语言文字应用研究丛书

⊙主编＼许嘉璐＼陈章太

中文信息处理

Zhongwen Xinxi Chuli

■傅永和＼著

广东教育出版社

图书在版编目(CIP)数据

中文信息处理/傅永和著 .—广州:广东教育出版社,1999.12
(语言文字应用研究丛书/许嘉璐,陈章太主编)
ISBN 7-5406-4080-4

Ⅰ.中… Ⅱ.傅… Ⅲ.汉字信息处理-研究
Ⅳ.TP391.12

广东教育出版社出版发行
(广州市环市东路水荫路11号)
邮政编码:510075
广东新华发行集团股份有限公司经销
东莞新丰印刷有限公司印刷
(东莞市凤岗镇天堂围区)
850×1168毫米 32开本 10.875印张 270 000字
1999年12月第1版 1999年12月第1次印刷
印数1—5 000册
ISBN 7-5406-4080-4/TP·11
定价:24.00元

总　序

当今世界，科学技术成了推动人类进步和社会发展的最重要的力量。一切学科，包括社会科学中的一些传统学科，都在思考自己如何投身到迅猛向前的科学技术的滚滚洪流中去。我们不能简单化地期望所有的学科都能为当今科技和经济的发展作出人们能明显感觉到的贡献：有的学科旨在探索大千世界深层的奥秘，有的是民族文化传承与积累的延续；同时也应该根据习惯和已有的知识结构，去研究学科内容可能对今后社会发展产生的作用和影响。

事物发展的客观规律是不以人的意志为转移的。当社会发展的速度加快时，特别是社会活动的中心——经济发生根本性变革时，生活就要向科学提出层出不穷、应接不暇的要求。能不能及时地意识到自身对于社会发展的使命，能不能主动地去满足这些要求，也就是主观世界能不能主动适应客观规律，将成为学科发展的决定性因素之一。

世界上的科学，经历了从综合到分析、从几门囊括古今天下事的学问分化为许许多多学科的过程。这个发展趋势符合人类对事物认识逐步深化的规律。如果从文艺复兴算起，这一过程已经进行了几百年，人类由此所获得的益

处几乎无法计算。现在人类之所以能“上天入地”、“说东道西”，宏观宏到宇宙的形成、反物质的存在，微观微到质子、基因，就是得益于学科的分化和分析的方法。但是学术越发展，人们越是感觉到分化和分析的局限性，无论是从宏观的角度，还是从微观的角度，越来越需要不同学科的综合。其实，综合与分析是不可分离的。就在学术朝着分化和分析的路上大踏步向前走的时候，也并没有完全抛掉综合，只不过不够自觉，不是主要的趋向，影响没有像现在这样无所不在罢了。学术的综合发展，已经并将继续促成许多新的学科的出现，人类从中所获得的好处将更是不可估量的。

与社会发展各个方面的关系越来越紧密和综合成为发展的主要趋势，当代学术的这两个特征，向学术界提出了挑战。在我国，过去几十年中所形成的一些思考问题的惯性和知识结构看来需要变一变了。

语言文字学历来被认为是基础科学，但是近年来其中的一些分支学科的应用性越来越受到人们的重视。现实生活的深刻变化，特别是信息技术的发展、社会的信息化进程，向语言文字学提出了解决种种问题的急迫要求，如一股巨大的浪潮，来势凶猛，出人意料。毋庸讳言，我国的科学技术界和决策层对此思想准备不足，语言文字学界则更是缺乏认识的准备，同时人才结构也不适应这一划时代的新需求。例如，在我国近千所有文科，特别是有中国语言文学学科的高等学校里，能够开设应用语言学课程的很少，设置相关教研单位、招收应用语言学研究生的就更少了。

应用语言学的研究范围，最初只限于对第二语言教学的研究。现在在我国，计算语言学、语言在特定行业中的使用（例如法律语言、广告语言、广播电视语言等）以及母语

(第一语言)教学,也包括在应用语言学之内。语言学界称前一种含义为狭义的应用语言学,称后一种为广义的应用语言学。1997年,原国家教委在修改学科目录时,将"应用语言学"从原二级学科"外国语言学"下的三级学科,改为一级学科"中国语言文学"下的"语言学和应用语言学"("外国语言学"则与"比较语言学"连在一起)。这一举措包含两种意义:应用语言学正式成为我国高等学校普遍培养多层次人才的一个学科,而不仅限于外国语言教学和研究单位;"应用语言学"的广义性得到了教育部门的正式认可。无疑地,这将对我国应用语言学的建设和整个语言文字学界的积极变革起到很大的促进作用。

为了加快我国应用语言学的建设,近几年来国家语言文字工作委员会支持筹建了中国应用语言学会,并举行了两次全国性的学术研讨会,组织了面向计算机信息处理的现代汉语词汇研究,开展了全国小学识字教学研究和交流,支持在应用语言学理论和法律语言、广告语言、广播电视语言以及语言与文化等方面的研究,加快了国家级大型现代汉语语料库的建设、有关语言文字规范和信息处理方面的规范标准的研制,开展全国性语言文字使用状况的调查。为了配合上述这些工作,配合应用语言学研究的逐步加强,我们编辑了这套《语言文字应用研究丛书》。

最初计划这套丛书共11种,除现在和读者见面的8种外,还有黄曾阳的《概念网络体系理论(HNC)》、于根元的《广告语言研究》和陈章太等的《叫卖语言研究》。黄先生因要集中精力进行该理论在计算机上的进一步完善和尽快使之产品化,所以为本丛书写作的事搁浅;于先生当时有广告语言方面的书问世,认为现在再写一本理论性著作的条件尚未具备,建议从缓;陈章太则因身体一时不适,不能

按时定稿。但是我们觉得，现在的这8种，还是基本上反映了当前我国应用语言学研究的现状和主要问题。当然我们仍然期待着黄、于、陈等人的著作能够早日出版。

我们希望这套丛书能够引起语言文字学界朋友们的兴趣，并从中得到一点信息，引发不同意见，一些问题可以由此而得以展开讨论。我们特别希望年轻学子能从中了解到应用语言学发展的必然性和到目前为止所取得的成绩、存在的不足，如果由此而激发了一些人对应用语言学的关心和从事这一“行业”的意向，那就更好了。

丛书中的每一种，都由作者先拟出提纲，我们阅后提出意见；书稿完成后，经我们通读，提出修改意见，再由作者修改定稿。即使如此，书中不够完善之处肯定还有，敬祈读者指正。至于文字风格等完全属于个人的习惯，则不强求一律。

在筹划和出版这套丛书过程中，一直得到广东教育出版社的大力支持，特别是社长黄尚立同志，亲自过问，给予关怀；责任编辑之一的曾大力同志具体策划、编辑，花了很多精力。在此一并表示由衷的感谢。

许嘉璐　陈章太

1997年12月1日

导 言

1642年，法国的帕斯卡尔(B.Pascal)用齿轮等机械装置制造了第一台能做加法和减法的计算器，用来计算税收。1694年，德国的数学家莱布尼兹(Leibunitz)改进了帕斯卡尔的设计，制造了可以进行加法、减法、乘法、除法四则运算的计算器。1946年，美国在艾克利(Eckert)和毛彻莱(Mauchley)领导下，研制成第一台数字电子计算机，叫埃尼阿克(ENIAC)。该机使用了18000个电子管、7000个电阻、10000个电容和6000个继电器，计算机的体积约3000立方英尺，重30吨，耗电150千瓦，占用机房面积170平方米，每秒可做5000次加法、500次乘法、50次除法。该机的缺点是：①没有存储器；②程序本身是用线路连接的方式来实现的，每当改换题目、更换算题程序时，需要重新连线，影响了计算速度。尽管第一台计算机有上述缺点，但仍被认为是现代电子计算机的始祖。1946年7月，冯·诺依曼提出了一份设计报告《初步探讨电子计算机的逻辑结构》，报告中提出了以下主要观点：

①存储器顺序编址，按地址访问每个编址单元；

②数据和指令都采用二进制，存放在存储器中，指令按执行顺序存储；

③计算机以运算器为中心。

目　录

导 言

1642年，法国的帕斯卡尔(B.Pascal)用齿轮等机械装置制造了第一台能做加法和减法的计算器，用来计算税收。1694年，德国的数学家莱布尼兹(Leibunitz)改进了帕斯卡尔的设计，制造了可以进行加法、减法、乘法、除法四则运算的计算器。1946年，美国在艾克利(Eckert)和毛彻莱(Mauchley)领导下，研制成第一台数字电子计算机，叫埃尼阿克(ENIAC)。该机使用了18000个电子管、7000个电阻、10000个电容和6000个继电器，计算机的体积约3000立方英尺，重30吨，耗电150千瓦，占用机房面积170平方米，每秒可做5000次加法、500次乘法、50次除法。该机的缺点是：①没有存储器；②程序本身是用线路连接的方式来实现的，每当改换题目、更换算题程序时，需要重新连线，影响了计算速度。尽管第一台计算机有上述缺点，但仍被认为是现代电子计算机的始祖。1946年7月，冯·诺依曼提出了一份设计报告《初步探讨电子计算机的逻辑结构》，报告中提出了以下主要观点：

①存储器顺序编址，按地址访问每个编址单元；

②数据和指令都采用二进制，存放在存储器中，指令按执行顺序存储；

③计算机以运算器为中心。

由于冯·诺依曼提出了电子计算机的一般原理，因此，他被称为现代电子计算机的奠基人。

从1946年第一台电子计算机问世到现在，电子计算机随着电子器件的发展，经历了四次重大的技术换代：

第一代(从1946年至50年代末)：以电子管为主要器件。主要应用于科学计算，为热核武器、核潜艇、洲际导弹、喷气式飞机的研制作出了重要贡献。

第二代(从50年代末至60年代前半期)：以晶体管为主要器件。应用重点开始从科技领域转向经济管理领域，使计算机不仅是科学实验的有力工具，而且逐步成为经济生活中不可缺少的协助者。

第三代(从60年代后半期至70年代前半期)：以中小规模集成电路为主要器件。计算机的外部设备开始多样化，图形显示和终端设备得以改进，方便了人机联系；计算机体系开始通用化、系列化，促进了小型机的发展。

第四代(从70年代后半期至现在)：以大规模集成电路为主要器件。微型机迅速发展，带来了计算机的大普及，计算机功能扩大。

目前，计算机技术正向第五代发展。新的一代计算机将模拟人类大脑的活动，提高学习、推理、判断的能力。在人机联系方面将更接近于自然方式，使计算机具有更强的语音识别、图像识别、自然语言理解等功能。

计算机由输入设备、存储器、运算器、输出设备、控制器等组成。输入设备的工作是把各种数据和程序送到计算机的存储器，常用的输入设备有键盘打字机、光电输入机、卡片输入机等。存储器相当于人大脑中的记忆部分，它的作用是把输入设备传来的各种数据和程序存放起来，以备调用。常用的主存储器有磁心存储器和半导体集成电路存储器，外存储器一般指磁盘、磁带等。运算器是计算机的核心部分，

专管计算，其运算速度极高。控制器负责输入设备、存储器、运算器和输出设备相互间的协调工作，就像人的大脑支配人的整体活动一样。输出设备的工作是把计算机运算的结果告诉计算机的使用者，显示器、打印机和绘图仪是常用的输出设备。

计算机的应用按其性质可分为以下几大类：

①科学计算，如机械设计、工程结构设计、系统工程设计、各种数值计算、气象预报、科学实验，以及对基本粒子、宇宙空间、动力学、气象学、天文学进行研究等。

②自动控制，如程序控制、信号控制、交通流控制、危险作业控制、仪器仪表控制、生产过程控制、顺序控制，以及对卫星、导弹、宇宙飞船等发射过程进行实时控制等。

③服务系统，如财会账务、商业统计计算、事物管理、自动售货、票据文件、火车座席预约、飞机座席预约等。

④通信控制，如信息交换、电子邮政、电视传信、专用数据网络、公用数据网络等。

⑤管理系统，如决策计划、财务会计、统计报表、劳动工资、物资设备管理、医疗保健管理、档案管理、新闻出版管理、图书文献情报检索、生产管理、科研管理、商务管理、金融银行管理、企业事业管理、交通运输管理等。

⑥计算机辅助设计，如飞机设计、船舶设计、半导体集成电路设计、大型自动系统设计等。

⑦人工智能，如密码破译、自然语言理解、语言翻译等。

计算机的上述应用，概括起来说，属计算机的数值计算和非数值运算。计算机的非数值运算就是信息处理。要使计算机处理语言文字，除使用 26 个拉丁字母拼写的文字可以利用通用键盘直接键入外，与 26 个拉丁字母有差异的其他拼音文字(有的比 26 个拉丁字母多几个字母，有的多十多个

字母，有的则使用非拉丁字母)以及表意文字(如一些民族文字)，则存在一个共同任务，即要教会计算机认识它们。只有教会计算机能认识民族文字，才能输入计算机。

至于如何把笔画繁多、结构复杂、数量庞大的汉字输入计算机，则是汉语信息处理必须首先解决的问题。

第一章 I

汉字的输入

汉字输入计算机的途径有两种：一是通过计算机的键盘人工键入；一是让计算机自动识别。计算机自动识别又分为两类：一是汉字的自动识别，即使计算机把手写体汉字或印刷体汉字通过光电阅读装置(光学字符阅读器)，利用光电扫描方法把一个个汉字识别出来；一是汉语的语音自动识别，即使计算机利用人们给它配备的“听觉器官”，自动识别汉语语音要素，从不同的音节中找出不同的汉字，或从相同的音节中判断出不同的汉字。其中，汉字键盘输入不仅现在是将汉字信息输入计算机的主要手段，即使在汉字识别输入和汉语语音识别输入达到完全实用的阶段，它也是不可替代的。在相当长的一段时间内，汉字键盘输入技术仍将具有生命力。

第一节　汉字键盘输入

在信息处理中，英文是以52个大小写字母作为文字信息处理的单位，采用通用键盘输入十分容易。而汉字是以一个整字作为文字信息处理的单位，加之汉字是一种二维结构的平面图形，其字形结构复杂且数量大，用键盘输入汉字复杂而艰难。

汉字键盘输入方法分成两大类：一类是整字输入方法；一类是编码输入方法。整字输入方法在中文信息处理技术的早期使用较多，目前使用范围较小。整字输入方法必须使用特殊的尺寸较大的汉字键盘，其造价较高，不能实现盲打。为了使大键盘面积尽可能缩小，便衍生出了笔触式(变击键为笔触)、滚筒式(变平面键盘为滚筒键盘)、翻页式(变一盘字为多盘字)、主键辅键式(变一键一字为主辅两键多字)等大键盘。20世纪70年代末以来，汉字编码输入方法发展很快，它多采用通用键盘作为输入工具。采用通用键盘作为输入工具造价较低，通用性强，易于实现盲打。

一、汉字整字输入方法

整字输入就是不用拆分汉字，一个键元表示一个汉字。这种输入方法的最大优点是：直观性强，不用记忆更多的操作规则，易于学习和掌握，也没有重码。它的缺点是：如果汉字量很大，而一个汉字又占用一个键元，则其键盘的盘面就得做得很大，不仅造价高，而且操作也很不方便，输入速率低。为避免上述缺点，一般只在盘面收入使用频率高的汉

字，称为盘内字。其余的汉字则不占用盘面位置，只另外编出汉字的代码表，需要键入时，则直接键入对应汉字的代码。这部分字称为盘外字。

目前，在一些应用系统中，笔触式大键盘尚有一定使用范围，如命令指挥系统、情报检索系统、数据库管理系统、物资仓库管理系统、中文排版印刷系统等。

二、汉字编码输入方法

汉字编码输入方法是运用某种编码方案、键盘设备及计算机资源，由操作者向计算机输入汉字的方法。要弄清这一问题，首先必须了解信息的编码。

所谓编码，就是用少量简单基本的符号表示大量复杂多样的信息。如用阿拉伯数码表示复杂多变的数字；用 26 个拉丁字母表示丰富多彩的英语词汇；用 12 个音阶表示优美动听的乐曲等，都是编码的例子。

一切信息编码都具有两个要素：一是基本符号的种类；一是基本符号相互组合的规则。在上述的实例中，0～9 是表示数字的 10 个基本符号，其编码规则就是十进制的“逢十进位”；A～Z 是表示英语词汇的 26 个基本符号，它们的编码规则就是英语的语法和语义学；0～7 是表示乐曲的基本符号。此外，还有一些专门的符号，它们的编码规则就是写谱规则。

当基本符号数量太多、种类特别复杂时，往往需要使用数量较少、种类简单的基本符号来表示，这种情况称为多重编码。汉字编码就属于这种情况。

几万个汉字是表示汉语的基本符号，但基本符号实在是太多了，而且汉字的笔画繁多，字形结构也较复杂，必须进行再编码，即用更为简单的、而且数量少的基本符号来表示