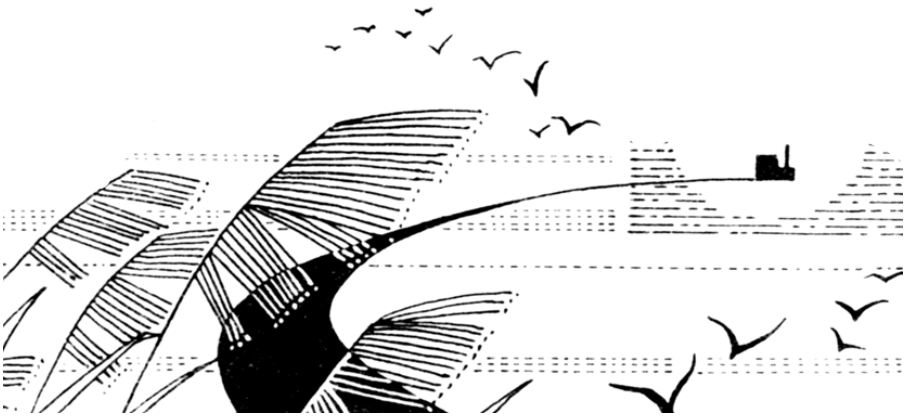
语言文字现代化（二）

马敏 主编



目　录

数字键盘汉字输入技术.....1
中日韩大字符集文字编码的比较研究.....7
智能化、标准化是五笔字型的方向.....13
使中文走向世界.....20
汉语指令电脑简介.....30
拼音号码输入法.....31
数字化内容与数字化工具.....32
智能技术与系统.....41
基于统计的中文词自动分类研究.....41
基于语料库的汉语字词相关性研究.....48
基于数据库的汉字构形学研究.....56
现代汉语词类相关性的统计与分析.....64
汉语盲文翻译的研究.....71
中文 Windows 环境下韩汉文字同屏显示.....83
日汉机译系统中日语分词技术.....89
日汉电子词典结构的研究与实现.....96
《现代汉语新词语信息词典》的结构.....102
CAI 更应以人为本.....111
对试题库建设的认识和做法.....122
基于网络的大学英语 CAI 课件.....131
新名词输入近代中国的文化内涵.....139
CJK 字根系统研究.....144
香港、澳门的音译转写问题.....151
看韩国的汉语、汉字研究.....158
信息技术和人性的异化与复归.....165

数字键盘汉字输入技术

一、引言

随着信息技术的飞速发展，计算机及其相关电子信息产品在向小型化、数字化方向发展，计算、上网与通信相结合。汉字输入不再局限于台式计算机上，手机、双向寻呼机、手持计算机（H/PC)、小型信息终端和遥控等仅有 10 个数字键的小型设备也需要输入或处理汉字。

尽管联机手写汉字识别、语音识别技术取得了很大的进步，但这两种输入方式总会有些人因识别率低而很难输入汉字，因此在小型化设备中用数字键盘输入是必不可少的。汉字输入一定要简单易学、实用快捷，上手能用。《信息技术数字键盘汉字输入通用要求》（GB/T18031-2000）国家标准已经发布，规定了一些便于通用掌握的基本要求，使数字键盘汉字输入规范化。

对于在通讯产品中使用汉字数字码，可以分为三个层次：字输入、词输入、连续语段输入（即智能输入)。选用哪一种层次要根据小型通讯产品中 CPU 速度和存储量而定。

汉字智能输入应用计算机人工智能技术，使操作更加简便，几乎不需要选字，尤其对于由 0-9 数字组成的拼音码（音码）和笔画数字码（形码)，重码率相对地来说是比较高一些，更加需要智能化，依靠汉语上下文关系计算机自动地进行同码字的选择，转换为汉字，输入者基本上不用选字。

利用全拼音数字码和全笔画数字码在 10 个数字键

上连续输入是一项新技术，它面向最普通的用户，做到不需培训，上手能用，并能够自动地转换数字输入码为汉字。我们的数字音码及形码智能汉字输入方法完全符合《信息技术数字键盘汉字输入通用技术要求》国家标准，且击键数最少。采用计算机人工智能技术解决了重码的选字问题。音码采用现行的汉语拼音，在标有拼音字母的数字键上输入，已会拼音的用户不需要再学习就可以使用。形码采用以数字为代码的笔画数字码，该编码符合《现代汉语通用字笔顺规范》，并有容错能力。因此该形码易学易用，也不需要培训。

小型化的通讯产品对智能输入技术提出更高的要求，因小型化的产品存储空间小，要求汉语上下文关系库不能太大。因 CPU 速度低对智能算法提出更高要求。这方面已经达到手持通讯产品能够接受的实用水平。

二、用拼音数字码输入汉字

小型电子通讯产品一般只有 10 个数字键，在这类键盘上可使用智能全拼音输入，下面介绍在数字键上利用汉语拼音输入汉字的方法。该拼音字母排列为 GB/T18031-2000 国家标准，键位分布如下：

1 ABC	2 DEF	3 GHI
4 JK	5 LM	6 NP
7 OQR	8 STU	9 VWX
*	0YZ	#

使用规则：声母字母用黑体表示的，用[#]号确认，

不是黑体表示的声母字母用[*]确认。这样就可以用拼音输入汉字了。

例字：李 LI 击 53，按[#]键确认，米 MI 也击 53，用[*]确认。

由于该方案拼音字母是按次序排列在数字键上，使用起来非常方便，不需要任何培训就能在数字键上输入汉字。

三、用笔画数字码输入汉字

音码有音的长处也有它的弱点，形码正补充了音的不足。我们的金笔画数字码---形码编码规则非常简单，符合国家语委颁布的《现代汉语通用字笔顺规范》，并有容错能力。

笔画数字码是以汉字的五种基本笔画为基础，利用笔画之间有无交叉的特性，将五种基本笔画扩展为 10 个数字而形成的数字编码法，每个汉字全笔画数字码的码长最大为 5。该方案成为制定 GB/T18031-2000 国家标准技术参数样本方案。

（一）笔画数字码编码方法

笔画代码规则：笔画	横	竖	撇	点	折
	一	丨	丿	丶	乙
无交叉	1	2	3	4	5
有交叉	6	7	8	9	0

笔画归并：提归一，竖钩归丨，捺归丶，各种折笔均归乙

（二）输入实例

清华大学

4416132807684443 46

笔画数字码是按汉字的笔顺次序取码，这与汉字的书写习惯保持一致。采用容易判断的笔画有无交叉的特征，又大大地拓展了编码范围。加上计算机人工智能技术的运用，使得汉字平均笔画数虽在 10 画以上，每个汉字最多取五个笔画就足以高准确率地连续数入汉字了。

四、数字键盘汉字输入的人工智能处理技术

数字音码及形码智能输入系统的智能处理，主要体现在自动处理重码的问题，不是靠人工选择同码的候选字，而是用马尔柯夫模型作为输入码到汉字的转换模型。

$S=<S_1S_2.....S_n>$为一句输入串，S_i 为一个词或者字的编码；

$T=<T_1T_2.....T_n>$为一可能的汉字串，T_i 为输入码 S_i 的一个词或字（即候选词或字）；

P(T|S)表示当输入是 S 时，输出为 T 的概率。

当 P(O|S)=MAX{P(T|S)}时，O 即为最佳结果。由 Bayes 公式得：

P(O|S)=MAX{P(T|S)}=MAX{P(T)*P(S|T)/P(S)}

在上式中，由于输入 S 已定，故 P(S)项不影响选择，可以不加考虑。当 T 在候选集之内时，P(S|T)项可以用 1 来代替。因此

P(O|S)=MAX{P(T)}(1)

把汉语语句看作是一个 Markov 源（即某状态的发生概率仅与其以前的状态有关），那么：

$P(T)=P(T_1)P(T_2|T_1)...P(T_n|T_1...T_{n-1})$ （2）

如果我们认为第 i 个字的出现仅与前面很少的 n-1 个字有关，则问题就会大大简化。这样的模型叫做 N 元语法。如果采用二元语法模型（bi-gram），即取 n=2，也就是说，在确定第 i 个字时只考虑前面一个字的出现情况，则可得下式：

$P(T_2)=P(T_1)P(T_2|T_1)$ （3）

对于以上情况，其参数项 $P(T_2|T_1)$称为二元同现概率，它们可以从对语料文本的统计计算中获得。可以用最大似然估计法（MLE）来计算上述二元同现概率，其计算公式如下：

$P(T_2|T_1)=N(T_1T_2)/N(T_1)$ （4）

在上式中，N(S)是字符串 S 在语料库中出现的次数。

采用最大似然估计法（MLE）方法来计算模型的转移概率，在训练语料不足或参数空间庞大的情况下，会遇到数据稀疏（DataSparsity）问题：即有许多合法的在未来的文本中要遇到的同现现象在统计语料中从未出现过，因而在遇到这种情况时，会出现零概率情况。合理地平滑处理数据稀疏的估值算法，目前有很多。解决此问题的一种较简单的的方法是使用二元和单字频率的加权平均，该方法是 Markov 模型的数据平滑方法，其基本思想就是：若统计数据不充分,确切地说不可信时，我们宁可回到 n-1 元组来计算。在实际应用中，是用它们的线形组合来实现的。在我们的模型中是使用了此方法来计算同现概率。

在我们的模型中用同现概率反映汉语文本中汉字间的相邻关系。为了获得汉语词词和字字的同现概率，我们对大规模语料进行了统计得到同现概率库。输入汉字

数字码时用动态规划法求最佳路径，得到最可能的汉字语段。

与台式计算机不同，对于小型化通讯设备或手持计算机来说，CPU 速度低，存储量小，要求同现概率库不能太大，对于统计得到的同现概率，要进行筛选、整理、压缩，并根据用户的实际需要对求最佳路径算法进行改进，以达到手持通讯产品能够接受的实用水平。

用《人民日报》、《中国青年报》及《参考消息》等社科方面的文章进行测试，自动转换汉字的正确率如下：

智能输入自动转换汉字的正确率：

全拼音（字母键）94.2%

（数字键）93.9%

笔画数字码（数字键）95.6%

例句 1：工作时间不能踢足球

全拼音

数字键输入：376308783343161862638306738

字母键输入：GONGZUOSHIJIANBUNENGTIZUQIU

笔画数字码

字输入：

121323112511442521132454255251232512416714

字词输入：12132251142132454251232512l6

例句 2：实现经济发展和社会进步

全拼音

数字键输入：83393164363432103163283238343618

字母键输入：SHIXIANJINGJIFAZHANHESHEHUIJINBU

笔画数字码字输入：

4454416715551514414208094513643673145241341146687421213

字词输入 :44541655154408095136731452434668721

五、结论

从对数字键盘汉字输入分析不难看出该方法的优良特性是：

1．使用 0-9 编码的汉字数字码输入技术，包括拼音数字码及笔画数字码输入系统。可以分为三个层次：字输入、词输入、连续句输入（即智能输入）。简单易学，上手能用，普通用户都能方便地驾驭汉字输入。对于装有数字码智能输入的系统，由于应用了中文上下文关系的智能技术，输入更加方便，连续输入基本上不需要选字。

2．经测试可以达到如下指标

汉字输入平均码长：

单字输入时，平均码长小于 6 键/字（包括确认键，字词混合输入时，平均码长小于 4 键/字。

重码字键选率：

单字输入时，键选率为 11%；字词混合输入时，键选率为 10%。

中日韩大字符集文字编码的比较研究

一、引言

随着 Internet 的飞速发展，信息的交流与传递不再受地域和国界的限制，人们在网上接触到的是全球化的信息，众多的信息资源共享使得全人类的智慧得以充分的融合。作为信息的主要载体之一的文字，同图像、声音相比，具有占用空间少、信息运载量大的优点。但无

论是进行网络信息浏览、检索还是机器翻译，均有多种文字输入、显示或转换的要求。然而，由于各种语言的编码不同和各国操作系统、应用软件的本地化使得无法显示或同时显示一些不同语种的文字。语言的编码冲突在某种程度上限制了人们的信息交流。因此，解决各国语言编码之间的冲突也显得日益重要。

由于 Internet 上语言编码分布各异，尤其是大字符集文字的编码分布更是错综复杂，中、日、韩编码之间的冲突就一直是语言文字处理的一大障碍。

二、问题的产生与分析

文化的差异导致了不同的语言文字的出现。对于西方文字，由于字符集较小，所以一般采用单字节编码（每个文字的编码长度为 8 位），而东方一些国家语言中文字数量较多，达数千个，通常称为大字符集文字，这些文字只有通过一个以上的字节才能实现其在计算机中的完全表示。

由于各国在对本国文字进行编码时，并未考虑到与其它各国编码的兼容性（同时处理的需要），所以产生了编码冲突现象。主要表现在以下几个方面：

（一）各国编码多样化

中、日、韩这三种语言是最具有代表性的东方大字符集语言，各有数种编码表示。

中文常见编码：

GB2312-80 中文简体国标码（汉字数：6763）

UNICODEWindowsNT,UNIX 等系统采用的编码(包括繁简体)

BIG5 中文繁体编码(台湾，汉字数：13053)

GBK 扩展中文 GB 编码

CJK 中日韩大字符集编码

（其中 GB2312-80 编码为现今常见编码）

日文常见编码：

S-JIS 日文系统常采用的编码

JIS

EUCODE

（其中 S-JIS 编码为现今常见编码）

韩文常见编码：

WansungCode(KSC5601-1987)

UnifiedHangulCode（在 WansungCode 上扩展的）

JohabCode(KSC5601-1992)

（其中 UnifiedHangulCode 编码为现今常用编码）

（二）外部冲突与内部冲突

外部冲突指的是不同种类语言之间的编码冲突。由于这三种大字符集语言编码，都是以两个字节为基本单位进行编码的。因此地址空间范围应该是在第一个字节是 0x00 到 0xff 之间和第二个字节是 0x00 到 0xff 之间的空间范围内。下面以这三种语言的常见编码中文 GB2312-80，日文 S-JIS,韩文 UnifiedHangulCode 为例画出的编码分布示意图。（此处未考虑日文半角文字编码）

（三）部分冲突和全部冲突

通过上图不难看出，中、韩编码有一块重叠区，日、韩编码之间有多处形成重叠区，还有一块由三种语言共同形成的重叠区，中文区完全与其它语言重叠。即中、韩编码之间存在着全部冲突，中、日之间和日、韩之间存在着部分冲突。这样就在计算机对文字处理的时候面临着一个难题：当有中、日、韩三种字符串资源时，如

何能在一台计算机上分辨出其所属的编码并将它们正确的显示出来。

三、现有的解决方案

（一）采用宽字符(WideCharacter)或多字节(Multibyte)编码方案

宽字符和多字节都是用多个字节来表示一个字符的方法，区别是宽字符中每个字符占有同样多数目的字节，而多字节中每个逻辑字符可以有不等长的字节编码。该类编码可以解决各国文字间的编码冲突问题，但它不能解决各国文字的兼容处理问题。CJK 是最新制订的针对中、日、韩三种语言的一字一码的大字符集编码方案，但是这种编码方案只能保证对中国大陆文字符号编码的兼容性，完全打乱了原日、韩文字的编码，因此对日本、韩国的本地化软件这种不兼容就是一个很难解决的问题。还有其他一些宽字符集编码方案（如：Unicode 等编码方案），Unicode 是一个世界范围的字符编码标准，Windows 在系统级字符和字符串处理中使用它，可以说 Unicode 在一定程度上改进了多国语言文字的处理能力。但是这种编码方案打乱了原来各国文字的编码，这种不兼容就是一个很难解决的问题，因为如果按照这个标准，那么原则上几乎所有与其不兼容的应用软件都应重写以适应固定大字符集编码方案的要求。这种重新编码而产生的不兼容性影响了其方案的推广应用。

（二）利用设置字体携带原编码信息的方案

该方法可以实现在同一编辑器(具有字体设置功能)内正常处理多种编码的文字，且每种文字都采用与该国语言标准编码兼容的编码方案。但在保证各国文字正常显示之前必须将各国文字的字体设置成与该国文字对应

的字体。它的主要缺点是不能在不具有字体设置功能的编辑器内（纯文本文件编辑器内）处理多文种文字以及无法同时在编辑器内正常显示两种或两种以上的编码文字。

（三）采用多文种处理软件

针对编码的冲突问题，各国均有适合本国特点需要的多文种处理软件，针对几种语言和编码进行显示和处理。该类多文种平台是通过用户对系统当前主语言编码页的人工切换来实现同一 Windows 下处理不同编码信息的，但这一类所谓的多文种处理软件，采用的编码方案是固定的，也存在着难以克服的不兼容问题，而且也无法对屏幕上的语言文字信息动态地进行识别，需要手动设置。这种处理方式使得某一时刻只能保证一种编码方式正常使用，而其他种类编码的文字信息显示出来的是乱码。

综上所述，一个固定的多文种编码方案存在着不可忽略的不兼容问题，无法正确识别出不同种类编码的文字，所以提出并采用基于智能识别技术的面向目标的动态文字编码方案，就是打破人们解决问题的常规思路，将平面编码空间扩展至立体编码空间去解决编码分配和编码冲突问题。这样，对用户来说不会造成编码混乱的局面。

四、基于智能识别技术的动态编码方案

（一）智能识别技术（全称为：各国文字编码的智能识别技术）

为了实现在同一屏幕下能够正确地显示不同编码的语言文字信息，首先必须得到该语言文字的编码信息，也就是确定该文字信息的编码类别，只有这样才有可能

实现对其的正确显示。而一段语言文字一般不会提供有关编码信息，(如果是以纯文本方式保存的文件，则绝对不会有编码信息)，如果采用 Windows 自带的记事本这种纯文本显示软件显示，它将按照系统缺省的编码方式来进行显示。那么当这种文字的编码与系统的编码方式不一样时，它就会被翻译成了一种新语言，也就是出现乱码。

由于每种语言的编码范围基本上都不相同，因此编码间可能出现非重叠部分，我们可以利用这种非重叠部分，通过对该应用程序字符串资源的识别，来确定该程序的国别和所用语言编码信息。

当然，这种识别过程对用户来说是一种透明式的操作，即在应用程序启动时，系统就自动取得其字符串资源并完成对其的语言编码识别和识别结果的登记。

（二）动态编码技术

一个固定的多文种编码方案存在严重的不兼容问题，不可能使各类编码同时显示。本系统的基于智能识别技术的面向目标的动态文字编码方案，就是将平面编码空间扩展至立体编码空间以解决编码分配和编码冲突问题。

在同一系统上采用多套编码（对于磁 8 盘上的数据可保证仅用一套编码)，而这些编码的选择使用在系统内部控制，根据不同的进程采用保证与该进程兼容的编码方案。即先对文字编码进行智能识别，根据识别出的编码种类选择相应的编码解释方案来进行处理。这样，即使在一句话中有两种编码，也可以保证将它们各自正常地显示出来，对用户来说不会造成编码混乱的局面，而且还可在同一系统下同时运行不同文种的应用程序。另

外由于每套编码又都是兼容某一种文字的多文种编码，所以在不同文种的应用程序中仍可进行多文种处理。

五、结束语

采用了基于智能识别技术的动态编码方案，使得对于不同编码种类的语言文字不必由用户预先设定主编码，即可进行多文种混合编辑，做到了在同一屏幕下正确地显示不同编码种类的语言文字，切实有效的解决了中、日、韩文字编码在中文平台上同时处理、相互转换的难题，在多文种文字处理、多文种信息交流等方面发挥了巨大的作用。

智能化、标准化是五笔字型的方向

尽管有越来越多的非键盘输入软件面世，但键盘输入仍然是文字录入的最基本方法。据统计，在汉字键盘输入法当中，五笔字型的市场份额约占各类输入软件的60%；为五笔字型输入法编制的软件，也多达十几种，比如王码五笔字型、陈桥智能五笔、万能五笔、世纪五笔输入法、五笔神码、五笔拼音、大五笔、GBK 五笔、晨曦五笔 GBK 等等。这些软件各有特点，但总的看来，还没有一个是集大成、各方面都能令人满意的。对于如何能让这个拥有大量用户的输入软件做得更好，下面谈谈我们的看法。为方便叙述，以下把中文 Windows 配置的全拼、ABC 智能拼音、郑码等输入法简称为系统输入法。

一、充分发挥计算机智能化的优点

五笔字型输入软件的设计者一开始就走了一条与拼音输入软件设计者方向相反的路。采用拼音编码的设计

者明白拼音码的先天不足--重码太多，于是就考虑如何利用计算机检索、判断速度快的优点在重码中选中预定的词。从高频先见、预设联想、以词定字、自动造词，直到如今根据句子自动选词，已经在很大程度上克服了因重码多而给选择字词所带来的困难，有效地提高了输入速度。

在那个计算机只采用GB2312-80字符集6763个汉字的年代，五笔字型曾是号称击键少、基本无重码的优秀输入法。但随着词库的增大，特别是到了中文 Windows 的 GBK 大字符集中，若采用五笔字型来编码，重码就会剧增。如果再加上一定量的词组，可以说大部分的字或词都会有重码。于是，五笔字型的设计者就考虑从编码的改进去补救。从不久前推出的王码 9801，就可以看出设计者的良苦用心。据介绍，该版本符合国家语委关于输入法的规定，字根的设定作了改进，多了几个组字能力颇强的字根如甫、气艮之类，减少了几个字根，还能够输入 BIG5 码。

从理论上讲，这样的编码应该是既方便学又方便用了，但实际效果却不尽如人意。首先是五笔字型的老用户不习惯，他们用原先的编码常常打不出预期的字。原先已经用得好好的了，如今要他们再学习新编码，谁愿意干呢？再加上相当一部分重码字的顺序也发生了变化，如部分二级简码反而排到了后面，大家都感到别扭。还有，用五笔码根本打不出 GBK 的字。能输入 GBK 的，其实还是拼音。据我们了解，愿意采用王码 9801 的老用户极少。老用户的情绪也影响到了新用户，从我们学院的学生最近学习汉字输入法的实际情况来看，没有一个选择王码 9801 的。

那么，面对新的形势，五笔字型输入软件的出路何在？那就是，走拼音输入软件的成功道路，利用计算机飞速增长的硬件资源，充分发掘计算机的智能。其实，早已有人用实践证明，这也是一条适用于五笔字型输入软件的成功之路。在这方面做得较好的，应该算陈桥智能五笔。以下我们就智能五笔等软件的设计，谈谈如何进一步改进。

（一）高频字词的调序

让调用过的字或词自动排在预选列的最前边，这并不是智能五笔的专利，拼音输入法在很久以前就采用了，应该说是最容易实现的。别看这么小小的改动，对于重码不太多的五笔来说，却十分有效：如果预选字词排在第一位，不用选，继续下边的输入它会自动上屏；如果排在第二位，可设为打空格选中，也毫不费事。估计能够解决 80%-90%重码问题。我们许多人愿意使用智能五笔，很大程度上就是看重它词组既多又能以按频率调序方式提高命中率的特点。这样看来，重码并不可怕，关键是怎么有效地解决它。

（二）输入信息的智能提示

智能五笔能检测到用户的输入码。如果用户没有使用简码或词组输入，它都能及时提示。这种设计对用户很有帮助，经过一段时间的使用，就会逐渐地掌握当前软件的各种输入码。

大多数字的五笔编码都好掌握，但也有少数字的编码不得要领，比如藏、舞、凹、凸。在这种时候用户往往会反复试，极大地影响了速度。我们设想，能不能参照拼音句输入的办法，当检獾接没Ль既胍桓鲎只虼首橛稚境　螅　12.锤　蓐洗位魅氲谋嗦爰扒懊嫖淖痔峁 ┌握