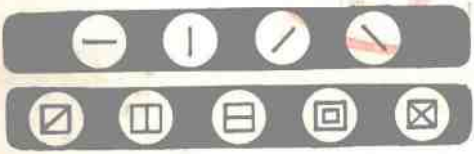
五笔字型计算机汉字输入技术

WUBI ZIXING
JISUANJI
HANZI
SHURUJISHU

王永民 著



河南科学技术出版社

五笔字型计算机汉字输入技术

王永民 著

河南科学技术出版社

内 容 提 要

计算机的汉字输入，是我国广泛使用计算机的一大障碍。国内外目前已先后研究出了数百种计算机汉字输入方案，而“五笔字型”方案是其中的佼佼者，已在国内外获得广泛应用。

本书作者为“五笔字型”方案的发明人。作者理论结合实践首先简要阐述了汉字编码的研究方法及各项原则，然后对本方案的编码规则、操作技术等作了详细系统的介绍，最后列出了“五笔字型”汉字编码国际一级字码本。大部分章节后还附有习题。

本书文笔流畅简洁，内容实用具体，可供各行各业从事计算机工作的人员学习使用。

“五笔字型”计算机汉字输入技术

王永民 著

责任编辑 马文翰 范云操

河南科学技术出版社出版

河南第一新华印刷厂印刷

河南省新华书店发行

850×1168毫米32开本 8.875印张 173千字

1985年10月第1版 1985年10月第1次印刷

印数：1—8,650册

统一书号13245·35 定价1.85元

汉字电脑的最新成就

（代　　序）

郑易里

从今天开始，可以说，汉字不能象拼音文字那样在二十六个字母键上打字、并在事业管理、科学研究等方面广泛应用的时代，已经一去不复返了。

我们知道，汉字不是拼音文字，但在汉字编码方面，人们总是常常把希望寄托在拼音方式上。这是由于拼音音素一般都不多，只用二十多个字母，便能把成千上万个汉字拼读出来。实践证明，汉字的读音常常不能从字形上观察出来，按字音拼成的字也不是一个个汉字本身，而是一批批读音相同的拼音文字。即令解决了同音字问题，人们还常常会碰到许多字不一定知道它读什么音，或读不准它的音。这是音码方面重大的基本问题，需要深入研究。

汉字楷化以后，字根便以笔画多少为准。一个字根或者一个由字根组合成的合体汉字，笔画多少不能直观，必须逐一认真合计才行。汉字字根，每字１～５个，一见便能立刻判断，十分直观。王永民同志在五笔字型编码方案中只抓着字中的１～４个字根，在字根中又只抓着字首字根的起笔一笔，牢牢掌握住了汉字核心信息的核心部分，收到了极其巨大的编码效益。

汉字无论在质变、量变方面，信息表现都很丰富。在汉字编

码方案中，汉字的信息采用得愈多，编码愈容易，但难记难用；汉字信息在科学概括方式下采用得愈少，汉字编码愈困难，但易记易用。极其易记易用的信息量是二十六个字母，但难度也最大。仅仅只用二十六个字母给近七千个之多的汉字进行字形编码，这在以前几乎是不可想象的。然而这样一个老大难问题，今天终于被王永民同志初步突破了。

王永民同志在他大学时代所学到的电子技术这一专业基础上，又刻苦钻研汉字规律，敏速地攻克了一座座难关，所以能在较短时期内，做出了空前所无的巨大成就。今天总结起来，约有以下几点：

一、创新码式和键式

一般码式都是按字母或数字逐一递增成序作码。总码数不多，还比较好记好用。若为二、三百个，那就难记难用了。加之这种按自然数排序的代码，弗晰范围较大，需要死记硬背，较难记熟。"五笔字型"码式是把二十五个主要字根按"横、竖、撇、捺、折"顺序分为五类，使弗晰范围缩到最小的适量程度，人们见字时便能立刻辨清起笔一笔，而不暇思索地把五个手指伺机操纵下的，只用五个数字表现的二十五组字根反映出来，实现了从未有过的击键的最高效益。

二、码长最短

王安字根1～99号，控制数字十个，每字击键一律六次。天龙字根共一百五十多个，每字最多击键五次。二者规律杂乱，科学意义较差。王永民同志的"五笔字型"方案把汉字字根高度概括为五类二十五种，各用一键控制，每字击键二至四次，码长最短而击键速度最高。

三、字型识别，配合巧妙

“五笔字型”方案把字型分为左右、上下、外内和单体四种，各用数字一个作代码，又取该字末笔一码互相配合，一共形成二十种之多的识别码，构思十分巧妙，把繁多而杂乱的汉字信息高度概括综合利用，大大提高了字与字间的分辨率，同时也大大提高了击键速度

四、间作与复种，密植与并株，手法翻新，电脑耕作，收益猛增。

大家知道，利用二十五个字母，按每码四个字母说，可以编三十九万多个码，若实编七千汉字，则电脑中还有三十八万多个空码可供利用。王永民同志在电脑汉字耕耘中，间作、复种、密植、并株，机灵活变，同时并举，使汉字电脑实现了闻所未闻的增产效益。他首先把“一、的、了”等二十五个高频汉字植入二十五个键位内，形成字间间作。继后，他又把其次的六百多个高频汉字植入六百多个行间的第一个空格内，形成行间间作。另外他又精选了两千多条词汇植入各行空位内，形成“并株”耕作法。二字词原需八码，三字词原需十二码，四字词原需十六码，现在每一条词汇的码数一律缩减为四码，使编码速度成倍成倍地迅猛增长。这只有在电脑中才能如此得心应手、机灵活变地实现这一收益极快极大的划时代的汉字耕种法。可以肯定地说，这是一项可贵的科学创造。它所蕴藏着的巨大的成就，一时还难以准确估计，还有待于进一步开发利用，为四化建设贡献力量。

写在前面

汉字经过了几千年中华文明的历史。它以读音简单，书写明洁，表意深邃和构词丰富的独具优点而久盛不衰，是当今世界上四分之一人口通用且有越来越多的国家关心重视的文种。在我国文字处理现代化的过程中，汉字又成为计算机应用中大家关注的研究对象。

汉字的计算机输入问题，是一个在语言文字学、信息科学、计算机技术、工程心理学的接壤处生长出来的边缘科学。它已成为影响计算机在我国普遍应用的“瓶颈”。研究汉字的理论、著述，古往今来，累积多矣，兹不详举。本文既不从文字学的角度论述汉字的渊源，也不去评价汉字的历史“功”、“过”，只是从多学科对汉字信息处理输入技术的一般要求出发，对我们的研究工作加以总结，描述一个字形编码方案研究的全过程，并试图建立一个字形编码体系。文中简要地介绍了我们的二十六键“五笔字型”汉字编码方案，以下简称“五笔字型”（WBZX）汉字编码方案。这里介绍的是1985年8月优化新版本。

几年来，我们在研究汉字字形编码的过程中，提出了一系列新的观点，创造了一些包括使用计算机辅助设计字形方案在内的实用方法（参见《字形编码设计中的机助方法》一文），从而将我们一九八〇年的六十二键“汉字层次分解编码方案”，改进为“三十六键六笔字型编码方案”；最后，又在一九八三年元月，

完成了《“五笔字型”汉字编码方案》。一九八三年十月至今，又完成了优化方案的新版本。

郑易里先生曾四次亲临南阳指导我们的研究工作。我们的“汉字层次分解编码方案”即参考了郑老关于汉字研究的科学理论。郑老创造的“一笔查字法”和《从人查字到机器查字》等著作，为我们研究和建立字形编码方案提供了关于汉字研究的重要参考。我们由衷地感激郑老给我们的教益。一机部自动化研究所扶良文工程师、山西矿业学院陈文熙教授对我们的研究工作，曾给予热情的帮助和及时的指导；我们也参考过其他有关同志的文献资料，在此一并致谢。

自“五笔字型”汉字编码方案发表、尤其是通过鉴定以来，广大用户曾给予了热情的支持和大力的协助，作者特向有关的厂家、研究机关、大专院校和用户表示衷心的感谢。

本书如实地反映我们的工作过程和粗浅的体会。其中第一章曾在一九八三年中文信息处理国际研讨会(ICCIP)上发表。“五笔字型”方案也有待于进一步优化和提高。书中难免谬误，恳请有关学者、专家赐教。

河南省委、河南省科委、南阳地区科委对该项研究工作的关心和大力支持，是该项工作不断取得进展的根本保证。

参加“五笔字型”汉字编码方案优化研究工作的还有：张道政、徐世营、常胜敏等同志。本书第三、四、五、六章与张一平合写。

目　录

第一章

汉字字形编码的原理与实践

第一节　汉字的字形编码原理

汉字源于象形符号，实际上是一种规范化了的图画文字。汉字的总数成千累万，常用字也有四五千个。这些字其所以独自成字，互不相同，也仅仅是因为它们有着不同的形。我国幅员辽阔，方言多杂，汉字字形却是统一的。古老的方块汉字所独具的种种优点，近年来已被越来越多的人所认识并加以研究和应用。因此，单依汉字的形进行编码，也是一种适合国情的编码方法。

本文所说的字形编码，是专指既不考虑汉字的读音，也不把汉字全部肢解为单一笔画，而是采用字根拼形组字的一类编码方法。

字形编码的基本思想是：把汉字视作一个图形。这个图形可由若干个因袭成习、相对稳定的部分（字根）组成。字形编码的任务是：根据构成汉字各部分（笔画和字根）的特征和它们之间的结构（字型）特征，为计算机编取唯一性汉字代码。这样作不仅是必要的，也完全是可能的。

信息论和计算机对汉字输入编码的基本要求是：键数少，码长短，效率高，重码少。

操作人员对方案的基本要求是：记忆量少，规律性强，规则

简明，好学易记，操作直观，敲键最好和写字相仿，键数少，最好能盲打。

汉字的字形编码设计，必须有其理论上的科学依据，并经过反复实践。关于汉字的理论是指对汉字历史地、系统而深入地分析和大量的基础研究之后得出的规律性。一般包括对汉字的笔画、字根、字型、字根优选、归并组合、键位排列的认识；实践就是按照计算机对输入方案的要求，研究其方案的过程。一般包括设计、编码、修改、考核、测试、上机验证、交用户使用等。

汉字字形编码的方法已有好多种。能否作到向计算机输入汉字，就跟写字一样，例如对于“萌”字，在键帽上刻上“艹”、“日”、“月”等常用字根符号，依次按键就完成了“萌”字的输入了。应当说，如能做到这样，那是再好不过的了。困难在于，当键位较少时，不仅无法将几百个不同的字根都刻在键面上，而且即使把它们按一定规律归并分组、联想记忆，也很难保证当它们共享同一键号代码时，能够共存共容。压缩键数的代价常常是重码的增加。在键数一定时，克服重码又常以字根搬家，破坏其规律性为牺牲。因此，一般的字形方案都需要有较多的键位。另外，字根越大，越是“无地自容”。所以，键数较少时，常又不得不忍痛“杀”掉某些字根作拆分处理。一般说来，研究键数少、码长短、字根大、重码少的拼形方案的困难盖来于此。当键数压缩到二十六个，每字击键限定在四次以内时，如不采取有效的特别措施，一、二百个字根是非常难于归并组合的。

书写汉字时，人们将一个个字根按一定的先后次序和位置关系拼排起来，就象搭积木一样，叫做拼形组字。基本字根的形象和它们在汉字中的先后次序及书写方法，是每一个有文化的中国人的既有知识和良好习惯。设法利用这一点，向计算机拼形输入

汉字，恐怕比让人们重新再接受一个新的（甚至是完全颠倒的）次序概念要容易得多，有意义得多。所以，我们主张：汉字字形编码时的取码次序，要与书写顺序尽可能保持一致。这样做，就可以结合语文教学，从小学一年级开始，教孩子们怎样在计算机键盘上拼形输入汉字。

第二节　字形编码方案的研究方法

一、确定方案的指导思想及要达到的主要指标。如：对汉字研究的理论，字根的定义及筛选原则，方案的应用对象，处理汉字的范围，键位数，允许的码长等。

二、制定一套希望达到上述目标的编码规则。规则要能适用于每一个汉字，使二义性尽可能小，特例尽量少。

三、同时考虑字根的组字频度及其实用频度，精心挑选汉字字根，组成方案的“基本队伍”。

四、按照既定的指导思想，将这些字根归并组合，赋予代号，形成草案。

五、人工编码。将所处理的汉字逐一分解赋号，同时记录其中的疑难（以备反过来修订取码规则，使之严格准确）。

六、进入机器处理（或继续人工处理）。将人工编码敲入计算机中，让计算机对方案的人工编码结果进行排序、拣重码，或按一定要求让计算机作部分归并组合实验。

七、根据第一次结果，调整草案的设计，修改机内数据代码，再运行，出现一个新的结果。反复数次，最后形成一个完整的单字码系列，并按国标号及自编码顺序打印出两种单字码本。

八、根据单字码的取码规则，进而确定简码、词汇码、高频

字码及重码的处理方法。

九、计算出每个代码在方案代码中的组码频度以及它们的实用频度（可由单字的实用频度算出）。据此按照键盘指法要求和键位代码的排列规律，将代号（或字根）安排在键位上。

十、计算出方案的各项理论指标。如平均码长、编码效率、键位熵值、键入速度、重码率、重码使用概率等。

十一、培训一定数量的操作人员，实测出实验指标。

十二、组织评议鉴定。

十三、推广应用，听取用户意见，再作个别或较大的修改，直至定稿。

第三节　汉字字形编码设计的基本原则

汉字字形编码的设计，不仅是一个理论问题，也有重要的方法问题。在理论与实践的结合中，作者在设计WBZX方案中，曾经遵循了下述基本原则。

一、基本字根优选原则

字形编码方案设计中选用的汉字字根越少越好，越精越好。少是指数量，精是指质量。

WBZX方案从六百多个字根中选用了一百三十个作为基本字根。字根的选取考虑了字根的两种频度：组字频度和实用频度。这些字根的合计构字频度约百分之七十，而合计实用频度约百分之八十。

汉字字根的组字频度，是某一字根在汉字的字种集合（比如《现代汉语词典》的全部一万一千个字头）中的出现概率（参见作者一九八一年五月《汉字基本字根的查频研究》一文）。掌握和应用字根组字频度的研究结果，就可以作到采用较少的字根拼形组成尽可能多的汉字。

实用频度，是某一字根在汉字的字数集合（比如《汉字频度表》统计的两千一百多万个汉字）中，也就是在现代汉语文字中实际使用的概率。汉字字根实用频度的研究结果，将提供字根优选以及字根在键面上合理排列的理论依据。

二、低频字根一法处理原则

根据不同的字数集合，不同的方法、不同的人会统计出不同的字根总数。如把单体结构都计入的话，有五百至六百五十个之多。据统计，有二百五十个字根已约占构成一万一千个汉字全部字根的百分之八十；有二百个字根已占全部字根实用频度的近百分之八十六。因此，减少字根总数对字根进行优选时，多数字根将“名落孙山”。

对于落选字根分别对待，一般需要较多的规则，因而不经济，也不必要。最有效的办法是：入选字根不再拆分，落选字根一法处理（比如，规定落选字根一律拆取某一两个笔画或字根参加编码）。这样，抓住了主要矛盾，既大幅度减少了记忆因素，使规则简化，又可以在编码过程中较少地遇到麻烦。

三、多字根共容原则

字形编码方案设计中，一个字根独占一个键位的方法已很少有人使用。因为这样作使键位增多，冗余惊人，浪费太大。一般

在选用一百五十至三百个字根时，可设计二十六至一百个键位，每一键位平均可放置二至六个字根。字根经归并组合共处一键，应考虑以下两点：

1.它们的组合使总的重码较少。

2.它们互相之间最好有某种共性或联系，以便记忆。

键位越少，字根越难共容；字根越大，越易相互排斥。字根共容一键的直接好处是使键位减少。在键位压缩到二十六个时，一百三十种字根平均每键五个。它们的共容还必须靠取码规则来保证。

例如："王"与"干"，加"氵"都是字，因此不能共容一键。但如规定再补取末笔，则它们便可能共容了。

四、多余信息删除原则

构成汉字的基本字根多少不一。有的一个字根即是一个汉字。有的汉字则由六、七个字根构成。例如"赣"，即由立、早、𠂇、工、贝构成。如依次把这些字根取码，则必然会因这些为数不多、使用频度不高的字而扩大了"最大码长"。这对整个编码效率可能影响不太大，但给机器的处理（例如建立检索表）却带来不便，以至存贮空间的浪费。所以，应当避免"全字根取码法"，采取"部分字根取码法"。例如，在WBZX方案中，对于多合字，规定取它的第一、第二、第三及末一个字根编码，余不计，删除"多余"信息。上例中"赣"，只取立、早、𠂇、贝；"慧"只取立、早、夕、心即可。

使用"多余信息删除原则"，成功地使最大码长限制在某一规定范围内。