

试论汉字编码的宏观要求及理想模式

On Macro-Requirements for and Ideal Pattern of Chinese Character Coding

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)

世纪之交,电脑正逐步介入我们工作、学习和生活的各个方面。电脑要进入千家万户,中国要建立信息高速公路……,然而,汉字输入“瓶颈”却依然存在,由于计算机产业的高速发展以及国民收入的普遍提高,电脑价格逐渐变得易于接受,而汉字输入“瓶颈”正上升为电脑普及的关键性制约因素。面对五花八门的编码输入法,广大电脑用户无所适从,更多的人只好举步不前。这一局面继续下去,可能在社会上造成一种普遍的心理障碍——对电脑望而却步、敬而远之。汉字编码万“码”奔腾的混乱局面已经并将继续对我国中文信息产业和现代化进程产生负面影响。针对此问题,本文以设立汉字编码的战略目标为导向,从分析汉字编码的矛盾入手,探索汉字编码的理想模式。

一、确立汉字编码的战略目标是当务之急

汉字编码至今不能走出万“码”奔腾的困境,一方面说明了汉字编码研究具有相当的难度,另一方面也可能是因为基础理论研究滞后,国家有关机构对汉字编码研究缺少宏观指导,至今没有确立汉字编码的战略目标。构成万“码”奔腾景象的是各编码人自由市场式的竞争。有不少编码方案都在某一方面具有独特的优点。遗憾的是,这些优点分散于不同的方案之中,所谓“各有千秋”。尽管如此,也足以说明,汉字编码的各种战术技巧已趋于成熟,但我们缺少的是高层次的战略构想,目标未定,主攻方向不明。时至今日,花去我们主要精力的还局限于速度之争、重码率之争、词库大小之争、部件多寡之争、形码音码之争、软件功能之争,还有市场利益之争。编码输入法的成败优劣主要看市场占有率,这是否反映汉字编码理论的缺少作为呢?编码研究要向更高层次发展,我们就要站在把握汉字的历史、现在和未来的高度,确立汉字编码的战略目标。

我们以为,汉字编码应该有利于中华民族继承

几千年汉字文化的历史遗产,有利于中华文明充分吸收和利用当今世界文明成果,增强汉字在信息社会的生命力。

具体说,一方面,中国拥有 5 000 年文明史,汉字本身也有 3 000 多年的历史(从甲骨文算起),一脉相承的汉字文化给我们留下了宝贵而丰富的历史遗产。如果我们不能在电脑时代很好地继承这些遗产,将是源远流长的汉字文明的损失。因此,汉字编码的最终目标应该让全部汉字自由进入计算机,也就是说,应着眼于全汉字集的计算机处理。另一方面,汉字长期不能适合电脑使用,无法充分利用当今先进的世界文明成果,在信息社会到来之际一直处于被动地位——本来就有人指责汉字“难学”,再外加编码,就更是“难上加难”。因此,如果我们仅仅以应付电脑使用为目的,再好的编码也会成为汉字的额外负担,汉字也将永远陷于被动。汉字要变被动为主动,编码就不能仅仅充当汉字进入计算机的“拐杖”,而更应成为利用电脑学习汉语文字的一种辅助工具。要使学汉字与学编码并行不悖,相辅相成,汉字入了门,编码也掌握了;要既会编码,又可反过来借助电脑智能轻松地学习更多的汉字。这样,编码将不再是汉字的负担,而成为直通汉字王国的桥梁,学汉字之路,自然也会“天堑变通途”。此外,阅读汉语辞书不能像学拼音文字那样直接翻页检字,这又是长期困扰中国人的一大难题。繁琐、不统一的检索方法浪费了我们大量的时间和精力,挫伤了有心通过查辞典掌握生字、生词的人们的学习热情和积极性。因此,汉字编码还要实现汉语辞书的直接翻页检字。

二、汉字编码遇到的两个矛盾

其一,常用字与非常用字的矛盾。几千年的汉字发展史,产生了几万个汉字,而现代中国人常用的却只有几千个常用字。据有关统计,3 755 个一级字占了 99.9%左右的使用频度。汉字编码当然首先

要重点保证常用字的快速输入。但是,却又不能将其它几万个汉字弃之不顾,尽管它们“不常用”。否则我们就不可能在电脑时代很好地继承几千年来极其珍贵的汉字文化遗产。因此,那些只以基本集为设计基础而无法适应字符集扩大的编码方案有其局限性。要解决常用字与非常用字的矛盾,难点在于常用字使用频率高而非常用字字量大,要对全部汉字实施编码,要么规则趋于复杂而难学,要么重码率高而影响速度。显然,几万个非常用字会“拖累”几千个常用字的简单、快速输入。

其二,字与词的矛盾。汉语以字为最小单位,但人们的思维和表达却是以词为最小单位,汉字编码输入要符合人的思维习惯并提高效率,就要在字编码的基础上“叠加”词编码。与几万个汉字相比,汉语词的数量更是天文数字,我们不可能、也没必要对所有的词实施编码。但是,从方便实用出发,建立一个几万条词语的大词库还是必要的。要实现符合思维习惯的、词为主导的输入方式,词库就要备有足够多的词汇,才不至于经常出现“漏词”而出错并影响速度。由于汉语词中包含不少单字词,且使用频率较高,字与词在形式上又无法区别,因此人们使用汉语时实际上是字、词混用(或者说是单字词与多字词混用)。这就给汉字编码出了另一道难题:既要满足方便字、词混用的要求,又不能让字码与词码发生冲突。否则,几万个词码“杀入”字码空间,对编码性能的影响将是毁灭性的:重码率将不可避免地急剧升高。字与词的矛盾,焦点在于既要避免字、词冲突,又要保持自然的思维习惯。大部分重码率较低的编码方案,尽管仅以基本集为编码字符集,其词库容量也很有限,原因就在于字、词冲突。而有些编码方案为了避免字、词冲突,采用加标识符或“字三码词四码”的方式,但这样“强行”划分字码和词码,用起来不自然,会产生一定程度的思维障碍,特别在高速盲打时会影响大脑的反应速度。此外,“字三码”也显得字码空间不足,导致字重码率过高(尤其当字符集扩大时)。

上述汉字编码的两个矛盾,制约着编码方案的性能。解决了这两个矛盾,才有可能产生理想的编码。否则,无论编码的细节如何“优化”,其先天不足是难以弥补的。

三、汉字编码应具的四项要求

第一,规范性。规范性的最低要求,是汉字编码不干扰中小学的语文基础教学,最高境界则是汉字编码可以作为一种规范汉字教学的辅助工具。只有符合文字规范的编码,才能反映和保持汉字固有的“美”,才能增强汉字的生命力。

第二,易学性。易学是普及的基础。很少有人愿

意在从小就学汉语文字的基础上,再花大气力去学汉字编码。易学性的最高境界是学汉字与学编码并行不悖、相辅相成,与现行语文教学规律高度一致,做到认字也即识码,编码利于识字。

第三,高效性。高效性反映在两个方面:首先,用于键盘输入时,具有可被普遍接受的输入速度(速度上限应能满足“专业”打字员的使用要求);其次,用于辞书检索时可以直接翻页检字。使用低效率的汉字编码必然影响全社会的工作效率。

第四,通用性。通用性有三个含义:通用于编码的各种应用领域;用于键盘输入时,通用于想打、看打和听打,并与国际通行的计算机设施兼容;通用于全汉字集。

四、汉字编码的理想模式

以实现汉字编码的战略目标为导向,在分析汉字编码两个矛盾的基础上,结合编码的四项要求,探讨汉字编码的理想模式,目的在于总结经验,开拓思路,减少同水平和低水平的重复研究。

汉字编码至今不能统一,直接的原因当然是还未找到理想的编码。什么样的编码是“理想编码”呢?有专家断言,“理想编码”就是“无编码”输入。近来,有人认定“拼音变换”的句输入方式是汉字键盘输入的方向。但“句”输入方式存在下列缺陷:首先,纯拼音变换对电脑来说很难做到百分之百正确;其次,退一步来说,即使电脑能做到百分之百的正确变换,以“句”为单位对人来说敲错键的几率也要大得多;其三,再退一步,就算人也不会敲错键,以“句”为单位的“想打”又是一个高要求,只有“出口成章”者才可能用得顺手;其四,无论任何原因引起对“句”的修改,都远不如以字、词为单位来得灵活、方便;最后,拼音变换对不认识的字无能为力。基于上述理由,笔者认为键盘输入仍以“字为基础、词为主导、智能处理”较为适用。还有人寄希望于语音输入能代替键盘输入,且不说语音输入能否解决全部问题,就现阶段来说,键盘输入是最经济的,并可能在相当一段时期占主导地位。因此,探讨“有编码”的“理想编码”,不失其现实意义。

现有几百种编码方案,就其主流来说,大致可归为形码和音码两大类。一段时间以来,不同阶层的人士对音码的肯定逐步升级,诸如:音码是国家“八·五”攻关重点;汉语拼音一级符号系统是最佳选择;日文同音词是中文的三倍,日本也不用字形编码;“拼音变换”是汉字键盘输入的方向……。形码方尽管已无还手之力,但并未气馁,他们坚定地认为:拼形优于拼音;历史上也只有“书同文”而没有“语同音”;汉字是一种只需要眼睛看就能思考和理解的文字……。显然,争论双方都有充分的理

由坚持己见,就如同双方均没有足够的说服力来驳倒对方一样。

只要我们不怀门户之见,冷静地分析这一问题,应该不难得出合理的结论:字形和字音是汉字的两个要素,二者缺一不可。形码和音码各有优缺点,而且二者形成互补关系:一方处于劣势的地方,正好是另一方的用武之地。音码的优势在于以词为主,输入汉字时符合思维习惯,没有“拆字”负担,缺点则是单字重码率太高以及难于输入不认识的字;而形码的优势恰好在单字编码重码率低,又不受字的认识与否的限制,但在音码占优势的方面则相形见绌。这说明,“单打一”的编码,无论是形码还是音码,都不可能成为“理想编码”。因此,“理想编码”的第一个思路,是要设法将音码、形码以及音形结合码等不同类型的多种编码组成统一体,依靠码元、键元、取码方式的科学组合,实现各类编码的不切换混用,使它们各尽所能、优势互补。如此,就可能解决汉字编码的两个矛盾:音码(或音形结合码)为主,用于输入常用字和词语,要求简单、快速;形码为辅,用于输入非常用字(可只限单字形式),也要求简单,但在速度(重码率)方面可降低要求,这样有利于大字符集乃至全汉字集的编码。显然,组合式编码符合人们使用汉字的自然习惯:认识的字或词,用字音输入;不认识的字,用字形输入。组合式编码的难点在于“不切换”。

在对汉字编码的四项性能要求中,规范性和易学性是前提,高效和通用只有建立在规范和易学的基础上才有意义。因此,“理想编码”的第二个思路,是要将编码完全建立在汉字基础知识之上,要充分地利用人们已经或应该掌握的如笔画、部首和拼音等汉字基础知识来实施编码,与语文基础教学相一致,实现识字、编码、查字、打字四结合。

就形码而言,除了码元的选取、汉字的拆分要符合文字规范和人们的识字习惯以外,影响编码易学性的另一个最重要的因素是码元与键元之间的对应关系。这种对应关系应力求简单、直接,不但易记难忘,还要反应速度快,不假思索。一般说来,部件“音托”(限于有音部件)最为理想,“形托”(有时显得勉强)次之,“笔画定位”由于需要一个思考过程,又次之。由此可知,最难处理的是大量的无音部件,值得重点研究。如果对无音部件的键位安排不能取得突破,形码很难走出困境。因此,“理想编码”的第三个思路是:形码部件“音托”为主,“形托”为辅,重点解决无音部件的键位安排,可采取灵活方式,例如,是否可以用笔画代替此类部件?

综上所述,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。在该系统中,音码、形码、音形结合码等各类编码可不切换混用,其中形

码部件以“音托”为主安排键位。

五、建议与设想

1. 加强语文基础建设,尽快统一汉字部首

汉字编码应该建立在汉字基础知识之上,但这个“基础”本身还有待于进一步统一。以部首为例,各种辞书所采用的部首的数量(种类)不尽一致,有些字的部首也不太明确,《新华字典》和《现代汉语词典》似乎很轻松地采用“多开门”的办法来解决,但编码却需要唯一性。事实上,汉字部首的不统一,也给汉字教学带来不便。统一汉字部首,无论对汉字的编码还是教学都将非常有利。传统部首是汉字的一个“传家宝”,但在流行的编码方案中,传统部首的作用却大多被忽略了。显然,传统部首本身的不统一以及一定程度的不确定性影响了它们在编码中发挥更积极的作用。因此,我们建议尽快统一汉字部首。据报道,《信息处理用 GB130001 字符集汉字部件规范》已通过审定,这是一个好消息。与一般部件相比,部首的问题似乎还要少些,既然部件可以规范,部首也应该可以统一。关键是我们对统一部首的必要性和重要性要有足够的认识。可以设想,规范了部件,再统一部首,必将极大地有利于汉字编码和汉字教学。

2. 开展对电脑辅助识字的深层次研究

语文现代化应包含语文教学的现代化。在如今一片“拼音变换”声中,也应该有人研究汉语文字的计算机辅助教学。这不但造福民族、造福子孙,有巨大的社会效益,也会有巨大的经济效益。21 世纪,中国经济必将崛起,与此相适应,汉字也理应发挥更大的作用。我们要有所准备。据说语文教育(识字当然是语文教育的重要内容)占中小学教育的 1/4,效果却不尽如人意。说得严重点,这是对整个民族智力资源的浪费。我国识字教学的改革试验已经取得了丰硕成果,出现了多种具有一定影响的“速成”识字法,遗憾的是还没听说它们与电脑技术结合。电脑辅助识字可能是历史留给我们解决“汉字难学”问题的最后机会,而汉字编码又应是电脑辅助识字的基础。电脑辅助识字当然远不止于识字、编码、查字、打字四结合。汉语文字独特的造字方法和构词方式本身隐含着“易学”而“有趣”的因素,符合文字规范的理想编码与电脑多媒体技术的完美结合,将可使学汉字在一定程度上变得如同玩电脑游戏一般轻松有趣。从现在起,电脑一定会在中国逐渐普及。而如果我们不但能够使普通大众不再为输入汉字犯难,还可以使学汉字变得如同玩游戏一样吸引我们的孩子,反过来又可大大加速电脑在中国的普及,这对于我们社会的发展与进步无疑将具有巨大的促进作用。

(责任编辑 蔡德诚)