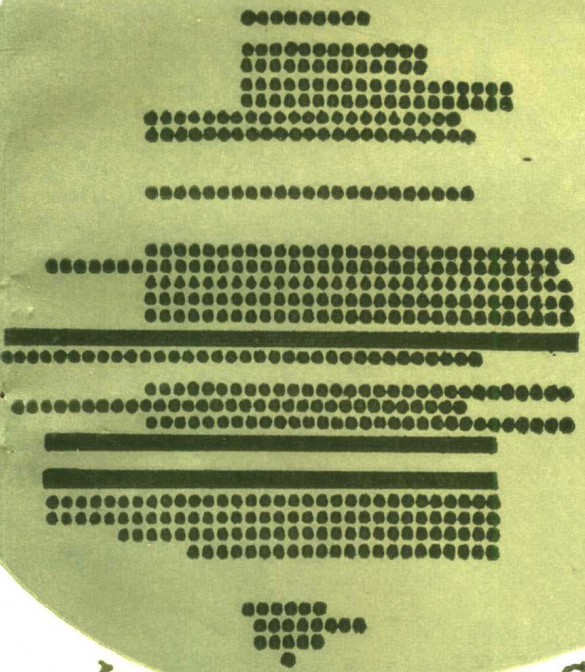
计算机科学丛书

汉字信息处理

陈增武　金连甫　编著



JISUANJI KEXUE CONGSHU

贵州人民出版社

计算机科学丛书

汉字信息处理

陈墉武　华建南　编著

贵州人民出版社

封面设计　石俊生

汉字信息处理

陈增武　金连甫　编著

贵州人民出版社出版发行

（贵阳市延安中路9号）

贵州新华印刷厂印刷　　贵州省新华书店经销

787×1092毫米　32开本　8.375印张　162千字

1988年5月1版　1988年5月第1次印刷

印数1—4,120

ISBN7-221-00097-2

N·02　　定价：2.60元

编 者 的 话

为了加速发展我国的计算机科学技术，在贵州人民出版社的大力支持和协助下，中国科学院软件研究所、复旦大学、吉林大学、浙江大学、武汉大学、南京大学、上海交通大学、山东大学、哈尔滨科技大学、西北电讯工程学院、贵州大学等有关方面的同志经过多次磋商，组成了《计算机科学丛书》编委会。

这套《丛书》的作者，大多是长期从事计算机科学技术方面的科研、教学工作并在近几年内出国考察或学习过的中年同志。他们既有丰富的实践经验，又对国内外计算机科学的进展有比较清楚的了解。《丛书》将向读者介绍现代计算机科学方面的进展及其理论、方法和应用知识，每本书的内容也都自成体系，独立成册，集中介绍一个专题。为了便于学习，部分书后还列有少量习题，可供读者练习。在写作上，《丛书》力求做到篇幅短，内容新，重点突出，适于读者自学，并使读者在较短时间内对每一个专题的动向和发展趋势得到较为完整的了解。

这套《丛书》可作大专院校有关学科的教材和参考书。

《丛书》以大学生、研究生为主要读者对象，也可供大专院校教师、科研工作者和计算机工作者参考。

我们相信，这套《丛书》的出版，将对广大读者了解和掌握计算机科学知识有所裨益。

《计算机科学丛书》

编　委　会

一九八六年一月

目　录

引　言

汉字信息处理是一门在近十多年来蓬勃发展的综合性技术学科，它和计算机的发展有着密切的联系。在我国，计算机产业的发展方针是面向应用、“以用立业”。为了将计算机广泛应用于事务处理、计划调度、办公自动化、情报检索以至印刷排版等方面，汉字系统是必不可少的。这种客观要求以及计算机和电子技术的进步，强烈地推动着它向广度和深度方面进军，出现了一批高水平的理论研究成果和有实用价值的汉字系统。

回顾汉字信息研究工作的发展，可以看到，它几乎和我国的电子计算机的历史一样长。早在五十年代末，我国研制第一台 104 大型电子计算机，就在机上进行了俄汉机器翻译工作。限于当时的技术条件，计算机的存储量有限，只能容纳少量汉字信息，汉字的输入输出技术也还没有解决，只能将中文译文用拼音字母打印输出。这是中文信息和计算机的最早的结合。

六十年代前期，因存储容量的限制，使汉字信息的研究

未得到大的进展。直至六十年代后期至七十年代，由于集成电路和磁盘存储器的出现，使汉字库的存储成为可能，给汉字进入计算机提供了物质条件。

在这一时期，国内外不少学者，对汉字的输入编码的研究百花齐放，专业或业余做了大量工作，并设计了多种输入键盘。

在打印技术上，也有很大进展。最早使用的针式打印机在精度和速度上都有较大提高。还使用了静电、喷墨、激光等新技术。每个汉字的点阵合成，已从初始的20×20以下，进到32×32直至100×100以上。精度高的字，它的美观清晰程度比一般铅印出来的字还好。近几年来，还根据用户的不同需要，做了信息压缩方面的工作，某些方面已取得出色的成果。

因此，无论从技术还是从经济角度来看，汉字进入计算机的条件已经成熟。但是，要做的事情仍然很多。这项工作涉及到计算机科学技术、汉字语言文字学和心理学等许多学科，是一门综合性的科学技术，必须依靠各个方面和部门的通力合作，才能达到真正实用的阶段。

那么."汉字信息处理"的确切含义是什么呢？众所周知，中文是中国的语言文字，特指汉族的语言文字。汉字信息处理包括对汉语书面形式和口语形式两种信息的处理，而不限于汉字的形式处理。

汉字信息处理的研究结果一般都形成为各种各样的系统。例如，属于汉语书面形式的系统有：各种汉字信息处理

系统；编辑排版系统；中文（文献）情报检索系统；外文汉字或汉字外文书面机器翻译系统；汉字书面语言理解系统；计算机辅助汉字教学系统以及汉字数据处理语言和汉字数据库等等。属于汉语口语形式的系统有：汉语语音识别系统；汉语合成系统；汉语口语问答系统；各种通讯系统；外汉或汉外口语翻译系统等等。还有介于二者之间的，为供盲人使用的自动阅读发声系统，供聋哑人使用的自动听写系统，或其它各种类型的人机对话系统等等。

近几年来，由于微型机的迅速发展，大规模集成电路的集成度迅速提高，价格大幅度下降，因此国内外研制成功了多种类型机汉字信息处理系统。它们目前大多数单独使用，也有少数已与（或正在配接）中小型计算机联接，作为这些机器的汉字信息处理终端机。这些微机汉字信息系统各有特色，在功能上互有差异，但其硬件的组成大致相仿，由汉字输入设备、中央处理机、磁盘存储器和汉字输出设备等构成，如图 1 所示。

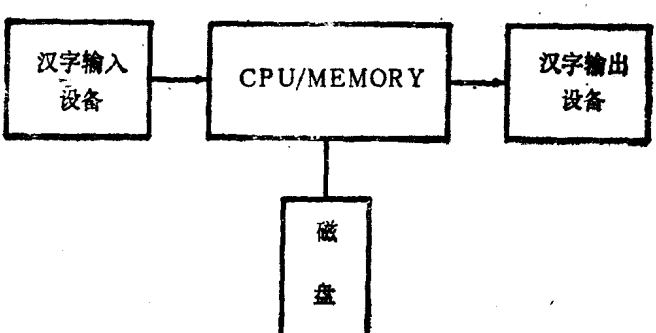


图 1

骤看起来，汉字信息处理系统与西文信息处理系统在组成上并无多大差异，但实际上，由于汉字是图形文字，由字组词，西文是拼音文字，词由字母构成，因此在输入输出方面汉字要复杂得多。汉字输入可分为键盘输入方式和自然输入方式两类，如图2所示。自然输入方式涉及到语音和字形的识别问题，难度更大。现在绝大多数系统均采用手工的键盘输入方式。

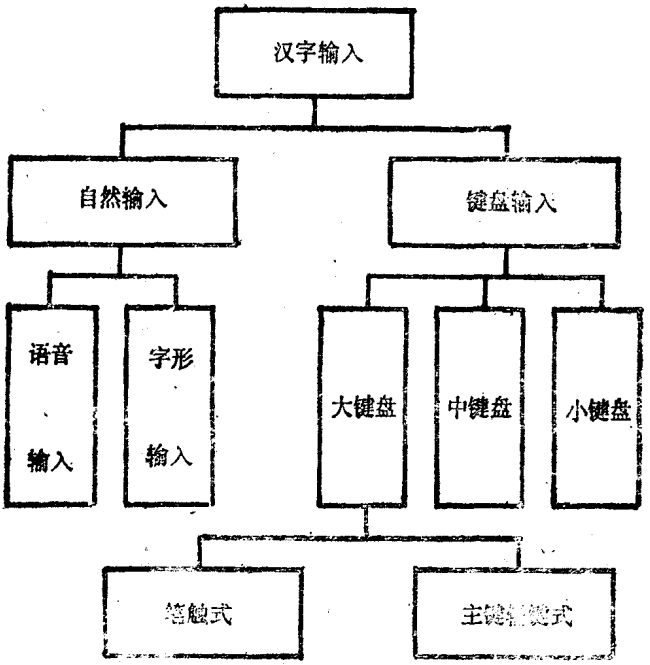


图2　汉字输入分类

在汉字系统研究的早期阶段，人们主要把精力放在汉字编码输入方案、汉字设备的研制上，这当然是一项必不可少

的需重点突破的工作。但随着研究的深入，人们更深刻地认识到，汉字系统的关键是计算机对汉字的数据处理，要从系统的整体上去把握和研究，使系统在处理汉字方面能和处理西文一样有效、一样方便。因此，在发展汉字终端的同时，各种汉字系统软件应运而生，汉字操作系统、汉字高级语言和汉字数据库等的研究正方兴未艾，推动着计算机在管理和其它许多领域中的应用。

在国际上，程序设计自动化已经进入了第三阶段，即对各种高级语言建立起程序设计的支撑环境，使用户能使用所提供的工具和库，方便地进行程序设计工作，提高软件设计的自动化和标准化程度，提高软件的可靠性。在程序设计过程中该环境给用户以友好的支持，使用户进行程序设计感到舒适而得心应手，能把精力集中到问题的逻辑方面，减少形式规定所带来的困扰。鉴于汉字进入计算机所带来的特有的困难，使计算机的处理复杂程度大为增加，更有必要提供全面的支撑，包括从汉字输入开始，到汉字数据处理语言以至输出各个阶段。这是一项基础性的工作，值得组织力量进行研究。

本书旨在综合近年来汉字信息处理的若干侧面的重要成果，向读者作一介绍。为了对汉字进行处理，首先必须了解汉字的特点，这是本书第一章所要叙述的内容。以后各章分述汉字的编码、输入与输出设备、汉字库的构成、汉字和汉语识别、汉字信息处理系统、汉字程序设计语言与支撑环境、汉字数据库，最后介绍一个具体的IBM-PC汉字系统。

由于汉字信息处理发展很快，加上水平和阅历有限，不当之处欢迎批评指正。

第一章　汉字的特点

我国是汉字的发源地，历史悠久，源远流长。如果从殷商时代（公元前十四世纪）的甲骨文算起，迄今已有三千五百年，在世界各国现行文字中，汉字是历史最久的文字。当今世界有近四分之一的人使用汉字。在漫长的历史长河中，汉字几经衍变，字形、读音和字数均几度变化和增减。但是，作为构成汉字的一些基本特征仍然延续下来。这就是：汉字是方块形的图形文字，以字为基本单位，用字组词，形、音、义缺乏有机的联系。因此，汉字和西方的拼音文字有着重大的区别。

要用计算机来处理汉字，第一个问题是如何将汉字输入到计算机内，计算机内以什么方式存放汉字，其实质是汉字的编码与识别问题。这是实现中文信息计算机处理的关键问题之一。只有解决好这个问题，汉字才能高效地、准确地输入计算机，顺利地进行中文信息处理。而为了做到这一点，必须对汉字信息的内容和特点有深入的认识与分析。

汉字信息的内容概括起来主要有四个方面，即：字形、

字音、字义和字频。字形是文字特有的，字音即其记录的语音，字义即表其达的语义。但是汉字的形、音、义三者的关系跟拼音文字不同。拼音文字是字形记录字音，字音表达字义，可以图解为：

字形→字音→字义

而汉字除了这样一条渠道以外，还可以用字形直接表达字义，可以图解为：

拼音文字的同音词（如no和know）也通过不同的字形表示不同的字义，但这究竟是个别的、特殊的现象。而汉字除极个别字外几乎所有的字都有同音字，都要通过不同的字形来表示不同的字义。这是普遍的，一般的规律。

汉字从结构上可分为两大类。第一类是不能从结构上读出音；第二类是可以从结构上读出音，称之为形声结构。汉字中占85%以上的绝大多数字是形声字。

在不带表音成分的结构的汉字中，它们的表意成分有象形、指事、会意三种。象形是用线条来描绘事物的形状，如“日”、“月”、“人”都是象形字。“指事”字用抽象的符号组成，或在象形符号上加上指定性的抽象符号，如“上”、“下”和“出”、“各”就是两种指事字。会意是用两个或两个以上的偏旁组合的字，如“从”、“北”、“明”、“林”等是会意字。上述三种结构方式的共同点是不带表音成分，依靠字形与所表示的词的意义发生关系，因而是纯粹表意的。这

三种结构方式中，象形和指事不带偏旁，会意则用偏旁组成。象形字与指事字是汉字偏旁的主要来源，有时略加变形化出各种新字。

形声字总是由形旁（或称义符）和声旁（或称音符）结合而成。形旁是汉字的表意成分，它主要来源于象形字。声旁是字的表音成分，声旁的来源是象形字、指事字、会意字，也有以形声字为偏旁的。“龚”，“效”，“霖”等都是形声字。

形声字的声旁有三个主要特点：

1.声旁所表示的读音是汉语的音节，所以不能把声旁跟音系字母等同起来；

2.声旁所表示的汉语音节是近似的；

3.声旁有约定俗成的性质，同一个音节往往用不同的声旁来表示。

形旁和声旁的配合主要有六种形式：

左形右声：蟥、胸、河

左声右形：鸽、颈、期

上形下声：草、尘、宇

上声下形：梨、婆、盲

内形外声：问、闻、辫

内声外形：病、近、阁

这六种部位可以概括为左右、上下、内外三种关系，其中以左形右声的形声字最多。

综上所述，汉字的形、音、义三者的关系当中，字形的