更艱難的挑戰等待著第三屆學生委員們去克服？我們很幸運地從今年暑假在中研院舉辦的第二屆語言學卓越營中，得到許多志同道合的伙伴加入，名單參見表（四）。其中有許多伙伴，如陳淳齡、龔書萍，與黃漢君等等，也都已經提供第二屆計算語言學研討會相當多的協助，非常感謝他們。第三屆學生計算語言學研討會的籌辦學生委員會，陣容將更堅強，也期待能更有組織，並且能發揮更大的功能。

表四：第三屆學生計算語言學研討會新增學生委員：

江振宇	交通大學	電信工程學系	（博士班）
陳淳齡	交通大學	資訊科學學系	（博士班）
鄭守益	交通大學	資訊科學學系	（碩士班）
簡嘉言	清華大學	資訊工程學系	（博士班）
張裕嘉	清華大學	資訊工程學系	（博士班）
龔書萍	台灣大學	語言學研究所	（博士班）
吳郁華	交通大學	語言文化研究所	（碩士班）
朱曼妮	清華大學	語言學研究所	（博士班）
黃漢君	清華大學	語言學研究所	（博士班）
李念貞	成功大學	外國語文學系	（學士班）

我們非常感謝張俊盛老師與曾淑娟老師的協助，使得下一屆 第三屆學生計算語言學研討會以及本學年度的研習會籌辦的提案，能在計算機語言學會的議案中，順利通過。下屆學生委員會籌辦的方向，如下表(五)。除了繼續籌辦學生計算語言學研討會之外，希望也能朝向推廣計算語言學、促進資訊科技與語言學等相關學術發展以及學子們之間的彼此交流邁進。希望能建立新聞群組，促進學生們彼此之間的交流與相關訊息的流通。也期盼能舉辦相關研習會，邀請老師，或者透過學生彼此相互教導的方式，將既有的研究成介紹給對計算語言學感興趣的學子們。

表五：第三屆學生委員會籌辦方向

<p>一、第三屆學生計算語言學研討會 時間：94年8-9月間與ROCLING大會同時召開 主題：計算語言學相關的各項議題</p> <p>二、新聞群組(newsgroup) 功能： (一) 意見交流：詢問、解答 (二) 訊息交流：公布研習會、研討會、各項與計算語言學相關的會議訊息、出版物訊息。</p> <p>三、計算語言學研習會 時間：93年至94年9月之間，定期召開。 研習主題： (一) 計算機、語言學基本術語說明 (二) 程式設計與軟體使用說明 (三) 語料庫使用說明 1. 書面語：如，Sinica (中), Livac (中), CANDLE (中、英), BNC (英)....等。 2. 口語、手語、南島語、閩語、客語、兒童、聽障、病句...等。 (四) 網路資源使用說明 WordNet, FrameNet, HowNet, VerbNet... Probank, Treebank... SUMO, BOW, Wordsketch... (五) 統計方法、軟體使用說明 (六) 語音合成/辨識相關軟體使用說明 (七) 運用語料庫研究語言現象的說明 (八) 其他建議項目</p>

新聞群組的籌建想法，來自於這二屆研討會籌辦的過程中，我們常需要對外公佈訊息，並且資訊所學生和語言所學生委員之間也常互相向對方請教一些「跨領域」的問題。我們也接到投稿的學生來信詢問問題。例如「什麼是 entropy？」、「Constituent 是什麼？」、「一個 chunk 有多大？」、「中文詞有多少個？」、「在哪裡可以找到相關詞彙的文獻？」、「分詞的標準是什麼？」、「如何自動斷詞？」、「X 語料庫如何使用？」、「有沒有語音的語料庫？」等等相關的問題。我們還常被問到「如何準備你們所上的碩/博士班的甄試/考試？」、「口試時要注意什麼？」、「落地簽要怎麼辦？」，或者被即將出國參加會議的人問到「那裡有什麼特產？」。這些問題，或者與研究相關，或者關乎升學、參加會議等大事。知道的人，或有經驗的人，也許很輕易就能解答對外行人/沒有經驗的人來說，像是山一般大的「難題」。希望我們能提供一個管道，讓大家能相互請教、切磋，雖然不保證所有的問題，都能得到解答，但也希望能

做到儘量將未獲得解答的問題，轉請教於其他方家。

研習會的構想，一方面承上來自於我們常問、常被問到的問題，另一方面來自於研究的需要。由於語言學研究所開設計算機或者程式設計課程的並不多，即使偶爾開課，一般也不是常態性課程。而資訊所開設語言學基礎課程的，也不多。但在研究上，可能都需對基本術語有所瞭解，才不致至於不解或者產生誤解。語言學研究所的學生，一般而言，對程式設計比較不熟悉。但是目前不論是分詞、句式剖析、字串搜尋、統計等，都已經有一些可開放申請使用的程式軟體。如果能先熟悉既有程式、軟體的使用，學習一些基礎的程式設計，或可有助於設計適合自己研究所需的程式。

此外，從事計算語言學相關的研究，語料庫是很重要的一項資源。目前中外語料庫，隨著資訊科技的發達，也日新月異。書面語的語料庫，不論是單、雙語，共時、歷時的語料庫都有。現代漢語有中研院的平衡語料庫 (Sinica Corpus)、北京大學的語料庫等；英文的有英國國家語料庫為代表。香港城市大學的共時語料庫 (LIVAC)，蒐集了大陸、香港、台灣、新加坡等地的現代漢語資料，為研究不同地區的現代和語的重要資源。中英雙語語料庫有清華大學的 CANDEL 語料庫，已有便捷的檢索介面，是研究中英對比分析的絕佳資源。古代漢語，也有中研院的漢籍文獻語料庫，是研究歷史語言學不可或缺的語料庫。

除了書面語之外，口語、手語、南島語、閩語、客語等等，以及兒童語言、聽障、病句語料庫等，也都有專家學者，已經建構，或者正在研究建構當中。這些語料，不論是對於語言學、語言教學或者計算語言學的研究，都相當重要。不過，這些語料庫各有各的建構目的與功能，因此，語料形式與檢索方式、各不相同。一般而言，除非研究者公開發表或展示其研究成果，或者有專人解說，否則，對初學的學生來說，有時要迅速便利地熟悉某個語料庫的資料結構或者使用方式，並不容易。如果能規劃有系統的語料庫介紹課程，相信對學生做相關方面的研究，會是很好的入門起始點，也可以縮短盲目摸索的過程。

目前網路上詞彙的語意及句式分析的資料庫也不少。例如 WordNet 對英語詞彙做了語意區分，是目前研究英語詞彙語意，以及從事語意區分研究的重要參考資料。除了英語之外，很多其他國家的語言，研究者也都正在努力建構自己語言的詞彙庫，可做為雙語或多語對比分析研究的參考資源。FrameNet 提供了英語詞彙搭配的語意角色和句式等訊息；VerbNet 提供了英語詞彙搭配的句式，以及句式轉換等訊息。HowNet 則提供了中文詞彙的詞語搭配訊息，並且以英語詞彙來定義。除了搭配詞語和句式之外，句構的解析，結構樹庫 (Treebank) 的建構也一直是研究的焦點。Sinica Treebank、Penn Treebank、Probank 都有中文句式庫的資料。此外，詞彙的語意階層關係，也是研究語言與自然語言處理，不可忽視的重要課題。SUMO、Sinica BOW 上已有豐富的語意關係訊息，是研究詞彙語意、語意衍伸的重要參考資源。Wordsketch 是目前相當熱門，也是對語言學、語言教學都可能很有幫助的資料庫。已包含將語料庫中，詞彙搭配詞語的分佈、例句作系統整理的資料。可以大大節省檢索者從龐大語料庫查詢資料後，還得逐一自行歸納句式、搭配詞語的時間。除了這些，還有許許多多有用的相關網路資源，如果能夠有系統地介紹，甚或整理出實用的使用操作手冊，相信將能對計算語言相關的研究，有所助益。

此外，統計方法，與現成統計軟體的使用說明，可能也是我們學生需具備的常識與技能，不過，並不是每個所都開有相關課程。如果能讓已經上過課程，熟悉使用方式的學生，作實

例示範教學，對有需要的學生而言，不啻是一大福音。

除了文本的語料之外，語音方面的資料，也是研究語音相關領域的學生引頸期盼的。在今年暑假舉辦的語言學卓越營中，請了專家來講授語音合成、辨識相關軟體的使用方式，也讓學生實際操作。但是因為實習課程的名額有限，學員們也紛紛表示希望能多開設相關課程。大型研討會的籌辦，所需人力、經費甚巨，可能不是每年都有。老師們平常都很忙，可能也不容易挪出時間來講授語音、文本語料庫等課程；而對初入門的學生來說，所需要的指引與需迫切處理的問題，大多可能偏向技術層面，如果能請熟悉該領域的同學來指導，可能可以幫助有需要的同學，解決燃眉之急。

可能也有剛接觸語料庫的同學，對運用語料庫研究語言現象的方法，對其限制、目前研究現況，以及未來發展，感到好奇或者疑惑。也許我們也可以安排相關的演講或者研究成果展示，一起與大家分享心得、展望未來。或者透過新聞群組的訊息交流中，也能機動安排大家所建議的交流學習項目。

◆ 結語

以上的初步構想，實有賴第三屆學生委員會齊心實現。也有待各方的支持，才能美夢成真，不至於幻想破滅。在此，再次感謝從以前到現在，一直鼓勵我們的師長們，也向未來將給予我們批評指教的師長、同儕們致謝。有您的陪伴，我們成長學習的路將不孤單、也將更寬廣舒暢！