

# 组合式汉字编码系统研究

Research on a Multiplex Chinese Character Code System

刘春华

(天津水泥工业设计研究院海外事业部,高级工程师 天津 300400)

单一类型的编码输入法已经有了充分的发展,在编码层次很难再有新的重大突破。音码、形码之间不存在简单的孰优孰劣之分,因为它们各自具有不同的优势及适用范围。认识的字和词语,用音码输入更符合思维习惯,不认识的字只能用形码输入。由于汉字编码输入技术长期未能取得根本性的突破,处于徘徊状态,人们似乎已失去信心,转而期待非键盘输入(如语音输入)方法以取代键盘。问题是,如果没有汉字编码,能完全解决中文信息处理及其相关问题吗?

笔者曾提出,汉字编码的理想模式似乎属于一种建立在汉字基础知识之上而与语文基础教学规律相一致的组合式编码系统<sup>[1]</sup>。所谓组合式汉字编码系统,就是要设法将音码、形码和音形结合码等不同类型的多种编码组合成一个统一体,依靠码元、键元及取码方式的科学组合,在汉字键盘输入时实现各类编码的不切换混用,使它们各尽所能、优势互补:用音码或音形结合码输入认识的字和词语,用形码输入不认识的字;使不同阶层的使用者以及同一使用者在其所受语文基础教育不同阶段均能用上与其语文知识水平相适应的编码,而所有种类的编码又都同属于一个统一的编码系统,各编码间的衔接具有与语文基础教育同步的、循序渐进的基本特征。

有专家早已指出,汉字键盘输入系统要由单一的输入方法发展到综合的输入环境<sup>[2]</sup>。组合式汉字编码系统即类同于一种“环境”,但重点强调这种“环境”要与语文基础教育相一致,并要统筹解决大字符集/全汉字集的键盘输入、汉语辞书的直接翻页检字以及为电脑辅助识字奠定基础这三个问题,而不仅仅只为用于汉字的键盘“输入”。

本文探讨编码组合的要求与方法。

## 一、编码系统的组成应与语文基础教育同步

汉字编码要建立在汉字基础知识之上,做到学

汉字与学编码同步,二者并行不悖,相辅相成,还要能循序渐进,即从初学汉字起,就可同时使用编码,而不必等到掌握了“足够”的汉字基础知识之后才能使用编码,否则利用编码和电脑学汉字就是一句空话。同时,汉字基础知识掌握得越好,编码用起来也应更能得心应手,效率也越高。显然,只有组合式编码系统才有可能满足这一要求,单一种类的编码则无能为力,因为不同的使用者以及同一个使用者在其接受语文基础教育不同阶段,其汉字基础知识水平处于变化之中,学汉字本身是一个循序渐进的过程,学编码要与学汉字同步并相辅相成,编码就应具有“应变”能力。

按现行语文基础教育进程,有关的汉字基础知识似可依次排列如下:汉语拼音,笔画及笔顺,部首及偏旁,部件……。

汉字编码要与识字教育同步,与语文基础教育相一致,让使用者在其接受语文基础教育的各个阶段均能用上与其知识水平相适应的编码,编码系统中至少应包含下列各种编码:全拼音码,笔画码,部件码……。

值得指出的是,汉语拼音是汉字的一种多功能辅助工具,它在识字教学中所具有的积极作用是任何其它编码不可能替代的。因此,编码系统中首先必须包括拼音码,并且必须是与汉语拼音方案完全一致的全拼码(类似用v代替ü的做法当然应该被允许),而双拼的必要性则值得怀疑。此外,用正确的笔画和笔顺书写汉字也是语文基础教育中最基本的要求之一。笔画码是形码中最为简单的一种,可用来作为一种辅助输入方式(例如用于搜寻不认识的字)。从循序渐进和辅助识字的角度来看,编码系统中备有笔画码也很有意义。有专家指出,要重视传统部首在汉字编码中的作用<sup>[3]</sup>。笔者赞同这种观点。部首概念在识字教育方面具有积极并有待进一步开发的作用,而且部首概念由来已久,具有根深蒂固的群众基础,因此,在字形编码中,充分发挥传统部首的作用也很有意义。

综上所述,编码系统要完全建立在汉字基础知

识之上,要与语文基础教育同步并保持一致,就要最充分地利用汉语拼音、汉字笔画及汉字部首等作为码元来组合编码。

## 二、正确确立形码的地位及用途

一段时间以来,形码受到诸多责难,以至有专家明确提出应放弃字形编码。特别是由于人们期待着非键盘输入方式能取得突破,字形编码似乎已少有人研究。因此,有必要正确确立形码的地位及用途。我们知道,在汉字的音、形、义三要素中,字形是使一个字区别于其它字的外形特征。不涉及字形的辞书检索局限性很大,即使按字音编排的辞典也大多要附“形检”法,如部首法或笔画法等,否则不认识的字就无从下手了(而不认识的字才最需要查字典)。识字教学中在指导学生如何准确识别和正确书写汉字方面也很费精力。此外,由于纯音码的单字重码率高以及不认识的字无法输入这两个缺陷的存在,也决定了在键盘输入法中加入字形因素是必要的。因此,毫无疑问,汉字编码要实现其统筹解决大字符集/全汉字集的键盘输入、汉语辞书的直接翻页检字以及为电脑辅助识字奠定基础这三个问题的战略目标,形码是必不可少的。

形码的困难主要集中在两个方面:一是汉字拆分与部件选取的随意性较大,不符合文字规范;二是编码规则较难掌握和记忆。仔细分析,导致形码陷入困境的原因也有两个,一是无章可循,二是用非所长。无章可循,是指长期以来没有统一的标准来规范汉字的拆分,导致编码者各取所需;用非所长,是指未正确发挥形码的优势,而企图以单一形码来满足键盘输入的全部要求,以牺牲规范性和易学性来换取少重码及高速度。《信息处理用 GB13000.1 字符集汉字部件规范》已经出台,无章可循的尴尬局面可以结束了。为了让形码发挥优势,首先应弄清楚形码具有哪些优点,适合于什么用途,再针对其特定用途提出具体的编码要求。

笔者以为,形码最主要的优点是:与字音无关,可处理不知读音的字。为发挥其优势,形码较适合于下列用途:(1)作为键盘输入的辅助方式,主要用于输入不认识的字;(2)作为汉字排检的主要方式,用于辞书检索;(3)作为电脑辅助识字的一种工具,以充分适应信息时代的现代化汉字教学要求。

针对形码的上述用途,似可提出编码要求如下:(1)汉字拆分与部件选取必须符合文字规范,并尽可能兼顾汉字教学的需要;(2)易学易记;(3)按码序排列的汉字应具有根据某种特征分类排序的效果;(4)通用于大字符集/全汉字集。

为使形码满足如上要求,一方面有必要降低其它方面的要求,例如:可允许稍高的重码率及码长;

可只含字编码而不含词编码,等等。另一方面,还应就形码的上述要求提出针对性的具体措施。

汉字拆分要符合规范,应遵循《信息处理用 GB13000.1 字符集汉字部件规范》。为兼顾汉字教学的需要,还应鼓励采用组合部件,对常用部首最好不再作下一层次拆分。

在满足规范性的前提下,影响形码易学性的最重要的因素是码元与键元之间的对应关系。显然,有音部件“音托”最为理想。同时,只要不牵强附会,对少数部件采用“形托”方式也可接受。最难处理的则是大量的无音部件,应予重点研究,寻求突破,否则形码将很难走出困境。为了达到易学易记的理想境界,形码部件似可按下列方式安排键位:(1)成字部件一律“音托”;(2)有习惯名称(俗称)的常用部首按其习惯名称中的关键字“音托”(如“丿”可对应 D 键);(3)极少数非成字部首可采用“形托”(如“乚”可对应 U 键);(4)其它非成字部件可采用“笔画替代法”,若为末部件时,用部件末笔代替部件本身,在其它位置时,则用部件首笔代替。

按照上述思路编制形码,似可满足形码四项要求中的三项,即规范、易学及通用于大字符集,但不能满足汉字排检的要求,因为由码序决定的汉字排列杂乱无章,并不具有根据某种特征分类排序的效果。编码用于辞书检索,难点在于汉字排序。除了规范和易学这两条共同的基本要求外,编码用于键盘输入与用于辞书检索时的要求并不一样,前者要求字和词能混用、多种编码能混用、敲键越少越好、重码率越低越好,后者则要求汉字最好能按照汉字本身的某种特征(如拼音、部首或笔顺等)分类排序,而并不关心字码与词码以及不同种类的编码之间是否发生冲突,对编码效率的要求也不必像键盘输入那样高,例如在保证易学易记的前提下码长稍长或重码率稍高都不影响使用,因为翻字典的动作远不如敲键盘那样快。如果我们在每个汉字的部件码之前先取首笔,即增加一位笔画码,就能使汉字排序效果大为改观,编码汉字基本上可按首部件(由部件首笔及部件读音两个因素决定)为特征分类排序,这样的形码可用于辞书检索。

最后,还要指出,少数码元的键位可能极大地影响编码的重码率,似有必要对它们作出特殊安排。但应注意两点:一是特殊安排的码元数量不能过多,否则就又回到牺牲易学性的老路上去了;二是要保持整个编码系统(包括音码、形码及音形结合码等)的统一性,并以最低的代价,换取最好的效果。首先,成字部件只能“音托”,不宜作特殊安排,否则会与音码及音形结合码产生矛盾,造成混乱。其次,作特殊安排的非成字部件应具有较高的动态和静态组字频度。部件的动态组字频度高,可频繁使用,容易记住;静态组字频度高,才可能对重码率

产生大的影响。

我们认为,《信息处理用 GB13000.1 字符集汉字部件规范》为形码的统一铺平了道路,因为它使该字符集中的每个汉字都具有了规范的基础部件系列,只要建立一种科学的、可普遍接受的码元与键元之间的对应关系,即可将这种基础部件系列转换成为一种形码。当然,为兼顾汉字教学需要,全部采用基础(末级)部件实施编码并不一定理想,因此还应研究在编码中尽量地、有规律地采用有利于汉字教学的组合部件及部首。

三、不同编码之间组合的要求

编码组合似应满足下列要求:

1. 不同种类的编码之间以及同一种编码中的字编码与词编码之间必须互不冲突,即应各自具有独立的、互不相交的编码空间。否则,编码系统中的各种编码就会“搅”在一起,陷入“无序”状态而没有使用价值(单靠智能处理无法解决这类问题)。

2. 同一种编码中的字编码与词编码的编码思路必须一致。否则易使人产生思维“障碍”而影响大脑反应速度。

3. 编码系统中码元的定义及键位安排必须统一。不同编码所用到的码元种类可以、也必然会有所不同,例如,拼音码只涉及字音而不涉及有字形;形码涉及部件和笔画,其中笔画码又只用到笔画;音形结合码既涉及字形又涉及字音,等等。但就整个编码系统而言,码元的定义及键位安排则必须是统一的,不同种类的编码所涉及到的同一种类的码元,如音形结合码中用到的部件及笔画与形码中使用的部件及笔画,其定义和键位都应该一致。

4. 不同种类的编码可根据其性质和用途对应于不同的字符集,但系统中至少必须有一种形码可通用于大字符集/全汉字集。

对上述前两点要求,若分别对待并不难做到,但要同时满足则有很大难度。事实上,大多数现有编码方案中的字编码与词编码的编码思路都是一致的,但它们却共用一个编码空间而产生字、词矛盾。因此,组合式编码的难点和关键即在于要实现不同种类编码之间的不切换混用,字、词以及不同种类编码之间必须互不冲突,在总编码空间之下的各编码子空间互不相交,而且这种空间分割必须以“自然”为前提,不得在字、词以及不同编码的输入方式之间转换时造成思维障碍,还要使编码空间的分配尽可能合理,即用于输入常用字及词语的主要编码(姑且称之为“主导码”)获得比其它辅助码更

大的编码空间。

四、实现编码组合的方法

编码组合,取决于码元、键元及取码方式的组合。下面探讨实现编码组合的思路和方法。

试以汉字的拼音、笔画、部首和其它部件为码元,以通用小键盘中 3×10 个键,即 26 个字母及 4 个字符“.,/;”共 30 个键为键元(包括 1 个万能键),统一对大字符集汉字、大词库词语及常用非汉字字符实施编码,要求编码输入以词为主、以音为先,包括音码、形码和音形结合码等不同类型的多种编码,可不切换地混用各类编码。

编码系统组成如下:

拼音码 字码:汉语拼音,6 位不等长码,主要以 3 755 个一级字为编码对象

笔画码 字码:笔画编码,5 位等长码,以基本集 6 763 字为编码对象

笔音码 字码:笔画与拼音结合,5 位等长码,主要以一级字为编码对象

音形码 { 字码:音形结合,4 位等长码,以 4.5~5 千字为编码对象  
词码:音形结合,4 位等长码,以 4.5~6 万条词语为编码对象

形码 字码:笔画与部件结合,5 位不等长码,以大字符集/全汉字集汉字为编码对象

简码 以音为主,1~3 位码长,以使用频率高的单字词、二字词和三字词为编码对象

字符码 “音”码(字符名称的“音”),3 位等长码,以常用非汉字字符为编码对象

我们将“.,/;”、“I”、“O”、“U”、“V”七个键定义为笔画键元。需要指出的是,七个笔画键元中,除 O 外,其它均不属于拼音首字母键,而以 O 为拼音首字母的字和词数量极少(一级字中 8 个、二级字中 6 个“O”音字,组词能力都很低),因此,可以认为,笔画键元集与拼音首字母键元集基本上不相交。部件的键位安排采用上文已说明过的形码部件的四条键位规则。

为了说明不同的编码如何能组合在一起而互不冲突,下面列出各种常规编码的编码规则(拼音码在后面说明),其中“音”代表字的拼音首字母,“字身”则指一个字去掉部首后的剩余部分。

|       |       |   |             |
|-------|-------|---|-------------|
| 音形码字码 | 单部件字: | 音 | +部首+字身首笔+末笔 |
|       | 二部件字: | 音 | +部首+字身首笔+字身 |

|       |         |    |   |     |   |      |    |       |     |     |
|-------|---------|----|---|-----|---|------|----|-------|-----|-----|
| 音形码词码 | 三部件以上字： | 音  | + | 部首  | + | 字身首笔 | +  | 字身末部件 |     |     |
|       | 二字词：    | 音  | + | 部首  | + | 音    | +  | 部首    |     |     |
|       | 三字词：    | 音  | + | 音   | + | 音    | +  | 部首    |     |     |
|       | 四字以上词：  | 音  | + | 音   | + | 音    | +  | 末字音   |     |     |
| 笔 音 码 |         | 首笔 | + | 次笔  | + | 第三笔  | +  | 末笔    | +   | 音   |
| 笔 画 码 |         | 首笔 | + | 次笔  | + | 第三笔  | +  | 第四笔   | +   | 末笔  |
| 形 码   | 单部件字：   | 首笔 | + | 本部件 | + | 末笔   |    |       |     |     |
|       | 二部件字：   | 首笔 | + | 首部件 | + | 末部件  | 首笔 | +     | 末部件 |     |
|       | 三部件字：   | 首笔 | + | 首部件 | + | 次部件  | +  | 末部件   |     |     |
|       | 四部件以上字： | 首笔 | + | 首部件 | + | 次部件  | +  | 第三部件  | +   | 末部件 |

从上述编码规则中可以看出,系统编码可分为“以音为先”和“以笔为先”两大类,由于拼音首字母(即“音”)键元集与笔画键元集不相交,因此两类编码的编码空间不可能相交。在以音为先的音形码中,词码的第三位码均为“音”,而字码的第三位码全是“笔”,显然,音形码中的字码和词码的两个编码空间也不相交。再看笔音码与笔画码,它们的末位码分别为“音”和“笔”,故这两种码的编码空间同样不相交。将形码与笔音码及笔画码比较,除首位码同为首笔外,其余各位码中,形码取部件(只有二部件字第三码取笔画),笔音码及笔画码取笔画(只有笔音码最后一码取“音”),由于以“音托”方式安排键位的成字部件及有习惯名称的常用部首在组字频度方面占有绝对优势,而以“笔画替代法”安排键位的非成字类部件的静态组字次数在所有部件的总组字次数中所占比例很小<sup>[4]</sup>,实际上部件码元对应于笔画键元的频次很低,两个以上部件同时对应于笔画键元的情况更是极少。因此,形码与笔音码及笔画码重码的概率微乎其微,它们的编码空间也几乎(如果不是绝对地)不相交。

最后讨论拼音码即汉语拼音。显然,拼音码属于“以音为先”一类的编码,它并没有自己独立的编码空间,少于4个字母的音节占用的是简码及字符码的编码空间,4个及4个以上字母的音节占用的是音形码的编码空间。如果说其它各种编码之间互不冲突是有意“设计”出来的,对拼音码则难以人为设计。因此,拼音码很有可能发生与音形码的重码。但是,不加声调的汉字音节总共只有417个,其中4个以上字母的音节又只占其中的1/3,与音形码的几万条编码及近54万编码空间相比显然是微不足道的,即使有极少数音节与音形码重码,只要把音形码置于重码组的首位就将会不会影响实际使用。

顺便指出,以音为先的音形码为4位码长,而以笔为先的笔音码、笔画码和形码则为5位码长(当然也可统一取4位码长,代价是增加重码,其得失尚有待斟酌),这需要利用电脑智能来解决码长不等问题。其实很简单,由于两类编码的首位码对应的键元集基本上不相交,4位码长的音形码可由

软件自动识别,不需要人工干预(不必加打空格键),仍可实现无间隔输入,只有以○为拼音首字母的字、词的音形码需要加打空格键,但这类字和词极少。汉语拼音最长有6个字母,而以音为先的音形码只有4位码长,这也可用软件的方法解决。

音形码是键盘输入的主要方式,它的编码空间占了4码长编码总空间的将近80%,其中词码空间又占了音形码编码空间的将近80%,符合“词为主导”的设计思想。笔音码和笔画码的编码空间不大,但它们是起辅助作用的,一是用于搜寻难字(如不知读音或不好确定部首的字),二是为初学认、写汉字的小学生提供一种与学汉字同步的简易输入法(与拼音码并列),作为使用音形码之前的一种过渡。形码采用5位不等长码,编码空间高达数百万,适合大字符集编码。虽然形码的码长偏长,但能适合其特定的三种用途,即键盘输入的辅助方法、辞书检索及电脑辅助识字。

下面是字、词编码实例(参看编码规则):

**音形码词码:**科学——KH XZ(科、禾、学、子)

计算机——JSJM(计、算、机、木)

**音形码字码:**碘——DSIB(碘、石、丩、八)

降——JE/1(降、冫、丩、斗),无音末部件“斗”用部件末笔“|”代替。

**形 码:**蚝——IC/M(丨、虫、丩、毛)

嗖——IK/Y(丨、口、由、又),无音部件“由”用部件首笔“丩”代替。

镱——/JLRX(丩、钅、立、日、心)

## 五、小结与讨论

1. 利用码元、键元及取码方式的科学组合,可以在汉字基础知识之上建立一个集音码、形码和音形结合码等不同类型的多种编码于一体的组合式汉字编码系统。组合式汉字编码系统的战略目标是要统筹解决全汉字集的键盘输入、汉语辞书的直接翻页检字以及为电脑辅助识字奠定基础这三个问

题。符合文字规范、实现不切换混用的组合式汉字编码可与语文基础教育同步并保持一致,做到学汉字与学编码并行不悖,相辅相成:小学生在入学之初,从开始学拼音、学写字起,就能同时用拼音码和笔画码打字;随着学生的语文知识水平的提高,掌握字的部首和部件后,即可改用效率更高的编码(主导码),而且这种改用属于“无级变速”,即已经知道读音和部首的字或词语,用主导码输入,还不能确定部首或不认识的字,仍用辅助码输入;学生识字量越大,对字的部首越熟悉,掌握的词语越多,汉字输入速度也就越快。同时,经常使用这样的编码系统,又能促进汉字基础知识(指拼音、笔画、笔顺、部首、词语等)的学习,对所学知识起到检查、复习和巩固的作用。编码系统中的形码应以大字符集/全汉字集为编码对象,要既方便输入不认识的字,又可用于电脑辅助识字,还适用于汉字排检。

2. 在组合式编码系统中,各编码子空间应互不相交,还要合理分配,以解决常用字与非常用字以及字与词两对矛盾。如此,可根据需要任意扩大词库而不会对字编码造成干扰;任意扩大字库(直至全部汉字)而不会对常用字的快速输入产生影响。各编码子空间的分割要以“自然”为前提,在字、词方式及不同种类编码之间转换时不会产生思维障碍。组合式编码系统将编码的基本问题都将解决于编码层次,从而为软件层次发挥水平奠定基础。例如,主导码和辅助码可采用不同的显示模式:前者在编码中途不显示,无重码直接上屏,有利于快速盲打;后者可作全程显示,以方便搜寻难字。还可利用笔画码在线查询字的部首或拼音,或者反过来用拼音码查字的写法或部首,等等。

3. 传统部首在识字教学及汉字排检中具有积

极作用,在汉字编码中也应受到重视。盼望语文界能尽快统一部首,不但统一部首的个数(种类),而且每个汉字的部首都应具有唯一性。此外,以传统部首作为汉字部件,在少数情况下可能会与《信息处理用 GB13000.1 字符集汉字部件规范》发生冲突,希望有关专家研究这一问题。

4. 在组合式编码实例中,笔画键元在编码空间的自然区分中起到了关键性的作用。此外,笔画还另有“妙用”:音形码字码的第三码只取字身首笔,大大降低了汉字拆分的难度(实际上音形码字码并不需要对汉字作严格拆分);形码在首部件之前先取首笔,使得按码序排列的汉字排序效果大为改观,除作键盘输入外,还可用于汉字排检,字典检索与电脑输入一致;采用“笔画替代法”可较好地解决无音部件的键位安排问题,使形码达到易学易记的理想境界。

5. 符合文字规范又适合汉字教学的汉字编码是电脑辅助识字的有效工具。从实现汉字教学的现代化角度看,研究这样一种编码也是必要的。呼吁语文教育界、电脑界及编码界紧密合作,也希望有关主管部门给予必要的重视。

#### 参考文献

- [1]刘春华,试论汉字编码的宏观要求及理想模式,科技导报,1997(7)
- [2]陈一凡、胡宣华,汉字键盘输入技术与理论基础,清华大学出版社、广西科学技术出版社,1994年6月第一版,P185
- [3]中国中文信息学会汉字编码专委会、中国计算机学会中文信息技术专委会评论员,汉字键盘输入技术现状及发展趋势,中文信息,1996(4)
- [4]李公宜等,汉字信息字典,科学出版社,1988年12月第一版,P1027~1058 (责任编辑 蔡德诚)

(上接第50页)

济起飞的前提性条件,也是科技起飞的前提。如果没有人力资本积累,不仅引进技术的技术效果甚低,而且会严重阻碍自主创新能力的发展。我国在这些方面的努力将直接决定科技起飞的时间,应努力提高R&D投入强度和人力资本的积累。

在实现科技起飞过程中,我国应选择什么样的技术创新道路呢?已有的典型道路有两种:日本的稳健型道路和韩国的激进型道路。日本在人均GNP只有385美元时,就把R&D投入强度提高到1%,然后,在近20年的时间里以大量使用外国先进技术并消化吸收它们为基本的技术创新方式,缓慢但扎实地积累技术实力,实现科技起飞。韩国在人均GNP接近2000美元时,研究开发投入才达到1%,然后,用5年的时间把该指标提高到2%,在改进技术阶段只稍事停留,给人以急于求成的印象。事实也证明,韩国的基础研究水平较低,甚至不及我国

和印度;技术开发存在急功近利倾向,特别喜好盲目仿效日本的新技术、新产品;对发达国家有依赖性,等等。

我国的区域技术、经济都有较大差距。技术创新能力最强的地区比最弱地区的技术创新能力指数平均高出近9倍。沿海省份和内地相比,沿海省份该指数比内地省份平均高出近3倍。因此,我国不可能选择韩国的激进型道路。日本的稳健型道路值得仿效。但基于地域差异大的原因,我国不能提倡企业以唯改进技术为最优选择。应选择总体上以改进技术为特征,局部上具有创造技术优势的技术创新道路。在重要产业、前景产业和重要省份,应大力推进技术创新能力的升级,即依靠国内外科技力量发展具有国际一流水平的产品技术、产业技术,从而带动全国技术创新水平的提高。

(责任编辑 肖庆山)