汉字输入技术与应用研讨会论文集

数字统一码原理与实践

徐万胥

东北师范大学电化教育系

【摘　要】本文系统地介绍了数字统一码汉字输入法。将汉字的基本笔画赋予顺序值和位置关系值，以笔画的顺序值与关系值作为汉字编码依据。数字统一码码元“1”～“9”，码长1～6码，编码字符包括ISO10646、GB18030中定义的全部汉字。数字统一码符合GB18031对数字键盘汉字输入法要求，具有易学、规范、高效的特点。

【关键词】统一码 ； 数字编码 ， 汉字输入法

1．引言

数字统一码是作者二十多年汉字编码研究的成果。数字统一码的设计思想是：坚持中、日、韩汉字编码方法统一；繁、简汉字编码方法统一； 字、词编码方法统一； 识字与识码规律统一。坚持编码的易学性，同时兼顾汉字的输入速度。

2．数字统一码的码元和键位安排

汉字是一个比较复杂二维图形，通常由若干个部件组成，而部件又由笔画组成。汉字的字形由笔画及其位置来决定。既使笔画完全相同，而笔画之间的位置不同，也可构成不同的汉字，例如，“于”和“亍”。

2．1 汉字笔画的顺序值

汉字的笔画有十几种，根据字典检字使用五个“一画部首”的惯例，以及数字编码的特点，我们取五种基本笔画：“横（一）”（含“提”）、“竖（丨）”（含“竖勾”）、“撇（丿）”（包括“啄”）、“点（丶）”（含“捺”）、“折（乙）”（包括左折和右折），并且将这五种笔画赋予顺序值，分别规定为：“1”、“2”、“3”、“4”、“5”。

2．2 汉字笔画之间的位置关系属性

笔画之间的位置关系属性是一个重要属性，可以作为汉字编码的依据。汉字中笔画之间的位置关系有：“相离”，如“八”；“相接”，如“丁”、“口”；“相交”，如“十”、“丰”等。实验表明，“相交”关系与前两者容易区分，并且较前两者带有更多的编码信息。因此，我们将“相离”和“相接”关系归并为“独立”关系。这样，笔画与笔画的位置关系就简化为两类：独立与相交。并且将这两类位置关系赋予位置关系偏移值：“0”和“5”。

2．3 汉字笔画代码

我们将独立的笔画“一”（如“二”、“王”中的起笔和末笔），“丨”（如“旧”、“四”中的起笔），“丿”（如“采”、“风”中的起笔），“丶”（如“广”中的起笔、“虫”中的末笔），“乙”（如“几”、“礼”中的末笔），按惯例，称为“横”，“竖”，“撇”，“点”，“折”。

将与其它笔画相交的“横”（如“右”、“木”中的起笔），称为“横交”。将与其它笔画相交的“竖”（如“丰”、“串”中的末笔），称为“竖交”。将与其它笔画相交的“撇”（如“独”中的起笔、“舟”中的第二笔），称为“撇交”。将与其它笔画相交的“捺”（如“又”、“文”中的末笔），称为“捺交”。将与其它笔画相交的“折”（如“又”、“力”中的起笔），称为“折交”。

笔画的代码由其顺序值与偏移值相加得到。因此，独立的“横（一）”、“竖（丨）”、“撇（丿）”、“点（丶）”、“折（乙）”的代码为：“1”、“2”、“3”、“4”、“5”。而“横交”、“竖交”、“撇交”、“捺交”的代码为：“6”、“7”、“8”、“9”。

照此类推，“折交”的代码应为“0”（5+5=10，取末位“0”）。考虑到数字编码的码元资源十分珍贵，少用一个码元“0”，对数字统一码在移动电话等数字设备上使用更方便。所以，不区分“折”是否与其它笔画相交，规定“折”无论是独立，或者与其它笔画相交，它的代码均为“5”。

2．4 数字统一码的键盘安排方式

数字统一码的码元为：“1”、“2”、“3”、“4”、“5”、“6”、“7”、“8”、“9”，共九个数字键。数字统一码的键盘安排方式，如图1所示。

图1 数字统一码键位图

3．数字统一码的取码方法

数字统一码规定一个汉字可以取一至六码，允许使用简码和词码，词汇码为等长码。

3．1 汉字的结构类型

汉字的结构类型决定了汉字的取码方法。全国信息技术标准化委员会在《汉字内码扩展规范（GBK）》中将汉字的结构类型为十二种：左右结构，左中右结构；上下结构，上中下结构；全包围结构，向下包围结构，向上包围结构，向右包围结构，向右下包围结构，向左下包围结构，向右上包围结构；嵌套结构。数字统一码将十二种汉字结构类型合并，归纳概括为四种基本结构。它们是：上下结构，左右结构，包围结构，嵌套结构。举例如下：

上下结构，如：字，花，冀。

左右结构，如：们，种，做。

包围结构，如：因，闻，函，区，庙，甸，起，进。

嵌套结构，如：申，央。

3．2 汉字的首尾切分

数字统一码对于上下结构、左右结构、包围结构字，采取二分法，把字分成“字首”和“字尾”两部分。切分字首、字尾的方法，与查字典取部首的方法相同。按取形完整的原则，将汉字一分为二。把上下结构汉字的上部构字部件，左右结构汉字的左部构字部件，作为字首，其余作为字尾。包围结构汉字的首尾，按书写顺序划分，先写的构字部件作为字首，其余作为字尾。对于嵌套结构字不切分。

3．3 数字统一码的基本取码方法

对于上下结构、左右结构、包围结构字，字首按笔顺取前一至三码，字尾按笔顺取前一至二码和末笔代码，整字最多取六码。对于嵌套结构字，按笔顺，取前一至五码和末笔代码。取码方法举例如下：

字445576　各359251　们32425　　种367257　闻425126　出52752

庙413251　起612515　进668454　　申25667　　右689　　丰6667

在输入汉字时，我们采取逐键提示的方式，每次键入，都有可选汉字提示，使用者即可选字，而不必等到输入全部代码。数字统一码是有重码输入法，与同类输入法相比，重码率低，同组重码字数少，在GB 2312 字符集中，同组重码汉字一般不超过10个，实现了一页提示行显示所有同组重码字。

3．4 数字统一码的简码

汉字的使用频度是不同的，仅“的”、“一”、“是”、“在”、“了”、“不”、“和”、“有”前八个高频字，就占汉字总出现次数的10%。因此，我们对一些常用字，不仅给出全部代码，而且给出“简码”，简码的长度分别为1、2、3码。并且，简码一定是全码的前1、2、3码，使用者不必记简码。下面例子给出了“是”、“在”、“了”、“不”、“和”、“有”等字的简码与全码：

是2　251124　　在6　682671　　不1　1324　　和3　367251　我3　367654

中2　2567　　正12　12121　　第31　314563　项121　121134

由于高频字简码和逐键提示相配合，仅输入前3码就可以有效地找需要输入的汉字。因此，使用者只要记住九个数字键位，按笔顺打3键，不必考虑如何切分汉字，也可以使用数字统一码。

3．5 数字统一码的词汇编码

词码为等长六码。二字词的编码由每个字编码的前三码组成，三字词、四字词以及多字词的编码取第一、二、三字每个字的前二码。输入时，字词混合输入，不必区分字码和词码。

4．多语种汉字处理

汉字不仅在我国使用，而且在日本、韩国等国家与地域的流通使用。由于汉字流通的国家与地域的不同，采用的字型与字符集也不同。而ISO/IEC 10646-1汉字符集正是源于中国的GB标准系列、中国台湾的CNS（Big5）标准、日本的JIS标准、韩国的KS C标准汉字符集。

为支持多语种汉字处理，数字统一码对ISO/IEC 10646-1：2000字符集中的中文、日文、韩文汉字统一编码。按汉字流通的国家与地域的不同，组成数字统一码汉字输入法不同的子集，即采取“统一码本分集实施”的方法。这些子集包括：中文GBK/GB 2312数字统一码汉字输入法，繁体中文Big5数字统一码汉字输入法，日文JIS数字统一码汉字输入法，韩文KS C数字统一码汉字输入法。采取“统一码本分集实施”的方法，符合汉字流通使用的实际，同时，也降低汉字编码的重码率和同组重码字数。

5．结束语

数字统一码具有规范、易学、高效的特点。数字统一码编码规则简单，拆字符合国家语委制定的汉字部件规范，取码顺序符合国家语委制定的汉字笔顺规范，各项技术指标达到国内先进水平。数字统一码申请了中国发明专利，专利号：00 1 1037.2。数字统一码的适用范围广，不仅适合于通常的计算机汉字输入，更适合于移动电话、视频点播、DVD、电子记事本、掌上电脑等输入汉字。

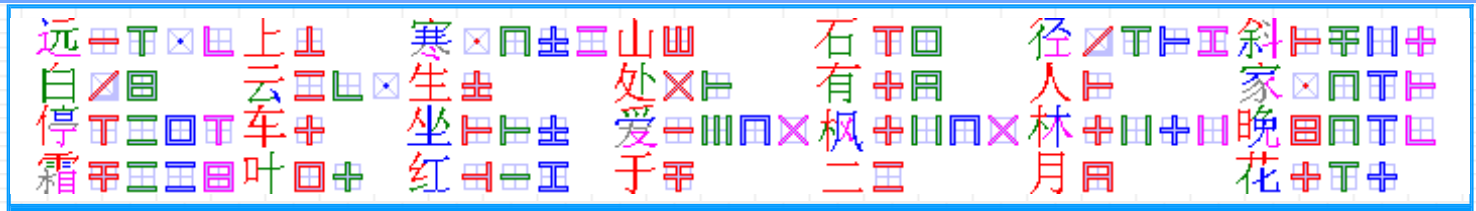
参考文献

[1] 中华人民共和国国家标准 GB/T18031-2000，信息技术 数字键盘汉字输入通用要求，标准出版社，2000年3月

[2]国家语言文字工作委员会，现代汉语通用字笔顺规范，语文出版社，1997年8月

[3]国家语言文字工作委员会，信息处理用GB13000．1字符集 汉字部件规范，语文出版社，1998年4月

[4] 徐万胥，计算机行知码原理与应用，东北师范大学出版社，1997年10月



井田汉字，独一无二的汉字结体构形理论，能够科学地解决数码时代汉字所面临的问题！