

内部讲义

计算机中文信息处理

第 一 册

赵 珀 璋 编 著

空 军 第 三 研 究 所

北京信息控制研究所

中国科学院研究生院

一九八四年七月

内 容 提 要

本讲义是讲述“计算机中文信息处理”的基本教程。全书共分十三章：一至五章讲述中文信息处理基本概念、现代汉语基础知识、汉字编码、汉字库字形和汉字编码检索；六至九章讲述中西文兼容的操作系统、程序设计语言、数据库和中文信息处理系统；第十章讲述中文信息通信；第十一章讲述汉字和汉语语言识别；第十二章讲述汉语自然语言处理；第十三章介绍有关中文信息处理的国家标准知识。全书约六十万字，分为四册出版，是一本比较系统、全面的讲述计算机中文信息处理的教科书和教学参考书。

本讲义适合于计算机工程与计算机应用等专业作为“计算机中文信息处理”课程的教材或参考书，也可以供从事中文信息处理研究与设计工作的工程技术人员作为参考书。

前 言

本讲义的第一稿是去年应中国科学院研究生院邀请讲授“中文信息处理”课程时编写的教材，印出后得到多方面的鼓励和指数。今年又应中国科学院研究生院和空军第三研究所之邀进行讲课，根据去年的教学实践和各方面提出的宝贵意见，对原稿在内容上作了较大的修改，去掉了一些过时或者在有关文献中容易查到的资料，增加了一些新的材料和实用的内容，同时也提出了一些新的看法，比如从信息编码角度来描述中文信息处理的模型等。这些看法并不成熟，目的在于探讨，请诸位专家、学者和读者提出宝贵意见。

“中文信息处理”是一门新兴的、综合性边缘学科，它涉及到数学、电子学、计算机工程与科学、心理学、语言学等学科。无论是理论方面，还是工程技术方面，都还处于其发展的幼年时期，远不成熟。

本讲义力图反映国内外中文信息处理的最新情况和各种流派思想，但由于各种条件的限制，难免挂一漏百；又因时间仓促，文中引用的材料来不及与原作者一一校对，如有错误，由笔者负责。欢迎原作者和读者提出批评指正。

计算机中文信息处理的流派很多，归纳起来不外乎两大派。一派主张突出中文信息的特点，认为应从系统设计上，包括硬件系统和软件系统都应该中文化，比如操作系统和程序设计语言都汉化，即走民族化的道路；另一派则认为中文信息处理与西文信息处理并无本质区别，完全可以实现中西文兼容处理，主张在原有系统上进行适当修改和扩充，使其满足中文信息处理的要求，走国际兼容道路。笔者赞同后一种观点。但在学术上，愿意参加前种观点的讨论。本讲义在这方面花了相当篇幅介绍中文化操作系统和程序设计语言的版本。

读者可以全面分析提出自己的观点。

目前，中文信息处理的理论研究和实践工作正在蓬勃开展，很多问题尚未解决；至于工程应用更是刚刚开始，还未形成一个完整体系。本讲义所介绍的一些资料难免有片面性和局限性，笔者愿与读者讨论、交流。

在我国发展计算机中文信息处理事业，需要建立许多标准，国家已经公布了一些；有些已完成制定工作，准备公布；有些还在制定中。本讲义力图将标准问题尽可能系统化，并将在最后册附上国家已经公布的标准GB1988、GB2311、GB2312，目的是使广大读者明白，关于汉字信息处理的国家标准是与国际标准ISO相兼容的。这对于从事中文信息处理的工作者，尤其是系统设计师，必须掌握和执行这些标准，并以此来指导自己的研制工作；对已制定尚未正式公布的标准，将尽可能详细介绍之，待国家公布后，如与国标有出入，以国标为准；对于正在制定中的标准，只准备介绍其制定思想和一些资料，供参考。

计算机中文信息处理涉及领域太广，笔者接触此领域的广度和深度都很不够，本讲义无论从材料选择和组织编排上，还有某些学术观点上，都难免存在许多缺点、错误，非常欢迎读者提出宝贵意见。

在第二稿编写过程中，得到了谢希仁教授、朱伯章所长、张佑鸿付总师、茅于杭、徐近需、舒文豪付教授、郭治方主任、王之燧、陈耀星、裘滨良工程师等热情指导和支持，在许多学术观点讨论中，得到了诸位专家、学者的指点，受益非浅，没有他们的指导和帮助，本稿不可能顺利完成，在此一并深表感谢。

在文稿付印过程中，张佑鸿同志承担了对全书的校核和编辑工作。刘文平、刘湘江、张建平、郝卫等同志在讨论问题，整理文稿和校对

中，都做了大量工作。没有他们的帮助，本讲义是不可能完成的。在此深表谢意。

由于笔者水平所限，加之时间仓促，第二稿仍然存在粗、杂、乱现象，笔者准备通过讲课实践再进一步压缩掉一些非必要的内容，进行修改精练提高，欢迎广大读者提出宝贵意见。

来信请寄：北京市 842 信箱，北京信息控制研究所。

赵珀璋

一九八四年六月十一日

总 目 录

第一分册

第一章 绪论

§ 1-1 汉字输入问题

§ 1-2 汉字存储问题

§ 1-3 汉字输出问题

§ 1-4 中文信息处理系统

§ 1-5 自然语言处理

参考文献

第二章 现代汉语基础

§ 2-1 现代汉语的特点

§ 2-2 汉字的形成、特点和结构

§ 2-3 语言

§ 2-4 词汇

参考文献

第三章 汉字输入编码

§ 3-1 关于汉字一维码的统计推断

§ 3-2 海曼公式及其推广与汉字编码时间

§ 3-3 汉字编码、排列法的心理学性质及其应用

§ 3-4 多作业对击键动作的影响

§ 3-5 汉字部件的分析方法和理论

§ 3-6 WBBX 汉字编码方案

§ 3-7 ZN9 号汉字 26 键输入码方案

§ 3-8 天龙中文编码法

§ 3-9 汉字树形结构中文字母编码法

- § 3-10 三角编码法规则
- § 3-11 汉字笔形编码法
- § 3-12 汉字基本笔画编码法
- § 3-13 汉语拼音方案输入法
- § 3-14 汉语普通话新文字电脑Ⅱ型
- § 3-15 “见字识码”编码方案
- § 3-16 声韵部形(UYBX)编码
- § 3-17 拼音中文
- § 3-18 拼音拼写文字
- § 3-19 汉字的联想输入方法
- § 3-20 计算机辅助编码
- § 3-21 汉字输入编码评测试行规则
- § 3-22 汉字输入方案比较法(附参考文献)

第 二 分 册

第四章 汉字字形库

- § 4-1 汉字字形库的基本要求
- § 4-2 汉字字形库体系
- § 4-3 汉字字形设计
- § 4-4 高分辨率汉字字形的压缩和表示
- § 4-5 计算机汉字字形设计系统
- § 4-6 汉字字形数据压缩
- § 4-7 一种汉字字形的笔画压缩法
- § 4-8 用哈夫曼树方法实现汉字字形库的压缩
- § 4-9 嵌套结构汉字字形库的压缩
- § 4-10 点阵字形的矩阵元素数的变换

§ 4-11 汉字拼形字库的设计

§ 4-12 汉字字库结构

参考文献

第五章 汉字编码检索

§ 5-1 顺序文件

§ 5-2 杂凑函数

§ 5-3 索引结构

§ 5-4 指针和关键字的压缩

§ 5-5 动态常用字库查找替换法

§ 5-6 通用汉字输入编码转换模块

§ 5-7 汉语拼音输入编码压缩技术

参考文献

第六章 中文信息处理操作系统

§ 6-1 CP/M操作系统简介

§ 6-2 中西文兼容 CP/M 操作系统设计要点

§ 6-3 中文操作系统 BZ-UCOS-2.0 (附参考文献)

第 三 分 册

第七章 中文信息处理语言

§ 7-1 中文 COBOL 语言

§ 7-2 中文 BASIC 语言设计举例

§ 7-3 程序设计语言“FORTRAV”处理汉字信息能力

参考文献

第八章 中文数据库

§ 8-1 数据库汉字信息的存储格式

§ 8-2 实例—GDMS-83-1 中文关系型数据库管理系统

§ 8-3 CPIM操作系统上的关系数据库系统

§ 8-4 数据压缩技术

参考文献

第九章 中文信息处理系统

§ 9-1 汉字终端

§ 9-2 汉字微型机系统

§ 9-3 汉字输入设备

§ 9-4 CKBD —— Ⅲ型汉字智能键盘

§ 9-5 汉字显示终端

§ 9-6 独立式的汉字显示终端举例

§ 9-7 汉字印字技术

§ 9-8 汉字制表

§ 9-9 中西文兼容计算机系统

§ 9-10 中文图形信息处理系统

§ 9-11 Z10微型机汉字处理系统

§ 9-12 汉字信息处理的人机会话

参考文献

第四分册

第十章 中文信息通信

§ 10-1 数据通信基础知识

§ 10-2 传输控制

§ 10-3 汉字通信有关问题

§ 10-4 通信规程转换器

§ 10-5 国家标准GB3453-82“数据通信基本型控制规程”

简介

§ 10-6 国家标准GB3454-82“数据终端设备(DTE)和数据电路终接设备(DCE)之间的接口电路定义表”介绍

参考文献

第十一章 汉字和汉语语音识别

§ 11-1 汉字的特征与表示

§ 11-2 汉字识别的预处理

§ 11-3 汉字笔画有向特征抽取方法

§ 11-4 印刷体汉字识别

§ 11-5 印刷体汉字识别的粗分类方法

§ 11-6 轮廓投影识别法

§ 11-7 手写体汉字识别

§ 11-8 模糊集合论简介

§ 11-9 汉字模糊笔画特征抽取方法

§ 11-10 汉字模糊笔画笔顺分析

§ 11-11 语音信号处理技术

§ 11-12 汉语的语音特点

§ 11-13 微型机语音识别接口

参考文献

第十二章 汉语自然语言处理

§ 12-1 语言工程

§ 12-2 汉语语音理解系统

§ 12-3 汉语拼音处理系统

§ 12-4 机器翻译概述

- § 12-5 机器翻译的两个主要方面
- § 12-6 机器翻译系统设计的几种方法
- § 12-7 机器翻译中目标语的生成问题
- § 12-8 自然对话智能机
- § 12-9 中文信息多维模型

参考文献

第十三章 几个国家标准

- § 13-1 国家标准“文字和符号成形设备的增补控制功能”
编制说明
- § 13-2 国家标准“信息交换用汉字编码字符集辅助集”
设计思想
- § 13-3 汉字编码字符集构成
- 附录 I 国家标准 GB 1988-80“信息处理交换用的七位
编码字符集”
- 附录 II 国家标准 GB 2311-80“信息处理交换用七位
编码字符集的扩充方法”
- 附录 III 国家标准 GB 2312-80“信息交换用汉字编码
字符集 基本集”

第一分册目录

前 言	2
总 目 录	5
第一章 绪论	18
§ 1-1 汉字输入问题	21
§ 1-2 汉字存贮问题	22
§ 1-3 汉字输出问题	22
§ 1-4 中文信息处理系统	23
§ 1-5 自然语言处理	24
参考文献	27
第二章 现代汉语基础	28
§ 2-1 现代汉语的特点	29
2-1-1 语音方面的主要特点	29
2-1-2 词汇方面的主要特点	29
2-1-3 语法方面的主要特点	30
2-1-4 现代汉语之要素	31
§ 2-2 汉字的形成、特点和结构	31
2-2-1 汉字是表意文字	32
2-2-2 汉字是代表音节的	33
2-2-3 用汉字记录汉语是不实行连写法的	33
2-2-4 汉字的结构系统	34
2-2-5 汉字的结构方式	37
2-2-6 汉字改革	39
§ 2-3 语音	40
2-3-1 语音的性质	40

2-3-2	汉语拼音方案	42
§ 2-4	词汇	46
2-4-1	词与词素的区别	47
2-4-2	词组	48
2-4-3	同音词	49
2-4-4	词典和字典	49
§ 2-5	汉字信息量	51
参考文献		52
第三章	汉字输入编码	54
§ 3-1	关于汉字一维码的统计推断	57
3-1-1	汉字码的概率分布	58
3-1-2	对码的均值、方差的推断	61
3-1-3	资料处理	63
§ 3-2	海曼(Hyman)公式及其推广与汉字编码时间	65
§ 3-3	汉字编码、排列法的心理学性质及其应用	68
3-3-1	字形编码心理学依据	68
3-3-2	汉字排列法心理学依据	69
3-3-3	从心理学角度,一个汉字编码方案的特点	72
§ 3-4	多作业对击键操作的影响	73
§ 3-5	汉字部件的分析方法和理论	74
3-5-1	现有汉字部件拼合汉字各种方法差异	74
3-5-2	汉字部件分析的理论基础	75
3-5-3	汉字部件分析基本方法	77
§ 3-6	WBZX 汉字编码方案	81
3-6-1	设计思想	81

3-6-2	汉字的笔画	82
3-6-3	汉字的字型	82
3-6-4	汉字的部件	82
3-6-5	部件、代号和键位表	82
3-6-6	取码规则	84
3-6-7	高频字码、简码	86
3-6-8	词汇码	86
3-6-9	键位排列及键位频度	87
3-6-10	主要指标	88
3-6-11	编码举例	90
3-6-12	小 结	91
§ 3-7	ZN 9号汉字 26键输入码方案	93
3-7-1	笔画种类及其序码	93
3-7-2	独体字的种类及其代码	94
3-7-3	合体单字	104
3-7-4	取码规则	105
§ 3-8	天龙中文编码法	110
3-8-1	中文字母	110
3-8-2	中文字母的分类	110
3-8-3	辅助字形	113
3-8-4	取码原则	116
3-8-5	注意事项	119
3-8-6	特殊字	121
§ 3-9	汉字树形结构中文字母编码法	123
3-9-1	汉字结构剖析	123

3-9-2	中文字母	128
3-9-3	汉字结构文法	129
3-9-4	“树”法	129
§ 3-10	三角编码法规则	130
3-10-1	取角原则	130
3-10-2	取形原则	134
3-10-3	同号字、异体号的处理	138
§ 3-11	汉字笔形编码法	141
3-11-1	指导思想	141
3-11-2	汉字笔画的分类和笔形代码	141
3-11-3	笔画的位置次序	143
3-11-4	部件的类型和部件的位置次序	145
3-11-5	笔形编码的简化	146
§ 3-12	汉字基本笔画编码法	148
3-12-1	原理	148
3-12-2	基本笔画键位设计	148
3-12-3	键的位置	151
§ 3-13	汉语拼音方案输入法	155
3-13-1	汉语拼音方案是汉字拼音化的基础	155
3-13-2	声韵调是汉字拼音化的要素	156
3-13-3	同音字和同音词	157
§ 3-14	汉语普通话新文字电脑Ⅱ型	161
3-14-1	标调法	161
3-14-2	构词法、联词法	162
§ 3-15	“见字识码”编码方案	168

3-15-1 编码的基本原则	168
3-15-2 汉字单字编码程序	169
3-15-3 字元的拼音和标记	169
3-15-4 词汇编码方法	171
3-15-5 单字拆成字元的规则	172
3-15-6 词汇码的应用	172
§ 3-16 声韵部形(UYBX)编码	173
3-16-1 声韵拼音编码	174
3-16-2 部形编码	174
3-16-3 举例	179
§ 3-17 拼音中文	180
§ 3-18 拼音拼写文字	183
§ 3-19 汉字的联想输入方法	185
3-19-1 汉字字形的联想	186
3-19-2 汉字字音的联想	186
3-19-3 汉字字义的联想	187
3-19-4 根据构词的联想	187
3-19-5 音词联想输入法	188
§ 3-20 计算机辅助编码	192
3-20-1 汉字编码计算机辅助设计步骤	193
3-20-2 综合调整设计方法	196
§ 3-21 汉字输入编码评测试行规则	198
§ 3-22 汉字输入方案比较法	204
参考文献	206

第四章 汉字字形库 (注:本章移入第二分册)

§ 4-1 汉字字形库的基本要求

§ 4-2 汉字字形库体系

4-2-1 字母式

4-2-2 数字式

§ 4-3 汉字字形设计

4-3-1 点阵字形特点

4-3-2 点阵字形要求

4-3-3 点阵字稿的技术问题

4-3-4 15·16点阵技术问题

§ 4-4 离分辨率汉字字形的压缩和表示

4-4-1 汉字规则笔画与不规则笔画的压缩表示

4-4-2 字形点阵的复原步骤

4-4-3 把一个向量转换成最近似的阶梯点的算法

4-4-4 在向量经过的阶梯点上写入二位标记的原则

4-4-5 把标记点阵转换成最终输出点阵的方法

4-4-6 汉字字形信息压缩性能指标

4-4-7 汉字扫描信息的线性增量压缩方法

§ 4-5 计算机汉字字形设计系统

4-5-1 ECCD汉字字形设计系统

4-5-2 汉字图形信息的冗余度分析

§ 4-6 汉字字形数据压缩

4-6-1 汉字字形的特性

4-6-2 二维予测编码

4-6-3 用 $m \times n$ 子矩阵信息法编码