信息处理与现代汉语问题探索

年玉萍 著

陕西人民出版社

宝鸡文理学院重点学科建设专项经费资助

信息处理与现代汉语问题探索

年玉萍 ◎ 著

陕西人民出版社

图书在版编目（CIP）数据

信息处理与现代汉语问题探索 / 年玉萍著 . —西安：陕西人民出版社，2011

ISBN 978－7－224－09884－6

Ⅰ. ①信… Ⅱ. ①年… Ⅲ. ①汉字信息处理—研究 Ⅳ. ①H127

中国版本图书馆 CIP 数据核字（2011）第 189022 号

信息处理与现代汉语问题探索

作　　者　年玉萍
出版发行　陕西出版集团　陕西人民出版社
（西安北大街 147 号　邮编：710003）

印　　刷　西安市建明工贸有限责任公司
开　　本　787mm×1092mm　16 开　13.75 印张　2 插页
字　　数　250 千字
版　　次　2011 年 10 月第 1 版　2011 年 10 月第 1 次印刷
书　　号　ISBN 978－7－224－09884－6
定　　价　28.00 元

序

21 世纪，跨入了信息时代，科技在迅猛发展，知识在不断更新，作为信息首要载体的语言文字不能不有所变化。

语言文字工作由方针政策的制定贯彻转入到立法、执法时段，推广普及普通话，推行规范汉字已是《宪法》和《国家语言文字法》定下的法制条款。随着计算机的应用发展，规范汉字，汉语词、句用法不仅是人们社会交际的需要，更是人机对话信息处理的迫切要求。而目前高等院校现代汉语教学远远落后于时代，教材内容陈旧，学生厌学，教师困教，这种局面不改变不行了，出路只有一条——必须改革。

如何进行教学改革？大家都在摸着石头过河。年玉萍同志的《信息处理与现代汉语问题探索》（以下简称《探索》）正是改革的一个尝试。

《探索》定稿后，作者要我为之作序。大概因为我曾当过语言教研室主任，退休十年至今仍任教学督导，对她了解之故。是的，几十年来，我是眼见作者由青年到中年成长变化的。她娴静文雅，貌美内秀，贤惠诚信，为人本里本分，干事认认真真，虚心、敬业、求实。现代汉语教学重视实践练习，师生共同研讨、释疑、解难，以身作则，教书育人。20 多年来从未有过哗众取宠和言过其实之辞。她给我的印象是实诚人，本打算就此着笔，及至照习惯仔细读完原作后，则大有刮目相看之叹！我必须在她“虚心、敬业、求实”六字之后郑重地添上“创新”二字！

《探索》“创新”何在？

20 多年来，面对现代汉语教学中的语音、文字、词汇、语法、修辞等内容，作者没有围着旧教材，依葫芦画瓢，原地团团转，而是站在计算机信息处理的高度，对它们逐一地进行了新的探索。全书既有计算机和现代汉语碰撞的概括论述，又有实例的具体分析，论证角度新颖，说理辩证客观，讲特点，抓要点，指

难点，既彰显着专业实践，又体现了时代的特色。

全书20多万字，分绪论、语音、汉字、词汇、语法、修辞、现代汉语教学等，各章都有作者客观而独到的见解。例如：

绪论概括介绍中文信息处理的过程和汉字字形识别、语音识别、汉字编码键盘输入三种输入法的优缺点，比较了中文信息和印欧语言信息处理的异同，阐发了中文信息处理与现代汉语研究的现状和发展趋势，让人们从当今信息处理的角度全面、客观、辩证地审视现代汉语。

现代汉语教学改革，提出教材要反映时代发展的特点，不能滞后于当前的语言研究和实际，体现时代性，结合计算机信息处理，增添字、词、句处理的内容，又针对学生和大学教学实践，采用启发、变换、对比式教学，以学生为主体，师生共同参与探索，利用网络及时反馈交流，突出实践的针对性，提高学生主动学习的积极性，培养学生发现、提出、探讨、解决实际问题的能力，转变教学理念，改进教学与评估方法，转变环境等等颇有新的创见。

语音中突出轻声、汉语拼音与信息处理；汉字中强调规范汉字形音义对实现汉字，汉语词、句处理的意义，从中文信息处理角度讲汉字的特点、笔画、笔顺与标准化的重要性，提出了在五笔字型输入法基础上减少字根笔画，以新代旧，保留汉字特点的汉字改革方案。词汇中指出词处理是信息处理的难关，分词研究和词类研究是关键，要立足于可操作，这方面还有大量子课题需要开发。语法中词法主要探索了量词，尝试用历时语言学的研究方法从普通话、方言、古代汉语三个角度及语法、语义、语用三个平面来探索量词，有一定的深度和广度。句法着重对“把”字句、兼语句、连谓句、存现句、紧缩谓语句五种句式作了深入详尽的探讨，颇有新义。最后，总结阐述了信息处理与现代汉语研究的情况，指出信息处理现代汉语的难关是排除歧义句。目前有两条途径：基于规则和基于统计。作者主张大词库小规则相结合，尽快为计算机提供现代汉语平衡语料库和有关现代汉语各种知识的数据库，努力探求好的分析方法，全面注重句法、语义、语用认知各方面的综合分析。修辞一章中主要从语音、词语、句式三个角度分析了修辞，视觉独特，最后宏观分析了婉约、豪放、华丽、朴素、繁丰等语言风格。

《探索》又是求实的结晶。每章列有参考文献，标明出处、依据，彰显当代专家研究成果。对度的把握客观公正，不夸大其词，无个人偏颇，而是一切从实际出发，实事求是。沿此路走下去，《探索》必将再上新台阶。

为什么能有《探索》面世？十月怀胎，必然要一朝分娩。作者勤奋求实，积20年教学的心得，积累经验教训，善于总结，这是内因。处于科技飞速发展的

时代，她又有好的机缘，其夫就是从事网页设计与制作、多媒体教学与应用等计算机教学研究的专门人才，中文信息和语言学界联手，使得她成为开设语言文字信息处理选修课的副教授，最终也促成了《探索》的问世。《探索》既属探索，必从求实出发，而创新归宿，首要看路子对不对，求全责备就不客观。全书言简意赅，点到即止，没有赘言谩语。

水稻杂交，小麦杂交，果树杂交可出换代的新品种，信息处理与现代汉语对撞必能产生反映时代特色的新学科。国家语言文字工作委员会原主任许嘉璐先生预言："一旦语言学家和计算机技术结合起来，所带来的不仅是中文信息处理事业的顺利发展，而且有可能引发语言学研究的一场革命，从而语言科学真正成为先导性的科学，走在科学技术发展的前列，受到全社会的重视。"《探索》的问世，标志着这场革命已经到来，愿语言学界人士和计算机专家更多地联系，迎接这场革命的到来，告别厌学困教局面，在共同探索中，让语言科学真正走在科学技术的前列，成为先导性的科学，一定会得到学界的青睐！

李慎行

2010 年 12 月 6 日

前　言

21 世纪是信息时代，计算机已经遍布于生活的各个方面，作为人类交际工具的语言当然不例外，与计算机的关系非常密切。计算机信息处理已经成为当今使用频率很高的一个词语。人们用计算机把汉字打出来，能根据具体的要求来改变汉字的字体、字号、颜色甚至动态。同时不仅能一个汉字一个汉字地打，而且还能按词语来打。人们为计算机的功能赞叹不已。但是，这些都还是“字处理”阶段，虽然 20 世纪 80 年代，计算机刚引入我国来时，汉字如何输入计算机成了当时的拦路虎，但是，在科学工作者的努力下，这一问题早已解决了，现在，对汉字的处理，计算机的功能已经很强大，基本上满足了人们的各种需要，所以说，“字处理”对于计算机来说还是最初级阶段。按照计算机的智能来说，完全能够像人一样来分析语言、理解语言。现在语言学家和计算机专家共同联合起来攻克计算机中“词处理”和“句处理”的难题，这是继汉字输入难关后的两个难点，其中遇到的问题比较复杂，任务非常艰巨，科学家们正在努力攻克这一难关。我们作为语言工作者及学习语言的人，当然责无旁贷。

这本书尝试将现代汉语和语言文字信息处理这两门课的内容结合起来，从信息处理的角度来探析现代汉语中的内容，使人们在信息处理的今天，对现代汉语有一个全面的认识。但是，由于中文信息处理这个任务非常艰巨，其中遇到的难题很多，很复杂，现在科学家们正在研究，再加上本人水平及材料有限，所以，文中对中文信息处理的情况只做了大体介绍，其中参考了学者们大量的相关资料。后边各章节都立足于现代汉语和现代汉语教学，对现代汉语中的语音、汉字、词语、语法里涉及信息处理的有关内容进行了分析、探讨，而且非常概括。由于本人水平有限，再加上信息处理的内容很尖端，文中存在的不足之处，敬请各位同仁批评指正。

年玉萍

2010 年 10 月

目　录

绪　论

第一节　中文信息处理概况

一、什么是中文信息处理

语言信息处理是指用计算机对自然语言的音、形、义等信息进行处理。即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工。

中文信息处理是用计算机对汉语的音、形、义等信息进行处理，也称“汉语信息处理”。

汉字信息处理是用计算机对汉字所表示的信息进行操作和加工。

中文信息处理是语言信息处理的一部分，或者一个分支。汉语信息处理解决的首要问题是汉字的信息处理，这是汉语的独特任务，也就是说，汉字信息处理是汉语信息处理的第一步，因此在很长一段时间“中文信息处理”的主要任务是“汉字信息处理”，有人干脆把“中文信息处理”称为“汉字信息处理”，现在比较多的称为“汉语汉字信息处理”。

中文信息处理大致包括以下一些科目：词的切分和频率统计；汉语句型和短语的研究及频率统计；汉语语义的研究；键盘和非键盘汉字输入技术及处理系统；汉语语料库的开发及应用；汉字的机器代码，程序设计语言的数据类型；汉语开放系统的接口规范；语声输入与合成；汉字识别；字形生成；汉语分析及理解；汉语生成；人机接口；机器翻译；情报检索；自动标引和抽词，自动文摘；全文检索；电子印刷出版系统；汉语辅助教学；电子词典等。

中文信息处理除了字处理阶段外，还有词处理阶段和句处理阶段，研究词处理和句处理是为了让计算机能理解人类的语言。用计算机处理自然语言是语言文字信息处理的重点和难点：人与人用自然语言交流之所以没有困难，是因为交流

总是在一定的环境中进行的，如果计算机系统实现了①人机会话②机器翻译③自动文摘（自动分类、文献检索、自动校对等）④能结合语境理解意思等语言信息处理功能，则认为计算机具备了一定程度的理解自然语言的能力。

二、中文信息处理的过程

计算机对自然语言的研究和处理，一般应该经过以下三个过程：第一，把需要研究的问题用语言学的方法加以形式化，使之能以一定的数学形式严密而规则地表示出来；第二，把这种严密而规则的数学形式表示为算法，使之在计算上形式化；第三，根据算法编写计算机程序，使之在计算机上加以实现。因此，研究计算语言学，不仅要有语言学知识，还要有数学和计算机科学知识。这样，计算语言学处于文科、理科和工科的交叉点上，是建立在语言学、数学和计算机科学这三门学科基础上的边缘性学科。

这些内容很复杂，任务很艰巨，涉及的学科比较多，现在正在研究中，但在完成这些任务之前，首要的问题是对汉字进行处理，这里主要讲一下汉字信息处理的过程：

汉字信息处理有三个过程：第一个过程是汉字的输入。即通过输入设备把文字信息转换成代码，并送入计算机。第二个过程是对汉字信息的加工和处理。即根据各类不同的应用，借助预先设计好的程序对输入的信息进行加工和处理，从而得出结果信息。第三个过程是汉字信息的输出。即通过输出设备，把以数据代码形式表示的结果信息复原成汉字。

下面，主要谈一下汉字的输入、汉字的存储、汉字的输入。

（一）汉字的输入

汉字输入是指利用汉字的形、音或相关信息通过各种方式把汉字输入到计算机中去的过程，汉字输入技术是汉字信息处理的关键技术。

汉字输入的方法有三类：

1. 汉字字形识别输入

（1）什么是汉字字形识别输入

也称汉字自动识别。即利用光学扫描方法将汉字的图形信息直接输入计算机，也就是用计算机自动辨别印刷或书写在纸（或其他介质）上的汉字。它属于模式识别和人工智能的范畴，是新一代计算机智能接口的一个重要组成部分，在应用上它是汉字信息处理系统高速自动输入的手段和根本出路，是汉语中文信息处理的一种好方法。

（2）汉字字形识别输入的类型

A. 联机手写汉字的识别。人一面写，机器一面认。这是最简单的一种汉字识别类型。

B. 印刷体汉字的识别。包括两小类：单体印刷体汉字识别，识别印刷在纸上的一种印刷汉字；多体印刷体汉字识别，同时能识别印刷在纸上的宋、仿宋、黑、楷等印刷体汉字。

C. 手写汉字的识别。包括三个小类：手写印刷体汉字识别；特定人写汉字识别和人机交互式手写汉字识别。

（3）汉字识别的基本思想与步骤

汉字识别的基本思想是匹配判别。

步骤：

A. 把需要识别的汉字集合中每一个汉字字符的字形特征存储在机器中，形成已知的汉字模。

B. 用图形输入板或光电设备（如图文扫描、光导摄像管扫描、激光扫描等装置）扫描输入一个未知的需要识别的汉字字符，抽取它的特征。

C. 将抽取到的代表未知汉字模式本质的表达形式（即各种特征）和预先存储在机器中的所有汉字特征一个一个地匹配，匹配用一定的准则进行。最后在机器存储的标准汉字模式表达形式的集合中，找出最接近输入汉字特征的那一个，该特征所对应的汉字就是识别结果，最后用相应的内部码来表示它。

（4）汉字自动识别的优点

汉字自动识别的方法有许多优点：

A. 实现了汉字的高速自动输入，大大减轻了人的脑力和体力劳动强度。

B. 突破了人工输入的速度局限性，彻底解决了汉字信息处理系统中手工输入效率低的问题。

C. 为办公自动化和下一代印刷技术的文字信息自动输入打下了基础。它还可以作为新一代计算机智能接口的重要组成部分。

D. 它有助于汉字文本高倍压缩存储和传输。

（5）汉字自动识别的研究进展

陈敏和王翠叶（1995）报道了我国汉字识别技术的进展情况。这项技术自20世纪70年代末起步，目前正向实用化发展，印刷体汉字识别是我国汉字识别研究的主流。1988年已有五六个系统基本达到实用化，并形成商品，它们的主要技术指标达到了世界先进水平。识别字数可达3755～4000个，识别速度为20字/秒左右，对中等印刷质量文本识别率达到95%～99%，识别字号为3～6号，

识别字体的宋、仿宋、楷、黑等，有一定版面分析和后处理能力，初步具备了适用的人机界面。

联机手写识别1988年已有几个初步实用的装置，其主要技术指标为：识别字数可达6763～12000个，识别速度与人书写的速度基本相当；初次使用的识别率为80％左右，经常使用可达95％。书写时要求笔形与笔画数比较正确，极常用的少数笔形与笔画可以连笔书写，笔顺不严格要求。

手写汉字识别1988年才开始认真研究，近几年进入高潮，全国已有几个实验性系统进行了鉴定。特定人手写识别已在小范围试用。识别速度用386微机为1字/秒。接近实用的交互式自学习手写汉字识别系统，可识别3755个字，其前10位候选正确率为80％～95％。手写印刷体汉字识别已从方法研究转向实用系统的研究。

（6）汉字识别存在的问题

识别的准确率受到各种限制：印刷质量、扫描时的位置等。

（7）汉字识别技术今后研究的主要方向

A. 人工神经网络技术用于汉字识别

人工神经网络技术具有高度的自组织、自适应和自学习能力。在我国手写汉字识别和印刷体汉字识别的研究中已得到了应用，今后将发挥更大的作用。

B. 语言学知识用于汉字识别

识别实际文本时，文中大部分字及其相邻字要受到词、句法、语义的限制，因而是相关的。识别系统利用这些相关性的知识，可改善孤立字识别时的性能。这样，把单字识别技术同语言学知识结合起来，能提高识别系统的水平。在已有的印刷体汉字识别系统中，后处理便利用了汉语的词进行自动纠错。今后将进一步利用词的上下文匹配和基本句法、语义的上下文匹配，来提高对实际文本的识别率。

2. 汉字语音识别输入

（1）什么是汉字语音识别输入

汉字语音识别输入就是通过“说”和“听”来和计算机交换信息，即利用声音识别技术，抽取汉字的语音特征，实现对汉语语音的自动识别。其目的是让计算机“听懂”用汉语语音所表示的汉字信息，以便通过口授将包含有汉字的程序、数据、命令、文稿等送入计算机。

（2）语音识别输入的优点

A. 输入速度快，说比写约快10倍，比打字约快4倍。

B. 工作强度低，使用最方便，将手解放了。

C. 使用最方便，不会受到编码规则对思维习惯的干扰。使用语音是人机对话的最自然的方式，也是名副其实的人机对话。

（3）语音识别的类型

语音识别的类型，按不同的标准分类，有如下三种不同的分类结果。

A. 按使用人分类

按使用人分类，有特定人语音识别和非特定人语音识别。

特定人语音识别指使用前由使用者对系统进行训练，让系统记住事先选好的字或词的发音特征，识别时由这个使用者将字或词读进系统。非特定人语音识别是供许多人使用的系统，使用者不用对系统进行训练。系统要能听懂任何人说的话，就必须让系统获取许多人说话的共性特征，并在处理中进行强化，使许多人说的同一语音的特征有极高的稳定性，对不同的语音有极大的区别度。

B. 按词汇量分类

按词汇量分类，有小词汇量语音识别、中词汇量语音识别和大词汇量语音识别。

小词汇量指几十个字或词，中词汇量指几百个字或词，大词汇量指几千甚至上万的字或词。

C. 按发音方式分类

按发音方式分类，有孤立词语音识别和连续语音识别。孤立词语音识别指识别时将字或词孤立地读进系统。连续语音识别指识别时将整个句子连续读进系统。要求系统既具备处理连续造成的同化、异化、脱落、换位等音变问题的能力，又具有通过语义、语法知识分析得出正确识别结果的能力。

另外，还有使用环境优劣的区分，也就是指噪声轻重情况。目前噪声下的语音识别还只是在初步研究阶段。通常所说的语音识别都是有较好的使用环境。

从技术上的难易程度说，上述每小类语音识别，后者都比前者难。如果将上述三小类排列组合起来，应该有 12 大类，即特定人小词汇量孤立词的识别、特定人中词汇量孤立词的识别、特定人大词汇量孤立词的识别、非特定人小词汇量孤立词的识别、非特定人中词汇量孤立词的识别、非特定人大词汇量孤立词的识别、特定人小词汇量连续语音的识别、特定人中词汇量连续语音的识别、特定人大词汇量连续语音的识别、非特定人小词汇量连续语音的识别、非特定人中词汇量连续语音的识别、非特定人大词汇量连续语音的识别。这 12 大类一类比一类难。

（4）语音识别研究的进展

我国语音识别技术经十多年的发展，目前已开始走向实用。以汉语全音节识

别的成绩最为显著。特定人大词汇量孤立词语音识别系统，具有较高的识别正确率和响应速度，有的已初步商品化，识别率基本能达到80%以上，有的还可达到95%以上。基于神经网络方法进行的汉语声母、韵母、声调的识别，已取得了可喜的成果，有些单位四声识别已达到近100%的水平。

利用声学信息进行的语音识别，有一些中、小词汇量的语音识别系统已投入实际应用，如口呼语音输入的自动查报电话号码系统、声控电话查号系统等。非特定人中、小词汇量孤立词语音识别已取得优异的成果，利用适合于汉语特点的概率统计模型对不同说者和话流速度的变异有相当强的适应性，目前正向大词汇量孤立词语音识别系统迈进。连续语音识别刚刚开始，特定人小词汇量的连呼识别，特别是连续数字串语音的识别在实验室里已做到实时识别，并有较高的识别率。连续语音识别的后处理工作，也取得了一定的进展。噪声下的语音识别已在做初步的方法研究。

(5) 语音识别研究今后努力的方向

加强识别方法和处理手段的研究，“提高语音识别的准确性”；加强非特定人、大词汇量、连续汉语语音识别的研究；开展建立语音库和语音特征库的研究；注意计算机听觉模型的研究;注意研究模糊数学理论在语音识别中的应用;加强人工神经网络方法用于语音识别的研究;研究语言学知识在语音识别中的应用。

目前语音识别的方法主要是利用语音信号中的声学信息和模式匹配来判断识别语音，语音识别的终极目的是语音理解。语音识别和语音理解一样，不能仅依赖于声学信息，还须依赖于语言学的信息。如利用汉语的词法、句法、语义和语用知识来解决语音信号多变性的问题。语音识别要达到高级水平，必须利用语言学知识。

3. 汉字编码键盘输入

其做法大多以原有西文计算机系统为基础，利用计算机所使用的ASCII字符来对汉字进行编码，使汉字符号化，并借助键盘输入计算机。这是目前语言信息输入的最主要方法。

(1) 汉字的键盘输入

如何在国际通用的小键盘上用不同的键位组合把6763个不同的汉字从字库里“检索”出来、“敲打”出来，这是汉字信息处理的首要问题，而键位组合的设计就是平常所说的“汉字编码”。1978年12月，我国召开了“第一届全国汉字编码学术交流会”，会上提出了汉字输入编码方案约40个。专家们介绍，当时内地第一种汉字输入编码叫做“支码”(唐旬，1995)。汉字与键盘匹配有很大的困难。后来人们通过拆分汉字字形来解决汉字与键盘的匹配问题。这就是汉字编

码中的字形码的起由。

汉字编码的类型可以按在编码方法中所使用的汉字主要属性来划分。目前较多地使用的汉字属性有字音、字形、字义、字频等属性，特别是字音、字形这两种属性。这样，汉字编码的类型就可分为形码、音码和音形码 3 种。

（2）汉字编码的类型

A. 拼音编码

按汉字的读音将其转换成汉语拼音的声母、韵母（或加上声调符号以及区分同音字的符号），或将双字母声母、复合韵母用单字母替代组成的编码。拼音编码可分为全拼音式和压缩拼音式。

拼音编码的优点是易学。缺点：一是重码多，导致输入效率低，令人不胜其烦，且极易造成视觉疲劳；二是对用户要求很高；三是难于处理生字。

B. 字形编码

将汉字分解为部件或笔画，并按照规定的顺序排列，用相应的字母数字等符号替代，按一定的规则取舍的符号组合，就是字形编码。属于字形编码的有笔形码、前三未一码和五笔画码。主要有两种：

笔画式编码：即将汉字分解为笔画，每种笔画用一个数字代替编码，每字取 6 码或 5 码。例如：将汉字的笔画分解为“横、竖、撇、点（捺）、折”等五种笔画，分别用 1、2、3、4、5 替代，按笔画的书写顺序排列，每个字最多取 5 码。

字根代码类：五笔字型汉字编码主要是字根码。录入人员根据《五笔字型键盘字根总图》中所确立的字根，按照每个汉字字根的排列组合顺序（根序）递次编码输入计算机。在键盘上用字根输入汉字，首先是字根的归类记忆难。把几百个字根归类到二三十个键位上，并牢牢记住，绝不是轻松愉快的事。其次是输入操作时的拆字难，拆字需要耗费较多的心理操作，增加了大脑的负担。

目前已问世的各种形码系统，一般是采用了“字根归类和拆分”的设计思路。

汉字编码发展到今天，形码仍然在为降低学习难度奋斗，音码仍然在为降低重码奋斗。这种情况，不适应中文信息处理技术的普及，特别是在 90 年代以后，电脑开始走入寻常百姓的家庭，电脑的使用者不是用它去“高速表达别人”，而是用它“述说自己的思维”。“说得出，就打得出”是他们最基本的愿望。让人们去背上百个字根确实困难；而音码全拼式又太慢、太费力。时代呼唤更科学、更简便、更合理的汉字输入方案。[①]

① 亢世勇：中文信息处理补充材料，汉字信息处理，http：//www. chinese. ldu. edu. cn。

（3）汉字编码的误区

汉字编码存在一些误区[①]：

A. 重码率越低越好，甚至追求无重码方案。实际上，“无重码”都是以牺牲易学性为代价的，邮电通讯中一直在使用的四码电码就是一种简单的无重码的设计。而我们完全可以发挥软硬件的优势，采用提示行选择、高频先见等方法在一定程度上容忍重码。

B. 速度越快越好。实际上不同的人员、不同的工作性质有不同的要求，最广大的一般用户要求不看键盘，以边想边打的方式输入汉字，对输入速度则只有一个最低要求，即只要每分钟输入 30 个字以上就可以了，而第一位的希望是越容易学越不容易忘越好。

C. 词库越大越好。进入词处理阶段后，各种输入方法纷纷关注词库的大小。因为词少了不行，往往打了词语码以后，词库中没有该词，又得退回来用字的方式输入。为了减少这种事情的发生，词库从 5000 条词发展到 1 万、2 万、3 万、7 万、9 万甚至更大。但是，词库越大，占的内存也就越多，而且，绝大多数人使用计算机都是在一定领域范围内工作的，他们所使用的词语也是有一定范围的。

所以，最佳的词库设计是：“通用词库”＋“专业词库”＋“个人词库”。

个人词库是个人自己生成的独用的习惯“词语”，现有词频统计结果表明，通用词库有 4 万条左右词已足够了，专业词语则各领域词语多少不一。因此，简单地认为词库的词越多就越好也是不全面的。

（4）汉字编码的原则

A. 社会学原则

汉字编码研究的目的之一是为了信息处理技术的普及，这一普及首先要面向教育。《全国中小学教学用汉字编码规范及计算机汉字输入系统》经当时国家教委批准已列入“八五”重点攻关项目。从 1993 年开始，计算机逐步列为我国中小学的必修课程。自此，计算机和语文、数学、外语一样成为青少年必须掌握的四个工具之一。汉字输入系统进入中小学课堂是培养跨世纪人才的根本大计，教学的规范化相应地要求汉字编码研究规范化，在规范化的前提下，将中小学的“识字、定字、查字、打字”教学统一起来。

汉字是我国悠久历史文化的一个象征，研究汉字的分解原则，必须联系我国

①张普：《步入信息社会的汉语和汉字》，汉语信息处理研究，北京语言学院出版社 1992 年版。