试谈利用电子计算机自动编制中文著作字索引

黄俊杰　陈恩泉　张　普

在中文系、计算机科学系党委和校科研处的领导下，我们尝试利用电子计算机自动编制中文著作逐字索引，并于一九七九年四月十九日取得初步成果。此项工作还具体得到中文系李格非教授的有力支持及数学系七〇九机站全体同志的密切配合，在此一并表示衷心感谢。

下面将我们的工作作一简要介绍。我们所以有勇气将这一尝试性的工作公之于众，见笑大方，并非因为它的编码系统和程序系统已经十分完善和成熟，而是诚恳希望借此机会得到有关领导和同志们的批评指正，以使我们的工作能进一步向前推进。我们愿与一切致力于我国语言研究工作现代化的同志共勉。

一、概　　述

二十世纪四十年代电子计算机的出现及随之而来的广泛应用，使许多学科发生了历史性的变革。特别是这一现代科技手段与文科的结合，使社会科学与自然科学相互靠近，产生了一系列新兴的边缘学科。

在语言学领域中，已经可以利用电子计算机翻译文字、制作索引、分析句法、合成语音、编纂辞典、教学语言等。这就构成了以计算机为主要工具研究语言的一门新兴的边缘学科——计算语言学。要探索语言研究工作的现代化，改变我国在语言研究方面的落后局面，不可不对计算语言学给予足够的重视。

利用计算机自动编制索引是计算语言学的一个部分。国外，利用计算机编制拼音文字的词索引和重要用语索引早已实现。国内，社会科学院语言研究所机器翻译组与计算技术研究所合作，也已于一九七六年实现了英、德等几种拼音文字的词索引的自动编制。然而，数量庞大、笔画繁多、结构复杂的汉字却使中文著作字索引的自动编制工作具有特殊的困难。方块汉字不便于信息化，这是所有需要中文信息处理系统的科学工作者遇到的共同难题。但是这一特殊困难也是可以克服的，一九七八年，我国正式成立了中国汉字编码研究会，这必将进一步推动汉字编码研究工作深入开展。

利用电子计算机自动编制中文著作字索引这一工作的重要意义是显而易见的。因为，编纂辞典和研究语言都必须以大量的第一手语言资料为基础，否则，就很难说是科学的、实事求是的。写一篇成功的语言学论文或编一部较有价值的字辞典，至少有三分之一以上的时间用于搜集整理资料；一个优秀的语言工作者，一生中至少有三分之一以上的时

间不是用于研究而是用于收集资料。这种单调、枯燥、重复的手工劳动的资料工作方式，我们一直沿用了一两千年。今天，计算机已经可以为我们收集、存贮、分析、比较、统计、检索大量的语言材料了；众所周知，它的工作效率是人所望尘莫及的。

因此，如果我们能利用电子计算机大量编制索引式资料，并提供极为方便、快速、多样的自动检索的话，将可能为语言研究工作的各个领域打开一个新的天地。专人专书的字词研究；汉语基本字、常用字的研究；音韵、语法的研究；断代及比较的研究，甚至各年级语文教科书的编写等都将因之有新的基础。

基于上述认识并考虑到我们的现有条件，我们于一九七八年六月选择了电子计算机自动编制字索引这一比较简单的课题开始进行探索。一九七九年四月十九日，电子计算机将输入的六三年版的《毛主席诗词》一卷自动分类、统计，分三种格式输出打印：

① 按部首（《新华字典》部首）排列的逐字索引；

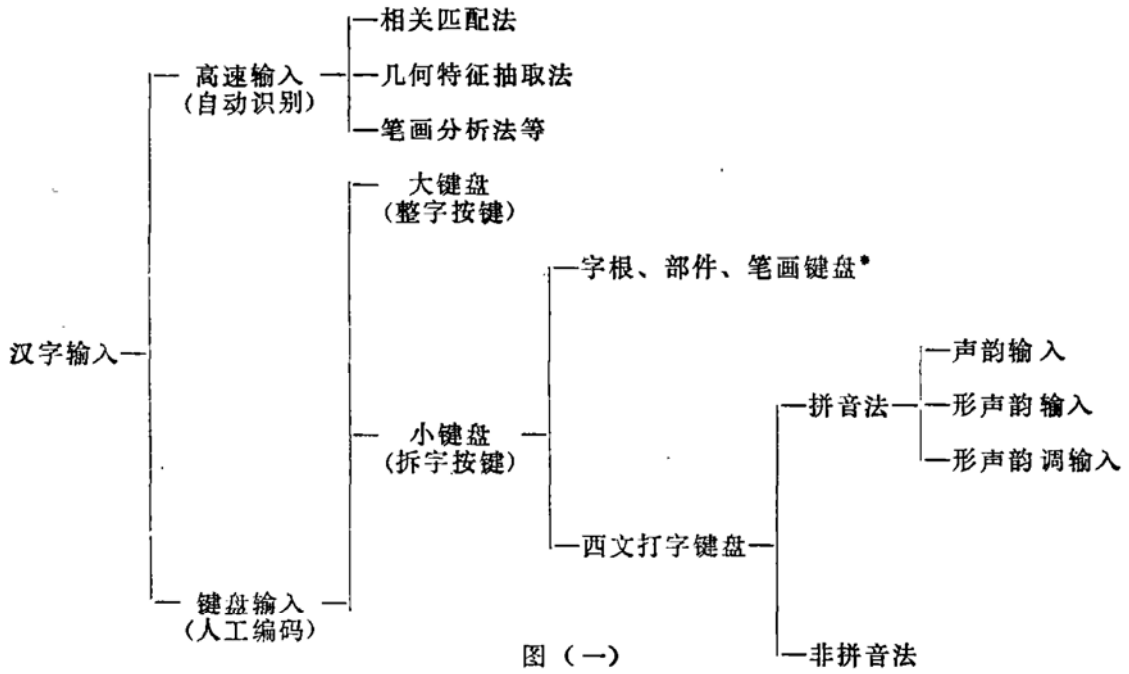
② 按音序（《新华字典》音序）排列的字频统计；

③ 按频度排列的字频统计；

简言之，计算机自动编制中文著作逐字索引就是将中文著作的方块汉字变为目前计算机可以识别的代码（称为数据）输入计算机，然后机器按照预先设计好的一套指令系统（称为程序）将该著作的"数据"按字分类、统计、登记页码、编排输出逐字索引。今后，只需将要制成索引的著作按照相应的编码系统送入计算机，机器即可按这套固定的程序将任何著作按上述三种格式自动编制输出。

二、编码系统

汉字输入电子计算机是目前国内外学者正在攻坚的科研课题。国内外都有不少可行的方法，各种方法都有优缺点。把这些方法综合分类，大致如图（一）所示：



图（一）

* 也有人称此为中键盘。

对电子计算机自动编制中文著作字索引来说，最理想的输入手段是印刷汉字的自动识别。因为用键盘输入，即使作到盲打，最快也只能每分钟输入字符六十个左右。而自动识别最慢每秒钟也可识别数以百计的文字。但在目前一段不太短的时间内，汉字的自动识别还不能很快实现，我们只有采用键盘输入的办法为汉字进行人工编码。

各种键盘输入的编码方案都是为了探求一种汉字输入的简便的操作方法、一字一码的编码技术和较快的输入速度。我们采用的是在汉语拼音之外附加"定字字母"以区别各组同音字的编码办法。这种方法被称为"形—声—韵—调"汉字编码，不过，在我们的编码系统中顺序是"声—韵—调—形"。采用这种以汉语拼音为主的编码原则无论从当前条件还是从长远影响看都是有利的，这一点，中国科学院声学所陈明远同志在《中文信息化与文字改革》(参见 1979．3．25《光明日报》)一文中已谈得很透辟了，这里不再赘述。

"声—韵—调—形"的编码设计，我们参照了欧阳文道同志的《拉丁化汉字编码方案》。这种方案主要是将汉字的编码分为两部分，一部分表示汉字的读音，一部分表示汉字的字形，以区分同音汉字。可以见字识码，也可以见码知字。例如：

he2kds(河)　he2dct(荷)　he2khz(和)　he2kdR(何)

其中 he2 表音，完全与汉语拼音方案相同。-kds、-dct、-khz、-kdR 表形。如果不加表形的"定字符号"，虽然用 1、2、3、4 区分四声，在《新华字典》八千五百字(包括异体，下同。)范围内，汉语一千二百音节中重码字组仍然太多了。在"定字符号"中，后两位，即"-ds、-ct、-hz、-dR"称为大码，表示汉字的部首。-ds 代表"氵"，-ct 代表"艹"，-hz 代表"禾"，-dR 代表"亻"。这样，"河、荷、和、何"四个同音字即可区分开了。但是，上述八千五百字经过大码定字后仍有五百多组字重码①，因此必须也有可能增加另一符号区别这五百多组重码字。这就是上述定字符中开头的一个字母，即"k-、d-、k-、k-"，称为小码或付码。经过小码定字，重码字组由五百多组减少为四十六组，其中常用字与常用字(或较常用字)重码的只有十几组，这就可以进行特殊处理了。

定字符中大码按《新华字典》的一百八十九部首分部。两个字符代表一个部首，使用《汉语拼音方案》字母表中二十一个辅音字母排列组合。为了便于记忆和使用，将一百八十九部首"据类系联"为二十一类。如：以 d 代表动物(兽类)，那么部首"马、牛、羊、鹿、犬、豕"的编码分别取各字声母加 d (下同)，即为"(马)-md、(牛)-nd、(羊)-Yd-、(鹿)-ld、(犬)-qd、(豕)-sd"；以 z 代表植物类，则"麦、豆、谷、禾、竹"等的编码分别为："-mz、-dz、-gz、-hz、-zz"等；以 x 代表形容词类，则"方、大、黑、青、小、赤"等的编码分别为"-fx、-dx、-hx、-qx、-xx、-cx"等，诸如此类。一百八十九部首代码表见图(二)(见下页)。

另外，为了便于记忆和避免按类排列中出现的重码，还采取了两个补充方法：

A、一些部首用传统称呼的声母缩写编码。如："竖心"(忄)-sx、"提手"(扌)-ts、"单人"(亻)-dR、"草头"(艹)-ct、"宝盖"(宀)-bg 等。

B、一些部首用倒码表示。如：R 类代表与人身有关的部首，即"鼻、目、舌、口、手、足、骨、血、齿、耳、皮、毛"等。其中"目、毛"重码，"舌、手"重码。我们将较常用

① 90% 以上的重码字组为两字重码，只有三组字重码在五字以上，最多为八字重码。

的部首用正码表示，而另一个用倒码。即：(目)-mR、(毛)-Rm、(手)-sR、(舌)-Rs。

	b	p	m	f	d	t	n	l	g	k	h	j	q	x	z	c	s	R	y	w	v
b			卜(卜) 贝(貝)						宀	疒				白				鼻	言		比
p	丬	八(丷)			疋(正)	丿	片								勹			皮	步		
m	米		黾(黽)	马(馬)	麻				门(門)		木	皿			麦(麥)	毋(母)	示	目		矛	
f				髟					口				缶	方		父				风(風)	
d	歹				犭	丶			門	亠				大	豆		彳	亻	刂	刀(刂)	斗
t	一								冂		土						扌			田	
n				鸟(鳥)	牛(牜生)							丨				女	纟(糸)				
l	卤(鹵)			龙(龍)	鹿	聿(隶聿)		里	力				耒(耒)	老							立
g	弓		艮(艮)	鬼		攵			工	广					谷		革	骨	瓜	戈	
k					廾					匚								口			
h					灬	虍	一			户	二	火		黑	禾		十				
j	钅(金)				攴	几(几)		斤	己(巳)			金	巾				角	臼	阝(巳)		见(見)
q	犭				犬								其	青					衣	气	欠
x	穴		辛		音	西(西)			心	山		夕		小(⺌)				血		彐(彐旦)	
z	彡	又(又)		隹	辶(辶)	止	乙(一丁乚)		自				舟		竹(⺮)	子(子)	爪(⺥)	足(⻊)			走
c	彳		辰	虫		艹		寸		厂			车(車)	赤	龟	臣		齿(齒)			采
s	石	彡	礻	鼠	豕	山	系	四	身	尸		水(氺)		忄(⺗)	巛	士	厶	手	殳	矢	食
R	阝(左)		毛						人			日				儿	舌	耳	阝(右)		
y	讠(言)		酉	鱼(魚)	羊(羊⺷)			页(頁)	业			月	礻	幺			羽		弋	雨	用
w	韦(韋)												瓦	文		王			攵		
v	尤																食		曰		

图（二）

定字符中，最困难的还不是确定一百八十九部首的代码，而是制定一个完全没有随意性的归部原则，以保证不同的人，对任何字进行归部都能取得一致。为了避免随意归部，必须坚决打破传统使用的“形义兼顾”的归部原则，实行彻底的按形归部(实际目前字词典的部首检字法也日趋按形归部)。一个字有多个部首时，取部首的顺序是先上后下；先左后右；先外后内；先复笔后单笔，与《现代汉语词典》的原则基本一致。例如：

鸿(取氵)　思(取田)　加(取力)　古(取十)　同(取冂)　意(取音)

含有多个部首的字归部要作到一致，首先是字形的切分要统一。现将我们的几种主要的字形切分举例于下：(划线部分为部首)

1. 上下分：


例字：思笛苗岚雾各


例字：想带笃荥架留

2. 左右分：


例字：加江柏休怕灶


例字：颖剁教飘都却

3. 外内分：


例字：国困团回因围


例字：匡巨匹


例字：凶凼画


例字：同闭闹

4. 斜分：


例字：勾句旬


例字：建这武戟徽赶


例字：府虎屑尾疾危


例字：右有友发考希

5. 三分：


例字：辫辩斑班瓣


例字：榔彬


例字：器鬲

6. 特殊切分：


(取山)　坐(取土)　头(取大)　 
(取王)　 
(取木)　 
(取口)

如果一个字按上述切分原则第一次仍然不能切出部首，就继续向下作第二级切分。切分第二级的顺序仍然是先左后右，先上后下，先外后内。如：

籮(取人)　糶(取米)　夥(取田)　蝨(取虫)　疆(取弓)

无法切分，或切分到最后仍无复笔部首的字则取基本笔画部首。如：

之(取丶)　曲(取丨)　不(取一)　飛(取乙)　年(取丿)

取笔画按最高笔取，无最高笔取左笔。这一点与一般字典中的取起笔不同。因为按起笔会使写同一字起笔不同的人判错归部，仍造成随意性。

小码取自除去大码部首的剩余部分，仍按大码的相同原则归部，不过，只取部首代码的第一个字母作小码。如：

怕 pa4bsx　　想 xiam3mxg　　江 jiam1gds

(待　续)