

*The Status and Progress of
Chinese Language
Processing Technology*

中文信息处理技术的现状与进展

通用中文代码國際聯合會 ACCC

Association for Common Chinese Code, International

中文信息處理技術的 現狀與進展

The Status and Progress of
Chinese Language Processing Technology

ACCC

通用中文代碼國際聯合會

1991 年 3 月

書同文

通用中文代碼國際聯合會

努力推進計算機化

夏作人書贈



主 编: 张轴材

副 主 编: 袁 琦 曹右琦 陈明源 陈群秀 李传槐 黄伟敏
刘汝林 孙玉方

编委、作者: (按章节顺序排列)

李传槐	刘连元	徐德江	张仲陶	黄伟敏	王 健
康立论 (<i>Lee Collins</i>)	陈碧云	孙玉方	张轴材	高利军	
李建中	<i>Ed Smura</i>	石云程	赖声煌	董汉忠	谢克中
曾勤勤	顾国安	郑伯顺	张光耀	高济慈	魏裕屏
贾晨翔	王月英	裴 杰	叶建明	<i>Mark J. Frederiksen</i>	
卜朝曦	张培玉	许世忠	陈群秀	陈明源	吴庆宝
鲍有新	段 祥	郑国梁	李 辛	林 宁	马应章
尹锦柏	周立群	朱 岩	常宝儒	梁南元	沈旭昆
郑延斌	吴文虎	张 普	夏 莹	张世平	曹右琦
向 华	赵国梅	康加深	茅于杭	张正权	郑艳群
何克抗	李秀兰	谢泳珪	张大为	倪光南	陈 燕
许 棣	屠 群	张报昌	陈冬成	杨紫达	许振露
于明江	戴瀛洲	吴洛仁	舒裕禄 (<i>Norman V. Shulman</i>)		
刘汝林	敖其尔	崔明洙	苏永勤	曾民族	张潮生
栾贵明	陶 沙	余帼媚	浜田真美		

翻译、译校: 张学孝 张忻中 郭 卫 侯怡波 金坚敏 赵 玲
任 晖 许戈辉 陈小荷 孙宏林 钱 江 陈 辉
石晓明

责任编辑: 王 健
编 辑: 王 霞 贾晨翔

编委会联系地址: 100044 北京西三环北路 22 号 ACCC 秘书处

Address: ACCC Secretariat
22, Xi San Huan Bei Road
Beijing 100044, China

FAX: (+861) 841 1363

TEL: (+861) 841 2606

1991 年版前言

本资料是根据 ACCC 1990 年全体会员大会和理事会的决议,由 ACCC 秘书处在 TC—C, TC—D 两个技术委员会协助下,组织邀请了海峡两岸及国内外数十位专家编写的。

与 1990 年版相比,91 年版有了若干重要的修订和补充:

- 第一.增加了“语言文字工作的进展”一章;
- 第二.字符集及编码、汉字内码两章都有了重要的补充,特别是反映了 UCS (ISO DIS 10646) 和 Unicode 这两方面的重要发展;
- 第三.增加了“多字符集汉字数据库”一章;
- 第四.将 GLYPH 与 FONT (暂译作字形与字型库)独立一章,以充分反映这方面的发展;
- 第五.汉字文书处理、汉语教学 CAI、情报检索、人脑电脑速记也各自成为了独立的篇章;
- 第六.中文输入技术一章,没有回避该领域百花齐放、百家争鸣的事实,作了具体化的描述和客观的评价;
- 第七.在 IT 领域中,目前最热门的课题,PostScript 和 Windows 在中文化方面的进展亦得到了较充分的反映;
- 第八.重新改写了“少数民族文字及各种多文种系统的实现”一章,力求更全面、更具体地反映这方面的进展;
- 第九.各会员、各公司齐动手,一起来编写,形成了内容翔实的“各公司的中文化产品”一章;
- 第十.读者会注意到,本资料的附件有了大幅度的增加,主要包括:
 - (1) 中国近几年有关字符集与数据类型的提案
 - (2) 日本 1989 年度关于情报技术领域日语机能的标准化调查报告书(由 ACCC 理事、富士通田中省三先生提供)
 - (3) 日本信息处理专刊(由 ACCC 积极会员日立公司提供)
 - (4) 韩文信息处理(由 UNIX 用户协会 PortSoft 会议提供)

尽管如此,在截稿时为止,我们仍感有许多重要的发展没有反映,有许多有成就的专家尚未邀请,在此,我们一方面表示歉意,另一方面也期望得到谅解,因为信息技术的发展是如此地日新月异,甚至扑朔迷离,囿于编者的眼界与水平,也只好勉为其难地编辑定稿了。

读过本资料样稿的朋友,一致地称赞其内容丰富,但也批评篇章体系的某些交迭和混乱。应实事求是地说明,内容之丰富,完全是各位在第一线科研开发的技术人员的功劳,而体系的缺憾,则是编者的过错,也部分地缘于时间的紧迫。我们希望今后的版本会得到改善。

本资料的出版,不仅得力于许多编者、作者、译者的辛勤劳作,也得到中国中文信息学会、机电部计算机与微电子发展研究中心、台湾中文微电脑推广基金会、AFII、Unicode 集团等单位、组织的鼎力协作。在此,ACCC 秘书处和编委会对他们表示由衷的谢忱!

编者

1991 年春节

目 录

1. 语言文字工作的进展	1
1.1 我国语言文字工作的进展	1
1.1.1 国家语委增设文字应用管理司	
1.1.2 我国主要语言文字规范和标准	
1.1.3 合理统筹语言文字工作与中文信息处理工作	3
1.2 国际民间学术组织——汉字现代化研究会	4
1.2.1 汉字现代化研究会成立 10 周年	
1.2.2 安子介先生的语文学术思想	5
1.2.3 袁晓园教授的“识繁写简”观	
1.3 中国历史上最大的字典八卷本《汉语大字典》全部出齐	5
1.4 当代中国第一部现代汉语大词典《语言大典》问世	6
1.5 第三届国际汉语教学研讨会	6
1.6 中国式的“托福”—HSK	7
1.7 日本追加人名用汉字 118 个	8
1.8 台湾国字整理小组的工作成效及历史语文研究所科研动向	9
1.9 两岸文字统一的呼声日见高涨	10
——台湾《联合报》文摘	
2. 字符集及编码	11
2.1 汉字编码国家标准	11
2.1.1 概述	
2.1.2 基本集	
2.1.3 第一辅助集	12
2.1.4 第二辅助集	
2.1.5 第三辅助集	
2.1.6 第四辅助集	
2.1.7 第五辅助集	
2.1.8 文本通信用编码字符集	13

2.2 少数民族文字编码国家标准	13
2.2.1 概述	
2.2.2 蒙古文编码字符集	
2.2.3 朝鲜文编码字符集	14
2.2.4 维吾尔文编码字符集	
2.2.5 彝文编码字符集	
2.3 中国台湾定义的汉字字符集	15
2.3.1 CCCII	
2.3.2 CNS 11643	
2.4 其它国家的汉字编码标准	16
2.4.1 日本	
2.4.2 南朝鲜	
2.4.3 ANSI/NISO Z 39.64—1989 简介	17
2.5 ISO 字符编码体系	17
2.5.1 ISO 2022 体系	
2.5.2 ISO/IEC DIS 10646 体系	18
2.6 Unicode	18
2.6.1 背景	
2.6.2 替代标准	19
2.6.3 方法与状态	
2.6.4 设计思想	
2.6.5 Unicode 字集	20
2.6.6 未来发展与字符登录	21
2.6.7 代码赋值	22
2.6.8 细目	23
2.6.9 Unicode 汉字	24
2.7 GB13000—90 (DP)	28
3. 汉字内部码	44
3.1 概述	44
3.1.1 内码概念的确定	
3.1.2 汉字内部码的标准化	
3.2 ASCII 体系的汉字内码	46
3.2.1 总的状况	

3.2.2 未占用 C1 区的编码方式	47
3.2.3 覆盖 C1 区的编码方式	51
3.3 EBCDIC 体系的汉字内码	53
3.3.1 采用 SO/SI 机制的编码方案	
3.3.2 “空档码”	54
3.4 汉字内部码推荐方案	55
3.4.1 汉字内部码规范的编制原则	
3.4.2 内码的编码原则	
3.4.3 ASCII 代码体系的汉字内部码推荐方案	56
3.4.4 EBCDIC 代码体系的汉字内部码推荐方案	57
4. 多字符集汉字数据库	60
4.1 概述—背景与渊源	60
4.2 基本软件支撑环境开发	60
4.2.1 意义	
4.2.2 系统结构	61
4.2.3 字库管理	
4.2.4 显示管理	62
4.2.5 键盘管理	
4.2.6 打印驱动模块	
4.2.7 数据库管理系统	63
4.3 数据制备与预处理：字型与属性	63
4.3.1 字型	
4.3.2 集属内排除重码	65
4.3.3 属性表填制	66
4.4 多字符集数据库的基本功能	67
4.4.1 认同/甄别的基本策略与规则	
4.4.2 认同/甄别的流程	68
4.5 归并结果	69
4.5.1 两岸归并格局	
4.5.2 CJK 归并格局	
5. 字形与字型库 (Glyph and Font)	73
5.1 字符、字形与字型 (Character, Glyph and Font Image)	73

5.2 SC18 及 AFII 在 CJK 图符登录方面的进展	75
5.3 点阵字模	79
5.3.1 栅格	
5.3.2 点	
5.3.3 点阵字模	
5.3.4 新颁布的标准	80
5.3.5 开发中的点阵字模数据集	
5.4 点阵式符号库制作技术	80
5.4.1 点阵库制作的现状	
5.4.2 点阵库的计算机辅助设计	81
5.5 汉字大字库及 3.5 万宋体 24 点阵字库的创作	82
5.6 长沙前进计算机研究所的 XS-1 高压压缩精密通用汉字库	83
5.6.1 引言	
5.6.2 XS-1 字库的性能特点	
5.6.3 中国计算机字型产业	84
5.6.4 高压压缩汉字库对个人计算机设计的影响	85
5.7 非点阵字模	85
5.7.1 矢量汉字库开发技术	
5.7.2 IKARUS 介绍	85
5.7.3 中文向量轮廓多种印刷字型自动产生系统	86
5.8 中文 PostScript	87
5.8.1 PostScript 的解释器 PUCScript	
5.8.2 中文 PostScript 的描边字型的发展	89
6. 操作系统	97
6.1 DOS	97
6.1.1 CCDOS 的现状	
6.1.2 CMEX 与 CSI 规格	99
6.1.2.1 台湾(财团法人)中文微电脑推广基金会	
6.1.2.2 CSI 规格简介	100

6.2 DOS 图形环境—Windows	105
6.2.1 Windows 为双字节字符集所提供的机会	
6.2.2 长青窗口汉字系统	108
6.2.3 Windows 中文化台湾资策会进展	109
6.3 OS/2	111
6.3.1 CCOS/2	
6.3.2 OS/2 中文化台湾的进展	113
6.4 OS/2 图形环境—Presentation Manager	114
6.5 UNIX 操作系统的中文化与国际化	114
6.5.1 引言	
6.5.2 UNIX 系统中文信息处理的基本原理	115
6.5.3 UNIX 中文信息处理系统的实现	
6.5.4 AT&T UNIX System V 的国际化努力	117
6.5.5 中文应用环境 CAE 的设计和实现	119
6.5.6 UNIX 国际标准化及国际化活动	
6.5.7 X/Open 本地化标准化活动	120
6.5.8 UniForum 国际化工作委员会概况	121
6.5.9 POSIX.1	122
6.5.10 PortSoft	123
6.6 UNIX 图形环境	124
6.6.1 机电部六所的实践与水平	
6.6.2 Sun 4 工作站上 X Window 中文化情况	125
7. 高级程序设计语言的中文支撑	126
7.1 概貌	126
7.1.1 ISO/IEC JTC1/SC22 有关决议	
7.1.2 ISO/IEC JTC1/SC22 的进展	
7.1.3 国际上增加程序设计语言汉字支持的困难	
7.1.4 国内概况	127
7.2 Fortran	127
7.2.1 国内 Fortran 语言中文支撑现状	
7.2.2 国内外 Fortran 标准中多字符集支撑能力扩充的进展情况	128
7.2.3 ISO Fortran 90 语言已增加的功能及不足之处	129
7.2.4 今后工作要点	

7.3 多字符集 FORTRAN 的实现	130
7.4 COBOL	131
7.4.1 日本建议	
7.4.2 CODASYL COBOL 委员会建议	132
7.4.3 中国对 COBOL 多文种处理研究之现状	133
7.5 BASIC	135
7.6 Modula-2	136
7.7 Ada 语言中文处理问题	137
7.7.1 Ada 语言当前标准对汉化的限制	
7.7.2 Ada 语言中文处理国内现状	
7.7.3 建议	
7.8 C 语言国际标准	138
7.8.1 环境考虑	
7.8.2 语言成分	
7.8.3 通用实用程序前导文件	139
7.8.4 本地化时间表示	140
7.9 ASN.1	140
7.10 数据库对中文信息处理的支持	140
7.10.1 当前情况概述	
7.10.2 汉字字符处理对数据库的基本要求	141
7.10.3 今后应继续开展的工作	143
8. 汉字文书处理系统	144
8.1 概述	144
8.1.1 文字处理机种类	
8.1.2 电子排版种类	
8.2 汉字文书处理特点	145
8.3 中外文文字处理机	147
8.3.1 文字处理机的功能	
8.3.2 文字处理机的发展动向	148

8.4 电子排版系统	148
8.4.1 电子排版系统的功能	
8.4.2 电子排版系统的发展方向	149
8.5 活跃的文字处理协会	149
9. 中文基础研究及应用技术	152
9.1 汉字属性	152
9.1.1 汉字属性的开发和利用是汉字信息处理技术深入发展的结果	
9.1.2 建立汉字属性系统	153
9.1.3 编辑汉字属性字典	
9.1.4 存在问题及今后的发展	154
9.2 词频统计与词语库	155
9.2.1 必须区分字与词	
9.2.2 词频统计的必要性	
9.2.3 词频词典的内容和结构	
9.2.4 词频统计和汉语信息处理	156
9.2.5 汉语通用词语库和专业词语库	
9.3《信息处理用现代汉语分词规范》与汉语自动分词	157
9.3.1 制订《信息处理用现代汉语分词规范》的目的和意义	
9.3.2 制订过程	158
9.3.3 国内外研究现状	
9.3.4 主要内容及难点	159
9.3.5 《分词规范》的验证和应用	161
9.3.6 汉语自动分词	162
9.3.7 汉语自动分词系统	167
9.4 汉语语音合成	169
9.4.1 汉语语音合成的现状	
9.4.2 汉语特点与合成策略	
9.4.3 汉语语音合成的新进展	
9.5 汉语自然语言理解与生成	170
9.5.1 自然语言理解国外研究现状	
9.5.2 汉语理解与生成国内研究现状	171
9.5.3 汉语理解与生成研究的进展及成果	
9.5.4 与国际上的差距	172

9.5.5 汉语理解与生成的难点和问题	
9.5.6 汉语理解研究的应用前景与发展策略	173
9.6 机器翻译	173
9.6.1 机器翻译国外研究现状	
9.6.2 机器翻译国内研究现状	174
9.6.3 与国际上的联系和差距	175
9.6.4 机器翻译研究中的难点和关键	
9.6.5 MT 研究的应用前景与几点建议	176
9.7 汉字信息处理词典	176
9.7.1 编纂目的	
9.7.2 词典的性质及读者对象	177
9.7.3 词典的收条	
9.7.4 词典的译名	178
9.7.5 词典的索引及附录	
9.7.6 词典的计算机辅助编纂	
9.8 汉字排序	178
9.8.1 字符集的字序与汉字排序	
9.8.2 汉字排序	179
9.9 简繁转换	182
9.10 汉字与盲文转译电脑化	182
10. 中文输入技术	183
10.1 汉字识别	183
10.1.1 中国汉字识别领域的现状	
10.1.2 中国汉字识别系统的配置和识别方法特色	
10.1.3 中国汉字识别技术发展的趋势	184
10.1.4 中国已鉴定过的汉字识别系统一览表	
10.2 汉语语音识别的新进展	187
10.3 四达—863A 型语音识别系统	188
10.4 汉字键盘输入方法与评测	190
10.4.1 概述	
10.4.2 键盘输入方法的分类	

10.4.3 邮电部向全国推荐使用的三种输入方法	191
10.4.4 键盘输入方法的发展	193
10.4.5 键盘输入方法的评测及发展	194
10.5 海峡两岸中文电脑输入技术表演赛	196
11. EDI	198
11.1 UN/EDIFACT	198
11.2 ISO/IEC JTC1/SWG EDI	198
11.3 中文 EDI	198
12. CAI 应用于汉语教学与研究	200
12.1 来华留学生 COA 能力的培养	200
12.1.1 COA 能力概说	
12.1.2 PJY 拼音—汉语变换系统	
12.2 对外汉语教学与 CAI 计算机辅助教学系统软件的研究	200
12.2.1 导语	
12.2.2 对外汉语教学	
12.2.3 计算机应用于汉语教学与研究	201
12.2.4 本系统的内容与流程	202
12.3 CEC 系列中华学习机汉语教学功能介绍	204
12.4 汉语拼音计算机辅助教学系统	205
12.4.1 基本技术和基本原理	
12.4.2 基本功能	206
12.4.3 技术特点	
12.5 机助汉语教学的基本软件开发	206
12.5.1 中西文多语种处理系统	
12.5.2 汉语语音合成系统	207
12.5.3 机器翻译系统	
12.6 汉语计算机辅助教学系统可实现题型的分类与设计	208
12.7 智能型计算机辅助汉语教学系统的设计与实现	209
12.7.1 系统结构与教学过程	
12.7.2 知识库	210
12.7.3 学生模型	211

12.7.4 教学决策模块	212
12.7.5 汉语语音合成及语音库	
12.8 汉语信息处理技术与对外汉语教学	213
13. 各公司的中文化产品	214
13.1 新华社新闻信息处理系统的多文种处理	214
13.2 联想集团中文化产品—联想式汉字系统	215
13.2.1 联想式汉卡系列产品	
13.2.2 联想式汉字环境软件	
13.2.3 汉字局部网和汉字操作系统	
13.2.4 汉字驱动程序	216
13.2.5 CAD 汉字系统	
13.2.6 联机汉字仿真	
13.2.7 联想排版系统	
13.2.8 PC—FAX 通讯系统	217
13.3 布尔的中文化产品	218
13.3.1 布尔专有主机系统—DPX 7000	
13.3.2 布尔全汉化超高速图文打印系统—MATHILDE	
13.3.3 布尔开放式系统—DPX/2	220
13.4 中国长城计算机集团公司 90 年在中文产品方面的新进展	220
13.4.1 GW—CVGA/24 高分辨率汉字图形显示系统	221
13.4.2 长城集成软件	223
13.4.3 GWART 3.0 交互式中文桌面印刷系统	
13.4.4 GWTOOLS 工具软件	224
13.4.5 MWE 多用户窗口字处理软件	
13.4.6 GWOCR 汉字识别系统	
13.5 HP 计算机的本地语言支持系统	224
13.5.1 什么是 NLS/NLIO	225
13.5.2 HP—UX 的 NLS/NLIO 的特点	
13.5.3 亚洲字符的输入输出	226
13.5.4 X Window 系统与 OFS/motif (TM)	
13.6 长江计算机集团的中文化产品	
缺	

13.7 DEC 公司汉字产品的策略、方向及应用	226
13.7.1 概述	
13.7.2 汉化的等级	227
13.7.3 汉字应用平台	228
13.7.4 开发工具	230
13.7.5 服务	231
13.7.6 今后的方向	
13.7.7 结束语	
13.8 IBM 在 NLS 领域中的成就和最新发展	232
13.8.1 IBM 的国家语言支持工作	
13.8.2 IBM 的汉字支持	
13.8.3 国家语言体系结构 (NLA)	234
13.8.4 综述	
13.9 NEC 公司的中文产品	234
13.9.1 中文处理特征	
13.9.2 中文终端机	236
13.9.3 中文激光打印机	
13.9.4 中文属性处理软件系统	
13.10 优利系统公司产品的中文化	238
13.10.1 简介	
13.10.2 中文化的实现	239
13.10.3 产品简介	
13.10.4 总结	240
13.11 从五笔字形到王码电脑	240
13.11.1 五笔字型输入和王码汉字输入体系	
13.11.2 “复线双轨”制的王码汉字输入体系	241
13.11.3 5.0 版王码汉卡操作系统	
13.11.4 王码电脑	242
13.12 四通文字处理机	242
13.12.1 四通 MS-2401	
13.12.2 四通 MS-2403	243
13.13 浪潮电子信息产业集团在中文信息技术方面的进展	244
13.13.1 浪潮微机汉化产品	
13.13.2 浪潮汉字显示传呼机	245

13.13.3 序列字根汉字信息处理技术	246
13.14 汉字终端进展综述	246
13.14.1 几个发展阶段	
13.14.2 汉字终端的技术进步	
13.14.3 汉字终端的应用发展	
13.15 KJDD 中日英兼容操作系统	247
13.16 多语种处理系统—QHML	247
13.17 CDC 计算机产品的中文化	248
13.17.1 CYBER 机在硬件体系上对中文信息处理的支持	
13.17.2 中文信息输入、输出设备的选择	
13.17.3 系统软件和公用程序的汉化	250
13.17.4 CDC 公司 UNIX 系统应用软件的汉化	251
13.18 SUN 公司的中文语言环境	251
13.18.1 概述	
13.18.2 国际化	252
13.18.3 区域化	253
13.18.4 本地化	
14. 少数民族文字及各种多文种系统的实现	254
14.1 概述	254
14.2 多文种信息处理中要解决的问题及处理方法简述	254
14.3 蒙古文信息处理简介	257
14.4 藏文信息处理简介	258
14.5 维吾尔文/哈萨克文信息处理简介	259
14.6 朝鲜文信息处理简介	260
14.7 彝文信息处理简介	261
14.8 壮文信息处理简介	261