汉语知识丛书

HÀNYǓ ZHĪSHI CÓNGSHŪ

中文信息处理与汉语研究

冯志伟编著

商务印书馆

汉语知识丛书

中文信息处理与汉语研究

冯志伟 编著

商务印书馆
1992年·北京

汉语知识丛书
ZHŌNGWÉN XÌNXĪ CHǓLǏ YǓ HÀNYǓ YÁNJIŪ
中文信息处理与汉语研究
冯志伟著
商务印书馆出版
（北京王府井大街36号 邮政编码100710）
新华书店总店北京发行所发行
河北香河县第二印刷厂印刷
ISBN 7-100-00440-8/H·167

1992年12月第1版 开本 787×1092 1/32
1992年12月北京第1次印刷 字数142千
印数 0—1 500 册 印张 7

定价：2.70 元

目　录

前　言

当代科技革命的主要特征，是以电子计算机为支持手段进行信息处理。语言是人类最重要的交际工具，是信息最主要的负荷者，语言文字的信息处理，是当代科技革命的一个重要内容。在这无比壮阔的科技革命的浩浩洪流中，仅有不到40年历史的电子计算机，正在冲击着已有4000多年文明历史的古老语言——中文。

在我国，将电子计算机应用于事务处理、计划调度、办公室自动化、印刷排版、情报检索、机器翻译、人机对话等方面，都离不开中文，因为所有这些方面的信息，都是以中文作为其载体的。

中文特指汉族的语言文字，也就是汉语的书面形式和口语形式。中文的书写形式是汉字，而且中文在语音、词汇、语法等方面，又有它独自的特点，这样，当用电子计算机来处理由中文记载的各种信息的时候，就自然会提出许多新的问题，对这些问题的研究，就形成了一门新兴的综合性技术学科——中文信息处理(chinese information processing)。

可见，所谓"中文信息处理"，就是利用电子计算机对汉语的书面形式和口头形式进行信息处理。

对于中文信息处理这个术语常有两种误解。一是把它的概念缩小，把中文仅只看作是汉字，使中文信息处理等同于汉

字信息处理；另一是把它的概念扩大，把中文看成是中国各民族的语言文字，从而把各少数民族语言文字的信息处理也纳入中文信息处理的范围之内。

汉字信息处理是中文信息处理的一个关键部分，它主要包括汉字的编码、输入、输出，汉字的识别，汉字的字频统计等，这些都是中文信息处理的重要内容，也是中文信息处理基础性的工作。但是，中文是一个多层次的结构，以书面汉语来说，汉字是中文最基本的层次，由汉字可以组成词，这是比汉字更高的层次，由词可以构成句子，这又是比词更高的层次。中文信息处理除了研究汉字信息处理之外，还要研究汉语词的自动切分和处理，汉语句子的自动分析和生成，以至于汉语语义的自动分析和加工等等。

少数民族语言文字的信息处理是自然语言信息处理的组成部分，中文信息处理也是自然语言信息处理的组成部分。不论是中文信息处理还是少数民族语言文字的信息处理，都是自然语言信息处理这个学科的分支学科，它们之间的关系是并列的，而不是彼此隶属的。因此，我们不能把少数民族语言文字的信息处理纳入中文信息处理的范围之内。

中文信息处理的研究在我国已有 30 年的历史了。早在 1956 年，我国学者丁西林就提出了中文电动打字机的问题。1956 年 10 月，钱文浩在《科学通报》上发表了《文字与通讯》一文，提出了用汉字编码方法把汉字转换成信息代码进行传输的理论问题，引起了国内外人士的关注。1958 年，新华社、邮电部、中央机要局合作设计了鼓轮式中文电传机。1959 年国庆前夕，中国科学院语言研究所和计算技术研究所合作，在

我国试制的第一台104电子计算机上进行了俄汉机器翻译试验，这是中文信息和计算机最早的结合。1969年9月，邮电科学研究院试制成功我国第一台电子式中文电报快速收报机，揭开了用计算机技术处理汉字信息的序幕。1974年8月9日，中国科学院、一机部、四机部、新华社和国家出版局向国家计委及国务院提出“研制汉字信息处理系统工程的请示报告”，9月24日，国家计委批准把汉字信息处理系统列为1975年国家科技发展计划。1978年，云南大学无线电系张其濬教授给中央领导写信，提出研究汉字编码的问题，经中央领导同意批转，于1978年在青岛成立了全国汉字编码研究会，这是我国中文信息处理方面的第一个全国性学术团体。1979年在上海嘉定召开了“信息处理交换用的汉字字符编码标准”讨论会，1981年3月，国家标准局发布了国家标准GB 2312—80《信息交换用汉字编码字符集——基本集》，并于1982年得到国际标准化组织ISO的承认。1981年6月27日，中国中文信息研究会在天津成立，选举钱伟长为理事长。接着陆续成立了几个专业委员会。1982年5月在厦门成立基础理论专业委员会，12月在南京成立汉字信息处理专用设备委员会，1983年3月在成都成立汉字编码专业委员会，5月在武汉成立汉字信息处理系统专业委员会与自然语言处理专业委员会，同月，在武汉召开了中国中文信息研究会第二次全国学术会议。1983年10月12日—14日，中国中文信息研究会与联合国教科文组织联合在北京召开中文信息处理国际研讨会（ICCIP'83），来自美国、日本、加拿大、澳大利亚、联邦德国、法国、英国等15个国家和地区的来宾77人，国内代表98人，会上宣读

了 77 篇论文；研讨会期间，与中国计算机技术服务公司、中国计算机用户协会联合举办了“计算机中文信息处理展览会”，展出了 38 项成果。截至 1983 年底，我国累计生产汉字终端 400 余台，汉字 15×16 点阵掩膜库 1000 套，汉字打印机 185 台，笔触式汉字键盘 320 台，年产值约 2500 万元。我国中文信息处理的研究出现了一片欣欣向荣的景象。

欧美和日本也有不少学者研究中文信息处理，美国有中文计算机学会(Chinese Language Computer Society)，日本有信息处理学会 (Information Processing Society of Japan)，都进行中文信息处理的研究。

另外，日本于1979年成立了非公开的“日文输入法研究委员会”，于 1981 年改为公开的“日文输入方式研究会”，进行日文假名和汉字输入方式的研究。1985 年开始，把研究领域进一步扩充，成立了“日本语文书处理研究会”。日本在汉字处理方面的研究与中文信息处理的研究有许多共同之处。

日文和中文都使用汉字，语法结构与印欧语有很大差别，因此，日文信息处理的研究，对于中文信息处理是很有参考价值的。日本在 60 年代中期到 70 年代中期，首先在对汉字有特殊需要的新闻出版部门使用了日文电子照排系统，日文输入主要依靠汉字键盘和假名键盘，输出技术使用活字和点阵，典型的设备是汉字电传打字机。70 年代中期到 70 年代末期，有了汉字终端，开始使用汉字-图形输入板、喷墨打印机和终端电子摄影技术，日文信息处理技术推广到了银行、保险公司、百货商店等部门，用于管理顾客名单和打印地址。80年代初期出现了日语文字处理机、日语小型商用计算机和日语个

入计算机，开始了假名—汉字转换输入、光学字符识别以及语音识别与合成的应用性研究，造出了一批产品，日文信息处理系统逐渐向办公室普及，进行所谓的“办公室自动化”(Office Automation，简称OA)，估计不久将广泛用于企业事业单位和家庭。此外，值得注意的是，日本的机器翻译研究近年来已逐渐走向实用化、商品化。

同日文信息处理相比，我国中文信息处理的研究还比较落后，我们应该努力学习国外的先进技术，结合中文的实际，把中文信息处理的研究提高到新的水平。

中文信息处理是一门新兴的综合性技术学科，它是语言学、计算机科学、自动化技术相结合的产物，中文信息处理的发展，有赖于这几个方面的专家的通力合作。但是，由于中文信息处理的对象是中文，因此，它与语言学有着更为密切的关系。语言学所揭示的关于语言结构的各种规律，是中文信息处理不可缺少的研究依据。例如，汉字结构的精细研究，有助于汉字编码和汉字识别的发展，实验语音学的研究以及汉语词汇、语法、语义及其相互关系的研究，为机器翻译和汉语人机对话提供了可靠的语言数据。可以说，中文信息处理的每一步发展，都离不开语言学的研究成果。另一方面，中文信息处理的研究，可以在很多方面对传统语言学和现代语言学的许多结论加以检验、订正和补充。例如，利用对语言文字各种要素的统计数据，可以从量的描述得出质的评价，从而丰富和发展传统语言学，机器翻译和人机对话研究中所发现的许多自然语言自动分析和理解中的问题，可以检验转换生成语法的结论。而且，在中文信息处理中涉及到的有关汉语的语音、

文字、词汇、语法、语义的各种问题，也必将成为汉语研究的新课题，促进汉语研究的现代化。

本书旨在探讨中文信息处理与汉语研究的关系，向语言学工作者介绍中文信息处理的基本知识，向计算机工作者介绍与中文信息处理有关的语言学问题。由于中文信息处理的研究对象是中文，本书所涉及的各种语言学问题均以汉语为主。目前已经开展的中文信息处理项目主要有汉字的输入与输出、计算机言语统计、计算机自动切词、机器翻译、人机对话等，这些项目与汉语研究的关系都比较密切，有着浓烈的语言学色彩，汉语语言模型是中文信息处理的基础性研究，与汉语研究更是息息相关，本书将着重地介绍这些方面的原理、方法以及主要成果。此外，我国近年来还研制了许多中文信息处理系统，在发展汉字终端的同时，还研究了各种汉字系统软件，如汉字操作系统、汉字高级语言、汉字数据库等，这些当然也是中文信息处理的十分重要的内容，但由于它们的技术比较专门，与语言学的关系也不如上述那些项目密切，本书中就略而不谈了。

中文信息处理与汉语研究的问题，是关系到我国语言学研究现代化的重要问题，涉及的范围很广，作者学识有限，孤陋寡闻，以蠡测海，难免有片面和不妥之处，敬请读者批评指正。

本书力图反映出我国学者，特别是汉语研究者在中文信息处理研究中的新成果，因此，在本书写作过程中曾参考过许多同志的论文和著作，没有这些同志的出色工作，没有他们的极为宝贵的研究成果，本书是写不出来的。本书在每章末均

列出了参考文献，在本书出版之际，谨向这些同志深表谢意。商务印书馆周行健同志对于本书的体例和写法提出过十分有益的建议，特在此致谢。

冯 志 伟

1986 年 5 月 1 日于北京

第一章　汉字的输入与输出

汉字是一个大字符集。公元100年许慎的《说文解字》就已收了9353个汉字，后来，汉字数目逐渐增加，公元543年顾野王的《玉篇》已收16917个汉字，公元1008年陈彭年等的《广韵》收26194个汉字，公元1615年梅膺祚的《字汇》收33179个汉字，公元1716年陈廷敬等的《康熙字典》收47043个汉字(增补以前的汉字数是42174字)，1959年日本学者诸桥辙次的《大汉和辞典》收49964个汉字，而最近出版的《汉语大字典》收的汉字已超过56000个。中文信息处理关心的主要是现代汉字，其字数要少一些，但仍然是形形色色、琳琅满目、错纵复杂的一个大字符集。计算机要处理中文信息，首先要解决汉字这个大字符集的输入输出问题。

第一节　汉字输入

汉字输入是中文信息处理的关键，不解决这个问题，中文信息处理就落为空谈，成为无米之炊。

70年代以来，我国学者对这个问题作了大量的探索和研究，到目前为止，提出的汉字输入方法大致有六类：编码输入法、整字输入法、拼音一汉字转换法、印刷体光学输入法、手写输入法、声音输入法。下面分别加以介绍。

1．编码输入法：

所谓编码输入法，就是给汉字规定一种便于计算机识别的代码，使每一个汉字对应于一个数字串或符号串，从而把汉字输入计算机。

一般地说，适合于计算机输入汉字的编码方案，应该具备如下特点。

① 简单易学：汉字编码方案应该有较强的规律性，规则简明扼要，操作人员不必经过很长时间的学习就可以使用。

② 重码率低：代码与汉字应该尽可能地一一对应，不能一个代码对应于两个或两个以上的汉字，产生重码。一般说来，一个实用的汉字编码方案，其重码率应控制在3%—1.5%以下。

③ 速度较快：操作人员经过短期培训后，可以实现盲打输入。

④ 覆盖面大：对于出现频度高的汉字都应该进行编码，以便覆盖绝大部分用汉字写的文章。

⑤ 成本低廉：汉字编码输入系统设备的成本不宜太高，要有较好的性能价格比，最好能使用普通电传打字机的小键盘，而不必另外添置大量的专用设备。

计算机汉字编码是目前举世瞩目的重大课题，学者们为此倾注了大量的心血，提出的编码方案已有500多个，其中上机通过实验和已被采用的编码方案已达数十种之多。这些汉字编码方案大致可分为四种：

(1) 形码：根据汉字的字形来进行的编码。如笔形编码

法、五笔字形编码法、三角编码法等。

笔形编码法把汉字的笔画分为一(横)、丨(竖)、丿(撇)、丶(点)、乛(折)、乙(弯)、乂(叉)、□(方)八类，分别用1、2、3、4、5、6、7、0等数字来代表，横、竖、撇、点为单笔，折、弯、叉、方为复笔。汉字代码是不等长码，最大码长为9码。

五笔字形编码法把汉字分解为部件，按部件进行编码，平均码长为4码，使用高频字简码和词汇码后，平均码长为2.8码。

三角编码法是对用于查字典的四角号码查字法改进而成的，它不使用汉字四个角上的基本笔画作为代号而采用汉字三个角的基本部件作为代号，由于基本部件符号的数目比四角号码查字法中的基本笔画符号的数目多得多，所以，重码字大为减少，不必采用四个角来编码，而只需用三个角来编码。

设计这样的形码时，必须对汉字的笔画及部件进行精细的分析。

汉字的笔画是很复杂的。传统上一直认为单笔有一(横)、丨(竖)、丶(点)、丿(提)、丿(撇)、乀(捺)六种，为了减少基本码字，笔形编码法把单笔分为横、竖、撇、点四种，而把“提”归入“横”，把“捺”归入“点”。这是因为：

第一，“提”和“横”的运笔方向相同，均是自左而右运笔，“点”和“捺”的运笔方向相同，均是自左上而右下运笔。

第二，“提”和“横”常常互相转化。汉字中末笔是“横”的独体字或部件，当它们出现在合体字的左旁时，其末笔的“横”就变为“提”，如“土、立、王”的末笔都是“横”，当它们出现在“块、堵、地、竭、端、站、玩、理、球”等合体字的左旁时，其末笔

的“横”都转化成了“提”。“点”和“捺”也常常互相转化。为了使汉字的结构方正美观，在许多情况下，“捺”要转化成“点”，如“泰、膝、漆、黍”等字中的“氺”是由“水”变形而成的，“水”的末笔由“捺”变成了“点”；还有“令、仓”的第二笔“捺”在“领、创”中均转化为“点”；“木、大”的最末一笔“捺”在“相、达”中均转化为“点”；而“奇、囚、困、菌”等字中的“大、人、木、禾”都是由“大、人、木、禾”转化而成的。

第三，汉字中没有以“提”起笔的字，因此，“提”和“横”虽可互相转化，但基本笔画取“横”不取“提”；汉字中也没有以“捺”起笔的字，因此，“捺”和“点”虽可互相转化，但基本笔画取“点”不取“捺”。

把“提”归入“横”，把“捺”归入“点”之后，单笔只有横、竖、点、撇四种。

汉字的复笔笔画形状多样，仅是“折”这种复笔笔画就有22种，其中，单折笔形14种：丿（竖钩）、冫（弯钩）、乀（斜钩）、㇏（卧钩）、㇂（竖弯）、㇄（竖弯钩）、亅（竖提）、㇀（横钩）、㇇（横折）、㇆（横折钩）、㇉（竖折）、㇈（横撇）、㇌（撇折）、㇎（撇点），复折笔形8种：㇇（横折弯钩）、㇅（竖折弯钩）、㇊（横折提）、㇋（横折折撇）、㇍（横撇弯钩）、㇎（横折折折钩）、㇇（横折弯）、㇅（竖折撇）等。笔形编码法把它们归并为“折”、“弯”两种，再加上“叉”、“方”，复笔笔画共4种。这样，便把复杂多样的汉字笔画归并为8种基本笔形。

设计形码还要分析汉字的部件。现代汉字的形体可以分为汉字、部件、笔画三个层次，笔画是最低层次，部件是中间层次，汉字是最高层次。由笔画组成部件，再由部件组成汉字，

因此，部件是汉字形体结构中的枢纽性单位，它上承汉字，下启笔画，起着承上启下的作用，部件比笔画完整，又比汉字简单，在现代汉字的形体结构分析中应该给以特别的注意。

从符号数目的多少来看，汉字层次最高，表示一个字只需用一个符号，但它所用的符号总数很多，如果有6万个汉字，就得用6万个不同的符号；笔画层次最低，表示一个汉字所用的符号数目最多，而它所用的符号总数最少，只有横、竖、点、撇、折、弯、叉、方等有限的几种；部件处于中间层次上，表示一个汉字所需的符号数目适中，而它所用的符号总数也适中。例如，五笔字形编码法把汉字的部件归并为664个，进行了认真的统计分析，为部件的优选和部件在键盘上的合理布局提供了理论根据。

(2) 音码：根据汉字的读音来进行的编码。音码一般以汉语拼音方案为根据，汉语拼音方案已有20多年的历史，在国内外广泛推行，它的合理性、群众性和科学性已得到公认，而且，它是以国际通行的字符集(拉丁字母)以及它们相近的发音为基础制定的，有利于国际交流，在计算机上使用汉语拼音，还有利于推广普通话，有利于我国文化教育事业的发展。

采用音码的最大困难是区分同音字的问题。汉字的音节共408个，而汉字的数目成千上万，这就必然导致大量拼音同音字的出现。赵元任先生曾写过一篇《施氏石狮史》的短文，全文如下："石室诗士施氏嗜狮，试食十狮；氏时适市视狮十时，适十狮适市，适施氏适市，氏视是十狮，恃矢势，使是十狮逝世。氏拾是十狮尸，适石室；石室湿，使侍拭石室，石室拭，氏始试食十狮尸；食时，始识是十狮尸，实十石狮尸。试释是