TUSHUGUANXUE
QINGBAOXUE

图书馆学
情报学

参考资料

5

书目文献出版社

图书馆学情报学

参考资料　第5辑

书目文献出版社编辑组编

书目文献出版社

北京

图书馆学情报学参考资料
第五辑

书目文献出版社编辑组编

书目文献出版社
（北京文津街七号）
秦皇岛市第二印刷厂印刷

新华书店北京发行所发行　　各地新华书店经售
787×1092毫米　1/16 开本 7.5 印张 100 千字
1984年5月北京第1版　1984年5月北京第1次印刷
印数 1—14,400 册　　定价：0.85元
图书分类号：G25　统一书号：7201·22·5

前　言

由澳大利亚国立大学图书馆、国家图书馆和香港大学联合发起的“中文书目自动化国际合作会议”于一九八二年八月二十九日至九月一日在堪培拉举行。参加会议的有我国、新加坡、马来西亚、新西兰、日本、美国、英国等国以及我国台湾、香港等地区的四十个单位六十余人。其中包括中国国家图书馆（即北京图书馆）、新加坡国家图书馆、马来西亚国家图书馆、日本国立国会图书馆、英国图书馆、美国国会图书馆、香港大学中文系、文学院和美、澳、新、日及我国台北一些大学图书馆、研究机构等。会议的主要议题是：（1）汉字信息交换码的编制和使用；（2）汉字信息处理技术的进展；（3）中文机读目录的编制及其进展；（4）关于未来国际合作的途径。与会专家、学者、图书馆工作者共发表论文与讲话近二十篇，本辑选择、翻译了其中十五篇。

众所周知，中国是汉字的发源地，使用汉字的人数最多。近几年来，我国在汉字信息处理技术的某些方面取得了一定进展，为图书情报界建立中文机读目录和文献检索系统提供了有利条件。但由于我们的基础工作薄弱，起步较迟，比较起来这方面工作的差距还是很大的。我们希望这个选辑有助于图书馆工作者、情报工作者和从事计算机工作的同志了解当前汉字信息处理技术，特别是中文书目自动化方面的进展情况。

附带说明一点：在台湾省与会人士撰写的论文中有些提法和引用的机构名称，凡我们认为不宜直译的，均做了必要的文字处理。但仍不免有所疏漏，加之专业水平有限，错误之处谨请批评指正。

编者

一九八三年三月

图书馆学情报学

参考资料

第 5 辑

中文书目自动化国际会议专辑

目　　次

中国标准汉字信息交换码与国际汉字信息交换码

华北计算技术研究所　陈耀星

众所周知，汉字是世界上最古老的文字之一，大约已有五、六千年的历史。在这个世界上每四个人中就有一个人使用汉字。据估计，每年全世界出版的书目约有50万条，其中16%来自使用汉字的国家与地区。预料这一比例还将继续增长。

由于电子计算机技术的飞速进步，汉字信息的自动化处理也日益发展。但是，由于各国、各地区的汉字处理系统所用的汉字代码不同，导致相互间不能直接进行汉字信息的交换。因而，就不得不进行繁琐费时的代码转换，既给使用与维护带来不便；又增加了汉字信息处理系统的研制周期与生产成本，极不利于国际间汉字信息的交换与汉字自动化处理技术的发展。故制定一套国际汉字信息交换码乃是汉字资料自动化处理的必然趋势，也是世界上所有从事汉字资料自动化处理工作者与汉字信息利用者的迫切要求和殷切期望。

为此，中国将首先制定国家标准汉字交换码作为尽快实现这一目标的第一步。

我国经多年努力，于1981年5月1日颁布了国家标准《信息交换用汉字编码字符集基本集》(GB2312—80)，适用于尚未推行简化汉字地区及少数用字量较大用户的第一辅助集与第二辅助集，目前正在制定中。

现首先就中国汉字标准交换码的问题，作一简要的介绍。

1.收字总数及各集安排

中国标准汉字交换码计划收字五万七千余字，基本上按使用频度分配在九个集合内，每个集合收八千字左右。第一个集合，即基本集，现已收6,763字，它们是进行汉字信息处理的基本用字。其余各集均为辅助集，所收字亦属辅助用字。其中，第一、二辅助集适用于目前仍然使用繁体字及用字量较大的用户；第三辅助集作为保留标准区域之用；从第四辅助集开始的以后五个集合为特殊应用编制。

2.基本集

目前已经制定的是基本集，即双字节GO集，其中收如下字符：

一般符号	202
序号	60
拉丁字母	52
日文假名	169
希腊字母	48

俄文字母	66
汉语拼音符号	26
注音字母	37
汉字	6,763
计	7,455

其中6,763个汉字参考使用频度分为两级，第一级为常用字，有3,755字，按汉语拼音字母顺序排列；第二级为次常用字，有3,008字，按部首排列。

2.1.汉字的选择

基本集中6,763个汉字是以中国文字改革委员会与中华人民共和国文化部于1965年联合发表的《印刷通用汉字字形表》(6196) 为基础进行选择。该表是中国文字改革委员会，中华人民共和国文化部，中华人民共和国教育部以及中国社会科学院语言研究所从1953年就着手研究的“中国通用汉字表”与“标准字形表”结合而成。它既进行了各种用字的调查统计，又经我国最著名的语言文字学家、教育家胡愈之、叶圣陶、吕叔湘、丁西林、黎锦熙以及其他若干专门从事语言、文字、排印等方面的专家根据各种资料采用各种方法，前后十几年，几经选择调整，多次在全国范围征求意见修改而成。该表自1965年发表后至今已使用了十多年，实践证明：基本上满足一般书面用字的需要。

在制定基本集时，根据汉字信息处理与交换的需要，又增加了五百多个科技、地名、姓名用字。因此，对我国绝大多数汉字信息处理系统来说，只要具有这基本集中的六千七百余汉字就能基本满足各种使用要求，而不必预备数万个汉字。无疑，这将有利于降低汉字信息处理系统的成本，有利于汉字信息处理与交换技术的应用与推广。

2.2.分级

基本集中六千七百余字的使用复盖率，根据我国1975年进行的查频统计，达99.99%以上。尽管这是中国现代社会的通用汉字，但它们的使用频度却是很不相同的。根据20年代—70年代的各种汉字使用频度的统计，三千一四千个字的复盖率达99.9%左右，因而实际上只要备有这三、四千个字就能大体满足一般应用之需要。所以无论从使用，还是从便于设备的分档制造，将这三、四千字作为“一级常用字”区分开来是必要的。在综合参考了频度统计、构词能力、实际用处等基础上，选择了一些具有代表性的常用字表，再进行重合字数统计，最后选择出3,755字作为一级常用字，其余3,008字为“二级次常用字”。

2.3.排列

汉字的排列方法，目前采用的是我国社会流行最广、使用最普遍的《新华字典》、《现代汉语词典》中所采用的“音序法”和“部首法”。

一级常用字一般均知其读音。为便于检索，按汉语拼音字母顺序排列，多音字取其常用音，同音调字以起笔笔形：横、直、撇、点、折为序；二级次常用字中许多字较生辟，不易读音，故按部首排列。部首与一般字典相同，略有改并。同部首字按除去部首以外的笔画数排列，同画数字按笔形顺序排列。

2.4.字体与字形

字体以中国文字改革委员会编印的《简化字总表》(不包括第二次汉字简化方案) 及中华人民共和国文化部和中国文字改革委员会联合发布的《第一批异体字整理表》为准。字形一

律以《印刷通用汉字字形表》为准。

2.5.编码

为便于国内外汉字信息的处理与交换，汉字信息交换码应与电子计算机和信息处理方面的有关国家标准或国际标准尽可能地一致与兼容。因此，中国标准汉字信息交换码以国家标准（GB1988—80）《信息处理交换用七位编码字符集》（与ISO646一致）为基础，按照国家标准（GB2311—80）《信息处理交换用七位编码字符集的扩充方法》（与ISO/DIS 2022一致）进行编制。根据这一标准，基本集中的每个汉字采用两个字节表示，每个字节为七位二进制位，并分别与GB 1988—80中的七位二进制位组，即：从0100001—1111110共94个代码相对应。汉字及非汉字图形字符就排列在这94×94个编码位置所组成的代码表中。这个代码表纵向分成94个区由第一字节标识；横向将每个区分成94个位置，由第二字节标识。因此，这个代码表最多可收8,836个图形字符，构成一个集合，即：双字节GO集，其在七位代码制中的转义序列为

ESC　2/4.4/1

3.第一、二辅助集情况

基本集颁布之后，国内外已有许多厂商按照这一标准生产了各种汉字信息处理设备与系统，并已投放市场及安装使用。据一年多来的使用表明，基本集中的六千七百余字已满足绝大部分用户的用字需要。但是，对仍然使用繁体字的地区及用字量特别多的少数用户，例如中国的台湾省，一些大城市的户籍处理系统等，则还不够。为此，我们根据一些实际使用的字表和典型辞书，经各种统计和处理，从五万余字中又选择了一万六千余字，分配在两个辅助集中。每集各收八千余字并分别作为《信息交换用汉字编码字符集》的第一辅助集和第二辅助集。

（1）汉字选择

第一辅助集中的八千余字包括中国目前使用最广泛的《新华字典》、《现代汉语词典》、北京图书馆的《中文图书目录检字表》与中华人民共和国邮电部编制的《标准电码本》中几乎所有未收入基本集的单字，以及《辞海》中的绝大部分单字。

第二辅助集中的八千余字，主要根据各字在《中文大辞典》、《中华大字典》与《辞海》中的义项数及《汉语大字典》(正在编纂中）中的部分新收字进行选择，然后再加适当调整。

第三辅助集暂作为保留标准区域之用。

（2）排列

由于辅助集中的绝大部分单字都很生辟、难读，故各辅助集中的字均分别按部首排列。部首除增加“　　”部外，其余与基本集相同。

（3）字体与字形

辅助集的制定主要便于仍然采用繁体字地区的汉字信息处理，所以不收简化字，也不再进行类推简化。为了便于印刷，除第一辅助集外，其他各集均用旧字形。

（4）编码

两个辅助集的编码结构与基本集相同。第一辅助集，即：双字节G1集，在七位编码制中的转义序列是：

ESC 2/4 2/9 4/0，用移出字符 SO 调用，

第二辅助集，即：G双字节 G 2 集，在七位编码制中的转义序列是：

ESC 2/4 2/10 4/0，用单移位字符 SS 2 或带锁移位字符 LS 2 调用。

第三辅助集，即：双字节 G 3 集，在七位编码制中的转义序列是：

ESC 2/4 2/11 4/0，用单移位字符 SS 3 或带锁移位字符 LS 3 调用。

对于一个用字量大于八千，少于三万二千字的系统，其内部处理码可采用八单位编码，即在两个七单位编码的最高位前再增加一位第八位，见下图：

第二字节

B2 8 B2 7 B2 6 B2 5 B2 4 B2 3 B2 2 B2 1

第一字节

B1 8 B1 7 B1 6 B1 5 B1 4 B1 3 B1 2 B1 1

各集与第八位的对应关系如下：

B2 8	B1 8	
0	0	——对应基本集 （G 0）
0	1	——对应第一辅助集（G 1）
1	0	——对应第二辅助集（G 2）
1	1	——对应第三辅助集（G 3）

当使用字数超过三万二千时，系统内部处理可用三字节编码或其他形式编码。这样，在进行汉字信息交换时，采用七单位双字节编码形式。但是在系统内部处理时，可以用七单位编码，也可以用八单位编码；可以用双字节编码，也可以用三字节编码，或其他形式的编码。

根据汉字使用的实际情况与有关信息传输理论，对使用频度高的字采用短码，使用频度低的采用长码，这有利于提高传输效率。

上述的辅助集正在制定中，现在还不是国家标准，今天在这里介绍是想趁此机会听听各位专家的意见。

第二、关于国际汉字信息交换码的问题。

世界上中国和日本是主要使用汉字的国家，而中国是汉字的发祥地。但是由于各受其本民族历史、文化发展的长期影响，日本使用的汉字与中国现在使用的汉字无论在字的形、声、义还是使用频度等方面都有所不同。例如：

选择	中国	選択	日本
图书馆		図書館	
团结		団结	

中国的“勉强”一词在日本是“学习”、“用功”的意思，“手纸”在日本是指“信，书信”而不是“卫生纸”；“假，售，掉”等字在中国是使用频度很高的“一级常用字，而在日本的汉字信息交换中却是“二级字”。

尽管如此，就整个五、六万汉字来说，极大部分汉字的字形与意思还是类似的。即使对两国已经制定标准的六千多个常用汉字来说，许多字的含义仍然基本一致。因此，制定国际汉字信息交换码不仅必要，而且完全可能。

下面就如何制定国际汉字信息交换码谈一点看法。

1.字的选择与分级

据统计，汉字约有五、六万。目前中国正在编纂的《汉语大字典》收字五万七千。如果加上日本等外族汉字，则汉字总数当在六万左右。但是无论中国，还是日本，实际经常使用的字数远远低于此数。所以我们认为，为了提高汉字自动化处理系统的效率和降低其成本，应根据汉字的使用频度进行大致分级。

如上所述，中国目前正在制定的国家标准汉字信息交换码并按照使用频度将六万汉字大体分成三类：一类有六千七百余字，即基本集（GO集）其为进行汉字信息处理的基本用字。GO集实际可扩充到七千五百余字；二类汉字有一万六千～二万四千字，其为进行汉字信息处理的辅助用字。中国目前正在制定的是其中的一万六千字，即：第一、二辅助集（G1、G2集）；其余的字统统为三类字，它们几乎都是很少使用的“罕用字”。作为国际汉字信息交换码，也应根据使用频度进行分类。从中、日目前已经制定的国家标准汉字信息交换码来看，也可以分成三类，而且重点应首先放在第一、二类汉字的选择与编码的统一上。在我国内部，由于历史和政治的原因，中国大陆和台湾省目前在汉字使用和交换码的编制上也很不同。但我们注意将台湾同行编制的《中文咨询交换码》和我们的编制原则有很大的一致性。我们非常乐于借这次会议的机会与这方面的台湾学者认真地交换意见。

2.字形

在汉字中，有些音、义完全相同的字在形上却稍有差别。例如：（爭，争），（速，速）（花，花）等，这造成了文字的混乱现象，增加了学习和使用的负担，对汉字排检法的改进增加了困难，也造成了印刷、打字在人力和物力上的浪费。为此，中国经多年努力，于1964年确定了汉字字形的规范，统一了六千多个通用汉字的字形。例如上例三组括号中，后面的字形都是规范字形，称之为“新字形”，前面的字形叫“旧字形”。由于字形的统一，减少了因字形差异而造成的混乱现象。所以我们认为在考虑国际汉字信息交换码时，至少应对使用频度较高的第一类字，采用规范化的字形，以利于汉字信息的交换及输入、输出等处理。

3.字体

（1）异体字

有些汉字的音、义完全一样，但字形却有较大差别，例如：胡（鬍），侖（崙崘）等。习惯上将括号内的字称为括号外的字的“异体字”。而且，有些字往往有多个异体字，这种异体字同样造成了汉字的繁难，增加了学习和使用的负担。为此，中国对异体字也进行了整理。1956年公布了第一批异体字整理表。计1，865字。经整理后精简掉1，055字，使常用汉字大为精练，实用。

作为国际汉字信息交换码，对常用的第一类字亦应本着“少而精”的原则，将音、义完全相同的字只收一个体。异体字一律不收。这无论是对汉字信息的交换，还是对设备的制造都是有益的。

（2）简化字

汉字的繁难不仅表现在字数多，字形多变，而且还表现在有些字的字形复杂，笔画数很多难以学习，不便使用。因而中国与日本都对常用汉字中较繁难的汉字进行了简化。特别是中国，从1956年起先后公布了2,238个简化字。并在全国推行。到目前为止，中国的汉字简化工作已接近结束。所以中国的国家标准汉字信息交换码基本集中的汉字一律采用简化字形，目前国际上采用中国简化字的还有新加坡。

作为国际汉字信息交换码，我认为第一类汉字也应采用简化字体，以利信息交换及汉字字形产生器等有关设备的制造。

4.字序与字音

汉字集合实际上是一个无定数的无序集合。各元素（单字）之间不能比较值的大小，没有一定的排序规则。所以汉字的排检法从东汉许慎算起，成千上万的学者研究了一千九百年，还没有一个被公认的标准方案。中国自1956年推广汉语拼音方案以来，许多辞书以及图书资料卡片采用了汉语拼音字母顺序排检法，即："音序排检法"，或音序与部首等混合排检法。

目前在中国使用最广泛的辞书，如《新华字典》、《现代汉语词典》，正文均按音序排列，同时附加部首检字表。所以中国标准中除基本集内易读音的一级常用字按音序排列外，其余一律按部首排序，以便检索。

我认为在收集编制中国际汉字信息交换码时，其排序方法参照中国标准中的排序方法，可能是比较理想的。

5.编码

为了便于国际间进行汉字信息交换与共享汉字资讯，国际汉字信息交换码在有利于信息交换，不影响使用效率的前提下，应尽可能根据有关国际标准进行编码。为此，我们认为国际汉字信息交换码以ISO646为基础，按照ISO/DIS 2022进行编码是合适的。

在大多数情况下，为省却信息交换时的转义序列，我们认为根据ISO2022将第一类汉字编入双字节GO集，第二类汉字编入双字节G1集、G2集和G3集，第三类汉字编入其他集合对汉字信息的交换与处理较为有利。

6.汉字控制功能与通讯控制规程

众所周知，汉字是表意文字，与表音文字相比，它有一些独特之点。利用现有的ISO646中的CO集与ISO/DIS 6429还不足以有效地进行汉字信息的传输与交换。所以我们认为在制定国际汉字信息交换码的同时，还应该制定有关汉字信息交换用的汉字控制功能与有关通信规程。

为了便于国际间的汉字信息交换，我们认为可以在ISO646的GO集与ISO/DIS6429的基础上再增加一些为传输汉字信息所必需的特有控制功能；而通信控制规程可参照ISO1745等有关国际标准进行制定。

如上所述，世界上主要使用汉字的国家是中国，其次是日本。作为国际汉字信息交换码，如果中、日两国能取得一致，则无疑对形成国际汉字信息交换码是一个巨大的促进。目前中、日两国均已制定国家标准汉字信息交换码，并且编码的方法都是根据ISO/DIS2020，

可见编码原则是一致的。但是，正如前述，由于中、日两国目前使用的文字无论在字形、字音、字义、字数还是使用频度等方面都有所不同，因而每个字，即使两个字形完全相同的字在代码表中的位置也不同，因而其代码也不同。根据目前已经制定和正在制定的国家标准汉字交换码的情况来看，中国使用的字数较日本多，如果比较一下中日两国的标准，则可以发现日本标准中的汉字除少数外，在中国标准中都有。所以我们认为，如果在中国标准的G 3集中补入该标准中没有的日本汉字，则中国标准汉字信息交换码就可以作为国际汉字信息交换码。

具体应用时可以将中国标准中的基本集，即G0集中的字形作为意义相同的繁体字，异体字，日本简化字等的代表字形，其编码可用于汉字信息的直接交换，亦可用于记录MARC磁带等介质上的信息。如果需要输出资讯的其他字形，则调用相应的辅助集即可。当然，作为一个完整的处理系统，为了达到上述这一点，可能还需建立各种汉字代码对照表，汉字字典，汉字管理器等有关支持软件。

诚然，中国是一个汉字母国。但由于是发展中国家，所以在用电子计算机处理汉字信息方面还刚开始。我们还没有什么经验，在制定汉字信息交换码方面也是如此。今天向各位介绍中国标准汉字交换码，同时谈了一点对国际汉字信息交换码的初步看法，主要是想借此机会向来自世界各地的专家、同行学习，听听各位朋友的意见。以便改进我们的工作。二是通过此次聚会增进我们之间的相互了解，以促进国际汉字信息交换码的尽早诞生。

（上接13页）

中，甚至在某些标准中。

CCCII第二卷于1981年2月14—19日以两卷集发表，并在台北国际图书馆自动化委员会中传阅，该卷收有16,162＋4,807＝20,969个汉字（包括10,793个异体字）。两卷书名分别为：

（1）CCCII第二卷的符号及汉字表

（2）CCCII第二卷的异体字表

CCCII的正式本将收有21,994个汉字及11,008个有关的异体字，共计33,002个汉字，将于1983年初发表。

收集汉字、确定每个字的笔形、确定每个汉字的异体字的工作以及许许多多有关分析、统计和体力上的工作都是由汉字分析组的同行们做的。由于他们的辛勤努力才有了今天这篇文章。

（乔凌译）

汉字信息交换码及其国际应用的可行性

台北铭传商学院　黄克东等

引　言

汉字信息交换码（以下用英文缩写CCCII表示）的设计工作主要包括两大方面：1．建立代码结构；2．将汉字组成代码。CCCII第一卷于1980年4月发表，显示了CCCII的主要特点，但是只容纳了4,807个最常用汉字。因此，第一卷仅适用于某些使用汉字数量有限的一般应用。

CCCII第一卷的工作只是整体工作的一部分。CCCII的最终目的是实现各种汉字信息的交换。

CCCII第二卷的正式本将于1983年初发表，将容纳21,994个汉字及有关的11,008个异体字，共计33,002个汉字。

CCCII的结构

1.1　兼容性

为了满足某些国际业务的需要，CCCII是按照ISO646的规则设计的，完全可以在7字位代码的情况下使用。中文图书及非书资料MARC格式就是一例。

根据ISO2022，在控制设备都保留与7字位结构相兼容的同时，CCCII还可用于8字位的情况。CCCII转义序列为ESC，2/4，4/2。这是一种用多字符集表示扩充图形的方法。在CCCII当中，每个汉字由一个3字节向量表示。这种结构在图2中有所描述。

1.2　对CCCII第一卷的某些修改

对CCCII第一卷在一些技术上有所改变：

（1）每个平面的第一段（如图3所示）将留给处理汉字串所专门需要的控制码使用。

（2）增添了Û和￥两个符号。

1.3　层

从第一平面开始，每6个连续平面组成一层。（见图4）第一层为规范汉字层。从第二层往后都用于异体字和将来的扩充。

二、对设计CCCII当中预料到的一些关键问题的想法

2.1　常用汉字

CCCII第一卷中的4,807个汉字为常用字。此字符集无论在任何应用当中都不可缺少。事实说明，在数字、写作和新闻印刷中，此字符集的使用率超过95%。

2.2　通用数据处理字符集

从1979年开始，进行了一项汉字调查与分析。这次调查的目的是想找到一个能满足台湾

当前业务数据处理所需（如：图书馆应用，人口统计应用等）的汉字字符集。这次调查的情况见表格1。这个字符集共收有约21,994个汉字，包括最常用汉字。

2.3 汉字异体字

去年，对第2.2段中收集的每个汉字的异体进行了一次调查和分析。一个字的异体包括它的繁体、古体、简体和某些地区使用的各种异体。就简体而言，这个字符集包括了古代中国、大陆中国及海外所使用的各种简体。

CCCII第一卷的草案包括了16,162＋4,807＝20,969个汉字及它们的异体（收有10.793个异体字）。此草案于1981年2月14—19日发表，并在台北国际图书馆自动化委员会中传阅。

目前规定，一个汉字的异体字数量不得超过6个，而收集的异体字总数为11,008个。正式的异体字符集将于1983年初发表。

2.4 汉字的分块

汉字根据其使用率分属不用的字符块。第一字符块用于4,807个最常用汉字。第二字符块表示第一块之外的通用数据处理所需的字符。其余汉字都归第三字符块。这种分布见图1。

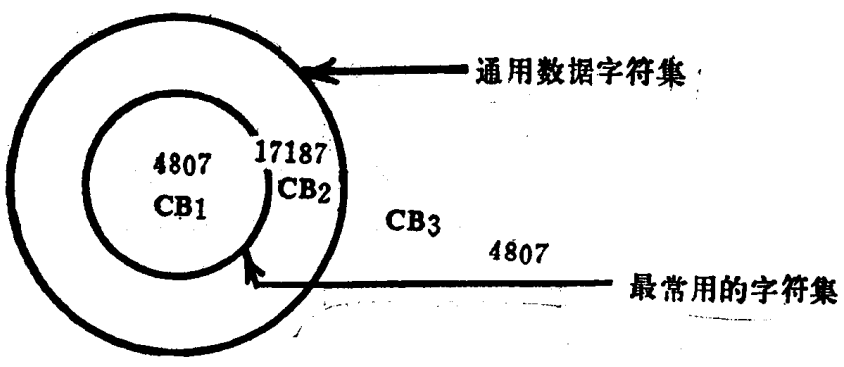

图1：汉字的划分

2.5 排序

在每个字符块中，首先根据康熙字典的部首顺序，然后按照笔划数进行排序。对那些具有同样部首和同样笔划数的字，再按照笔划顺序排。笔划的先后顺序如下：

（1）点 （2）横 （3）直 （4）撇 （5）捺

2.6 异体字的处理

一个汉字有异体字，这或许是中国语言文字的一个独具特征。一个汉字和其异体往往具有完全相同的发音和字义，但笔划形象不同。通常在书写时可通用。但是当异体字被来区别人名、地名和事物时，它们被看作是不同的字，不可混用。在CCCII中，异体字代码具有相似的两个最靠右边的字节——B_2，B_1，作为其对应汉字的代码。换句话说，异体字被放在与其对应汉字的同段同位上，只是不在同一平面。这种安排可以单独检索异体字，并且在需要的时候可容易地互换编入程序。表格2举有异体字范例。

2.7 人物姓氏及字符集

根据程衡雄教授最近在台湾进行的一项关于中国姓氏的调查和统计，中国有9,177个姓氏。

任何一个希望得到国际间承认的计算机代码，和希望使汉字代码标准以适应书目处理和人口统计数据处理用途的汉字字符集至少应当包括足够的处理所有中国姓氏所需的汉字。否则，就可能面临两个严重问题。

（1）因为缺少某些字而不能进行数据处理。

（2）如果所缺汉字暂时被添进了字符集中的用户控制空间，也不便于计算机之间的通讯，因为该具体代码在互换各方的协议中原先没有。

为解决以上提出的问题，完整的CCCII希望容纳现存所有汉字，据估计有80,000多个。

2.8 CCCII的字符子集

CCCII希望成为一个汉字信息交换的国际工具。因此，中国语言及信息处理中用到的所有符号都将收入并编成码。然而，从用户的角度看，许多用途中不需要这样完整的字符集。CCCII在设计中考虑到了这个问题，使用户可以选择一个适合其具体需要的CCCII子集。

CCCII子集有2个子代码字节长。(见表格3说明)在这种情况下，通过互换各方的同意，只需2字节就可以表示一个汉字。请注意，CCCII的所有这些子集的代码结构都与ISO646完全兼容。

2.9 内存94基数代码

第一层的前6个平面（见图4）共有94×94×6=53,016个位置。由于此数小于2^{16}=65,536，因此，利用94基数转换，第一层中的3字节代码可以压缩成一个16字位或者两个8字位字节，作为无符号二进制数。这种转换是一种一对一、双向、单独转换。因此，如果所有汉字是属于第一层的，此代码可作为大批数据储存的内码，以节省储存空间。

国际应用的可行性

3.1 缺少一个单独的数据处理系统的问题

目前，国际上已有许多中文数据处理系统。但是每个系统的数据处理能力都相当有限。导致这种现象的原因之一是缺少一个识别计算机中汉字的有力工具。

通常，一个计算机可提供的汉字检索键只有一种。而且这种检索键直接关系到中文输入的方法、或者计算机系统所使用的汉字内码的排序序列。例如，一个计算机使用汉字的语音代码作为输入方法，那么机器就可以向你提供语音代码作为检索键。如果想通过汉字笔划数来查会怎样呢？那么就根本行不通，除非计算机的记忆储存进行了所有的笔划信息。计算机最有可能用笔划数作为排序规则来制定汉字内码，这样的话就可以提供以笔划数来检索汉字，解决信息检索需要。由于大多数中文数据处理系统的输入方法也同制定内码的排序规则有密切关系，因此只能允许通过一种键检索汉字。公共服务系统对于这种状况是无法容忍的，因为用户们可能有各种大量的问题必须用不同的键进行工作。况且，仅仅因为计算机系统的服务功能有限而禁止用户使用他所熟悉的某种正常方法来输入或处理汉字也是不理想的。

关于数据库结构的设计逻辑，CCCII不仅为各种异体字保留了足够的空间，而且还可以简便地同各种文档连接，编制发音索引。

本文的主要内容就是提出一个解决这一问题的方法。

3.2 汉字索引的特点

汉字索引的问题就是怎样通过汉字的各种专门信息（如发音、笔划、部首、电报码及三

角号码等）识别、检索或称呼一个汉字的问题。显然处理罗马字的经验在这里无法直接应用。当我们查一个汉字时，往往通过同该字有关的某些信息来查。因此这不是一个简单的代码对应问题。而是一个连接检索问题。为此我们必须储存所有必要的信息以保证能够进行连接检索。

3.3 相互参见数据库的组成部分

组成相互参见数据库的信息是在编制汉字索引中常常被用来作为检索键的汉字的各种专门信息。我们所收集的信息包括以下几种国际上使用的项目。

（1）CCCII

（2）部首

（3）笔划

（4）笔顺

（5）字源

（6）语音代码。包括：国语，韦氏、耶鲁、拼音及刘氏

（7）电报码

（8）三角号码

（9）四角号码

（10）各种中文数据处理系统的内码

将内码包括进去的目的是企图向现有数据处理系统提供对那些信息的相互参见。并以此命名了数据库。

3.4 数据库的结构

在刘大健进行的一项语音体系的比较研究基础上，建立了汉字索引相互参见数据库的框图。完全由于其有效性，此系统的一部分已被一些公司用于小型计算机，甚至微型计算机。

3.5 文档的数据结构

图7描述了这些文档中每个记录的数据结构。据估计，这个拥有约50,000汉字及15,000异体字的数据库所需的总面积还不到15兆字节。所有文档中之最小者是发音索引文档，仅需100千字节就可容纳所有常用的发音系统。这部分是根据的刘大健的一项语音体系的比较研究。由于其有效性，此系统的一部分已被一些公司用于小型计算机，甚至微型计算机。

中文相互参见数据库磁带档的注释

4.1 注释

这个注释阐明了汉字相互参见数据库检验数据库中三个文档的格式及记录结构。尽管此检验数据库仅存有4,807个汉字，其格式及记录结构将被用于即将发行的以下各个数据库。

4.2 磁带

这个磁带中有三个文档，即相互参见表，汉字图象文档，和拼音表。每个文档及其记录都将按照以下格式进行描述。三个文档均记录在九磁道磁带上，记录密度为1,600BPI；无标号非组块式。

文档名称

记录数量

记录方式

记录长度

每个记录的字段数

记录格式，以图表示

4.3 第一文档

相互参见表

4,807个记录

汉字方式

39个字／记录

9 个字段

记录格式：

1	2	3	4	5	6	7	8	9
CCCII	康熙字典	笔　划	笔　顺	三角号码	四角号码	电报码	林氏序列码	连接发音索引
Z6	D4	D2	D5	D6	D4	D4	D4	D4

图例说明：Z表示十六进制，如Z 6 表示 6 个十六进制数；D表示十进制，如D 4 表示 4 个十进制数；A表示ASCII，如A 9 表示 9 个ASCII符号；N表示ASCII空格，如 N21 表示ASCII代码中的21个空格。

4.4 第二文档

汉字图象文档

4,807个记录

二进制方式

128个字节／记录

32个字段

记录格式：