



即学即会电脑丛书

谢柏清 谢新洲主编

中文输入 无师自通

张化瑞 编著



東方出版社

中文输入 无师自通

東方

391.14

ZHR/1

图书在版编目(CIP)数据

中文输入 无师自通/张化瑞编著. - 北京:东方出版社,1997.6

ISBN 7-5060-0806-8

I. 中… II. 张… III. 汉字输入系统 - 基本知识 IV. TP391

中文输入 无师自通

张化瑞 编著

出 版: 东方出版社(北京朝内大街 166 号, 邮编: 100706)

责任编辑: 吴玉萍

印 刷: 北京展望印刷厂

经 销: 新华书店

开 本: 787×1092 毫米 1/16 印张: 9.125 字数: 219 千字

版 次: 1997 年 6 月第 1 版 1997 年 6 月第 1 次印刷

书 号: ISBN 7-5060-0806-8/T·7

印 数: 1 - 5000 册

定 价: 16.1 元

即学即用电脑丛书

前言

计算机的迅速发展,使其应用已经遍及到社会生活的各个方面。计算机不仅广泛应用于工业、农业、交通运输、银行、金融财会、新闻印刷等各行各业,而且已经开始普及到家庭。现在千家万户购买微型计算机已经形成热潮,出现了计算机热。目前我国使用计算机的人员不仅是专业人员或对计算机懂的较多的非专业人员,而且包括千家万户的普通使用者。这些人的专业各异,文化水平也有较大差别,正是这批人推动着电脑技术向家庭化、实用化、多用途的方向发展。

如何使千家万户的普通使用者快速、方便地掌握计算机的知识和计算机的使用,编写适合初学者学习的通俗读物,是非常迫切的任务。

目前计算机的书籍很多,大致可以分为以下几种类型:

第一类是计算机的使用手册,这是针对某种类型的机器或软件写的。对于没有接触过计算机技术的人来说,学习起来是十分困难的。大量抽象的概念、命令语句的格式和操作规定,不好理解,也不容易记忆和掌握。初学者对这种书往往望而生畏,即使买上几本厚厚的书,能看懂的也没有几页,使用电脑还是不知从何下手。

第二类是计算机的教材,这类书针对高等学校学生学习计算机的基本知识和基本操作系统地进行讲解,要求学习的人有一定的数学基础,如对数的进制、逻辑运算、国际单位制等有初步的了解,对于文化水平较低的普通用户甚至受过高等教育的读者学习起来也有一定的困难。

第三类是翻译的各类国外电脑书籍,有手册型的也有傻瓜型的,这类书是普通用户的参考书,但有些译者翻译时在专有名词上只引用原文,而不说明原意,使初学者还是不得要领。

第四类是普及型的电脑读物。这些书五花八门,但真正能快速引导读者入门和提高的书还不多见。

通过多年计算机教学的实践和接触各类人员学习电脑过程中发生的问题,我们认为学习计算机的使用有两种办法,一种办法是学习某个软件,如有的大专生想学习汉字编辑软件的使用,可以选择 WPS 软件也可以选择 Windows 的书写器或 Word 软件进行学习;有的人想学程序设计,可以选择 C 语言、Pascal 语言或 FORTRAN 语言进行学习和编程;又有的人想学习绘图,可选择一种绘图软件进行学习。这种学习是针对某个软件来学习。另一种学习方法是用电脑干工作,即学即用。譬如一个教师,要写讲义或教材,就要学会用电脑写作,如果家中只有一台 286 的机器,最好选择使用 WPS 软件进行写作,如果有一台 386 以上的电脑,又有 8M 以上的内存,就可以考虑使用 Windows 的 Word 软件来进行写作。于是就可以有针对性地学习这两种软件,这就是用电脑写作。若此教师要处理学生的成绩,就可以学习使用电子表格软件或数据库软件,也就是用什么学什么,即学即用。实践证明对于多数的在职人员或电脑的初学者

者使用后一种方法进行学习,见效快,收获大,易产生兴趣,也能迅速见效,提高工作效率,产生效益。对于大、中、小学生学会用这种方法学习,可以作为前一种方法学习的补充,能增加对问题的深入了解,熟练地操作和使用计算机。

基于上述思想,我们编写这套丛书,能从完成某个任务入手,掌握电脑的各种软件及硬件的功能,力图能将软件的功能使用得巧妙、纯熟,举出生动的实例,给读者以启迪。同时可以按照书中的实例进行操作,学习并完成自己要用电脑做的事。

即学即用电脑丛书的选题的考虑是:一方面是要适应当前各类读者学习、使用电脑的需要,另一方面也考虑到较全面掌握和学习计算机的基础知识和技能的各个方向。目前先选择了十个有代表性的题目,今后还将不断地补充和扩展。这十个题目是:

- 一、电脑写作
- 二、电脑绘图
- 三、电脑游戏
- 四、家用数据管理——FoxBASE⁺数据库管理系统的使用
- 五、巧用 DOS 命令
- 六、即学即用 Windows
- 七、实用软件工具
- 八、电子表格——Excel 5.0 中文版简明使用指南
- 九、家用电脑购买前
- 十、中文输入 无师自通

即学即用电脑丛书的特点是突出实用性,强调通俗易懂、即学即用,提高读者学习使用计算机的兴趣,完整配套,形成系列。这是与目前的手册类或教材类计算机图书明显不同之处。我们力图从读者使用计算机要干什么入手,告诉读者怎样干,即包括应有什么样的硬件配置和使用什么软件,软件如何使用。若读者已经有了微机,由于机器的档次不同,建议使用不同的软件来达到完成工作的目的。在每个选题的讲解中举出大量的实例,说明如何使用各种软件,读者可以在学习这些实例的过程中举一反三,完成自己的工作任务,达到即学即用,掌握计算机的基本知识和技能。

本系列丛书的作者有的是具有多年计算机教学经验的、能熟练使用个人电脑的教授、副教授、讲师、工程师,也有的是熟练使用电脑的研究生和高年级大学生。书中讲到的各种问题和方法都是经过实验的,可供读者借鉴,并尽可能将实验的方法也教会读者。古语说得好:“熟能生巧”,这些作者在使用电脑的过程中,已经积累了丰富的经验,所写的内容是熟练使用电脑的结晶——技巧。将这些经验奉献给读者,和广大读者一起继续学习和使用电脑,达到更纯熟的境地。

欢迎计算机专家、学者以及各行各业的读者、电脑爱好者对本套丛书提出改进意见,提出批评。

主编 谢柏青 谢新洲

1997 年 1 月

目 录

即学即用电脑丛书前言

引言 沧桑巨变·····	(1)
第一章 沙场点兵——汉字的字音和字形·····	(3)
第二章 巧夺天工——自通中文序列输入法·····	(9)
§ 2.1 先睹为快·····	(9)
一、“新生事物的成长”·····	(10)
二、“有朋自远方来”·····	(10)
三、“千里共婵娟”·····	(11)
§ 2.2 略作说明·····	(11)
一、零声母·····	(11)
二、韵母 ü 和 üe·····	(12)
§ 2.3 输入界面·····	(12)
一、输入窗口·····	(12)
二、选择确认·····	(13)
三、修改拼音·····	(14)
四、输入英文·····	(15)
五、功能按钮·····	(15)
§ 2.4 综合示例·····	(16)
一、《荷塘月色》选段·····	(16)
二、《北风跟太阳》·····	(22)
§ 2.5 输入技巧·····	(24)
一、适可而止·····	(24)
二、当机立断·····	(25)
§ 2.6 声韵双拼·····	(25)
一、键位对照·····	(25)
二、方式切换·····	(26)
三、动态提示·····	(27)
§ 2.7 标点对照·····	(28)
§ 2.8 抛砖引玉·····	(30)
第三章 曲尽人意——智能 ABC 中文输入法·····	(31)
§ 3.1 先看示例·····	(31)
§ 3.2 略作说明·····	(33)
§ 3.3 熟悉界面·····	(33)

§ 3.4 残途同归	(35)
一、拼音输入	(35)
二、笔形输入	(38)
三、辅音以形	(43)
四、双打输入	(44)
§ 3.5 断句造词	(45)
§ 3.6 往者可追	(49)
一、朦胧回忆	(49)
二、强制记忆	(51)
三、频度调整	(52)
四、光标跟随	(53)
五、词头词尾	(53)
§ 3.7 别有便道	(56)
一、以词定字	(56)
二、一级简码	(56)
三、数词量词	(57)
四、符号输入	(58)
五、英文输入	(59)
六、中文标点	(59)
七、动态键盘	(60)
§ 3.8 知己知彼	(64)
一、了解词库	(64)
二、按词输入	(65)
三、因人而异	(66)
第四章 自成一统——常见形码输入法	(67)
§ 4.1 表形码输入法	(67)
一、观形定码	(67)
二、笔画分类	(68)
三、部件分类	(69)
四、编码规则	(75)
五、功能设置	(85)
§ 4.2 郑码输入法	(93)
一、体系概观	(93)
二、笔画分类	(93)
三、字根分类	(94)
四、分解规则	(107)
五、取码规则	(109)
第五章 随心所欲——自制和改造输入法	(113)
§ 5.1 自制新法	(113)
一、特点概观	(113)

二、程序安装	(113)
三、码表格式	(113)
四、码表实例	(115)
五、码表转换	(118)
六、创建过程	(119)
七、实际应用	(119)
§ 5.2 改造旧法	(126)
一、重码选择	(126)
二、编码清除	(130)
§ 5.3 兼及其余	(132)
一、UCDOS	(132)
二、TWAY	(133)
结语 回归自然	(136)
参考文献	(137)

引言 沧桑巨变

传说中仓颉造字,“天雨粟,鬼夜哭。”

历史上蔡伦造纸,刀简废,笔墨兴。

文字是记录语言的工具,传播文明的载体。文字形成以后,人类才摆脱了蒙昧,进入文明时代。汉字的成形,即使以相当成熟的甲骨文作为标志,距今也已有三千多年。在世界四大文明古国中,只有中国的汉字得以延续和发展,一直保持了表意文字的形态。

汉字的发展,也历经沧桑。殷商初具雏形,春秋五霸擅盟,战国七雄纷争,遂使汉字的字形也处于混乱状态,分国而治,各行其是,“文字异形,言语异声。”到秦始皇统一中国,匡正文字,据说是由李斯订立小篆,才结束了春秋战国时代文字的混乱状态,出现了汉字字形的国家标准。尽管始皇的皇位的继承充其量不过是接二连三世,但其统一的文字的流传却达成百上千年。

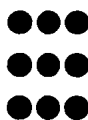
汉字形体的变化,从甲骨文到金文、大篆、小篆,都是线条结构;从小篆到隶书的演变,是汉字形体最大的一次变化。曲笔变为直笔,线条变为笔画,使书写速度显著加快,当然这也是和书写工具的改进联系在一起的。此后的变化,到草书、到楷书、到行书,或为进一步提高速度,或为使形体更加规范,基本都在这一框架内。到唐代,楷书作为正书的地位就固定下来,此后一千多年,没有再发生变化。

每次书写工具的变革都伴随着文字使用方便性的巨大飞跃。甲骨文时代,只有占卜时才作文字记录;到竹简的出现,“汗牛充栋”、“学富五车”并非夸大其辞;造纸普及,抄书又成了专门的职业;活字印刷,才真正使书籍大众化,文字的功用日渐彰显。

时至近代,西学东渐,西方的科学技术随同其载体拼音文字一起传入中国,引起了学者们对汉字的反思。由于拼音文字是直接表音的,学习起来比较容易,因为在学习文字之前就已初步掌握了语言,熟悉了发音,所以基本上没有什么困难。而汉字是表意文字,知其音未必能识其形,还必须从头学起,建立起字音和字形之间的联系,这无疑加长了识字教育所用的时间。新文化运动的主要矛头是对着传统儒家思想的,当然也包括其载体文言,倡导白话,甚至波及汉字,主张拼音化。有人甚至认为,汉字不废,中国难兴。现在看来,当时的主张未必全是正确的,尤其是在对汉字的看法这一问题上;但从当时来讲,倒应该说是必要的,在两千多年的封建传统面前,如果只说些模棱两可的话,恐怕不会有什么大的效果。汉字识字难,这是事实;方法得当,亦可化繁为简,但主要还应看其用起来是否方便。汉字是方块字,也就是平面结构,或称二维结构,对空间的利用率高,辨识直观。拼音文字则全是一字儿排开,是线性结构。举个例子来说,我们要数一群小鸟的个数,如果小鸟是排成一队的,便要数上一会;如果小鸟排成一个方阵,那一眼就能看出来。如下图所示:



排成一队



排成方阵

表意文字和表音文字的区别也与此类似。汉语是联合国工作语言之一,同一份文件,用几种语言刊行,汉语版本是篇幅最短的。

方圆不过八万里,上下何止五千年。世界在越变越小,历史却愈延愈长。机不可失,时不我待。操纵电脑“洋为中用”,驾驭汉字“古为今用”,乘汉字之舟,借西来之风,载华夏文明驶向明天的辉煌。

第一章 沙场点兵

——汉字的字音和字形

汉字在殷商时代雏形初具,到春秋战国羽翼渐丰,秦统小篆,汉变隶书,代代相传,生生不息,汉字的队伍便不断壮大。甲骨文中,现在能辨识的不同汉字已达数千个;东汉许慎作《说文解字》,收字九千多;《康熙字典》容纳的字数已超四万;九十年代出版的《汉语大词典》收字个数达到五万多。这也还没有囊括汉字家族的所有成员。汉字的总个数到底有多少,有人说六万,有人说八万,没有一个确切的数字。但有一点是肯定的,这就是,在不管几万的总汉字中,只有几千是常用的,现代通用的汉字也不到一万。其余的大多数字,我们很可能一辈子也碰不上一次,更不用说去使用了,除非你是研究古文字、古文献的。因此,本书以后讨论的汉字输入问题,针对现代汉语中的通用字。

在现代汉语中,通用字有多少呢?国家公布的通用字表包括七千字。这七千字能通用到什么程度呢?我们不妨来看看对从三亿多字的书籍、报刊中抽样的两千多万字的文字材料的统计结果:

两千四百字就能覆盖统计材料中汉字总出现次数的百分之九十九;
三千八百字就能覆盖统计材料中汉字总出现次数的百分之九十九点九;
五千二百字就能覆盖统计材料中汉字总出现次数的百分之九十九点九九;
六千六百字就能覆盖统计材料中汉字总出现次数的百分之九十九点九九九……

金无足赤,字无止数。在剩下的千分之一、万分之一、十万分之一中,大都是那些生僻的地名用字、人名用字和古汉语用字,不认识这些字对于我们理解语义基本没有影响。从上面的数字中可以看出一个大致规律:每多认识一千四百个字,就可以在统计材料中汉字总出现次数的百分数的小数点后面再加一个九,把不认识的字的比例降低一个数量级。这说明,越靠后的字,对阅读起的作用越小,七千以后,就几乎可以忽略不计了,因此,七千通用字,已经足敷日用了。

同是通用字,不同的字的地位也是不一样的。有的字总要用;有的字经常用;有的字时而用;有的字偶然用;有的字极少用。如“的”字,差不多每二三十个字里面就要用到一次;而“曩”字,读几十篇文章也不见得碰着一回。根据使用频度的不同,又可把通用字分为最常用字、常用字、次常用字、不常用字和偶用字五个等级。现代汉语中最常用的十个汉字是:

的 一 是 在 不 这 我 主 人 们

统计使用的材料不同,后面的次序可能略有差异,但“的”字当头,后面的字难以望其项背,其统治地位无法动摇。这也是现代汉语的特点决定的,在古代汉语中,岂有“的”之市场,应是“之”的天

地。

国家颁布的常用字表,收字三千五百个。又按其常用程度分成两级。

在计算机中,通常使用的是信息交换用汉字编码集基本集,也分为两级,一级字 3755 个,二级字 3008 个。我们后面讨论的汉字输入问题一般就限制在这个范围之内。

每个汉字都包含了字形、字音和字义的信息,甚至还记载着几千年来字形、字音和字义发展变化的历史,可以说每个汉字都是全息。汉字发展至今,我们只见今形、唯闻时音,对于古形、古音完全可以置之不理。古形尚有书可稽,古音便只有靠学者们去旁敲侧击。至于字义,则无法割断联系,因为有的字某一义现在不用了,但它却在成语中传袭下来,如“慢”字,现在只有迟缓的意思,在词语“傲慢无礼”中,却表示不恭,对不可慢待之客,便要说“有失远迎”。而且字义的分歧也多,一般的字都有几个义项,多的至于十几个,确定性不强,因此通常不采用字义作为输入汉字时的编码因素。

字形和字音也存在一字多形和一字多音的问题。

一字多形是指有的字有繁体、异体,国家已经审定过异体字,颁布了简化字总表,繁体字和异体字都是非规范字,按规范字输入,就不存在一字多形的问题了。近年来由于港台及海外的影响,繁体字又有死灰复燃、沉渣再起之势,这都是汉字规范所不能允许的,《人民日报(海外版)》也已改用简体字,繁体字是不会再有市场的了。

多音字在现代汉语中尽管相当普遍,但其所占的比例并不高,大约为十分之一左右。在多音字中,又以两音字为主,三音以上的也是只占其中的十分之一左右。所以,总音数只比总字数多出九分之一左右,相差无几,便于处理。

汉字是表意文字,其中五分之四是形声字,形旁表示义类,即其意义和哪一类事物有关,声旁表示读音。由于汉字的音、形、义都是随着时间的推移不断变化的,而字音的变化最快,字义次之,字形自隶变以后在漫长的封建社会中几乎就没有改变过,所以就出现了音、义和形脱节的问题。现在形旁仍和意义相关的约占形声字的七分之六左右,与意义不相关的不到七分之一。如“杨”为木本,“油”归水类;“路”在足下,“顶”于头上。

在所有形旁中,最常用的十个是:

讠 艹 口 扌 木 钅 亻 虫 讠 土

在形声字中,声旁与本字的读音完全相同的(即声母、韵母、声调都相同)将近五分之二,如“芳菲”;基本相同的(声母、韵母相同,声调不同)将近五分之一,如“梅花”。这两项合起来就超过一半,“秀才认字念半边”,多半还是对的。另外有四分之一声母或韵母一项相同,如“英”和“央”合声,“松”和“公”押韵。还有五分之一声母和韵母全不同,如“海涛”。

在所有声旁中,最常用的十个是:

各 非 古 令 肖 干 乍 者 圭 青

常用的形旁和声旁,它们的组字能力都是相当强的。我们来看一下拿前十个最常用的形旁和声旁进行组合的情况:

形/声	各	非	古	令	肖	干	乍	者	圭	青	总计
讠	洛		沽	冷	消	汗		渚	洼	清	8
艹		菲	苦	苓				著		菁	5
口	咯	啡	咕	吟	哨		咋		哇		7
扌		排		拎	捎				挂		4
木	格		枯	桉	梢	杆	柞	楮	桂		8
钅	铭		钴	铃	销			锺		锒	6
亻		俳	估	伶	俏		作		佳	倩	7
虫			蛄	蛉	蛸		蚱		蛙	蜻	6
讠		诽	诘		诮	讪	诈	诸	洼	清	8
土											0
总计	4	5	8	8	8	3	5	5	7	6	59

由上表可以看出,在最常用的形旁和声旁的所有组合中,一半以上是有意义的(即存在对应的汉字)。

利用声旁和读音之间的关系,可以比较有效地确定一组字的读音,帮助分不清某些声母和韵母的人纠正发音。尽管偶有例外,规律性仍然很强。如:

在常用字中,要区分 zhi 和 zi 这两个音,可以利用它们各自对应的常用声旁:

发 zhi 音的字的声旁多为“支、直、旨、执、只、止、至”,如:

支	枝 肢 吱	
直	值 植 置 殖	
旨	指 脂 酯	
执	挚 贲 繫 鸷	
只	织 职 帜 咫 枳	(chi 炽)
止	址 趾 芷 祉	(chi 耻 齿)
至	致 侄 桎 室 蛭	(shi 室)

后面括号中为常见的发另外的音的字。

发 zi 音的字的声旁多为“次(ci)、子、兹”,如:

次	资 姿 咨 恣	(ci 瓷)
子	字 仔 籽 孜	
兹	滋 孳	(ci 磁 慈)

要区分 shi 和 si 这两个音,可以利用它们各自对应的常用声旁:

发 shi 音的字的声旁多为“寺(shi)、式”,如:

寺	诗 侍 恃 峙*	(chi 持 zhi 痔 峙*)
式	试 拭 轼 弑	

在右上角加星号的为多音字。

要注意,“寺”的本音 si 与作声旁时的读音 shi 是容易混淆的。

发 si 音的字的声旁多为“斯、司”,如:

斯	嘶 嘶 厮
司	伺 伺* 嗣 (ci 词 祠 伺*)

要区分 li 和 ni 这两个音,可以利用它们各自对应的常用声旁:

发 li 音的字的声旁多为“里、利、丽、离、历、厉”,如:

里	理 厘 鲤 狸 哩 狸 俚 锂
利	梨 犁 莉 俐 痢 痢
丽	鹧 丽 骊 酃
离	璃 篱 漓
历	沥 雳 析
厉	励 砺 蛎

发 ni 音的字的声旁多为“尼、(繁体儿)”,如:

尼	泥 呢* 妮 呢 倪 (ne 呢*)
(繁体儿)	霓 倪 颀 貌 睨

要区分 fu 和 hu 这两个音,可以利用它们各自对应的常用声旁:

发 fu 音的字的声旁多为“(副字旁)、付、府、夫、孚、复、甫、(数字旁)、弗、伏、父、卜”,如:

(副字旁)	福 富 副 幅 辐 蝠 匐	(bi 逼)
付	符 府 附 咐 拊 跗	
府	俯 腑 腐	
夫	扶 肤 芙 麸	
孚	浮 俘 孵	
复	腹 馥 馥	(bi 復)
甫	辅 脯*	(pu 浦 铺 脯* 匍 bu 捕 哺)
(数字旁)	傅 缚 敷	(pu 溥 bo 博 搏 膊)
弗	拂 佛* 氟	(fei 沸 费 狒 fo 佛*)
伏	袱 茯	
父	斧 釜	
卜	赴 讣	(bu 补 pu 扑 朴* 仆)

发 hu 音的字的声旁多为“胡、户”,如:

胡	湖 糊 葫
户	护 沪 扈 (lu 芦 庐 炉)

诸如此类,皆可为鉴。

对不认识或发音不准的字,就要通过查字典来解决。但一般人不愿意查字典,除自身因素外,字典难查也是一个很重要的原因。一般都是不认识的字才去查,这时音序检字就起不了作用了,只能用部首检字法或笔画检字法。

用部首检字法,首先要定出部首来,部首在字的左边、上部、外面的还好办,在右边、下部、里面的便要稍事斟酌,非左非右、不上不下、亦里亦外的就难上加难了。这且不说,定出了部首,还要数部首的画数;数出了画数,再从同画数的部首中按照起笔顺序找到要查的那个,记下该部首的序号或在检字表中的页码;到了这一部首的麾下,还要重复一遍刚才的过程:数出要查的字除了部首以外剩下的那一部分共有几画,再从同画数的字中挑选,首笔相同看次笔,次笔相同看季笔,如此反复,直至查出。如果在上述过程中有一步定错,如归什么部首,部首有几画,除了部首又有几画,哪儿错了哪儿返工,归部不对从头再来。

笔画检字,过程比部首检字简单些,即省去了定部首这一步,并把两次数笔画的过程合二为一,变成只要数出整个字的笔画数,再从同画数的字中按笔画顺序查找。

不管怎么查,都要先数笔画,而数笔画这一工作,实是极啰嗦的,在你一笔笔数的过程中,就已经把每个笔画都默念了一遍。为什么不把默念的每个笔画都直接用来查字呢?就象查英语单词那样,看到了第一个字母就可以向词典中一定的页数翻,或前或后,略作调整,再根据第二个字母进行修正,逐步缩小范围,直至准确定位。事实上,常常是不等辨到最后一个字母,这个词就唯一确定了,后面的字母根本未用。先数笔画数再查汉字,就跟先数出一个英语单词有几个字母再从相同字母数的单词中查找一样,都是没有必要的。笔画数所包含的信息量是很小的,其区分作用还不如两个笔画大;但要确定笔画数却是相当费时费力的,一是要整个字过一遍,二是有些结构的笔画数很难确定,三是容易数错。

有人可能要说,在分部首或笔画数的情况下,从相同画数的字中查找尚且很困难,何况不分部首和笔画数呢?问题正在这里。在一般的字典上的部首或笔画检字表中,部首检字表是只列部首和除部首以外的笔画数,这两项相同的便排在一起,内部的次序,说是按起笔、次笔依次排,但并不列出,比如你要找除部首外以撇作起笔的字,还得从头检阅,这岂不麻烦。由于汉字是方块字,具有平面结构,易于辨识,略难检索;但只要把汉字按笔顺转化成笔画序列,五种笔画,横、竖、撇、捺、折,分别用数字 1,2,3,4,5 来代表,以笔画序列作为索引,汉字就和英语单词一样易于检索了。如:

汉字	笔画索引
札	12345
一	1
二	11
三	111
十	12
工	121
大	134
人	34
不	1324
是	251112134
的	32511354

按笔画索引的顺序排出来,就是:

1	一
11	二
111	三
12	十

121	工
12345	札
1324	不
134	大
251112134	是
32511354	的
34	人

当然这里只是简单的举了一个例子,实际上在这个例子中,前三个笔画已经能起到完全的区分作用。实际编排时,可以把一组字的笔画索引中前面相同的若干位提出来,列在前面,如:

35251211	鱼
15	鲢
21	鲇
25	鲤
41	鳙
43	鲜
51	鲫

在技术上,当然也可以排得紧凑一些,如:

413	广
1234	床
1234	麻
13	磨
21	靡
3112	摹
3115	麾
32	魔
35	麼
43	糜
55	縻
441	彳
1234	沐
1234	淋
2	滑
25	湘
33	浙
342	漆
345	淞
4	沐

这样查起来还是很迅速的。而不必纠缠于笔画数和部首,一律按笔顺来。原来数笔画的功夫,现在就把字查出来了。不是按部首排,笔画索引又按常用部件分段,兼得部首检字之妙,又无确定部首之苦。

第二章 巧夺天工

——自通中文序列输入法

“自通”中文序列输入平台(AUTOWAY)是一种面向广大普通用户开发的、以语言规范为标准的无编码中文输入系统,采用拼音输入中文时其智能测度达到96%左右,基本上不需要选择同音字。AUTOWAY的核心是一个高度自动化和高度智能化的系统,它能最大限度地发挥计算机非凡的计算能力,对用户输入的低冗余序列自动进行语词切分和序列译码,从而减少用户使用时的思维负担,使学习更为容易,使用更加轻松。

“自通”中文序列输入平台(AUTOWAY)是由北京隆光威尔公司开发的,发明人为蒋子刚。目前推出的是1.0版本,刚刚经过了两千万字的训练,现在已经显示出不同凡响的智能。在后续版本中,随着训练量的增加和软件的改进,将会具有更高的智能程度,最终可能达到99%左右,也即,每输入一百个字,只需要选择一次。基本上就是你只要连续地输入拼音,系统自动给转换成汉字,如同拼音文字的输入一样,而且不需要加空格,有过之而无不及。这样就可以使十几亿使用汉字的华人从繁重的录入工作中解脱出来,在信息高速公路上奔驰。

“自通”内部采用的马尔可夫过程和广义哈夫曼码,都是在信息领域久攻不下的阵地,一经突破,自显威力。采用马尔可夫过程,即把输入的汉字序列看成是前后有关联的,而非孤立无援的。这样就可以综合考虑一个字所处的环境,来具体确定这个字,而不需要更多的信息,这就称为低冗余序列输入,就是说,你输入的信息几乎都是有用的,重复的成分很少,不做无用功。

汉字输入从单字输入,到词语输入,再到序列输入,是一个逐步前进的过程。采用词语输入,已经考虑和利用到了词语内部字与字之间的联系,但还没有利用词语和词语之间的联系。采用序列输入,则是把整个句子中字与字之间的联系都用上了,同时还包括邻近的几个字之间的联系。

序列输入技术用于汉字的键盘输入,是其第一个突破口,然而其作用远不止于此,用于语音输入、手写输入、印刷体识别,都大有可为。更进一步,还可用于多语种之间的自动翻译,前景是非常广阔的。

§ 2.1 先睹为快

百闻不如一见,事实胜于雄辩。

不妨先看几个例子,增加感性认识。

一、“新生事物的成长”

要输入“任何新生事物的成长都是要经过艰难曲折的。”这句话,只需键入:

renhexinshengshiwudechengzhangdoushiyaojingguojiannanquzhede.

中间不需要加任何空格。原封不动地把每个字所对应的汉语拼音键入即可。

当你输完几个字的拼音以后,系统就会把前面的拼音转换成汉字。如输入上面那句话时,当你打完了“任何”的拼音 renhe,再输“新生事物”的拼音 xinshengshiwu 时,不等你把 xinshengshiwu 打完,系统就已把 renhe 转换成了“任何”,显示在 renhe 原来占的地方上,跟后面尚未转换的拼音 xinshengshiwu 连为一体,如下所示:

任何 xin sheng shi wu

当你接着输入“的”字的拼音 de 时,“新生事物”就已转换过来,变成:

任何新生事物 de

不断重复这一过程,就得到了整个句子。我们看一下系统逐步转换的中间结果:

任何 xin sheng shi wu

任何新生事物 de cheng zhang

任何新生事物的成长 dou shi

任何新生事物的成长都是 yao jing guo

任何新生事物的成长都是要经过 jian nan qu zhe

任何新生事物的成长都是要经过艰难曲折 de

任何新生事物的成长都是要经过艰难曲折的。

转换的时机是,当系统认为一个拼音序列无法通过再增加一个音节变成另外一个有意义的组合时,就进行转换。

二、“有朋自远方来”

输入“有朋自远方来,不亦乐乎?”

只需按拼音输入:

youpengziyuanfanglai,buyilehu?

系统就自动转换成对应的汉字,逐步转换的过程如下:

有朋 zi yuan fang lai

有朋自远方来,bu yi le hu

有朋自远方来,不亦乐乎?

需要注意的是,尽管在输入拼音的时候不要求加空格,但系统在显示拼音的时候还是按音节显示的,这是由于系统自动进行了音节的划分;在输入拼音的时候也不要求在词与词之间加空格,但系统在作转换的时候还是以广义的语词(即常用的语言片断,与词语的概念不同,如本例中的“有朋”、“自远方来”、“不亦乐乎”)为单位进行的,系统在内部自动进行了语词的划分。