计算机

汉字系统的设计与实现

J·S·J·H·Z·X·T·D·S·J·Y·S·X·

周浩华 著

华南理工大学出版社

计算机汉字系统的设计与实现

周浩华　著

华南理工大学出版社

内 容 简 介

本书对汉字系统及汉字信息处理的基本组成及其应用、发展作了详细的介绍，并对信息处理交换用的汉字代码、汉字库、汉字编码、汉字输入、汉字输出、计算机汉字系统的软硬件配置及系统的设计与实现等作了深入浅出的论述，并融进了作者十多年来在这方面的研究成果及心得。

本书可作为大专院校的本科教材，亦可作为成人高校的教材，对从事计算机汉字处理的有关技术人员也有很大的参考价值。

【粤】新登字 12 号

计算机汉字系统的设计与实现

周浩华 著

*

华南理工大学出版社出版发行

（广州·五山·邮码 510641）

各地新华书店经销

华南理工大学出版社电脑排版室排版

广东德庆印刷厂印装

开本：787×1092 1/16 印张：16.25 字数：386 千

1992 年 12 月第 1 版 1992 年 12 月第 1 次印刷

印数 1—3000

ISBN 7—5623—0317—7/TP·21

定价：4.25 元

前　言

用计算机处理汉字信息是信息时代的需要，是汉字信息处理现代化的一个重要标志。另一方面，计算机能够处理汉字信息，大大加速了计算机在我国的推广应用，并对我国的文化事业、出版业、管理、办公室自动化、通信等领域产生深刻的影响，使这些领域发生重大的变革。

计算机处理汉字信息的重要性和迫切性促使广大汉字信息工作者努力去探索、研究和开发。经过十多年来的艰苦奋斗，汉字信息技术的各个领域取得了重大进展，并迅速成为我国计算机界的一个重要的、非常活跃的新领域。各个大专院校相继开设了“计算机汉字信息处理”的课程，极需一本教材，不但讲述汉字信息处理的原理，而且讲述进行汉字信息处理的计算机汉字系统的设计和实现的方法，以满足计算机专业的教学需要。笔者从事计算机汉字处理的研究和教学工作十余年，积累了一些研究成果和心得体会，收集了一批资料，撰写成这本“计算机汉字系统的设计与实现”。

全书分为六章。第一章概述了计算机汉字系统的研究内容、意义、应用领域以及所取得的进展。第二章讲述我国在信息处理方面的一些编码标准，特别是汉字的有关标准，比较深入分析汉字内码的问题。第三章讨论了汉字库的标准化、点阵汉字库、压缩汉字库的原理和生成方法，最后介绍了作者提出的智能汉字库方法，实现统一汉字库的研究。第四章除介绍汉字本身信息、几种较典型的输入方案外，着重介绍键盘编码输入方案的设计以及在计算机上实现的方法。考虑到自然语言（文字和语音）作为人机接口和交流的重要性，将会成为汉字信息处理的主要研究课题，书中在这方面作了较详尽的介绍，特别对实用的语音识别作了探讨。第五章阐述汉字输出的三种形式：显示、打印、语音。对各种屏幕格式显示汉字的原理和实现方法、各种打印机打印汉字的原理和实现方法、语音输出的合成原理和实现方法、汉字通信的问题等作了详细的论述。第六章从整个计算机汉字系统的角度，讨论了系统的硬件、软件的配置、汉字系统选择的方法，特别重点阐述了实现中西文兼容的汉字系统的方法，并以 CP/M、Apple Ⅱ、IBM-PC 系列为例作了详尽的分析。

在本书的撰写过程中得到华南理工大学计算机系领导和同事们的支持和鼓励，特别是郭荷清副教授对本书的结构和内容提出了许多有益的建议，在此谨致深切谢意。

由于水平所限，书中恐有疏漏谬误之处，敬请读者指正。

作　者

1990 年 6 月于广州

目　录

第一章 绪 论

汉字是一种古老的、历史悠久的文字，对中国文明的发展起了重要作用。汉字是世界上使用广泛的文字之一，有十多亿人口在使用。汉字又是一种自具特色的象形文字，与其它的拼音文字不同，汉字是以形为主、数量庞大的文字。

文字的发明曾对人类文明的发展起到特别重要的作用。到了目前高度文明的时代，信息以很快的速度增加，以至有的学者称之为"信息爆炸"的时代。英国科学家詹姆斯·马丁有个推测，人类科学知识在19世纪每50年增加一倍，20世纪中叶每10年增加一倍，到20世纪70年代每5年就增加一倍，估计21世纪大约每3年增加一倍。现代科学知识90%是1950年以后积累起来的。

据联合国教科文组织统计，全世界现有重要报纸8000种，发行4～5亿份，杂志5～6万种。

面对这样庞大的信息量，信息科学的技术革命势在必行。

70年代以来，微电子技术和计算机科学取得了重大突破，通讯技术和网络技术也取得长足的发展，形成了现代信息处理技术环境。作为信息中的重要角色的文字，纳入这种新的处理格局是理所当然的，拼音文字以其拼音的良好条件，由少量字母组成，率先进入这个现代信息处理领域。

汉字由于其以形为主、数量庞大的特点，采用现代处理技术，具体地说，采用计算机处理比较困难。这引起不少专家学者的关注，引起了很多人的担心。以至直到80年代初，很多学者惊呼："汉字信息处理是信息处理的瓶颈。"

经过十多年的不懈努力，我国汉字信息处理技术得到了飞速的发展。各种中西文兼容的计算机系统在市场上涌现。汉字信息处理进入现代信息处理的环境。

§1-1 汉字信息处理系统的基本组成

一个基本的汉字信息处理系统主要组成如图1-1所示。

一个基本的汉字信息处理系统(简称汉字系统)主要由四大部分组成。

(一) 汉字输入部分

汉字输入部分主要解决将汉字信息送入计算机。汉字输入依送入信息的形式和采用的方法不同大致可分成三大类型：

(1) 键盘编码输入

键盘编码输入就是将汉字用各种方法编成代码，由键盘打入计算机。依键盘的类型不

同，又可分为：大键盘、专用键盘（中键盘）、标准西文键盘（小键盘）。

大键盘多半采用一字一键的形式，关键在于生产一种寿命长、价格低和尺寸小的键盘。

专用键盘根据输入方案采用的常用部件来设计专用键盘，对照每个汉字的组成部件击相应的键，将汉字编码输入计算机。

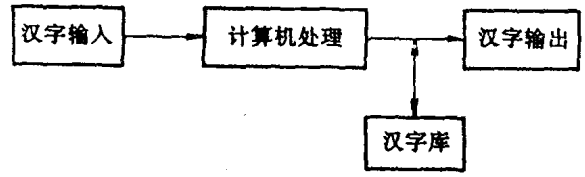

图 1-1　汉字信息处理系统的基本组成

最常用的是用标准键盘的编码输入方案，这种类型的编码方案已有五六百种之多。根据编码原理，大致可分为：字形码、笔形码、拼音码、联想码、混合码等。

（2）自然语言输入

自然语言输入包括文字和语音输入两大类。文字输入就是直接写汉字的输入（联机手写输入）或者整页输入（印刷文稿或手写文稿）。

语音输入就是直接向计算机输入语音信息，其中又分认人识别（专人）或不认人识别。

自然语言输入具有良好的人机接口性能，是人类追求的与计算机接口的最好的形式。但是，它的技术比较复杂，目前仍处在发展阶段。

（3）交换码输入

已经联网的计算机可以用交换码传输汉字信息。

（二）计算机处理部分

计算机汉字处理部分是汉字信息处理系统的核心，它的主要功能为：

（1）实现汉字的输入输出

实现汉字输入是将三种类型输入的汉字信息识别出相应的汉字，并且能够将汉字字形显示或打印输出。

（2）实现汉字信息处理

即实现中西文兼容，汉字信息和西文信息一样，能够进行存贮、运算、传输等处理。

（3）实现西文软件的汉字功能

在管理方面的应用中，数据库系统等软件能扩展有汉字功能是非常有用的。

（4）能开发和运行各种中文应用软件

计算机处理部分是由计算机的硬件和软件共同完成的。

（三）汉字库部分

汉字库是一个汉字系统的基础，它存放着汉字的字模。每个汉字系统必须要有一个汉字库，而汉字库的类型，存放汉字的数目，字模的种类由具体系统的应用需要选定。

（四）汉字输出部分

汉字输出有三种形式。

（1）汉字字模点阵

输出汉字字模点阵的使用最多，最常用的是作屏幕显示和制表打印。此时，要配置符

合汉字要求的CRT和打印机。另外，对于具体应用，要配置专用的设备，例如，对排照印刷系统，需要专门的数字照相机等等。

(2) 交换码输出

网络系统的计算机之间，汉字相互传输是用交换码的形式进行的。此时，要将汉字信息用交换码传输，而不是传送字模点阵。

(3) 语音输出

直接将汉字用汉语输出，即"说话"输出。此时，要将汉字信息转换成语音信息，由扬声器说出。

§1-2 汉字信息处理系统的应用

汉字信息处理系统有广泛的应用范畴，大致有如下几个方面：

(一) 管理信息系统

随着生产的发展，各种管理信息系统在国民经济各个部门广泛地应用。从政府各部门对国民经济的计划、统计、管理系统，到各个企业、工厂、银行的管理信息系统，甚至个人的事务系统，都要作汉字信息处理。很难想象，不采用现代化的计算机管理信息系统如何能管理好一个大城市的车辆；不采用计算机系统如何管理全国航空的预订票工作；不采用计算机系统如何能及时做好全国的人口普查的统计工作。反过来，采用计算机系统能够及时准确了解本部门的信息，作出正确的判断和决策。显然，在我国的条件下，汉字信息处理技术是必须的。

(二) 通讯

通讯中广泛应用电报和电传。我国电报采用的电报码已有一百多年的历史，还是在1880年开通我国第一条电报线路，请外国工程师设计的。由于长期采用手工编码和译码，四位数字代表一个汉字，工作人员需将常用字背诵，费时失事。如果采用汉字系统处理，将会大大加速整个通讯过程，减少人为失误。

电传是一种快速传递信息的工具，可以设想，如果记者们使用和计算机连接的汉字电传，消息从各地直接将新闻送入总部计算机中让编辑部进行编辑，新闻的出版将会更快。如果用在商业活动中，将随时变化的商业、金融信息传回企业，将会带来良好的经济效益。

计算机网络的飞速发展，使计算机可以享用网络内的信息资源，各种资料库可以随时查询，当然也包括大量用汉字记录的资料，例如，可以查询专利文献。甚至，可以查询销售各种货物、日用品的各个商店的价格。要达到这种汉字资料信息的广泛使用，就要增加相应的汉字记载的资料信息库和实现汉字信息的通讯。

(三) 文字出版

文字出版业是文化的基础。到现代，计算机出版系统充分显示它的快速、高质量、省人力、全自动的优点。

汉字出版业必然也会走上这条现代化的道路。经过多年的艰苦努力，我国也研制出这种系统。

计算机照相排字印刷系统可用在印刷业及出版社，作大量出版书刊报纸的编辑、照相排版工作。采用这样的系统，由计算机终端输入书刊的内容，在屏幕上进行编辑、排版、校对等工作，每页内容存储在机内，需印刷制版时，由存储磁盘调出，输出到照相排字设备，照出每页的底片，由底片制版。这种流程由于编辑校对均在屏幕进行，省却清样校对、人工捡字排版等工作，将会大大加速出版的过程，减少人为的错误，提高书刊的质量。

计算机轻印刷系统是用激光打印机打印出纸型，可作为学校、机关印刷教材讲义、学习辅助资料、文件等。由于激光打印机可以打印出精美的文字和图形，因此甚至可以用来为企业打印图文并茂的产品使用说明、产品广告。

可以预料，随着计算机汉字照排系统和轻印刷系统的推广应用，我国出版业将会发生重大的变化。

（四）办公室自动化

办公室自动化的设备将会由文书处理设备、计算机管理系统和通讯网络组成。文书处理将解决公文来往、总结表格、可用中文打字机或汉字终端。它的计算机管理系统，对企业单位来说，不仅管理企业的日常事务，还管理企业的生产、财务、销售、设备和人员。通过各种统计数字和信息，不仅可以对整个企业进行指挥和调度，还可以为企业作出决策。

（五）人工智能方面

汉字信息处理与人工智能结合，将可产生出巨大的成果。

汉语和汉字的计算机识别是一个很重要的领域，语音和文字的识别是解决人与计算机、人与智能机器人之间的重要界面。如果能解决人与计算机的自然语言的沟通，即解决自然语言识别和自然语言理解，则人同机器的界面变得更加友好，人对机器的使用和指挥更加容易、更加简单。也可以研制出盲人阅读机。

机器翻译也是很有发展前途的，如果解决自然语言接口，直接将语言和文字送入计算机，由计算机将语言理解后，按语言规律翻译成指定的自然语言，如英译中或中译英等。

专业文献的翻译系统将会首先应用，解决大量科技文献资料的翻译困难。

总而言之，计算机汉字信息处理系统的广泛应用，对汉字文化和国家的现代化，将产生重大影响。

§1-3 汉字信息处理技术的发展

汉字是一种古老的文字，最早的汉字编码、分类，大概可追溯到春秋时代《尔雅》一书，首次提出“部首”的概念和做了归纳。随后，各个时期的字典、词源、词海等，都要采用不同的汉字排列和查找方法，这也就是汉字编码查字典的实践。

“汉字信息处理”一词是近十多年流行采用的，是指用计算机进行汉字信息处理。

50年代末，在研制103、104计算机时，就开始了俄汉机器翻译的工作。只不过由于受当时机器的限制，和受中文信息基础研究的限制，没能取得进展。

1974年8月我国由多个部委组成联合小组，开展了“748工程”的研究，对汉字频度进行了非常有益的基础研究，其后，诞生了我国汉字研究的第一个国家标准GB2312-80。

1978 年 12 月在青岛召开了全国汉字编码学术交流会，开始了汉字信息处理研究的新一页。

1980 年开始筹建中文信息研究会，并于 1981 年 6 月在天津正式成立，并先后设立了五个专业委员会：

(1) 基础理论专业委员会；

(2) 汉字编码专业委员会；

(3) 汉字信息处理系统专业委员会；

(4) 自然语言处理专业委员会；

(5) 汉字信息处理专用设备专业委员会。

从此，汉字信息处理研究逐渐形成高潮，国内外学术会议相继召开，各种研究成果不断涌现，各种计算机汉字处理系统进入市场，各种汉字处理的应用系统投入使用。

汉字信息处理的主要研究成果有如下几个方面。

(一) 基础研究

汉字信息处理早期没能取得进展，除了机器设备的原因外，基础研究还没有进展是一个重要因素，只有基础研究取得进展，才可能使整个学科飞速发展。

(1) 频度研究取得进展

字频的研究主要统计汉字使用的频度，它奠定了汉字库字数的设置。在字频研究基础上，确定了 GB2312-80《信息交换用汉字编码字符集　基本集》的 6763 个汉字、《第二辅助集》的 7237 个汉字、《第四辅助集》的 7039 个汉字。

一般计算机装设基本集的一、二级常用汉字。其它应用按照需要再装设第二辅助集的汉字、第四辅助集的汉字。

部件频度的研究主要统计组成汉字的各种部首偏旁的使用频度，为采用字形编码原理的编码方案提供依据。

词频的研究主要统计词组使用频度，作为确定常用词组的依据。汉字的一个特点是组词能力特别强，即词的个数比汉字数多。选取常用词组，对降低占用存储量非常重要。“七·五”项目中词频的研究取得了成果。

(2) 对汉字处理系统的基础研究

究竟如何将一个西文计算机系统扩充成中西文兼容的系统，如何设计一个汉字系统，经过多年的摸索，找到了有效的方法。

其中关键之一是进行汉字内码的研究，研究汉字在机内的表示方式。同时，研究了如何在操作系统一级扩充汉字功能的方法，研究了大、中、小型机实现汉字的方法等等。

(3) 汉字本身信息及编码方法的研究

汉字本身存在哪些信息，如何利用这些信息进行编码，是汉字编码方案的基础，了解这些信息，灵活利用这些信息对编码技术是很重要的，正是近年来重视这些信息的分析和利用才涌现出大量的编码方案。

(4) 编码方案评测规则

客观地评价编码方案，建立一套建立在科学基础上的评测规则，排除人为的因素，来评测编码方案。这些评测规则本身就是一项基础研究。几年来，对它们进行了一些研究，

初步提出了评测规则的试行方案。

(5) 自然语言的基础研究

对文字识别的方法进行了很多的研究，对印刷体和手写体汉字的研究，取得了许多基础的成果。

在语音识别的研究方面，在音素的识别和音节的识别方法和原理方面也有了一些进展。在汉语语音合成方面，取得了很大成功，可以合成高质量的汉字语音。

自然语言理解方面，对文字的语法规律和自然语言理解的实现软件方面都进行了研究工作。

(6) 标准化

上述 GB2312-80 不但规定了汉字库的汉字数量标准，还规定汉字信息的交换码标准。根据 GB2312-80 规定的一、二级常用汉字，按区位号顺序研制了我国 16×16、24×24、32×32、48×48 的标准固化字库。还在制定汉字设备使用的汉字控制功能码国家标准，即文字和符号图形设备的增补控制功能等等。

(二) 编码输入方案百花齐放

编码输入方案已达五六百种之多。专业打字员最高的汉字输入速度已达 205 字/分(自选样本)。可以讲，汉字研究在输入方案方面百花齐放。在众多的方案中，形码占主导地位。

但是，由于汉字本身规律的复杂性，输入速度快的方案，编码规则比较复杂，难为大众所掌握。相反，规则简单的方案，输入速度比较慢。在众多的方案中，仍然还没有公认的、压倒的方案。特别是缺少真正易学、易记、易用的方案，因此出现多种方案并存的现象。

1986 年 4 月在北京举行了全国首届汉字编码输入方法的定性评测，评出 A 类方案 11 个。但是，实际状况并没有明显变化。可能要在实践中优胜劣汰，或者在总结现有方案优点的基础上产生更好的方案。

不管怎样，近年来，编码输入方案的确取得了重大进展，对推动汉字处理技术的发展和新系统的应用，的确起了很大的作用。

(三) 中西文兼容计算机及汉字处理产品

近年来，我国研制的中西文兼容计算机系统大量涌现，解决了大、中、小型机的汉字信息处理问题。研制出一批中西文兼容的微型计算机，不但将进口的各种微机汉化，更可喜的是建立了长城等多个国产计算机企业，生产具有中文特色的计算机、中文终端、中文打字机等系列产品和各种照排系统、轻印刷系统。

(四) 汉字系统广泛应用

汉字系统的广泛应用标志着汉字处理系统在技术上逐渐完善，意味着汉字处理系统走上大量应用的前景。特别是在管理上的广泛应用，对改善机关、企业、事业单位的管理素质，跟上现代化的步伐起了很大的作用。汉字产品的应用对改善我国的汉字通讯，促进我国印刷业的现代化将发挥越来越大的作用。

(五) 汉字信息处理技术的发展

(1) 继续加强汉字信息处理技术的基础研究

前面提到，几年来由于加强了基础的研究，因而汉字处理技术和应用都得到了巨大的

发展。目前汉字的编码方案正是利用汉字本身的信息原理构成的。可以讲，有新的原理就有新的编码。

汉字信息处理技术，除了继续完善基本系统的性能外，研究的重点将会逐渐转移到整页汉字的识别、汉语语音的识别、自然语言理解和机器翻译等课题。这些课题将与人工智能、模式识别等学科结合在一起，并考虑汉字信息的特点进行研究。

预期整页汉字识别系统将从目前的实验室试验系统，在提高识别率和稳定性后将慢慢走向实用阶段。

语音识别单音节的认人识别系统，识别少量汉字的实用系统，将会步入市场。由于汉语的同音字问题不仅严重影响拼音方案，而且更加影响汉语语音识别的继续发展。这将是一个重要的课题。

前一阶段利用汉字的信息，下一阶段必然会利用汉语的句法、词法、语义等信息。因此，很有必要从现代信息的观点对汉语的上述几种信息进行研究，这需要汉语学家和计算机学者共同努力。

(2) 促进汉字信息处理基本系统的标准化

重视汉字信息处理技术标准化和标准的制定，逐步完善标准体系，有利于汉字设备的生产和推广应用。

当前比较关键的是制定汉字内码的标准。由于汉字内码对汉字应用软件的兼容性影响极大，制定统一的标准有利于软件兼容和机间通信。

此外，对汉字的文字、图形的控制方式，即打印机的控制码，图形显示的控制码，图形、汉字显示方式等应作明确的、标准化的规定。

最困难的问题恐怕是编码输入方案的统一，强制性的统一可能是行不通的。进一步完善评测标准，定期进行权威性的评测，诱导编码的发展方向，可能还要经历一段在实践中自然淘汰的过程，多种输入法并存，又有相互竞争的局面会慢慢达到共识。

(3) 开发国产化的汉字产品

国内生产汉字产品有了良好的开端，汉字产品应该要把汉字功能作为产品的一个重要功能来设计，而不是将汉字作为一个附加功能贴在西文产品上。既然，汉字库是汉字产品的基础，又需占用大量存储容量，如果像汉字终端那样设计内装式汉字库，汉字功能操作与西文功能一样命令化，相信生产出来的主机与外设性能会更好。

注意优选机型，做好型谱系列化工作，扩大批量生产，加速研制新品种。

(4) 加速推广应用

推广应用促进国民经济的发展是非常重要的工作。目前已有条件在通讯、管理、印刷出版、办公自动化等方面大力推广应用。

大量的应用，反过来又促进汉字产品生产的发展，促进汉字处理技术的发展。

可以预期，汉字技术将会走向成熟，将会有更多的成果、更多的产品、更广泛的应用。

第二章　信息处理交换用的汉字代码

信息系统由信息采集、信息加工、信息存贮、信息传输和信息利用几部分组成。信息在每部分都以代码形式存在。

汉字信息采集依不同的汉字信息类型有不同的形式。以语音形式输入汉字信息，声波通过转换成电信号送入计算机，而键盘编码输入则以英文字母、数字或符号键的代码组成“外部编码”(简称外码)。

汉字内部处理码包括存贮码、运算码和传输码。存贮码是存贮汉字信息内容的代码。运算码是参与各种运算处理如分类、合并、增、删、改的汉字信息代码。传输码分两种：内部传输码和传输通讯码。内部传输码供系统内部信息传输用，如将处理结果送显示器和打印机等等。传输通讯码是指系统之间的汉字信息交换用代码，这种码留待汉字通讯讨论。

存贮码、运算码和内部传输码最好能统一起来，减少内码之间的转换，并与国家标准交换码有简单的换算关系。在本节中着重讨论字符和汉字的一些国家编码标准，讨论区位码、国标交换码和汉字内码。

§2-1　GB1988《信息处理交换用的七位编码字符集》

GB1988是各种字符(控制字符和图形字符)的国家代码标准，它是根据国际标准化组织(International Standards Organization)的标准ISO646《信息处理交换用的七位编码字符集》制定的，它是我国计算机专业的基础标准。

GB1988这个字符集用7位编码定义了128个信息处理交换用字符。7位编码的高三位$b_7b_6b_5$组成8列，低四位$b_4b_3b_2b_1$组成16行，第0、1两列定义了32个控制字符，第2～7列定义了94个图形字符，另加间隔字符SP(2/0)和抹掉字符DEL(7/15)(参见图2-1)。

本标准的32个控制符定义和94个图形字符的名称及相应的代码详见GB1988文本。

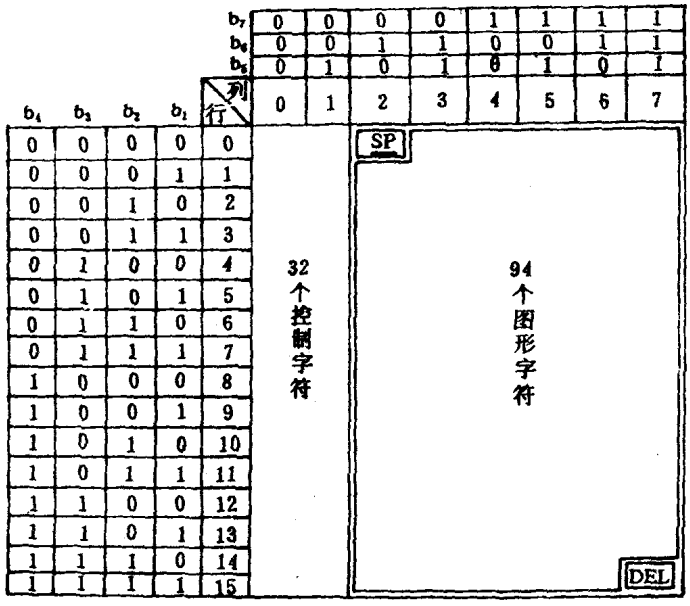

图 2-1　GB1988 字符代码定义图

§2-2　GB2311《信息处理交换用七位编码字符集的扩充方法》

GB2311 国家标准是根据国际标准 ISO2022《七位与八位编码字符集的扩充方法》制定的。

GB1988 定义 128 个字符及其编码，只能满足西文信息处理的需要。对于汉字信息处理系统要处理成千上万个汉字，需要一些特殊的控制功能，必须要制定一套编码扩充的方法，GB2311 就是这种编码扩充的国家标准。

GB2311 是以七位编码字符集为基础进行代码扩充的。它首先将 GB1988 的 32 个控制符定义为 C0 集，将 94 个图形字符集定义为 G0 集。SP 和 DEL 定义为单个控制字符。

增加 CS 为单个控制字符，由 32 个控制字符组成的增补控制功能集：C0 集或 C1 集。由 94 个图形字符组成的多字节图形字符集：多字节 G0 集，G1 集，G2 集和 G3 集。

如果用二个七位编码为例，则每个图形字符集能对 $(2^7-34)(2^7-34)=8836$ 个汉字进行编码。如果对三万个汉字进行编码，则要四个图形字符集。对五万个汉字进行编码，就要六个图形字符集。

扩充了控制字符集和图形字符集之后，使用时要指明和调用它们，规定了相应的控制字符：

(1) 移出字符 SO(0/14)：调用 G1 集，系统改变原来状态。

(2) 移入字符 SI(0/15)：调用 G0 集，系统改变原来状态。