電腦在語言學裡的運用

S-Y William Wang

王士元

UC Berkeley &
National Tsing Hua University

加州大學柏克萊校區暨清華大學

Proceedings of ROCLING I(1988)
R.O.C. Computational Linguistics Workshops I pp 257-287

中華民國第一屆計算語言學研討會論文集 257-287 頁

電腦在語言學裡的運用

王士元

(0)摘要

電腦在語言學裡主要有三種不同的運用方向，由於近年來硬體及軟體都在日新月異地進展，電腦研究，尤其是電腦語言的研究，跟語言學的關係，一定會越來越密切的。

（一）一般講來，電腦的早年運用，就是作計算。在語言學裡也是如此。例如三十年前，我們把英語裡輔音的使用頻率，用統計方法來處理，解答了一些抽樣方法上的問題。目前這個計算功能，當然還是很有用。現在我們正在使用電腦分析語言間的親屬程度，因爲需要作很多的計算，電腦是最關鍵的工具，並且還可以用到它的製畫功能，畫出準確的樹圖，來說明語言中的關係。

（二）後來電腦的容量越來越大，微電腦的容量很快地就從幾千個 bytes 增加到幾千萬個 bytes，我們就能存入大批的語言材料，以備系統地整理及分析。研究漢語音韻的 *DOC* 就是爲此而設計的。 它可以幫助我們找出種種不同的音變規律，同時也可以找出一些正証或反証。近年來，臺灣學者又自動化了不少漢語的史書、土著語言以及閩方言的資料。同時市場上又見到了許多處理大批材料的軟體，如 dbase、Clipper，種種人工智慧系統，又配上漸漸好用的一些中文系統，這些語言學的工具，一定會對中國語言學發生很重要的推動力量。

（三）第三種運用，最近才開始正規地發展，也是最能發揮電腦功能的一個方向。這就是用電腦來模擬人的語言行爲。這裡包括的有(a)合成語音，(b)合成句子，(c)識別語音，(d)了解語意，和(e)翻譯語言。要是我們可以用電腦的模擬系統成功地模仿語言能力，那麼這個電腦系統或程序可以說是對這個能力的一種科學解釋。進一步說，這也是對人提供了一種部分的科學解釋。因爲語言能力是人的最顯著的特徵。

電腦在語言學裡，是件最有啓發性的中心工具。很多有意義的假設由它而起，沒有它，很多先進的研究是作不成的。中國有很多的年輕學者，在這方面有興趣，有成就。這對中國語言學講來，是個非常吉祥的預兆。

(1) 前言

八月裡，陳克健老師和黃居仁老師到清華來看我，要我在這個討論會說幾句話。那時候我有點猶豫不決，我跟他們說，我在計算語言學方面已經成了一個dinosaur，一條恐龍了。

因爲我專心學計算機是在做研究生的時候，離現在已有三十多年。那個時代的確是這門學科的原始時代，用的是眞空管的機器，打洞的卡片，開起來又慢又鬧，跟現在的 15 mega hz的微形機比，確實是天淵之別。尤其是最近五六年日新月異的進展，我也沒有好好去掌握它，所以自己知道是非常落伍，跟很多硬件軟件已經脫節很久了，就怕沒有什麼可講的。

可是後來想想，一來不該辜負兩位老師的好意，二來這倒是個難得的機會跟些年輕的同學們聚在一起，說說我們共同的興趣，而且這又是 ROCLING 第一屆的會議，應當來湊湊熱鬧，所以一下硬了頭皮，就決定來了。我想最起碼我可以帶來一些歷史的經驗，以備大家參考，也提出幾個不成熟的想法拋磚引玉，向專家們請教。

依我看來，計算語言學是可以從好幾個不同的角度來著手的。我們先從比較抽象的方面說起，講一些語言的幾種不同的體現，以及這些之中種種的關係。

(2) 自然語言跟電腦語言

大家都知道，世界上的語言多得很，可是究竟有多少呢？一般語言學家是認爲最少也有五六千個，大概不超過一萬個，其實這些數目也是很不準確的，因爲語言跟語言之間的界線，是很難畫分的，而什麼是語言？什麼是方言？也是很難說的。給語言分界的標準，往往會互相衝突。比方說瑞典話跟挪威話都是北日爾曼語言，分化並不久，最多也不過是一千多年，並且又因爲這兩個國家共同的邊界很長，兩國人民的關係比較密切，所以用這兩個語言來通話，絲毫沒有問題。趙元任先生也提過(1959)，他在加州大學教課，有一次他的班裡有個瑞典學生、丹麥學生，也有個挪威學生，這三個人經常各用各的語言討論課裡的東西，可是這到底是不同的國家，因此大家就把他算作三個語言。值得注意的是，這裡用的標準不是語言內在的距離，而是政治上的界線把它畫作三個語言的。

我們再舉個相反的例子，依史書的記載，周朝末年有些貴族人物帶了一批人馬，從中原地區移到太湖一帶。我們不知道兩三千年以前太湖一帶原有的居民是說什

麼語言，可能是漢藏系裡苗瑤支的語言，或者是南亞系，甚至是南島系的語言。無論如何，周朝人帶來的語言就跟這些原有的語言互相影響，同時也從中原語言隔開而個別變遷。這只不過是史書裡記載的最早的一次移民。要是我們在史書裡仔細地去查，就會找到許多次不同形式不同分量的人口移動(周振鶴跟游汝杰在這方面做了很有用的分析，*1987*)。

久而久之，太湖這帶的語言就形成了現代吳語。那麼吳語跟官話是不是兩個不同的語言呢？還是一個語言裡的兩個方言呢？那就看我們用的是甚麼標準了。我們知道中國有"十里不通語"這句話，一個寧波人跟一個西安人各用各的家鄉話是絕對不可能談話的，這語言的內部距離比瑞典話跟挪威話要大得多，因爲它們隔開的時間也比瑞典話跟挪威話要久得多。那麼爲甚麼我們把相似的、距離近的瑞典話跟挪威話算成兩個不同的語言，而反而把寧波話跟西安話納入同一個語言呢？這當然是由於許多歷史傳統、文化、文字等等的很多因素，而又不是由於任何語言內部的標準了，所以我們應該知道，五六千個語言這個數目其實是個含糊的數目，沒有很多科學根據的。

這一百多年來語言學最實在的成果就是把這幾千個語言比較系統地、客觀地分析了一下，當然，不是每個語言都分析得同樣的精密仔細。人口多，歷史久的語言，研究的材料就多，敘述的也就深入。人口少而最近才發現的一些語言，材料就很少，有時候就只有幾十個詞而已。可是總而言之，我們可以說，目前對語言的輪廓大致是摸清楚了，並且可以說，這幾千個語言從某些客觀的條件上看來，都是大同小異的。就是說，人類的所有語言，基本上都是很相似的。

這自然不是說語言中沒有差異，漢語裡的聲調就是很多語言裡沒有的。句法上來講，漢語的動詞放在句子當中，有的語言卻把動詞放在句前，如一些南島語言，還有的語言把動詞放在句後，如一些阿爾泰語言。從構詞上來講，漢語的名詞系統可以說是沒有性也沒有格的，可是非洲的 Bantu 語中，各詞往往是分七八個性的，北歐的芬蘭語名詞是有十幾個格的，可是這些都是極端的個別現象。總而言之，把語言整個系統跟整個系統比，相同的地方比不同的地方總是多得多。並且語言跟語言之間沒有過渡不了的隔離，只要時間夠，任何一個語言系統都可能演變成任何一個別的語言系統。所以有人說過，人類的語言雖然表面上看來好像是千變萬化的，可是基本上卻都是同一個模型變出來的。

其實語言這麼相似，想起來也應該是很合理的。因爲使用語言主要是有兩個生理上的條件，一個是大腦的神經組織，另一個是發音的器官。據生理學的研究，所有的人，不管是甚麼種族或者是甚麼樣的社會結構，在這兩條件的具備是一樣的。

同時呢，在語言發展的早期，一些原始人群的生活環境也是大同小異。因此語言上的需要也是比較一致。既然內在的生理條件跟外在的環境需要都相同，那麼所產生出來的語言當然也就會差不多一樣的了。

(2.1) 語言內外跟編譯解譯

現在我想談談外部語言和內部語言的關係。外部語言就是說出來讓別人聽到的話，這是人類交際最基本的工具，可是每個人的內部思想顯然也是建立在語言上頭的。在一般日常情形下，我們的思路已經搞得很熟，以致完全自動化，我們就感覺不到這些思路大多是用語言溝通的。往往一個新的問題太複雜了，要想得很精微的時候，語言就又會露出頭來，那時候，一個人就會自言自語，甚至於自己給自己打手勢，不過這內部語言，即不說出來的語言，到底是不是跟外部語言完全一致的呢？要是不一致，那麼兩者之間是個甚麼關係呢？我們目前還沒有頭緒。

也許我們可以做個比喻來幫我們初步地了解這些問題。大家都知道，在計算機編程序的時候，一般用的語言是所謂的高層次語言。三十年前，有兩個最常用的高層次語言。一個是 IBM 推行的 FORTRAN,這個字是 FORmula TRANslation 的縮寫。另一個是 John Kemeny 在Dartmouth 大學編成的 BASIC。這個字是 Beginners All-purpose Symbolic Instruction Code。這兩個語言主要的功能都限于計算。因爲這門學科剛起頭的時候，大家都認爲計算機唯一的用處就是幫我們作計算，幫我們 compute，因此才管它叫 computer，或計算機。可是這四十多年來這門學科有驚人的進展，它的功能已經遠遠地超過本來的意料，而在人生的極多方面都變成少不了的工具。連一些本來想像不到的學科，如文學、美術、音樂等等都用上計算機。美國有一個學報，叫 Computers and the Humanities，就是專門討論這方面活動的。計算機的功能越多，爲它設計的語言也就跟著多了起來。目前一般計算中心所設備的高層次語言,除了Fortran 跟Basic 之外,還有C，Pascal，Lisp，Cobol，Prolog，Forth 等等，多不勝舉。這些高層次語言有兩種好處，一來是它們跟我們日常用的語言比較接近，用的是看得懂的詞,如 IF，THEN，ELSE，GOTO，END 等等，二來有不同的語言來適應不同的工作需要。

比方說，公司裡要算大批的帳，語言要比較簡單易學而並不需要太多功能，這是一類語言，Cobol 可以說是這一類的原始型態。而做人工智能方面研究的時候，語言就得要靈活得多，並且它處理的數據，往往不是數目，而是一長串一長串的詞，就是很多的 lists。而它需要的語言，要能處理這些長單，process 這些 list，這就是 John McCarthy 早期設計 Lisp 的動機。Lisp 就是 list processing這兩

個詞所拼出來的。Lisp 的一個特點，就是它程序的內部結構，是可以用樹圖來理解，就好像是傳統句法用的樹圖。因此這類的語言，包括由Lisp所產生出來的 Prolog，都跟我們日常用的語言是比較接近的。

自然語言，如漢語、英語、日語，我們知道都是日新月異地在變化。那麼人造的計算機語言，既然不是自然語言，是個別的人所創造出來的，它們變不變呢？有趣的是，它們也老是在變，並且在變中，往往也需要定標準。比方說，Lisp 的種類很多，就是說它的方言很多，只是在最近，才肯定了一個標準語言，管它叫common Lisp，就是普通 Lisp。就好像大陸把一些漢語方言裡的特徵融合在一起，管它叫 common speech，就是普通話一樣。更有趣的一個例子是 Basic。 因爲它設計得比較早，三十多年前就在 Dartmouth 大學開始教學生用了，所以它的方言種類特別多，有 IBM 的 Basic 及 Basica，有 George Washington 大學的 GWbasic。 很多計算機甚至於根本就把不同的Basic設計在機器的 ROM裡頭。Basic 的發明者，John Kemeny，是個數學家，也做過一陣子 Dartmouth 的校長，它對於這些不同的方言，屢屢表示不滿，覺得他們自己的 Basic 才是堂堂正統的 Basic，別家的都是些旁門左道，給他貶爲 street Basic 。因此近年來，他們就發行了一套自己的軟體系統，叫作 True Basic， 就好像在說別人的東西都是假的。

其實這裡頭，有個相當有意思的問題，自然語言變化是有一些一定的基本原則，有的是生理上的因素，有的是社會上的因素。那麼計算機語言的變化，這些人造語言的變化，是不是也有些基本原則呢？這些原則跟自然語言變化的原則又有些甚麼關係呢？我希望會有專家把這兩種語言變化的原則來作比較，相提並論，可以寫一篇挺有啓發性的文章。

這些高層次語言，對於使用機器來編程序的人，的確是很方便的。可是對一台機器講來，是很不適合的。因爲機器最基本的語言，只有兩個單元，就是 *1* 跟 *0* 。因此在操用機器之前，先得把高層次語言，一步一步地編譯成機器能懂的指揮。編出來的一些數據，就很不像原來的高層次語言，更不像自然語言。因爲這些數據是爲了不同的CPU 所設計的，要讓機器能直接地施行這些數據，速度上自然也要非常地快。

編譯高層次語言，一般分爲兩種方式。一種是一句一句地解譯，原來的老Basic就是這樣，所謂 interpret，就像在聯合國開會一樣，某國大使說一句，翻譯人就給他翻一句，這樣就顧不到後頭還沒有說出來的話。這樣也有他的好處，就是比較有伸縮性，可以隨機應變。前頭沒說清楚或說錯了，後頭可以補充或糾正。編程序的時候，前幾條行不通，後頭可以再加幾條來補它的缺。但是interpret 語言的最

大弱點，就是這種方式比較慢。

另外的一種呢，就是把整套的程序全部編譯，同時也把這一套裡使用的所有單元完全聯系起來。這就有點像把那位大使的講演，從頭到尾作一個整套的翻譯。全部翻譯完之後，才把它播送出來。對計算機來說，這就是純粹的 compile。實行這樣編譯出來的程序，當然就快得多了。同時呢，這種程序的內部結構跟原來的高層次語言在形式上距離就比較遠，要是我們隨便把機器停下，就很難知道機器在程序裡已經走到那個步段了。

這個區別，就是 interpret 跟 compile 之中的區別,是很有啓發性的。我們可以連想到，口語跟思想之中會不會是有相似的區別呢？我們每天都有些一整套一整套的行爲，一點都不用思考地去作。比方說，早上起床、洗臉、梳頭、吃早飯，成了習慣了，就一樣一樣地去作，A 自然而然地引到 B，B 自然而然地到 C 等等。

有時候我在想思與言之中的關係，就不知道這個比喻打得對不對。 就是說，一大部份的思路可能就像從日常語言 compile 出來的數據，雖然形式上已經跟語言很不一樣---至少在我們的感覺上,可是這些思路卻是從語言"編譯"出來的一群現成的整套的程序。遇到新奇或特別複雜的情形,那些現成的整套的程序用不上了,我們就得又從語言開始,製造一些新的程序來分析和解決這些新的問題。從一個compile 語言變到一個 interpret 語言，一句一句的解譯，不是整套的來了。這只不過是個比喻。用思想的工具，也就是人的大腦來比機器，總是不會完全適合的，因爲目前最最先進的計算機還是遠不如大腦那樣多功能。可是把一般已經完全自動化了的思想看成整套由語言編譯成的數據，也許會起些間接的啓發作用的。

(3) 電腦作計算

前面提到，電腦最初的用法就是作算術，因爲那開始時代,大家還沒有想到它其實是有很多別的功能。1958年我正在密西根大學念語言學，想到了一個很有趣的問題，就是一個語言使用它的音位是有規定的頻率的，這些頻率跟音位的發音部位是不是有什麼規則呢？發音的器官，有的是大而遲緩，有的是小巧玲瓏。比方說用舌尖發音該是最快最方便的，那麼舌尖音的使用頻率是不是比別的部位音都要高呢？基本上看來，這是一個統計問題，是須要作很多算術的，爲了要答覆這個問題，我就開始用上計算機了，算出來的結果，還寫了篇文章(Wang and Crawford, 1960)。

我記得那時候用的是一台非常原始的電腦叫 IBM-650，機器裡是大批的眞空管，還沒有用上半導體，又大又熱，機外打的是硬紙卡片，要用它的時候就得先把程序跟資料打成一套卡片，然後把這一盒盒的卡片送到計算中心的櫃台上排進隊，隔一陣子再到櫃台來拿算出來的結果，所以從頭到底，機器根本就不讓別人動手的。有時候一不小心一張卡片上打進了個極小的錯誤就會糟蹋很多時間，因此那些卡片要對了又對才能放心。比起現在，那時候的電腦要是說它是事倍功半，還是算不錯的呢。

近年來，我們還是經常地用電腦來作計算，主要的是想了解語言中的關係，怎麼樣把它量化，剛才談到語言總是在變化的，同一個語言分化出來的幾個方言，時間越久，這方言中的距離越遠，就越不容易互相通話。在史丹福大學，我有個好友，是位遺傳學家，他告訴我，他們研究人群移動的時候，面對同樣的問題，他們測量的是人體特徵，就是血液裡一些基因，遺傳學家發明了一種計算方法，就是把基因上的距離跟人群在空間上的距離用個相當簡單的數學方式表達出來，我們倆就開始合作，看看是不是能把這測量基因的方法搬到語言學裡去。

我們決定研究語言的詞彙，同一個基因在不同的人群裡會有不同的表現，譬如髮色的基因有的人是黑的，有的人是黃的，有的人是紅的。同樣的，一個詞意在不同的語言裡也會有不同的表現，譬如國語裡的"茄子"，廣州話叫"矮瓜"，蘇州話叫"綠素"，福州話叫 "紫菜"，要是我們收集大量的詞，就可以測量出一批語言之中的詞彙相似度，然後就可使用這些相似度的數據來計算它跟空間距離的關係了。

當我們把這些方法用到一群南太平洋的南島語的時候，算出來的結果，可以說是很成功。參看圖(1)，這張圖是取自 Cavalli-Sforza跟我寫的文章，我們那次研究了17個小島上的語言，X-維是島跟島中的距離，Y-維是語言間的同源詞的百分比。詞彙跟空間距離是蠻有規則的，曲線的彎度我們在原文裡也提出了解釋，主要是由不同詞意的變化速度不一致所造成的。這樣從語言特徵算出來的成果是跟遺傳學從基因特徵算出來的成果很相像的，同一套計算方法可以在兩個不同學科裡運用。

可是當我們用這個方法來處理漢語的詞彙，就沒有成功。用的資料是北京大學出版的漢語方言詞彙(1964)，裡頭有將近一千個詞意，該是夠的了。可是因爲中國歷史久，移民的方式又非常複雜，所以算出來的圖裡，一點也看不出任何規律來，顯然遺傳學所用的這個方法只是在歷史地理條件比較簡單的時候才是有用的。這些條件複雜的時候，只能在統計學裡找別的方法了。

其實近一二十年來，分類這個學題發展地也很快，1973年，Sneath跟 Sokal 就

出版了一本頗有影響綜合性的書，叫作數值分類學，北京已有中文版。因爲世界上的東西，多部份都有分類的需要，包括所有的動物及植物。所以發展這個學題的學科，除了統計學，還有人類學、生物學等等學科的注意。我們這幾年就是在研究怎樣運用數值分類學來分析語言中的關係，可是要在這樣的工作，沒有電腦是不行的。基本上是用它來作計算，所以電腦裡裝了一個特製的co-processor，就方便的多。同時我們也用它把算出來的語言中的親屬關係，用樹圖畫出來，可是這些是量化的樹圖，跟一般傳統的樹圖不一樣，這些量化的樹圖裡，每一條樹枝的長短，都是由距離大小決定的。

總而言之，電腦的開始階段主要功能是計算，目前語言學裡，還是少不了這個功能，尤其是近年來作種種語言實驗，這個功能是越來越重要的。

(4) 電腦作資料庫

電腦的功能變成如此的多彩多姿是有好幾個因素的。其中一個關鍵因素就是它的容量增加得特別快。這對語言學講來，又是很有用的。我記得廿年前我跟幾個朋友在討論語言變遷的問題，那時候，鄭錦全、陳淵泉、謝信一，他們都還在柏克萊，我們覺得談變遷的時候不能只舉三三兩兩的例子，就算夠了。要了解一個變遷的整個過程，要敘述它的來龍去脈，應該把這個變遷所牽連到的詞，一個一個地都查出來，才能把情形搞清楚。這又得用上電腦了。

那時候我的實驗室裡已經買了一台電腦，是台 DEC 作的 LINC-8。雖然價錢貴得要命，可是容量倒小得可憐，總加起來，它的 RAM 只有四千個 BYTES。可是我們可以說是初生之犢不畏虎，也沒有怎樣顧慮到資料多、工作量大、電腦不夠用，就開始把漢語方言的材料電腦化起來，這個資料庫，我們管它叫Dictionary on Computer，簡稱爲DOC 。並且把在語言研究方面的幾點理想寫成了一篇文章，登在1970年的一份學報裡。

我們起初用的是北京大學編的漢語方音字滙，裡頭除了中古漢語的音類，還有十七個方言，每個方言有兩千四百多個字的發音，我們就決定用廿二個 BYTES來代表每個字的發音，後來我們又加上了一些別的資料，包括上海音、中原音韻、日語裡的漢音及吳音、朝鮮音、越南音等等，統統加在一起，是有一百三十多萬 BYTES 的一個資料庫，現在看起來，並不怎麼樣可觀，幾個軟盤就存得下，可是那時候給一個小小的 LINC-8 來作，實在是太勉强了，所以不久就搬到大些的電腦上去了。

這一套電腦化的漢語方言，可能是語言學裡最早的一個電化資料庫，對我們講

來，是很值得作的，那時候我們正在辯護一個歷史語言學的理論，就是詞彙擴散理論，因爲有這麼大的一個資料庫，能找出很多的証據來支持我們的論點，起了肯定的作用。讓中國語言學衝進了以前是清一色 Euro-centric 的理論圈，這也是理論語言學第一次認眞地注意到漢語方面的歷史材料。

語言變化時有許多現象是須要比較大量的材料才能看到它的傾向的，以前語言學家討論語言歷史的時候，多是以爲音變是全部的絕對照規律的。事實並非如此，例外是非常多的。要是我們用達爾文演化理論來看這倒是很合理的現象，因爲語言總是在變化，而變的時候,最典型的表現就是一個詞同時有兩三個不同的發音，這些問題，鄭錦全、陳淵泉、謝信一，都在不同的文章裡討論過，漢語方言那麼複雜，所謂字的文白異讀只不過是個初步的分析，歷代裡人群的移來移去，城市也跟著興旺或頹廢，因此一般方言裡的層次還不止兩個。這方面作得最細的研究，大概要算是連金發的1987博士論文分析的閩語詞彙。

DOC 作成後，鄭錦全跟我曾經提供一些粗略的材料，1971年的那篇是講漢語裡的聲母。1987年又有一篇講聲調的，圖(2)就是取自那篇，圖的左邊是講中古漢語(MC)的聲調在三個閩語裡的分佈，第一行的 lun，就平聲裡的不送氣的清聲母，在 DOC 裡一共是有364個字。這些字百分之九十三都變到廈門話裡的陰平調裡去，百分之二是陽平，還有百分之二到上聲(廈門話的上聲是不分陰陽的)，及百分之三到陽去調裡去。

我們做到潮洲聲調的時候，就發現一個很有趣的現象，在圖上可以看到中古漢語的陰去調 (就是3un 跟3ua)大部份都是變到潮洲的陰去調(就是3a裡的85%跟88%)這是完全合理的，可是中古漢語裡的陽去調(就是全濁的3vo及次濁的3vs)，卻是沒有那麼守規則，圖上可以看到全濁的139個字，只有44%在3b裡，其餘的有40%都到2b去了,次濁也有相似的傾向，133個字中卻有52%到2b去了，據我們所觀察，像這樣50%與50%的半斤八兩式的對比，就是一半在3b一半在2b是個極少見的情形，連金發的文章就是把造成這個情形的一些因素最仔細地分析出來，他的研究對了解方言接觸是很有啓發性的。

圖的右邊是把這17個方言的調類，作了個總結。我們可以看到最低的一排，就是廣州方言，一共有九個聲調，粵方言裡的聲調是漢語裡最多的，然後我們朝北方走，越走就聲調越少，我想這是因爲聲調變化的趨向，多部分是合併，少部分是分裂，北方方言是歷代來的政治文化中心的語言(國都絕大部份是在北部:長安、洛陽、北京),所以語言變得快，聲調也合併得早，因此就比東南岸方言的聲調少。

像這樣用大量的資料，統計的方法分析，又是沒有電腦就做不成的，DOC 這類的資料庫還有些別的可取的地方，就是作出來之後，大家都可以用軟盤寄來寄去，是非常方便的。這樣就變成了一個學術裡公共的工具，用它的人越多改良它的機會就越多，一方面可以盡量地刪除它裡頭的錯誤，另一方面又是可以一步一步地摻進新的資料，要它豐富起來。連金發最近在輸入很多新的閩南語的資料，就是一個例子，作研究的資料及方法，總是希望越公開才能越客觀，越有累積成果的可能。

近年來好多位台灣學者都在作電腦資料庫。聽說中研院的歷史語言研究所跟計算中心合作，已經把很多的史籍方面的材料，存入機器，將來會讓一些有關的學術結構運用，大陸方面也是相似的發現，見《文學遺產 2.141,1988》。這當然是比翻那些舊老的 INDEX或CONCORDANCE 要有用的多，一定會推動漢學研究的效力，搞語言學的人更能用這個資料庫來分析語法或語意的變遷，希望這系統完成之後，會吸引很多學人的興趣跟合作。

並且這方面的活動不限于漢語，而許多台灣的土著語言資料也在電腦化 (參看李壬癸 1986-7)。這些語言在南島系語族裡佔有非常關鍵的地位，台灣很可能是這一大語族的一個重要起源地，這個資料庫一定會起很大的作用。同時我們知道中國境內，至少有一百多個不同的語言，這些語言之中的關係至今還沒有用過什麼量化的方法處理，台灣土著語言的資料庫，可以算是爲中國非漢語研究開了一條新路。

(5)模擬語言行為

今天要談的第三種的電腦運用，是用它來模擬人的語言行爲。這方面的研究，以前不是沒有人作過，比方說，遠在公元1779年前，Kratzenstein就用一套振動的簧片，切斷氣流，合成了言語裡五個基本的元音:i,e,a,o,u。他的發明並且獲得了聖彼德皇家學院的獎學金。可是現在用電腦模擬語音比那時候的簧片及木匣子，要方便靈活得多了。目前用電腦來合成語音可說是一件相當簡單的事情，基本講來，是由兩種不同的方法---analog及parametric。

Analog方法是直接模擬發音器官的一些動作及體積，這就是Kratzenstein用的方法。我們知道，說話的時候，人的喉嚨及口腔就像一根150多毫米長的管子。這根管子有的地方比較粗，有的地方比較細，看我們正在發的是甚麼音，要是我們用X光分析出來某個元音的形狀，我們就可以用電腦來模擬這根管子，然後我們就可以用上電機工程裡一些現成的transmission line 的理論發出音來。

近年來，這方面用來合成語音的理論已經了解的相當徹底了。我們不但有能力

來從管子的形狀推測出它的聲譜，合成相當好的語音；我們也能作相反的推測，就是倒過來，從聲波的頻率及幅度，我們可以算出管子的形狀來。

另一種合成語音的方法叫 parametric---就是把語音分析成一套共時的參數，通常是十幾個參數 ---parameters。比方說，*F0*是一個必要的參數，它是代表聲帶抖動的頻率，在合成漢語時是特別重要的，因爲它是聲調的主要成分，另一個是*F1*，它是第一個共振峰的頻率，另外還有 *F2*,*F3*等等。我們要把一秒鐘的語音存在電腦裡，如果是用兩萬個樣品，並用*12*個bits的信噪比，那麼這一秒鐘語音的信息量將是*20000x12=240,000bits*。

可是在把這語音變成參數後，這信息量就大爲減低，要是我們用*12*個參數，而每個參數平均用*6*個bits來敘述，那麼每套參數只須用*72*個bits。這些參數的數值變得很慢，所以在傳達語音時，先把語音變成參數，傳達後，再由參數合成語音，這樣可以大大地壓縮帶寬，減少信道信息量的條件。近年來，語音學主要就是在研究這些參數之中的關係及其重要性。語音學裡最基本的問題，就是怎樣把傳統的抽象的語音特徵(phonological feature) 跟具體的參數聯結在一起。

我們既然把語音分析成一套參數，就可以人工地改變這些參數，把改後的參數再合成語音，就會產生出一種很奇怪的現象。就是這樣造成的語音，是半眞半假的。因爲它一部分是原有的人說的話，而另一部分卻是人工編進的一些數值，不是人所發出來的，或是人根本就不能發出來的。

圖*(3)*跟圖*(4)*是用電腦製出的一些語音分析材料圖。*(3)*有三部份，上面是我說的一個短句:"語言與語音"，這句話先說在話筒裡，然後經過一台模擬數字轉換器把聲波存入清華大學語言實驗室裡的電腦，這幾台電腦都備有一套比較完整的分析波形的軟體，叫作ILS(Interactive Laboratory System)。圖*(3)*的中部的曲線是這句話聲波裡的基頻率*(F0)*跟振幅。圖*(3)*的下部是電腦劃出來的聲譜圖，因爲這句話就是一串濁音，所以共振峰一直就沒有間斷，只是在末了一個音節，由于"音"字裡的鼻音，*F2*中斷了一段。

圖*(4)*是把句子前頭的元音，用ILS把它劃成三維的聲譜圖，可以讓我們把共振峰的幅度看得更準確，去年北京有一位聲學家馬大猷，寫了一本關于語言信息的書，是個很好的介紹。

辨別語音要比合成語音難得多，所以雖然目前合成語音已經能作得很完美了，但是辨別語音卻是還有很多基本問題，無法解決。主要是因爲人在聽話的時候，可

以利用很多聲波之外的信息,而這恰是目前電腦作不到的，要是我們願意接受一些限制，譬如說,只辨別一個人的語音，不顧慮到別人的語音，或者說,發音人在每個字或每個詞之後都停頓一下，並且只限于一個固定的詞庫，那麼這件事情就好辦得多,可是這就跟人的語言能力差得很遠，因爲我們日常用語言,能毫不費力地跟好幾個人同時交談，並且我們也不需要跟我們說話的人,在每個字後要停頓片刻。可是就這樣人人都很容易做到的事情，電腦還遠遠不如呢!尤其是在臺灣這社會裡,方言情形非常複雜，有的人說的是標準國語，有的是江浙國語，有的是帶了閩南口音的國語，有的是客家口音的國語;總而言之，在這多采多姿的語言環境裡,要一個電腦去辨別語音了解語意，照我們目前的知識講來，是絲毫沒有希望的。

(5.1)翻譯

人的語言能力最精彩的表現就是在翻譯,因爲在一個理想的翻譯過程中，一般是要包括以下的五步程序:

(一)把源語模擬似的聲波分析成一個數字式的表示系統,這立刻就牽涉到剛才我們說的，電腦還沒有希望做到的語音辨認,其實語音辨認還只是這部分開頭的一段，音素跟詞分析出來之後,另外還要用構詞法跟句法把這些詞中的關係用樹圖表示出來。

(二)求出源語的句法樹圖後,我們就得從它算出一句話一句話的語意。語意學是整個語言學裡最不發達的一門,這大概是因爲意義是非常模糊的東西，很難把它編成一個合理的系統,所以目前我們對於怎樣來代表一個句子的語意，還沒有什麼大家都同意的方法。可是我想,語意的結構一定或多或少有它不同層次的，而樹圖是表示分層結構的好辦法,所以也許語意可以用樹圖來畫。

(三)求出語意，畫出語意樹圖之後，我們才能向目標語的方向出發，這一步的目的就是要求出目標語的那一個語意樹圖是相對於源語裡的那個語意樹圖。這裡的困難，不只是語意學不發達的問題，現在還要牽涉到語用學上面去,光是semantics還不夠用，一定還須要用上更難研究的,跟社會，文化等等語外因素有密切關係的pragmatics。有時候可以聽到人說， 某某人說話沒有分寸,不懂禮節，不知道說話的輕重。最顯著易見的就是稱呼要看年齡，地位，場合。這是外國人到中國來，或者中國人到外國去，很容易搞錯的語用問題，可是這是最簡單的，在兩個文化接觸的時候，的確是有很多更複雜的語用衝突會發生。

求出目標語的語意圖之後，下兩步就容易得多。第四步就是要把語意圖變爲句

法圖，第五步是把句法圖，利用數字模擬轉換，變成目標語裡一個句子的聲波。最後這一步，就是剛才我們談的合成語音，現在電腦可以做得很好了。

翻譯工作有兩種形式，就是口譯和筆譯。口譯的時候，就是把源語的話，一邊說就一邊翻，大家可能在電視上看到過聯合國開會時，某國的大使在講演，同時就有不同的翻譯者把他的話立刻就翻成種種的語言，譯者跟著源語跟得很緊，有時候甚至源語的句子還沒有說完，目標語的話就開始了，就好像是譯者能夠預料到那句句子會是怎樣結尾的，並且譯者時常會一邊聽著源語的話，一邊說出目標語的話。口譯要作得好，是件很不容易的技巧，需要相當的訓練，也需要一種特殊的天才，這的確是個很奇妙的語言能力，從電腦模擬的角度看來，現在還是一件差得遠的事情。

筆譯就比口譯簡單得多，因爲譯者沒有時間的壓力，可以參考句子前後的材料，反覆沉思，求出最理想的翻譯。同時我們知道，有些材料比較容易翻譯，譬如理，工，數學方面的文章不怎麼依賴著語言跟文化，翻譯起來不太困難；而文學材料，尤其是詩詞，特別難處理。因爲一首詩的美，不止是思想玄妙，往往還有歷史上有趣的典故，而它同時也有聲音上的美，諧聲，押韻，都是使它好聽的辦法。所以唐詩三百首，不知道翻譯成英文多少次了，每一版本都不同，我也看過了至少十幾個不同的翻譯，但是沒有一個有原文那麼美，這可能就是原則上根本就做不到的事情。

電腦翻譯是在1950年代開始的。那一陣子，句法理論，語意理論，語用理論，都還沒有走上軌道，一般人都沒有領會到一個語言的結構有多麼複雜，所以大家都很樂觀，(也許可以說是天眞)，以爲電腦既然發展得那麼快，不久就會變成萬能的，人所能做的事，電腦一定會做得更好，美國的國防部也深深覺得須要把翻譯電腦化，因爲他們每天都有成千成萬的外文材料，需要分析，翻譯後才能知道哪些材料是國防有關的，而哪些是無關的。但是要翻譯那麼多的外文材料在美國的人工，特別是從軍事觀點來說是可靠的人工，是絕對做不了的，因爲工作量實在是太大了，所以那一，二十年裡投資了好多億的美金去發展電腦翻譯，一開始是讓華盛頓的Georgetown大學去作俄英翻譯，後來一些別的大學以及一些大的電腦公司，如IBM，也都參加這項工作，俄文的規模最大，後來又加了德文，法文等等，我們在加州大學，起頭的一段也是作的俄英翻譯，可是作了兩年之後，就把我們在加大的研究重點移到中文方面，主要的任務是想用電腦把一些中文的科學文章，特別是生物化學跟物理全部自動地翻成英文。加大的這項工作前後收到美國科學基金會及國防部的支持，作了七八年之久。

電腦翻譯作得很起勁的時候,也鬧出一些笑話來,這些笑話無論它是眞是假,還是可以讓我們想到這項工作裡的各方面的困難,尤其是一些俗語,特別不好處理,據說聖經裡有一句話是:

The spirit is willing ,but the flesh is weak.

它帶有宗教性的教訓,就是說有些心靈裡願意作的事而被我們的肉體之弱所限,可是翻成俄文的時候卻走了樣了,spirit 還有一個較爲古老的意思,就是烈酒,如俄國人喜歡喝的VODKA之類,英文裡flesh是跟meat相對。Flesh 比較抽象,多部分是講人體的,譬如說一個人的親屬是他的flesh and blood。 Samuel Butler 也寫過一本著名小說The Way of All Flesh,是講世上的肉體誘惑。而meat 是講吃的肉像牛肉之類,可是有些語言裡就沒有這個區別,flesh 跟 meat 都是肉。這次的電腦翻譯,找錯了詞,把一句宗教性的訓語翻到食物的方面去,變成:

Vodka 還可以,不過肉是餿了。

還有一個俄語英語的笑話也是由一句俗語,就是:

Out of sight, out of mind.

意思是說,多次不見面,濃厚的感情也會消滅的。國語裡也有一句相似的俗語,就是人在情在。可是這句話翻成俄語的時候,竟變成了瞎了眼的瘋子。這個有趣的錯誤也是理解得到的,因爲俄語裡瘋子這個詞是сумасшéдший。它的兩個詞根是 -ум- 就是心靈,-шéд- 就是走出,走出心靈的人,就是個瘋子,沒有視覺,離開了心靈的確是個瞎了眼的瘋子。

其實翻譯上的不恰當,在雙語的場合裡,很容易找到,前幾個月我在香港,就見到許多。一個是馬路上的一個牌子,是叫開汽車的人在小路口要特別小心,注意過路的人,可是牌子上寫的是: Beware of pedestrians.

可是仔細地想一想覺得這個牌子有點可笑,因爲英文裡 beware of X 這樣句子裡,X是會傷害你的東西,所以你得提防它,譬如最常見的是人家家門口的 Beware of dog,或者是山上開車時,會看見的 Beware of falling rock。或者是大都市熱鬧地方的 Beware of pickpockets,就是小心扒手。我不知道國語裡"小心行人",是會給人一種什麼樣的感覺,可是我知道英文裡,beware of pedestrians 是不恰當,它的含意是,行人在你不提防的時候,對你會有不利的,也許會來個行人咬你一口。這個例子也是說明一個詞的意義,有時候很細微的差別,還是會很重要的。

有些例子是讓我們看到句法上的不同。翻譯的時候絕對不能把源語的句法，勉强地套在目標語的句法上，有時候剛開始學中文的美國學生，會說出這種有趣的句子：

我不會說中國話很好。

一聽到這句子時，覺得這個學生既然在學中國話，爲什麼又要說他不會說中國話很好，可是一經分析，就看得出來，他用的是中國話的詞，硬套上了英文句法，就是：

I can't speak Chinese very well.

其實這句話裡的否定詞在語意上是否定 very well 這個狀語的，要是這樣的句型，就跟國語比較接近了。

I can speak Chinese NOT very well.

我　中國話說得　　不　很　好。

可是英語句法裡有條規則，就是在某些情形下，把否定詞提到句子前頭第一個動詞的後頭，可是這個學生卻把英語的那條規則錯用在國語的句子，造成這樣的笑話。

同樣的英語句法也常把一些所謂的 WH- 疑問詞像 who, what, when,where 提到句子前面。比方說英語不說 you live where ,而說 Where do you live。鄭秋豫女士有一次跟我說，她聽到某位先生說英文確實地說出這麼一句話：

You ask me, me ask who go?

這個例子跟剛才那個恰好相反，這次的翻譯毛病是把中文的語法套在英文句子上了，這個句子是詞對詞從中文翻過來的：

你問我，我問誰去。

這個句子最顯著的句法毛病，就是沒把 who 提前，變成：

You ask me, who do I ask?

同時我們知道往往一個句子是可以有很多不同的分析的，因爲它裡頭的詞跟詞

可能有不同的句法或語意的關係,這種句子人聽到的時候,平常都不會出問題的,人跟人之間都有很多相同的經驗知識、感覺,幫我們來了解語意。可是一台電腦就沒有這些設備,並且目前也不可能有這些設備,因此每遇到一個多義的句子,就很不容易辨別出那個語意是對的,哪些是可能的,哪些是不可能的。比方說有這麼個英文句子,要翻成中文:

THE POLICE WAS ORDERED TO STOP DRINKING BY MIDNIGHT.

我想一個作翻譯的人,看就會懂它的意思,就是命令警察在午夜前禁止喝酒,句裡有很多沒有說的東西,比方說因爲 STOP 後沒有賓語,我們就不知道被禁止的是誰,雖然由于我們對社會組織的了解可以猜想這個命令不是要警察自己停止喝酒,可是這個猜想不是從句子本身所得出來的。這樣利用信號之外的信息來了解信號,剛才在談語音辨別的時候也提到過,這是人的語言能力的本色,是很不容易教給電腦的。

這句子當然還有很多別的解釋,我們猜想禁止喝的東西是酒,因爲酒能傷人,DRINK 是可以指任何飲料的。我們猜想午夜是禁止的時候,可是它也可能是發出命令的時候,甚至于 MIDNIGHT 也可以是一個發令的人或行政單位,或是一種禁止的方法,還有我們猜想這個命令是講每天午夜後不准喝,白天可以喝,但是它也可能是說從那天的午夜後永久就再不准喝。總而言之這句子有很多種的不同解釋,我們能輕易地作個恰當的選擇,是因爲我們有無限多的信號之外的日常知識,幫我們了解這句子的背景。況且了解了這個句子,又能幫我們了解下一個句子,這樣的語言行爲現在我們不容易用電腦來模擬。

可是就像辨別語音一樣,要是我們把翻譯的條件能安排得很嚴格,那麼這是可以作出一點成果來的,圖(5)就是取自十五年前,我在Scientific American 裡發表的一篇文章。那時候還沒有電腦中文系統,所以中文句子都是要用電報號碼打進去,在柏克萊幫我作這項電腦翻譯的,主要有鄒嘉彥跟幾位語言學系跟電腦學系的研究生。圖上看得出我們把漢語的詞類分得非常細,名詞,動詞及形容詞,每個都分好幾十類,找出相對的英文詞後,除了調整英語詞的屈折之外,往往還得移動某些詞組、詞串,才能變得像個英文句子。這方面的工作,我們在*1973*年的 Linguistics 作過報告,這篇文章也有人譯在北京的*1979*年的語言學動態裡。

模擬語言行爲是件極有趣味的研究,也是非常難作的一件事,因爲它牽涉到的知識範圍太廣,目前寫電腦程序還不容易知道怎樣下手,近年來越多的研究小組都是跨學科性的,語言學的也有,電子系的也有,資訊系的也有,這是一個好現象,

同時語言學家也開始認眞地把語意跟語用，摻進語言理論。有這樣的進展，雖然題目是非常難，我想遲早還是能作出些成果來的。

(6)中文系統

我們學術界裡的人，主要的任務是探求眞理，幫人類增加智慧，了解我們的環境，讓我們的生活變得越來越幸福，越來越有意義。因此我們同時也應該注意到社會上的需要，尤其是一些在專業上有能力爲它效勞的需要，在這方面我們可以談談中文電腦的發展。

在十月六日的聯合報上(第十一版)，就是兩個多禮拜以前，有一篇短篇新聞，標題寫的是中文電腦人才荒，文章裡頭說到全國商業總會對這個問題一個粗略的估計，臺灣現在已經電腦化的，或即將電腦化的公司和工廠，最少有十五到二十萬家以上。據報上說，要是平均每家需要兩名操作電腦人員，那麼臺灣在未來的一年之內，就會需要三四十萬專業或兼職電腦操作的人員，現在這方面的人才遠不能滿足社會上的需要。

這說的只是臺灣，需要這類人才的地方還有香港，新加坡跟大陸，總加起來，這個市場會是驚人之大。而我覺得，ROCLING 裡的人物，在這市場的發展，是會很有作用的，因爲近年的這些活動只能代表一個萌芽時代。依我看來，現在我們所用的很多中文系統，譬如倚天，國喬，以及大陸的華達等等，都是些比較簡單的系統，從一條工程事件觀點上來看，事情的確可以做了，字是可以輸進去，可以打出來了，可是從語言學跟心理學的觀點上來看呢，那是有許多可以改良，值得改良的地方，因爲目前的一些系統，大部分是把一群漢字字庫套在現成的軟體上，MSDOS也好，WORDSTAR 也好,可是還沒有下多少功夫爲用户的方便著想，也沒有把漢字的特徵，漢語的一些構詞方法，加進系統去，在這些地方努力，是特別需要的，就是要把語言學裡的一些理論及方法跟電腦學裡的一些理論及方法融成一體，而這也就是我對 ROCLING 所懷的一個希望,一個期待。

(6.1)靠音還是靠形

今天我們自然不能仔細地研究怎樣改良漢語輸入的每個細節，只能輕描淡寫地提出兩個問題,第一個是輸入方法是靠音好呢?還是靠形好呢? 這個問題，短期的答覆跟長期的答覆可能不同，短期的答覆就是要根據目前的電腦的使用，據我所了解，目前絶大部分的時候都是有了現成的手寫的稿子，交給專業的人打的。這些專業

的打字員，是受過一段專門的訓練，他們是眼看著寫好了的字，一個一個打進去的。在這種情形之下，可能靠形輸入會比靠音快，可是這也不一定，不知道有沒有人把它作過科學的比較。

可是要是我們用長期的眼光來想它，那也許靠音輸入有幾個優點，一個是比較容易學，無論我們用的是漢語拼音，或者是ㄅㄆㄇㄈ的注音符號，反正漢語的音節極爲簡單，因此可以很快就學會，尤其是用拼音的時候，拉丁字母在一般的鍵盤上都有了，那麼在國際的場合裡，我們自然是希望儘量地去推廣電腦的運用，這對我們國家全盤的進展是很重要的。並且將來電腦使用大衆化之後，音輸的中文系統，還能起一些推廣標準國語的作用。

就從我個人的英文經驗來說，我覺得電腦文字處理對我的幫助，實在是太大了。八九年前，我也是用手寫，或是用舊式的打字機打，用剪刀剪，用膠水貼，然後最後的稿子要是能找到人幫我重打，那麼算是容易的。往往最後的稿子，都要自己從頭到尾再打一次。近年來我用些文字處理的軟體發現我可以坐在電腦前一邊想一邊打，要改的時候一點也不費事，可以一大段一大段的移來移去，這比以前動不動就得要人幫忙，的確好得多了，所以這幾年來，我雖然東奔西跑的，但是我總是帶著我的小電腦，一個兩三百美元的 LAPTOP。時常在飛機場裡，甚至於在飛機上寫東西，用 LAPTOP 的 BASIC 編些小程序。我希望漢語的文字處理，不久也能夠發展到如此的方便。

除了學習的時間短，還有點好處就是往往有的少用的字，一時忘記怎麼寫了，可是還是可以照樣地打進去，會說而忘了寫，這是常見的事，會寫而不會說，是比較少的，這也是靠音輸入的一個優點。

可是我們自然不能不注意方言對發音的影響，其實這個問題，是每個語言都有的，只是程度不同罷了。比方說，英國的很多方言，包括所謂的標準音，就是以一些貴族學校裡教的 RECEIVED PRONUNCIATION (RP English)爲基礎的標準音，在元音後頭的 r 音都早已消失。因此，這些詞裡雖然寫上 r 的字母，可是講起來都絲毫沒有r 的音，如burr, burp, bird等等，這些方言叫作無r 方言(r-less dialect)，在美國，澳洲的一些地方也是常見到的。保留 r 的方言在這點是比較保守，它們叫有 r 方言，也叫 r-ful dialect，可是無論是有 r 無r，他們用起電腦來輸入文字，沒有任何困難。

國語裡的捲舌音，跟這個情形，有一點相似的地方，雖然國語裡有一整套的捲舌音，就是ㄓㄔㄕ這三個聲母，漢語大部分的方言裡都沒有這套音了，所以漢語裡

也是有 r 方言跟無 r 方言。一般的情形是,這些捲舌音跟舌尖音合併了，在無 r 方言裡，有些國語裡分開的詞,就變同音了。比方說，國語的"三"和"山"，國語的"數"和"素"，上海話都用 s 。我想歷史的潮流是偏向著無 r 方言，再過幾十年，或一兩百年,可能所有的漢語都會變成無 r 方言。

還有一個相當普遍的音變,就是國語裡 i 元音後的舌根和舌尖韻尾在很多方言裡，也在合併。在這些方言裡，"彈琴"跟"談情"就變成同音了,像這樣有 r ，無 r ，n和ng的問題,對電腦來講，不應該是很難解決的，最起碼的方法，就是把用户方言裡合併的詞，全部在螢幕上現出來讓用户去選擇。

(6.2)用詞作單元

我想談的第二個問題，是比這個靠音靠形的對比更爲重要的。說起文字處理，大家第一會想到的，就是方塊字。漢字在中國人的心裡有很悠久的傳統,所以不知不覺地,漢字就變成一般電腦系統的基本單元。可是我想，這個問題值得從頭考慮一下，漢語有個很顯著的特徵，就是這個等式:

一個字=一個詞素=一個音節

在這方面，漢語的結構是相當整齊，在口語裡，雖然有些詞尾已經漸漸地弱化成個單獨的輔音，黏在前頭的詞根上，如 tam，wod，huar(他們,我的，畫儿)等，可是在話說得慢的時候，還是照著漢字一樣,說出完整的音節。

相反的例子自然也有，就是一個詞素有多個音節，或多個漢字，這裡多半都是些借音的外來詞,如早期借入的胡同,葡萄,玻璃,幽默.....,以及比較新的瓦斯，高爾夫....。可是漢語是比較不大用外來詞的，所以例子不是很多。

有些詞素可以單獨說出，也就是個別的詞，例如我，你，他，水,飯,山.....等等，可是很多詞是兩個或三個詞素串起來構造成的,這種多詞素的詞，國語裡好像是比任何別的方言都要多，例如火車，吝嗇，打火機，黑板，大方。有的時候還用上些常見的詞尾，例如磚頭，枕頭，上頭.....的"頭";孩子，帽子，麥子....的"子"。由於環境總是在變化，所以很多詞跟它本來的關係越來越遠，把詞素的意思加起來，往往就得不到整個詞的意思，打火機是比較容易了解還是用火的，可是現在火車一般都不用火了，黑板往往不是黑的，雖然它的功能沒有改變。而一個大方的人，很可能是一個也不大也不方的人。

我想我們使用語言的單位,往往不是詞素，而是詞。詞是詞素組成,的它的數目當然要比詞素多的多，可是要是電腦的容量充分的話，那麼把詞當作找字的單元，是會有很多方便的。現在我們可以看幾個具體的例子。我覺得近年來一些中文系統，倚天是做得比較好用的，所以用倚天的1.45版來討論這些個人的經驗。

譬如說，我要打"電腦"這兩個字，那麼我得先按四個鍵，就是ㄉㄧㄢ跟第四聲的ˋ，螢幕上就會給我一排字讓我選擇，"電"字是這排的第九個，我就要再按個9鍵，那樣"電"字就可以打上。然後我得按三個鍵,就是ㄋㄠ跟第三聲的ˇ，這次我所要的"腦"字是第三個，所以就要按個3鍵,那麼打"電腦"這兩個字的過程是八個鍵:ㄉㄧㄢˋ9ㄋㄠˇ3等八個鍵。

值得注意的是，每打個數目的時候，我們得移動視線，從書面,或鍵盤,或螢幕的上部移動到螢幕的下部。做選擇的時候，我們得注意到排裡的每一個字,要是第一排裡的十個字裡沒有我們要的字,那麼就要按 SPACE到第二排裡去找，還沒有的話，再按 SPACE 到第三排去找。還有的呢,就是想打字打得快一點，字在第一排裡沒看見，就到第二排,第三排,別的排裡去找,最後還得回到第一排來，反而花掉更多的時間,所以在字排裡找字的這一步，是最慢的一步，況且還容易走錯路。

那麼在排裡的次序，就變得很重要。讓我再舉個例子，前幾天，我在給朋友寫封信，順便練習中文輸入，我想打的是"意思"這個詞，按了個一的元音跟ˋ第四聲後，第一排就出來了好幾個我不認得的字，第一個字是"弋"字，查了字典後才知道好像這是個什麼跟彎弓射箭有關係的東西。第二個是"刈"，是一種割草用的鐮刀，第三個字明明是衣裳的"衣"字，是第一聲,爲什麼要擺在第四聲裡頭呢?後來想一想，大概是因爲在古漢語裡，"衣"字當作動詞的時候的去聲吧。"意思"的"意"應該是常用的字,可是第一排裡反而沒有，要按了 SPACE 後，現出了第二排，找到第八個字，才終於找到了。

同樣的，在同一封信裡我要打"機會"這個詞，"機"字也是個日常都用得到的詞，可是這個字，我要到第三排的第三個字才找到。換句話說，它是排後到第三十三個位子裡。這裡的問題跟剛才的 "意思" 是一模一樣的，就是沒有注意到字的使用頻率，往往把一些很少見的字排在前頭,或者把一些有特殊唸法的字也放在前頭，這些低頻率的字，在找常用字的時候，都是一種干擾，至少是一種不方便。所以在進一步設計中文系統的時候，我們必要顧慮到使用頻率，甚至於有十幾個極常用的虛詞,如"的"，"了","個"，"不","是"等可以特別處理，用單鍵或SPACE 來叫它，如"d"，"l","g","b","s"等，現在"是"字尸丶裡的第十四個字，第一字卻是"王士

元"的"士"字。這方面讓我個人用起來雖然很高興，可是這個"士"字，日常講來，用頭不能算大。

國語裡有一千多個不同的音節，要是不算聲調的話，那麼就只有四百多個音節，據我的感覺，好像一般的情形是一個音節裡只有一兩個常用的字，其餘的都是比較少見的字。要是能把每個音節的常用字總排在第一位，並且把 SPACE 作爲選第一位字的信號，那麼使用者就可以經常用大姆指來按 SPACE ，不移動視線，不被選擇這部分過于耽誤了。

可是更重要的是，我們應當儘量地把詞作單位。原則是 A 可能是一個音節，有很多字；B 可能是一個音節，有很多字，可是 AB 在一起的時候，它就會有自動的選擇。就像語言給我們預先就準備了一個篩子，只利用極小的一部分，爲了說明這個道理，讓我們再看"電腦"這個例子，在倚天1.45版裡，DIAN 這個音節一共是39個字，第一聲有11個字，第三聲有7個字，第四聲有21個字；可是國語裡第二聲的 DIAN 字一個都沒有，這個缺可以很容易地用音韻學來解釋，就是第二聲是陽平調，而國語裡的陽平調，塞音聲母都得是送氣的。類似地，我們也可以說明國語裡爲甚麼沒有bianˊ，jianˊ，zanˊ，跟zhanˊ這些音節。

NAO 這個音節在第一聲裡，一個字都沒有，這也有它的音韻學解釋，因爲第一聲是陰平，聲母一定是要帶聲的。第二聲裡有12個字，第三聲裡有3個字，第四聲裡只有2 個字，加在一起是17個NAO字。那麼39個DIAN字乘上17個NAO字，一共是663對字，數目是很大，可是絕大部分都是國語不用的。我們可以很容易地看到，實在常用的詞，的確是非常少，也許就只有我們說的"電腦"這一個詞。設計中文系統時，這樣語言學心理學很基本的知識實在是必當參考利用的。

所以我覺得，最理想的是讓我們一連串就按五個鍵，就是ㄉㄧㄢㄋㄠ，然後就按 SPACE，然後螢幕上就現出三兩個詞，最常用的詞就用SPACE選，這樣就省事得多，打一頁字，就可以少花很多時間。

再進一步的辦法，就是讓電腦自動地隨時地跟著使用者的輸入字的頻率調整它在螢幕上出現的順序。一個詞用了幾次後，就自動地把它的位置提前，這樣作來，在同一篇文件裡，電腦就變得越來越聰明，速度也是跟著很快地增加。

我說這些話的用意不是要批評中文系統的發展，站在一旁說風涼話。相反的，我覺得這方面的進展，在短短幾年中就有今天這樣的成果，很不容易，我們是可以爲它驕傲的。可是這方面的發展，對於電腦在中國語言學裡的效力，實在是太重要

了。有些學者，已經很注重這些問題 (參看謝清俊,1986)。我衷心的希望,在未來的幾年中,中文系統可以達到更高更完美的水平。使它不只是一個交給別人替你打出漂亮稿子來的系統,而使它變成一件常在手旁，不得不用的工具，幫著我們整理思想，處理資料，提高我們工作的效率。在這方面，英文電腦目前是好用得多，常有美國朋友對我說，他們現在用慣了電腦，就覺得以前拿著筆寫文章,實在是個想像不到的原始方法。用慣了電腦，他們工作的效力增加了好多，事半功倍。我希望不久，我們也可以用中文電腦來很輕鬆，很容易地做很多事，增加中國學術界的水平，這是我們要把電腦中國化的一件很迫切的任務。

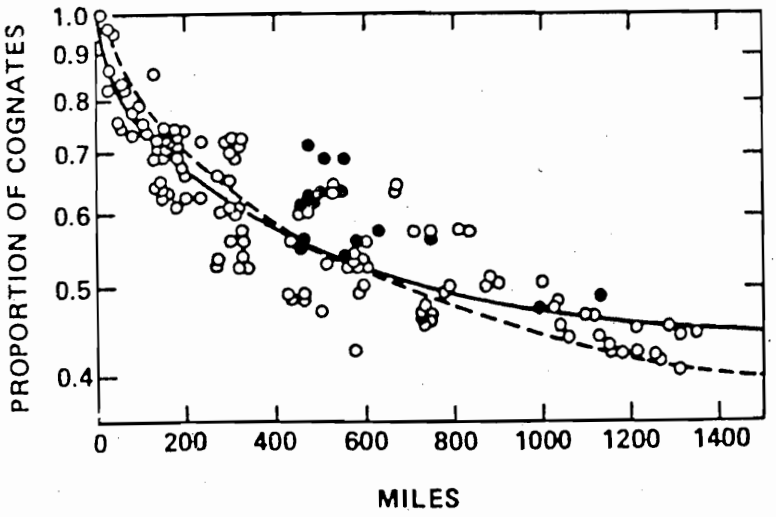
(7)結語

最後我想稍微談談計算語言學這個詞，就是說 Computational Linguistics。人家問我的時候，我曾經說過，依我的看法，計算語言學不是一門獨立的學科。現在應該補充說下去的是，計算語言學就是語言學，而它的特點，就是運用一些最先進的工具，來研究語言學裡的很多問題。有些問題是語言教學的問題，現在美國語言學界裡所謂 CALL，就是Computer aided language learning,是個很時髦的教學方法；有些問題是針對著語言歷史的，我剛才也談到了這方面的活動；有很多問題是跟人工智慧有關，就是電腦翻譯，語音識別等等。因爲人跟人交際的主要工具是語言，而人跟電腦，以及將來的電腦與電腦之間的交際，主要的工具也都是種種的程序語言，這種種的語言的目標都是相同的----就是傳達信息。因此，研究程序語言與自然語言當中的關係，一定會對了解自然語言有很大的啓發性。

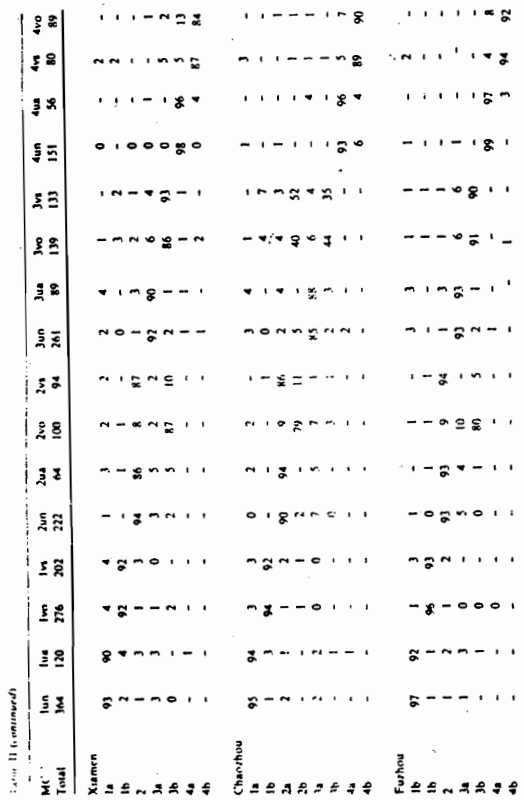
總而言之，我是覺得，如果我們要把語言學變成一門科學，電腦就是一件我們不能缺少的關鍵工具。所以計算語言學這個名字，是有一點兒冗餘，因爲利用電腦來做研究，就是語言學裡很中心的一部分。沒有電腦，很多先進的研究工作根本就做不起來，所以我說，計算語言學不是什麼可以擱在一旁的玩具，而是用電腦的力量來推動我們這門學科的主流活動。

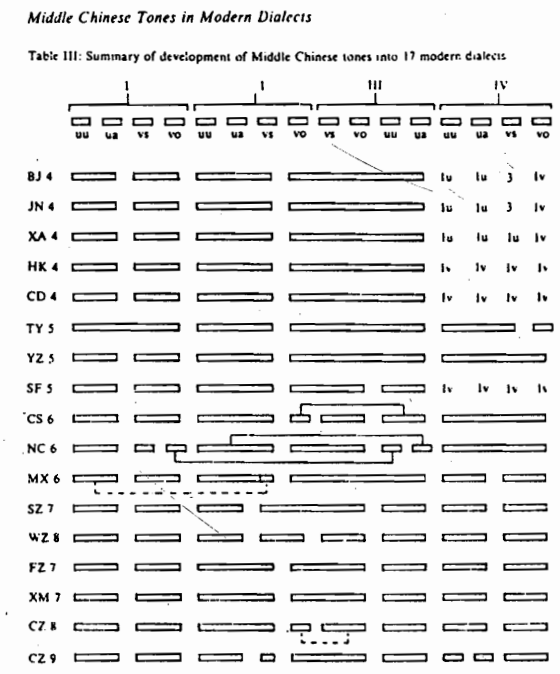
過去的兩百年，語言學的核心，一直是在歐美國家，所以目前的一些方法，理論，大部分都以歐美的語言，歐美的學術做出發點。因此有時候，再謹慎的學者，講話時都免不了帶些 Euro-centric的毛病。可是,近年來，電子工業的發展，東亞不見得比歐美慢，電腦教育呢，據我這兩個月在臺灣的經驗來說，也是非常普遍，而且速度還在增加。今天 ROCLING這個會議，就是個很吉祥的預兆，我的熱烈的願望，就是能在這條路上，儘快地朝前跑，運用這股從電腦得來的新力量，把中國語言學推進到一個領先的地位去。

(8) 圖解說明

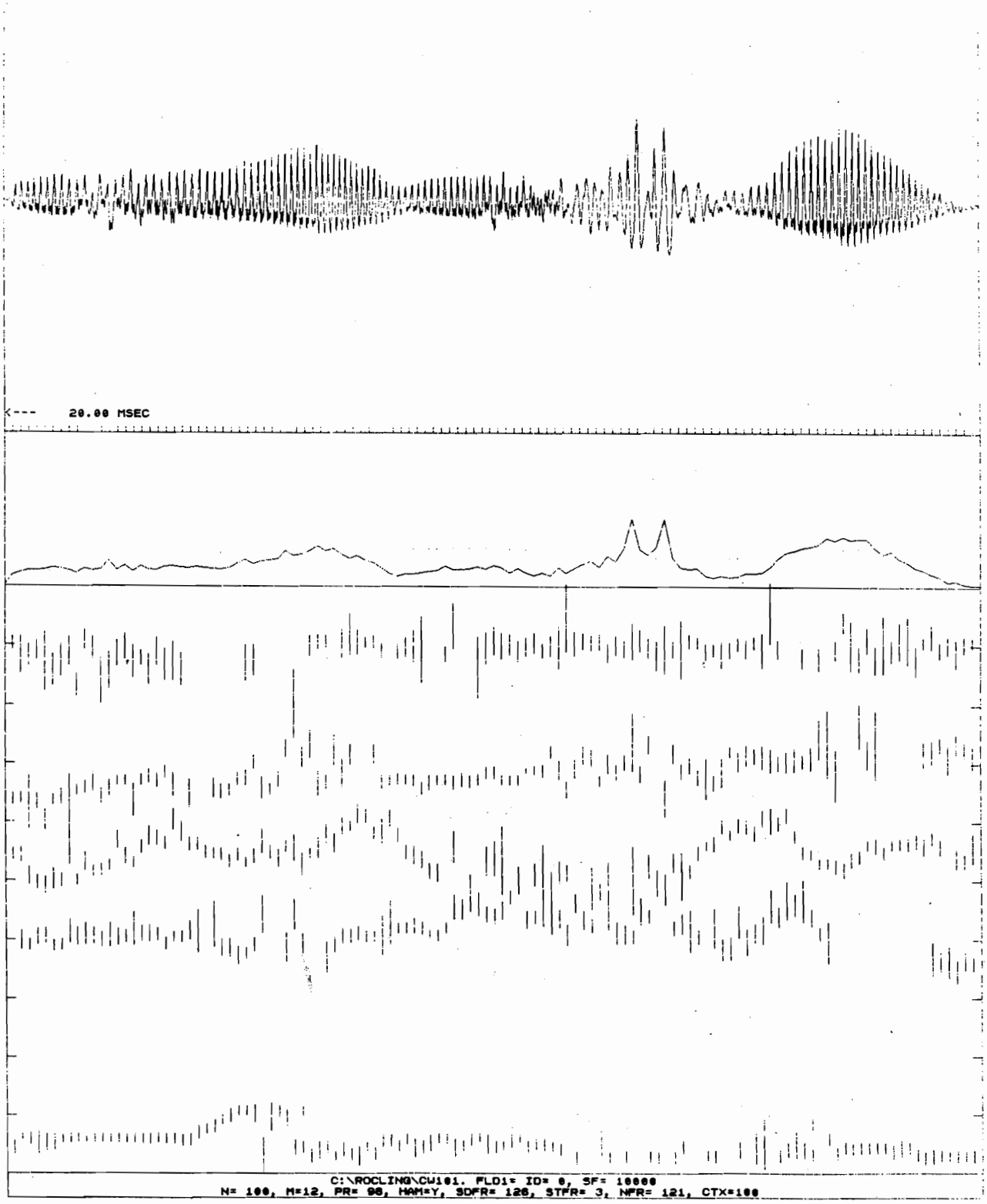


圖(1)說明詞彙的相似度跟空間距離的關係，取自 Cavalli-Sforza and Wang, *1986*.

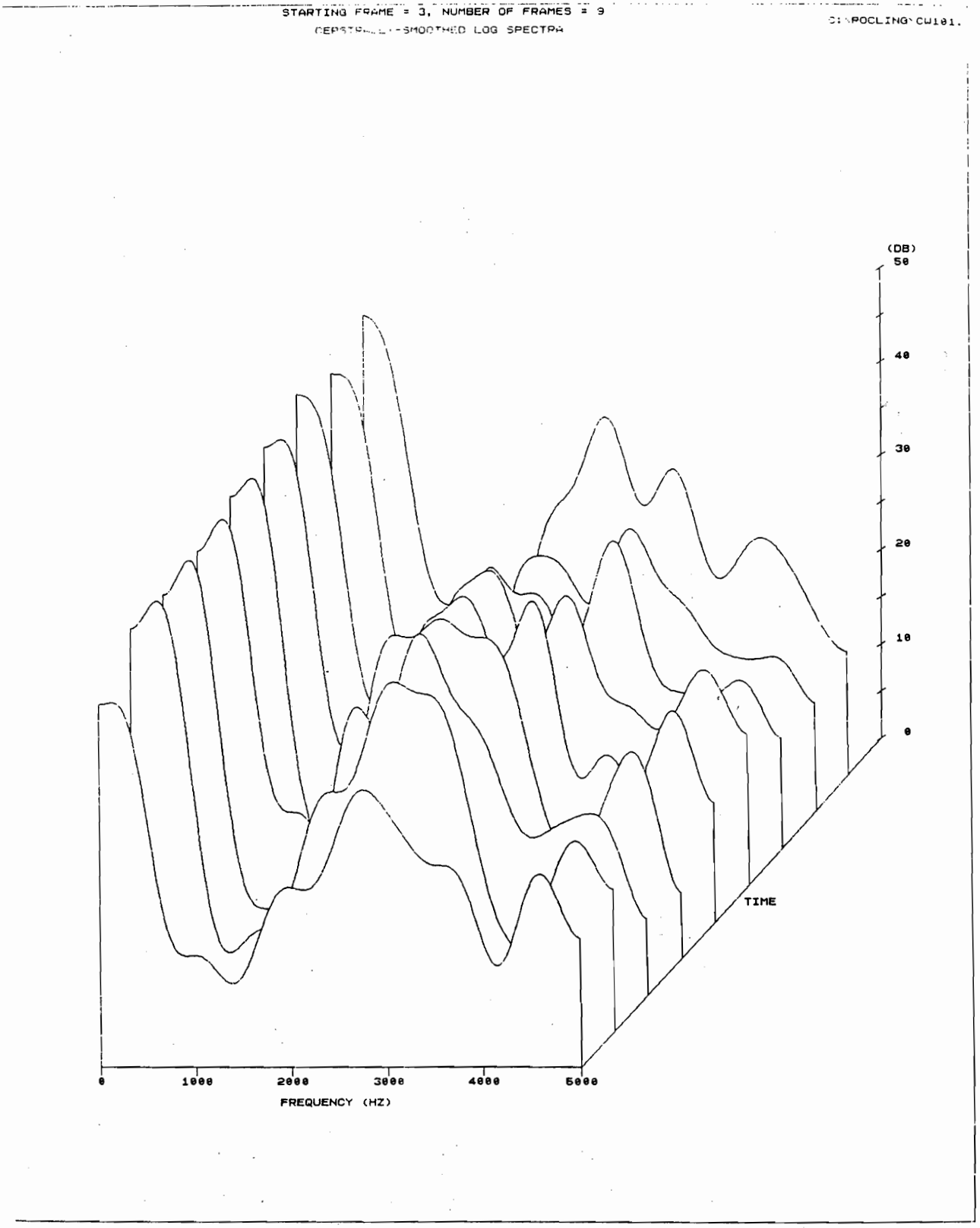




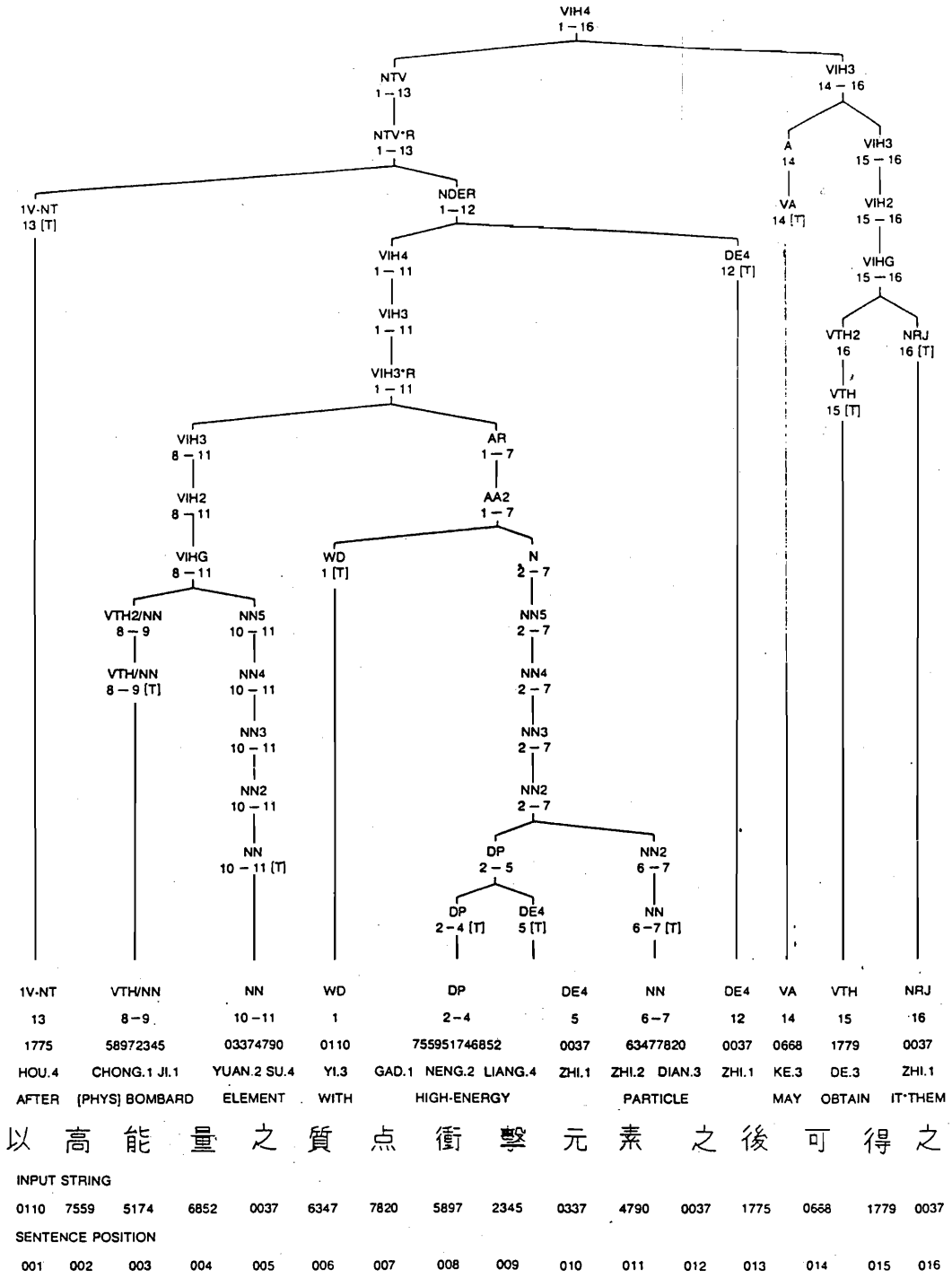
圖(2)是用 DOC 所查出來的聲調分佈，取自 Wang and Cheng, *1987*. DOC 最早討論是Wang, *1970*.



圖(3)上部是作者說"語言與語音"的聲波，中部是這句話的基頻率及振幅，下部是這句話的聲譜圖，平均可以看到五條共振峰.



圖(4)是從"語言與語音"的開頭元音做出來的三維聲譜圖，畫圖(2)及圖(3)的軟體用的是清華大學語言實驗室的 ILS (Interactive Laboratory System).



COMPUTER TRANSLATION of a Chinese sentence from a scientific text produces a reasonably accurate and understandable result. The string of Chinese characters is entered into the computer using a numeric code for each character. The position of each character in the sentence is also entered. The computer searches its memory for the meaning of each character and then performs syntactic analysis of the sentence. In converting the sentence into English the computer makes permutations of the word order. The sentence means: "It may be obtained after bombarding the element with high-energy particles." Research into computer analysis of Chinese is being conducted by the author and his colleagues at the phonology laboratory of the University of California at Berkeley.

圖(5)是分析漢語句子時，電腦自動產生的一個樹圖，包括多種名詞類及動詞類，取自Wang, *1973*.

*(9)*參考資料:

Joseph D. Becker. *1984*. Multilingual word processors. In Language, Writing and the Computer (Wang, *1986)*.

M. Y. Chen and W. S-Y. Wang. *1975*. Sound change: actuation and implementation. Language *51.255-81*.

C. C. Cheng and W. S-Y. Wang. *1971*. Phonological change of Middle Chinese initials. 清華學報 *9.216-70*.

L. L. Cavalli-Sforza and W. S-Y. Wang. *1986*. Spatial distance and lexical replacement. Language *62.38-55*.

J. Kemeny and T. Kurtz. Back to BASIC. Hanover: True Basic Inc.

S. E. Levinson and M. Y. Liberman. *1981*. Speech recognition by computer. In Language, Writing and the Computer (Wang, *1986)*.

C. F Lien *1987*. Coexistent tone systems in Chinese dialects. PhD dissertation, UC Berkeley.

L. R. Rabiner and R. W. Schafer. *1978*. Digital Processing of Speech Signals. Prentice-Hall.

Sneath and R. R. Sokal. *1973*. Principles and Practice of Numerical Classification.
趙鐵橋譯 *1984* 數值分類學 北京:科學出版社.

W. S-Y. Wang. and J. Crawford. 1*960*. Frequency studies of English consonants. Language and Speech *3.3.131-9*.

W. S-Y. Wang. *1970* Project DOC, its methodological basis. Journal of American Oriental Society *90.1.57-66*.

W. S-Y. Wang, S. W. Chan and B. K. T'sou. *1973*. Chinese linguistics and the computer. Linguistics 118.89-117. Also in Proceedings of US -Japan Computer Conference *1972*.
北京:語言學動態*4.19-27*, *1979*.

W. S-Y. Wang. *1973*. The Chinese language. Scientific American, February issue. Reprinted in Language, Writing and the Computer (Wang, *1986*).

W. S-Y. Wang, ed. *1986*. Language, Writing and the Computer.W. H. Freeman and Company.

W. S-Y. Wang. *1987*. Representing language relationships.pp.*243-56* in Biological Metaphor and Cladistic Classification.(Hoenigswald and Wiener, editors) University of Pennsylvania Press.
武漢:中南民族學院學報*3.106-112, 1985*.

W. S-Y. Wang and C. C. Cheng. *1987*. Middle Chinese tones in modern dialects. In Honor of Ilse Lehiste pp.*513-523*. Foris Publications.

李壬癸. *1986-7*. 臺灣土著語言資料自動化. 漢學研究通訊 *5.165-167*, *6.144-147*.

謝清俊. *1986*. 中文資訊處理的現況與展望. 中央研究院計算中心.

聯合報. *1988*. 中文電腦人才荒. 十月六日星期四第十一版.

趙元任. *1959*. 語言問題.臺北商務書局.

周振鶴, 游汝杰. *1987*. 方言與中國文化. 上海:人民出版社.

英漢雙解計算機軟體辭典. *1987*. 臺南:金川出版社.

馬大猷. *1987*. 語言信息和語言通信. 上海:知識出版社.