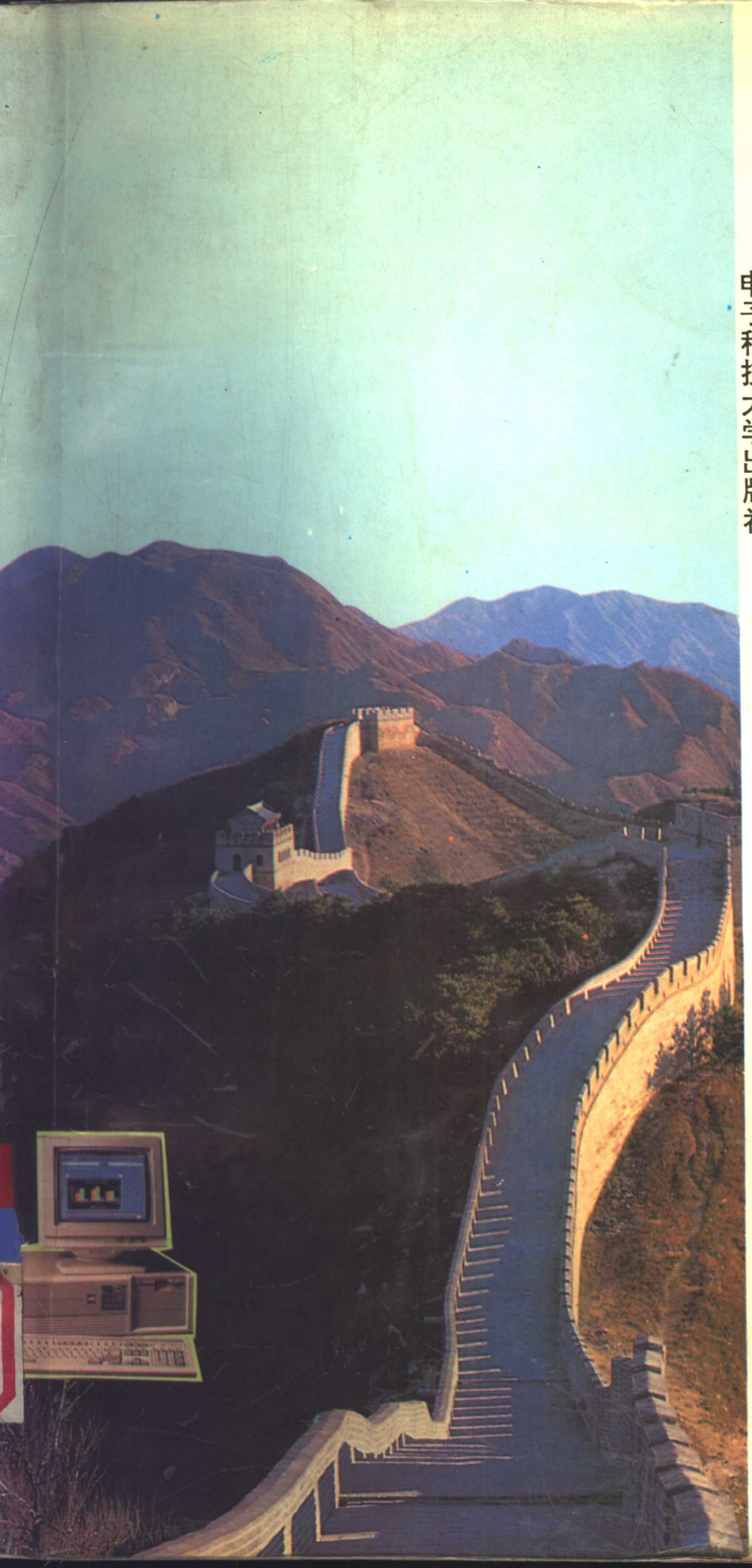

钱培德 编著

计算机中文信息处理技术

电子科技大学出版社

计算机中文信息处理技术

钱培德　编著

电子科技大学出版社

计算机中文信息处理技术
钱培德 编著
*
电子科技大学出版社出版
（中国成都建设北路二段四号）
电子科技大学出版社激光照排中心照排
四川省平武县印刷厂胶印
四川省新华书店经销
*
开本 787×1092 1/16 印张 16.625 字数 400 千字
版次 1992 年 9 月第一版 印次 1992 年 9 月第一次印刷
印数 1-8000 册
中国标准书号 ISBN 7-81016-368-X/TP・28
[川] 016（15452・168） 定价（压膜）7.40 元

序

由苏州大学计算机工程系钱培德教授编著的《计算机中文信息处理技术》一书与大家见面了，我十分高兴地向广大读者推荐这本内容丰富、深入浅出有关中文信息处理技术方面的书籍。我认为，它既是一本大专院校计算机、自动化、电子学等专业适用的教科书，又是一本面向广大计算机及语言文字处理工作者的读物。

众所周知，从60年代初开始，计算机的应用已由数值计算扩大到数据处理。由于计算机性能不断提高，数据处理的范围正在日益扩大，文字信息处理便是数据处理的一种典型形式。文字信息处理的应用十分广泛，在办公室和管理自动化中的事务处理、文字处理和文字翻译、情报资料与图书的自动编目和检索、报刊杂志的自动编辑和排版、计算机辅助教学以及公用咨询服务等领域的各种应用，都与文字处理有着密切的关系。由于语言文字的计算机处理将改变社会的管理、组织形式以及社会的文化教育生活，因此，文字信息处理技术的发展水平将是一个国家和社会走向现代化的一个重要标志。在我国，推广信息处理技术应用的关键在于解决对中文信息的处理。首先它要解决汉字的输入、加工和输出，进而要解决汉语的计算机处理。由于汉语和汉字的固有特点，使得中文信息处理是一门涉及计算机科学、信息科学、语言学、文字学等自然科学和社会科学等多种学科的交叉学科，中文信息处理每前进一步都会遇到和汉语文字有关的许多难点，因此，中文信息处理技术的进步必将大大推动信息科技和汉语文字的发展。我想以上这些就是国家十分重视中文信息处理的原因。

为了在我国推动信息处理技术的发展以及计算机应用的普及，广大从事计算机工作的科技、工程技术人员迫切需要介绍计算机中文信息处理技术方面的书籍。80年代，中文信息处理技术得到了很大的发展，在一些大专院校也开始开设这方面的课程，但至今还不能在大专院校中普遍开设中文信息处理基础、中文信息处理系统的设计与技术以及优化技术的课程。我们应该创造条件，让中文信息处理技术成为计算机专业的一门主要课程。如果这样做，必然会对我国的信息处理事业产生深远的影响。

我深信，本书是能胜任上述任务的。本书写得概念清楚，层次分明，叙述流畅。它以中文操作系统为目标对象来阐述中文系统主要模块的设计，是很有特色的。本书作者几年来一直在信息处理技术的教学和科研第一线工作，积累了丰富的教学和实践经验，使本书较好地做到了理论与实践的结合。愿本书能为我国经济建设和文化建设作出贡献。

我祝贺《计算机中文信息处理技术》一书的出版！

王尔乾

一九九一年八月于清华园

王尔乾教授系国家教委高校理科计算机科学学科教学指导委员会副主任、清华大学计算机科学与技术系主任。——编者

前　言

在我国建设四个现代化的过程中，计算机技术获得了较大的提高，计算机的普及和应用也有了较快的发展。目前，计算机已在我国的各行各业中发挥着巨大的作用。信息处理是计算机应用的一个非常重要的方面，我国是一个使用中文的国家，计算机处理的信息中大部分是中文信息，所以，中文信息处理技术在我国计算机的普及和应用中具有重要地位。国家对计算机中文信息处理十分重视，目前在这一领域中，我国在不少方面处于世界领先地位。

由于中文信息处理在我国计算机科学和技术中的特殊性和重要性，所以，我国许多高等学校的计算机专业自80年代开始，相继开设了《汉字信息处理》课程。该课的传统内容是讲授汉字信息处理的基础知识和基本原理，目前这门课的教材也是按照这一要求来编写的。因此，从这门课的内容来看，是比较抽象和枯燥的，再加上该课的学时不多（多为36学时至54学时），实践环节少，所以学生普遍对这门课缺乏兴趣，学后反映收获欠丰。从以上情况来看，按照传统方法开设的《汉字信息处理》课程，只能使学生了解汉字信息处理的基本原理，缺乏对实际系统设计技术的了解，更缺乏对这方面实现方法的了解。

我们认为，中文信息处理技术是一项能够反映我国特点的计算机技术，所以有必要对《汉字信息处理》课程的内容进行改革，以提高学生对这门课的兴趣，丰富他们的中文信息处理知识和技能。我们提出这门课的教学目标为：要求学生掌握计算机中文信息处理的基础知识和基本原理，在此基础上，要求学生掌握中文信息处理系统的设计技术和实现方法。很明显，这门课的内容不再是抽象和枯燥的了，而是比较实际的技术和方法。这样，就有必要编写一本能符合上述教学目标的新教材。

我们系从1989年开始在计算机应用专业，按照以上新内容开设《中文信息处理技术》课程，取得了较好的效果。虽然这门课开始时列为任意选修课，但是接连几届学生都是全数选修该课。在学习过程中，学生们普遍兴致很高。

笔者于1989年承担了开设《中文信息处理技术》课程的任务，并编写了教材。为了突出技术性，需要选定一个典型的中文信息处理系统作为讲解蓝本，以便具体介绍其设计技术和实现方法。本人选择了中文操作系统作为这个蓝本，因为中文操作系统是中文信息处理系统的基础，它技术全面，内容广泛，而且其技术和方法具有普遍性。另外，《中文信息处理技术》课程一般在四年级开设，在其前面已开设过《操作系统原理》课，故通过对中文操作系统的讲授，能加深学生对操作系统原理的理解。在讲授过程中，本人不断优化和充实了教材内容。

在电子科技大学出版社的支持下，笔者根据讲授《中文信息处理技术》课的体会和经验，在原讲稿的基础上，重新编写了这门课的教材，并取名为《计算机中文信息处理技术》。本书共有十一章，可分为三个部分。第一部分包括第一章至第四章，介绍中文信息处理基础，其主要内容为：中文信息处理技术的发展，中文系统的组成，汉字的属性，汉字代码体系，汉字的输入与输出原理，以及中文系统的用户界面等。这部分内容注重于基础知识和基本原理。第二部分包括第五章至第九章，介绍中文信息处理系统的设计技术和实现方法，其主要内容为：中文系统软件设计基础，系统总体设计技术，系统自举模块的设计，显示输出模块的设

计，键盘输入模块的设计和打印输出模块的设计。这部分内容以中文操作系统为蓝本进行全面的阐述，并注重于基本方法和基本技能，突出技术性和实用性。第三部分包括第十章和第十一章，介绍中文系统的高级设计技术，其主要内容为：VGA 显示卡的程序设计，VGA 显示输出模块的设计，词输入处理技术，联想输入处理技术，汉字库结构的设计，SPOOLing 子系统的设计，以及古文字信息处理系统的设计。这部分内容注重体现我国在中文信息处理系统中使用的先进技术，它也是对第二部分内容的优化和提高。

本书的每章末均附有习题，全书共提供了近 200 道习题。这些习题可以分为三类：一类是基本题，适合于学生作为作业完成；另一类是实验题，适合于学生上机实验用，这类题一般需要编制程序；还有一类是思考题，这类题目具有一定难度，适合于学生课外思考，或者由教师作出提示后，再由学生完成。由于习题数量较多，故教师只要选用其中部分习题即可。

本书是按照 72 学时编写的。根据本人讲授这门课的体会，建议对第一部分内容采用讲授与自学相结合的方法（比如，第四章可以完全由学生自学），有的章节还可以简讲；对第二部分内容则要精讲，并要注意突出重点，还要配上相应的上机实验（实验题可选用习题中的相应题目）；第三部分内容可采用选讲的方法，讲授其中的部分章节，给学生一个启发，其它内容留给学生以后自学。如果按照 54 学时讲授本书的话，则应免讲部分章节的内容。

笔者在写作过程中，得到了清华大学计算机科学与技术系主任王尔乾教授的支持和帮助，他专门为本书撰写了序，谨在此向他表示衷心的感谢。我还要感谢南京大学计算机科学系的王绪龙副教授，他对本人的写作给予许多指导和启发。

复旦大学计算机科学系副主任钱乐秋副教授、上海交通大学计算机科学与工程系副主任候文永副教授和航空航天部 8359 研究所郑智光高级工程师审阅了本书稿，并提出了许多宝贵的意见和有益的建议，在此向他们表示诚挚的谢意。

本书的形成过程中，始终得到了电子科技大学出版社沈凤鸣老师和吴艳玲老师的热情支持和帮助，她们为本书的出版付出了大量的劳动。我向这两位老师和电子科技大学出版社的领导同志表示由衷的感谢。

在写作本书的过程中，广泛参考了国内外的有关资料，从中得到了许多启发，并且汲取了不少营养。在此向这些参考资料的作者深表敬意。

最后，希望使用本书的老师、同学和广大读者，能对本书提出宝贵的意见和建议，以便本书再版时对内容进行修正和充实。

钱培德

于苏州大学计算机工程系

一九九一年八月

目　录

第一章　中文信息处理概论

第一节　绪　　言

一、中文信息处理的必要性

中文是我国的通用文字，严格地讲，中文应包括我国各民族所使用的文字。由于汉字在我国使用的广泛性，汉民族人口众多，在我国具有重要地位，汉字和汉语就自然成了我国的特定通用文字和语言。由此可知，汉字在中文中具有特别主要的地位，因此，在不少场合，中文信息处理就体现为汉字信息处理。

现代社会是充满信息的社会，对信息进行处理和管理是社会的需要。由于社会信息日趋庞大和复杂，如果仍用传统的人工方法来实现对信息的存储、传递和处理，则要花费大量的劳动，而且由于信息繁多和人脑工作的固有特点，往往使这些工作不能达到令人满意的结果。

计算机的问世，特别是微型计算机的大量涌现，使得利用计算机进行信息处理已成为可能，又由于计算机所具有的优点，使得其完全能够胜任这种工作。所以，利用计算机进行社会信息的处理已经势在必行。

计算机最初是由西方国家研制出来的，所以计算机的内部机制对西方文字具有很好的适应性。从而，利用计算机进行西文信息处理是相当方便之事，许多西方国家早就利用计算机在作这件事了。

我国社会中要进行处理的信息主要是中文信息，其中绝大部分是汉字信息。然而，现有计算机的内部机制对汉字不具备较好的适应性，所以，汉字信息的输入、输出与处理，均要比西文信息的相应处理困难得多。当然，这与汉字自身的特点也有关系。显然，必须把汉字引入计算机，才能使计算机在我国获得广泛应用。也就是说，如果不能很好地解决在计算机上进行汉字信息处理这个问题，那么就不可能在我国推广计算机的应用，各行各业也就不可能实现现代化的管理。

随着计算机技术的不断发展，计算机系统的功能也不断增强，计算机的应用领域也在不断拓宽。汉字信息处理的涵义和涉及的范围也大大扩展了，现在已包括情报资料和图书的自动编目与检索；书刊和报纸的自动编辑与排版；事务处理和企业管理；办公自动化与数据通信等。因此，解决计算机的汉字信息处理问题，已到了刻不容缓的时候了。我国是汉字的发源地，对于汉字的研究最深入，因此对汉字结构的特性及使用情况最熟悉。同时，我国对发展计算机汉字信息处理技术的要求最为迫切，得到的收益也最大。所以，我国理应在计算机汉字信息处理领域中走在世界的最前列。

二、中文信息处理技术的发展

早在50年代末期，我国就在国产的104计算机上进行由俄语到汉语自动翻译的研究工作，并研制成了俄汉机器翻译模型样机。从这时起，汉字已开始和计算机结下了不解之缘。这就是我国计算机中文信息处理研究的开端。

到了60年代后期，我国开始对汉字信息处理技术进行进一步的探索和研究，并成功地研

制出了汉字电报译码机。这种机器能以点阵方式在纸上输出汉字字形，为以后大量使用的汉字点阵式打印机提供了基础。

从70年代开始，我国开始系统地研究和开发汉字信息处理技术，在国家有关部、委的支持下，于1974年制订和组织开展了我国第一个大型汉字信息处理工程项目的研究，并定名为“748工程”。这项工程项目包括三个研制任务，它们是：精密型汉字编辑排版系统、汉字情报检索系统、汉字通信系统与汉字终端设备。这三项任务均取得了重大成果，把我国的汉字信息处理水平提高了一大步，并且获得了相当多的技术条件和研制经验。

从70年代末开始，由于大规模集成电路存储器和成套的微处理机芯片进入我国应用领域，因而在很大程度上促进了汉字信息处理技术的发展，不仅使原有的一些技术得到更新，而且研制成了一些新型的汉字输入与输出设备，在技术指标、可靠性和实用性方面，均有极大的提高。从那时开始，我国已能用国内自己研制的汉字设备与计算机配置成多种应用系统，特别是以微处理器为基础的汉字信息处理系统发展更为迅速。

进入80年代以来，我国的汉字信息处理技术更加蓬勃发展，这方面的学术研究和学术交流更加活跃，各种学术团体和组织纷纷成立。1981年成立了中国中文信息学会，由著名的钱伟长教授担任理事长，下面设立了基础理论、汉字信息处理系统、汉字编码、汉字信息处理专用设备、自然语言处理和汉字字形等专业委员会。中国计算机学会也设立了中文信息处理技术专业委员会。这些专业学术团体组织了大量的国内和国际学术交流活动，有力地推动了我国的中文信息处理技术的发展。

在这段时期内，国家对中文信息处理技术极为重视，先后颁布了一系列的中文信息处理标准，有力地支持和推动了这项技术的发展。国家颁布的中文信息处理标准如下：

(1) GB1988-80，“信息交换用的七位编码字符集”。

(2) GB2311-80，“信息处理交换用七位编码字符集的扩充方法”。

(3) GB2312-80，“信息交换用汉字编码字符集（基本集）”。

(4) GB3453-82，“数据通讯基本型控制规程”。

(5) GB3454-82，“数据终端设备（DTE）和数据电路终端设备（DCE）之间的接口电路定义表”。

(6) GB5199·1～5199·2-85，“信息交换用汉字16×16点阵字模集及数据集”。

(7) GB5007·1～5007·2-85，“信息交换用汉字24×24点阵字模集及数据集”。

(8) GB6345·1～6345·2-86，“信息交换用汉字32×32点阵字模集及数据集”。

(9) GB5261-86，“文字和符号图形设备的增补控制功能”。

(10) GB7589-87及7590-87，“信息交换用汉字编码字符集第二辅助集和第四辅助集”。

目前，国内在进行汉字基础理论研究的同时，已制订出了汉字信息处理设备与系统的研制和生产规划。我国的汉字信息处理系统已由试验阶段发展到了成熟阶段。在汉字信息处理系统的配置中，除了提供必要的汉字设备和接口外，最重要的是软件配置，而其中以汉字操作系统最为重要。我国在汉字操作系统的研制和开发方面作了不少工作，一般是对已有的西文操作系统进行扩充和改造，使其成为能处理汉字信息的汉字操作系统。例如，在微型机的CP/M操作系统中加入了处理汉字输入输出的模块，形成了能支持高级语言和应用程序处理汉字信息的新系统。对于VAX系列的小型计算机系统，也已完成了把它的VMS操作系统扩充为具有汉字处理功能的CCVMS操作系统。对于著名的UNIX操作系统，国内也已完成了对其多种版本及变种的汉化工作。IBM-PC微型计算机系列是我国的主流机种，我国成功地开发了

其主操作系统 PC-DOS（MS-DOS）的汉化版本，这就是著名的 CC-DOS 汉字操作系统。

总之，由于汉字信息处理系统的推广应用工作愈益得到政府各部门和各类业务部门的重视，故可以肯定，在今后几年内我国的汉字信息处理技术将会以更快的速度向前发展。

我国是一个多民族国家，许多少数民族也拥有各自的民族语言和文字，少数民族文字信息处理与汉字信息处理具有同样重要的意义，它们是整个中文信息处理的重要组成部分。随着少数民族地区生产和文化的发展，这些地区的社会信息也会越来越庞大和复杂，所以有必要研究和开发少数民族文字信息处理技术。

我国少数民族文字信息处理技术已取得了令人鼓舞的成绩。微型计算机上的蒙文、藏文、维吾尔文、柯尔克孜文、哈萨克文、朝鲜文、壮文和彝文等少数民族文字操作系统已先后问世，在这些操作系统支持下的许多应用系统也已研制成功，并且投入应用。例如，藏医诊断系统、电视台蒙文节目编制与合成系统、彝文激光照排系统等。由此可见，我国的中文信息处理技术已经获得了全面的发展与提高。

三、中文信息处理系统的组成

中文信息处理系统由硬件和软件两大部分组成。硬件包括计算机硬件、字库、输入设备和输出设备。软件包括中文操作系统（系统软件）和应用软件。图 1-1 为中文信息处理系统的组成示意图。

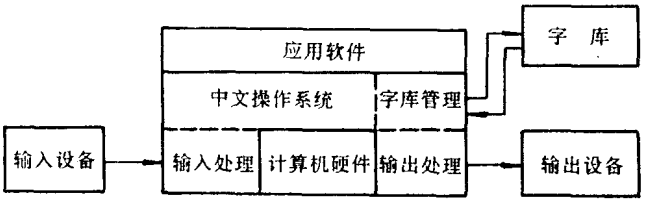

图 1-1 中文信息处理系统的组成

1. 计算机硬件

中文信息处理系统中的计算机硬件一般并无特别之处，通常就是普通的西文计算机硬件部分。这部分主要包括处理器、总线逻辑、内存储器、外存储器（磁盘、磁带）和 I/O 接口等。根据系统的需要，这部分可以选用大、中、小型计算机，也可以选用微型计算机。国内开发的中文信息处理系统大多采用微型计算机。

2. 字库

字库中存放中文字形信息，以实现中文信息的输出。字库可以根据系统的实际情况存放在不同的介质上。如果字库存放于内存或磁盘上，则称为软字库；如果字库固化于 ROM 中，则称为硬字库。软字库一般均放在主机内，而硬字库则可放在主机内，亦可放在输入设备和输出设备中。主机内的字库由操作系统的字库管理模块进行管理。

3. 输入设备

这里说的输入设备是指用于输入中文信息的输入设备，它实现把中文信息变换成中文信息在系统内部的表示形式，送入主机内。目前常用的输入设备有键盘、语音识别器、字形扫描识别器等。键盘通过对中文信息编码，敲击与编码符对应之键，把中文信息输入主机内，这是最廉价的输入设备，也是目前使用最广泛的中文信息输入手段。语音识别器通过接收和识别中文信息的读音，把它们输入主机内，这就是让计算机直接“听懂”人的声音，这种输入设备现在已接近实用阶段了。字形扫描识别器通过对中文信息的字形进行接收和识别，把它

们输入主机内，这就是让计算机直接“识字”。被识的字有印刷体和手写体，对前者的识别已接近实用水平，但对后者的识别尚处于研究阶段。

4. 输出设备

这里说的输出设备是指用于输出中文信息的输出设备，它实现把中文信息在系统内部的表示形式变换成中文信息的字形或语音。目前常用的输出设备有显示器、打印机或语音合成输出器等。显示器能实现把中文信息的字形显示在屏幕上，如果显示器内带有专用的硬字库，则可大大加快显示速度。打印机能实现把中文信息的字形在打印纸上印出，如果打印机内带有专用的硬字库，则可大大加快打印速度。语音合成输出器能合成中文信息的读音，也就是让计算机直接“讲话”，目前这种输出设备已处于研究阶段的后期。

5. 中文操作系统

中文操作系统是中文信息处理系统的系统软件，它是中文信息处理系统的基础，因此它在系统中具有特殊的地位。一台西文计算机配上相应的中文操作系统后，就成为一台中文计算机，由此可见中文操作系统具有多么大的“神通”。中文操作系统通常包含字库管理模块、输入处理模块和输出处理模块，它们分别完成对字库的管理，对输入、输出设备的管理和驱动。中文操作系统向用户提供了一个中文信息处理的界面。例如，它向应用软件提供了输入、输出和处理中文信息的功能。

6. 应用软件

应用软件是指在中文操作系统支持下运行的实用程序和应用程序。根据中文信息处理系统的性质和用途，各种系统都要有相应的应用软件。对于一些典型的应用软件，应提供商品化的应用软件包。由于中文信息处理系统种类繁多，设计各种应用软件的工作量较大，因此有的系统就采用把相应的西文应用软件改造成中文应用软件，以迅速扩大中文信息处理的应用范围。

第二节　汉字的属性

汉字在中文中具有重要地位，欲研究中文信息处理，一定要研究汉字信息处理，也就有必要研究汉字自身。汉字的属性是指汉字所具有的性质和特点，其中最基本的是汉字的字形、字音和字义。对这三个基本属性的研究，有助于对汉字信息输入和输出处理的研究与开发。本节将对汉字的字形、字音和字义分别进行阐述。

一、汉字字形

汉字字形是汉字形体结构的图像，汉字的字形是呈方块形的。在汉字信息处理中，为了获取汉字的字形信息，常常要在不同层次上对方块形的汉字进行分解，以满足不同的处理要求。目前还没有对汉字字形进行分解的统一方法和标准。从目前常用的字形分解方法来看，大体上有单字、形素、笔画和字根四个层次。下面介绍笔画和字根分解法。

1. 笔画分解法

笔画是人们书写汉字的步骤，每次从落笔到提笔，便是一个笔画。一个笔画所形成的轨迹，就是笔形。根据笔画对字形进行分解，一般可有30多种笔形。随着分解方法的不同，笔形的数量也不同。目前最广泛选用的是五种笔形，它们是：横、竖、撇、捺（点）、折。汉字“札”就包含了这五种笔形。据统计，横、竖、撇、捺、折是汉字中使用最多的笔形，它们在汉字中的相对使用频率分别为：横28%，竖18%，撇15%，捺13%，折7%。其余各种笔形