为何汉字形码输入法难以走出“难”的困境？
---谈谈一些技术上的欠妥观点*

张小衡
香港理工大学中文及双语学系
ctxzhang@polyu.edu.hk

摘　要：近年来汉字形码输入法市场萎缩，举步为艰。这固然与拼音输入法的兴起和发展有关，但根本的原因在于形码输入法的难学与难用。而这“难”又与部件编码技术上存在的种种欠妥观点有着千丝万缕的关系。本文指出了八个这样的观点，并逐一进行分析讨论，以求解脱良策。这些观点包括：一：部件是具有组配汉字功能的笔画组合；二：基本部件是不能再切分的部件；三：信息处理和语文教学需要两套不同的汉字部件体系；四：应尽量限制部件数量；五：部件输入法需要列出“部件-代码”对照表；六：单部件字需特殊编码；七：部件分组应尽量离散，注意热键；八：部件编码难，笔画编码长，难以两全其美。

关键词：汉字输入法、字形编码、技术观点

Technical Misconceptions in Component-Coding Chinese Character Input Methods

Zhang Xiaoheng
Department of Chinese & Bilingual Studies, Hong Kong Polytechnic University
ctxzhang@polyu.edu.hk

Abstract: In the IT domain of Chinese character input, an overwhelming majority of both research workers and research products belong to the branch of component-based coding. Yet the influence of this branch is dwindling. The reason is twofold: externally due to the aggressive development of Pinyin input methods, internally due to the existence of some handicapping traps. The paper is concerned with the later, focusing on the introduction and discussion of eight technically misguiding points of view.

Key Words: Chinese character input method, Component-based coding, Technical misconception

1 前言

*本研究得到香港理工大学的资助。项目号码：香港理工大学 1-9827.

从 1978 年 7 月 19 日上海《文汇报》报道支秉彝教授的编码方案到今天，汉字形码输入法已经走过近四分之一个世纪。这对于信息技术时代来讲不得不说是一段漫长的历程。在这历程中形码输入法，尤其是部件编码输入法，形成了千军万“码”的阵容，为现代社会做出了有目共睹的功绩。

但是，随着拼音输入法的兴起和发展，形码输入法日显步履艰难，甚至连当好拼音输入法的“助手”都不容易。其主要原因在于形码输入法难学难用又难记。为什么会这样呢？这与汉字庞大的数量和复杂的字形固然有关，但与“指导” 编码方案设计的各种欠妥观点或看法也不无关系。

张普教授关于汉字键盘输入的三个误区（即“重码率越低越好”、“速度越快越好”和“词库越大越好”）的分析（张普，1992）在社会上曾引起巨大的反响，有效地纠正了人们头脑中的一些偏见。这些误区观点是具目标性的，主要讲述一个研制得“好”的输入法应该是什么样的；而本文讨论的欠妥观点则侧重于技术方面，主要讲述“好”的输入法应该如何建立。之所以不用“误区”而用“欠妥”是因为这些观点往往都有其正确的方面，只是对于汉字输入法的研制来说存在着不够全面、不够具体、难以操作或事倍功半等欠佳的地方。

形码输入法大多属于部件编码输入法，以规定的基本部件（即字根）为码元。本文将围绕部件与基本部件的技术处理展开讨论，内容是根据笔者多年来在香港这个繁简体汉字并用的地区从事输入法的教学、研究和软件制作的亲身体会总结出来的，在此与有关专家讨论，不妥之处，请指正。

2 一些技术上的欠妥观点

2.1“部件是具有组配汉字功能的笔画组合”

这种定义在学术上和语文教学上是完全可行的，但对于计算机信息处理来说却不易操作。问题出在短语“组配汉字功能”（简称“组字功能”） 。首先是难以判别：一个笔组结构应该在哪个字集中组成多少个字以上才算具有组字功能呢？如果字集定得太小，则用途不大，许多集外字处理不好；如果定得太大，则用户难以操作，因为他/她只认识其中一小部分字。至于组字数量的定界，更是难以找到一个较为理想的标准。定为一个字的话，那就等于没有限制，因为在任何一个汉字中划分出来的任一个笔组都满足这一要求。如果定为二，又有什么理由说明这样比三、四等更优越呢？况且不论定为多少，都会出现同一笔组的部件资格在不同的参照字集中有不同的判别的混乱局面，而且字集变化越大，问题越严重。要是还要考虑字的使用频率及其行业地区差别的话，那笔组的部件身份就更难判别了。

其实文字专家从特定字集中划分出来的部件许多就只出现一次（请参看《汉字信息字

典》(李松宜、刘如水，1988)和《汉字属性字典》（傅永和，1989）两书中的基本部件频度表），这进一步证明了在实践中部件的定义和划分不能完全依靠“组字功能”的有无。

2.2 “基本部件是汉字中不能再切分的部件”

这个定义的内容也没有错,但不够具体，“不能再切分”需要一个便于操作的说明。

“不能再切分”首先应该指如果再分就不是部件了，但是这牵涉部件的定义，进而涉及到上文所述的 “组字能力” 判别难题。即使能顺利判别一个笔组是否有“组字能力”，也不一定万事大吉。请看，如果根据 “组字能力”的有无（或强弱）将结构类似的部件对“示-元”和 “黑-杰” 中的前者（“示”和 “黑”）定为基本部件而将后者（“元”和 “杰”）定为非基本部件，那就破坏了基本部件的形体规律性，增加了学习的难度。

“组字能力”的路不可靠，只好用附加规则来规定“不能再切分”的内涵，例如《信息处理用 GB13000.1 字符集汉字部件规范》（国家语委, 1997）规定笔画交重的部件不拆，这是很好操作的。但该规范同时又指出相离相接可拆，这就带来了新的问题，因为“可拆”同时意味着“可不拆”，什么情形下不能拆呢？回答这一问题往往又得用到“组字功能”。这是一个很值得重视的问题，因为汉字中笔画相接相离的情形远比较重的多。

由于难以用规则的手段为基本部件找到一个既全面又好操作的定义，而基本部件的无二义性的判断对于汉字输入又是至关重要的；所以人们只好将自己认可的对应于某个字集的所有基本部件一一列出，硬性规定下来。然而， 由于缺乏强有力的规则作基础，基本部件的取舍没有一个统一严密的标准，由此形成的基本部件表常常五花八门，规律性差，意味着用户得死记硬背。这种“列表定义法”通用性也不好，“基本部件表”会随着相应字集的变化而变化。因此，靠基本部件表来编码的形码输入法当然就难以不难了。到了这种境界，本节标题中关于基本部件的定义也就失去了意义。

2.3 “信息处理和语文教学需要不同的汉字部件体系”

如果信息处理和语文教学采用两套不同的汉字部件体系，则在一定程度上承认两者的不可协调性。同一个笔组结构是否为基本部件，是否为独体字，两套标准可能会作出不同的判断。例如，根据《信息处理用 GB13000.1 字符集汉字部件规范》，“非”是基本部件，“韭”是基本部件的组合；而作为教育部教材的《现代汉字学纲要》（苏培成，2001. p88）的看法却恰恰相反。这样不仅会影响语文教育，而且会增加汉字输入系统用户的负担。幸好《信息处理用 GB13000.1 字符集汉字部件规范》较好地考虑到语言教学的需要，大大减少了两者冲突的机会。

认为两套部件体系难于合二为一的主要依据是：信息处理用的基本部件数目不能太大，所以部件颗粒度应该比较小，因此需要专用的部件体系。关于信息处理用的部件是否非得少而小，我们将在下一节讨论。即使真的需要小部件也并不意味着一定得有别于语文教育

的部件体系。费锦昌教授（1996）为我们提出了两全其美的解决方法：统一部件系列，粗细兼供，任君选用。

采用小部件的另一个理由是：一个汉字的编碼部件数和部件的大小成反比，如部件太大，则汉字码长过短，影响键选率。例如，如果按照语文教学的做法，将大量的形声字只划分为两个部件，用两个字母表示，则会出现大量的重码。其实，一个汉字的部件少并不意味着码长必定短，因此不一定会影响键选率。换句话说，我们还可以通过其它的办法来保证较好的鍵选率。关于这一点，下文 2.6 和 2.8 节将有进一步的讨论。

2.4 “应尽量限制部件数量”

一般来讲，一个给定汉字集所需的编码部件集的大小同部件颗粒度成正比。而部件颗粒度太小了则会把字拆得过于零碎，这样，不仅需要较大码长，而且可能破坏汉字的结构规律。因此较有影响的部件编码输入法所使用的部件数目都过百。然而，人们同时又普遍认为部件多了也不好。主要理由有二：（1）部件越多越难认难学，（2）部件越多越难于合理分组安排到有限的（26 个）键元上。

从 2.2 节的讨论中我们已经看到增加部件个数不一定增加学习难度。如果将部件对“犬-太”和“黑-杰”中的前者（犬黑）收为基本部件，而将后者拒之门外，恐怕还不如一视同仁把它们都收进去好掌握。因为两组部件都具有某些共同的形态特征。又如，如果把“日、木、水、土”定为需要死记的基本部件，还不如将“日、月、金、木、水、火、土”都收进去，因为后者是一个人们所熟悉的“系统”。可见，元素数量多的部件集不一定就比元素数量少的部件集难学难记，要看是否有规律可循，是否有效地利用了人们的固有知识。当然，如果待学的部件无规律可言或所提供的规则太过繁复，用户就只能逐个死记，则部件数量越多就会越难学难记。但实际上，汉字输入法中选用的编码部件是完全可以做到有规律或基本有规律的。再说，减少部件数量有时反而会影响规则的建立，减弱其完整性和覆盖面，形成恶性循环：“部件太多--减少部件数--破会部件识别规律和汉字（拆分）结构—需要死记的例外部件增多--部件太多”。结果是得不偿失。

大部件集也不一定就比小部件集难以合理分组安排到有限的键元上。部件分组方法的主要流派是“音托”和“形托”（张普，1997, p26），两者各有难处。音托的困难是：许多部件无读音或无统一读音；形托的困难是：许多部件难以找出与英文字母相似的规律。因此出现许多牵强处理的部件，它们的组别需要特别记忆，因此部件数不能多，越多越难处理。然而，除了传统的“音托”和“形托”之外，我们完全可以找到一些规律性较强，较好掌握，其使用难度又与部件多寡无关的分组方法。例如，多笔部件可根据自身的头两笔分配到 25 个键元上。

2.5 “部件输入法需要 ‘部件-代码’对应表”

翻开《汉字键盘输入技术与理论基础》（陈一凡和胡宣华，1994）这本著作，您会发现其中介绍的众多部件编码输入法都拥有一个“特征元键位分布表”。这是一个“部件-代码”对应表，列出该输入法的每个部件码元和与之对应的键元字符。一个部件通常用一个键元表示。

需要列表往往意味着没有简单明了的规则可用，要求用户记住每个部件及其对应的键元代码，因此难度较大。

另一个缺点是，一个部件-键元对应表只适用于一个汉字集（及其子集）。增加或改变字集元素意味着部件集也可能需要修改，因此对应表也得修改。而且字集越大对应表中的部件码元越多。如果还要考虑修改后的部件集的键位分布均匀性，则对应表的调整与改动可能会更大。这对于输入法的稳定性和易学性都是不利的。

第三个缺点是单部件字往往需要特殊处理。这将在下一节作进一步讨论。

2.6 “单部件字需要特殊的编码规则”

纵观各种部件编码输入法，还可以发现它们对单部件字常常采用特殊的取码方式。例如仓颉输入法规定某些编码部件在单独成字时应拆分编码，例如“士”（在合体字中）的部件码是 G，但独立成字时则必须拆分为“十(J)、一(M)”。表形码的单部件字处理方法是：除该部件的编码外，还要增加补充码 KK，以及该字第一个声母，例如“厂”的部件码是 J，单字码是 JKKC。这样处理无疑增加了编码的复杂性和学习难度。

单部件字（独体字）需要特殊处理，这是因为根据码表一个部件往往只用单个键元字符表示，因此单部件字需要“加长”编码，以防键选率过高。其实一个部件对应一个键元的做法并不是天经地义的，一个汉字的基本部件少并不意味着码长必定短。这将在 2.8 节作进一步解释。

顺便指出，采用小部件也是为了解决码长过短的问题（以保证较好键选率）。这也是“一部件一码”的误解所造成的。

2.7 “部件分组应该尽量离散，并注意热键”

“分组离散”是指把部件较为均匀地分布到每个键元上，这样既能降低键选率，又能平衡手指负担。“注意热键” 是指为较易操作的键位适当增加负担，以利快速击键。

然而，分组离散和重用热键常常会削弱编码的规律性，增加例外。例如在安排单笔部件“横、竖、撇、点、折”的键位时，如果重点考虑离散和热键，则由此产生的部件-键元对应关系从语言科学的角度来看是无理的，需逐个记忆；如果将这五种基本笔画分别用其普通话拼音的首字母编码，即“横-h、竖-s、撇-p、点-d、折-z”，结果既合规范又容易学习。权衡利弊，恐怕后一种处理方法更可取。

众所周知，全拼音汉字输入法的键位设计无论在分组离散或重用热键方面都没有专门

考虑，且单字重码率很高，但因为它符合语文规范，易学易用，所以很受青睐，其影响度已远远超过各种在键位分布和热键处理方面费尽苦心且重码率极低的形码输入法。使用最为广泛的（ASDFGHJKL）英文键盘的字母键位设计也是不大有利于英文快速打字的，但由于历史悠久，习以为常，一直是大家所公认的“标准国际键盘”。

因此分组离散和重用热键不应该以语言规律和习惯为代价。

2.8 “部件编码难，笔画编码长，难以两全其美”

在形码输入法中，部件编码和笔画编码常常被视为两类不同的输入法，井水不犯河水。然而它们的优缺点是互补的：部件编码复杂难学，但码长较短，汉字输入速度快；而笔画/笔形编码简单易学，但（由于编码颗粒度小）码长较大，输入速度慢。

其实，我们可以考虑将它们结合起来，取长补短。可以根据笔顺笔形关系来定义部件，来为部件分组编码。这样能提高编码方案的规律性，而且一个部件的码长不需局限于一个建元字符，单部件字符的输入也不需特殊编码。

3. 结束语

上文分析讨论了汉字形码输入法，尤其是部件编码输入法，在技术上存在的八个欠妥观点。这些观点的消极因素相互联系，形成阻碍形码输入法健康发展的“羁绊网”，使其难以走出“难”的困境。

对于信息处理来说，一个字的部件划分恐怕应该多从该字本身的形体结构来考虑，只有在这种做法鞭长莫及或严重违反语言规律的地方才借用其它手段。因此，应该充分重视分隔沟（苏培成，2001, p75,76）和单结构(中国大百科全书出版社,1994, p164)的作用，并注意参照 GB 部件规范的有关规则。

应该尽量通过有高度概括性的合理的规则明确定义基本部件及其键盘代码，注意灵活性和可扩展性，不要局限于某个汉字集，应将一切可能（即符合汉字形态规律）的汉字都考虑进去，做到一劳永逸。即使需要同时列出基本部件表，也应该做到简明的文字定义能覆盖该表的绝大多数内容，以减少用户的负担。

我们利用上面的思路，设计了一个无传统部件-代码对应表的与具体汉字集无关的部件编码输入法，并成功应用于 GB13000.1 大字集。效果是令人鼓舞的（张小衡，2003）。

由于作者水平和文章篇幅所限，上述讨论挂一漏百，不妥之处，请批评指正。关于上述每个欠妥观点的更深入更全面的分析讨论，恐怕都足以自成一篇文章。

参 考 文 献

[1] 张普. 走出汉字输入的三个误区. 《中国计算机报》, 1992 年 2 月 11.(又载于 张普著《汉字编码键盘输入文集》. 北京：中国标准出版社.)

[2] 李松宜，刘如水主编.《汉字信息字典》. 北京:科学出版社, 1988.

[3] 傅永和主编.《汉字属性字典》.北京:语文出版社, 1989.

[4] 国家语委. 信息处理用 GB13000.1 字符集汉字部件规范. 北京：国家语言文字工作委员会, 1997.

[5] 苏培成.《现代汉字学纲要》.北京：北京大学出版社, 2001.(p88;p75, 76)

[6] 费锦昌.现代汉字部件探究. 《语言文字应用》. 1996 年第 2 期 (p25).

[7] 张普. 首选中文输入法的补充 - 字根式. 见 张普著. 汉字编码键盘输入文集. 北京：中国标准出版社, 1997. p26.

[8] 陈一凡，胡宣华.《汉字键盘输入技术与理论基础》.北京：清华大学出版社；南宁：广西科学技术出版社, 1994.

[9] 中国大百科全书出版社. 《语言文字百科全书》. 北京：中国大百科全书出版社, 1994. 〈p164〉

[10] 张小衡（2003）. 正易全：一个动态结构笔组汉字编码输入法. 《中文信息学报》, 2003 年第 3 期。

鸣谢：作者的同事张群显博士，苏詠昌博士，刘镇发博士等在工作上给了不少帮助和支持。特此鸣谢。