文章编号: 1003－0077(2004)04－0060－06

# 《信息处理用 GB13000.1 字符集汉字部件规范》在输入法应用中的难点讨论*

张小衡

(香港理工大学 中文及双语学系, 香港)

摘要:《信息处理用 GB13000.1 字符集汉字部件规范》对于规范汉字形码输入法具有非常重要的意义。然而,在实际运用上却存在着部件数量太大,部件定义难以操作,部件拆分组合不易掌握等难处。造成困难的原因主要有:(1)基础部件主要靠列表来确定,(2)部件强调按理切分和成字组合,(3)过多依赖"组字能力"的判别,(4)过分注重部件数量的限制。要走出"难"的困境,应该在现有规范的基础上根据汉字的形态特征制定出简便可靠的部件识别规则和切分规则。实验证明,这种方法是行之有效的。

关键词: 计算机应用; 中文信息处理; 汉字输入; 汉字部件; 规范

中图分类号: TP391 文献标识码: A

# Difficulties in the Application of "Chinese Character Component Standard of GB 13000.1 Character Set for Information Processing" for Chinese Character Input

ZHANG Xiao-heng

(Department of Chinese & Bilingual Studies, Hong Kong Polytechnic University, Hong Kong)

**Abstract**: Chinese Character Component Standard of GB 13000.1 Character Set for Information Processing is an important document for the standardization of Chinese character input methods. Yet, when employed to the design and implementation of a nontrivial Chinese character input system, the standard encountered a number of difficulties: the hard-to-remember large number of coding components, the difficult-to-maneuver definition of basic components, and the poor rules for component disassembly and assembly. The sources of these difficulties include (a) definition of basic components by enumeration, (b) disassembly and assembly of components based on etymology and formation of characters, (c) reliance on the judgment of character-forming capability of candidate components, and (d) over-emphasis on the restriction of the number of basic components. To escape from this difficult position, we urgently need convenient and reliable rules for component identification and segmentation, which can be built up on the basis of the existing component standard by taking full advantage of the form features of Chinese characters. The feasibility and effectiveness of the proposed methodology have been verified by the successful development of the ZYQ Chinese character input system.

**Key words**: computer application; Chinese information processing; Chinese character input; Chinese character component; standard

## 1 引言

《信息处理用GB13000.1字符集汉字部件规范》[1] 于1997年12月5日公布,开宗明义地指

* 收稿日期: 2004－02－18
基金项目: 香港理工大学研究资金资助项目(1－9827, G－T957)
作者简介: 张小衡(1958－), 男, 助理教授, 博士, 主要研究领域为计算语言学和计算机辅助教学.

出:本规范主要用于中文信息处理,特别是汉字键盘输入。六年多过去了,但在"中国期刊全文数据库"网站上却难以找到关于该规范在输入法应用方面的专题报道或讨论。笔者近两年来研制繁简体兼容汉字形码输入法,选用 GB13000.1 字集,自然地用到了该规范,在实践中体会良多,望与业内人士商榷。

《信息处理用 GB13000.1 字符集汉字部件规范》(以下简称《规范》)给出了 GB13000.1 字符集的《汉字基础部件表》及其使用规则,对于规范汉字形码输入法,尤其是部件编码输入法,具有重要的指导作用。《规范》以形为主,立足现代,而且规定相交笔画不可切分为更小部件,有效地避免了一些混乱局面。此外,基础部件表收录了许多常用部首,符合人们的语言习惯。当然《规范》的成绩还有很多,但本文将重点讨论它在汉字输入法应用中遇到的一些困难及其处理方法,目的有两方面:(1)请教于有关专家,(2)促进《规范》的进一步完善。

## 2 《规范》在实际应用中遇到的困难

### 2.1 基础部件太多

《规范》的主体是《汉字基础部件表》,其部件数多达 560。《规范》还规定基础部件不得拆分成更小的部件。我们知道,一般来讲,一个给定的汉字集所需的编码部件(也称码元、字元、字根等)的数目是与部件的颗粒度成正比的,编码部件越大所需要的部件越多。可见,在 560 个部件的基础上,不切分,单靠合并是难以减少编码部件总数的,况且《规范》所允许的部件合并只限于成字的组合。要用至少五百多个部件码元通过 ASCII 键盘上的二三十个键元来为 GB13000.1 字符集的两万多个汉字编码,恐怕是一件相当艰难的事,由此产生的输入法恐怕用户也不易掌握。况且当汉字集进一步扩大时部件数也会随着增加。

其实,研制《规范》的专家学者在限制基础部件的数量方面已经作了很大的努力,然而,560 这个数目对于用户来讲仍然是一个相当沉重的负担。

### 2.2 部件的定义难以操作

严格来讲,编码部件多并不一定难以掌握,要看是否配有简明易用的部件识别规则。可惜,《规范》没有提供这样的规则。

《规范》将汉字部件定义为"由笔画组成的具有组配汉字功能的构字单位"。这种定义在学术理论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:一个构字单位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(例如 3500 常用简体字),则用途不大,许多集外字不好处理;如果定得太大(例如 GB13000.1),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合理的标准。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画组合都满足这一要求。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,以及人们的识字能力差别的话,那笔组的部件身份就更难判别了。

实际上,语言文字专家从特定字集中划分出来的部件许多就只出现在一个字中(请参看《汉字信息字典》[2]和《中文信息处理》[3]两书中的基础部件组字统计表),这进一步说明部件的定义和划分不应该过多依靠"组字功能"的有无或强弱。

### 2.3 基础部件的文字定义也不够具体

《规范》将基础部件定义为“最小的不再切分的部件”。满足什么条件的部件才算“最小的不再切分的”呢？定义中没有说明。最小的不再切分的部件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，进而牵涉到上文所述的“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，如果根据切分后各成分的“组字能力”的有无(或强弱)，将“示—元”和“黑—杰”两组字中的前者(即“示”和“黑”)定为基础部件而将后者(“元”和“杰”)拒之门外，那就削弱了基础部件的形态特征规律，增加了学习的难度。

《规范》规定的“不得切分”的部件有两种：交重笔画和基础部件。前者有效，但只适应于少数部件，而且存在几个笔画搭挂的例外情形；后者与原基础部件的定义构成循环定义。

无识别规则指出基础部件有别于其他笔组或部件的形态特征，意味着要求用户根据《汉字基础部件表》死记五百多个部件，这恐怕是难以接受的。况且，基础部件表还会随着相应字集的变化而变化。

2.4 部件拆分和组合的规则不易掌握

《规范》规定，拆分部件时，“交重不拆”，这是很好操作的，但同时又指出，“相离、相接可拆”，这就麻烦了，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关于组字功能或基础部件的判断，两者都不易处理。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交重的多。

《规范》还规定对多部件的汉字进行拆分时，应先依汉字组合层次做有理据拆分，直至不能进行有理据拆分而仍需拆分时，再做无理据拆分。按理据拆分需要字源知识。一般人对常用字的字源理据都只是一知半解，何况那些难字僻字和外国汉字！然而，作为音码的补充，形码输入法的一个主要任务就是处理这些生僻字。

《汉字形义分析字典》[4]是有字源理据说明的收字较多的现代辞书，但也只处理7000个通用字。作者还在书的前言中特别指出“有些字的形义联系聚讼纷纭，争论不已，迄无定论。”著名的文字学家都如此认为，怎么可以要求普通的中文输入法用户对GB13000.1的20902个古今中外汉字的字源理据都有一个唯一正确的了解呢？

即使能按字源切分也不一定能收到最佳的效果。例如，人们常常把“章”字分解为“立 &早”，而不是“音 &十”，尽管从字源上看，后一种切分法才是合理的(《说文解字》中“章”从“音”从“十”)。这一事实说明，笔画相接与相离之间的形态差别直接影响到人们对汉字结构的分析和表达，但是《规范》的部件拆分规则却没有顾及到这一点。

又如，赢、碧、修、疆等字《现代汉语规范字典》[5]和《21世纪汉字学习字典》[6]等普通语文辞书都根据字形把它们简单地处理为“上(中)下”结构或“左(中)右”结构。如果严格根据这些字的字源理据(许多人都不大了解)来拆分，将“赢”字二分为“贝”字和余下部分，将“碧”字二分为“白”和余下部分，将“修”字二分为“彡”和余下部分，将“疆”字二分为“土”和余下部分，将繁体字“虽”二分为“虫”和“唯”，这样做未免太古板了，与“信息处理用”的性质很不相称。而且允许这种“边角切挖”在字形上大大增加了切分的可能方式，增加汉字编码的难度。

即使我们能把每个字的字源都弄清楚(其实是不可能的)，而且假设每个字都只有一种理据解释(其实无法做到)，部件切分仍然存在不确定的问题。例如，形码输入法常常需要将合体字分为两部分来编码(如“二笔”、“仓颉”和“正易全”的做法)，然而某些字根据字源在同一个层次上却有多条切分线，例如左中右结构的“街”字、上中下结构的“暴”和“衷”等，二部切分时应该选取哪一条切分线呢？对此，《规范》未作规定。此外，字源分析往往要追溯到繁体字，甚至篆、金、甲骨文字形上去，其结构分析结果对某些字形迥异的简体字来说可谓鞭长莫及。例如，

对繁体字“农、兽、书、  、壳、亲”等字的字源分析并不能帮助我们对相应的简体字“农、兽、书、县、壳、亲”作部件切分。

关于部件组合,《规范》规定基础部件可以组成成字部件使用,但不得组合成非成字部件使用。这样可以限制许多中间部件,避免某些混乱,但其代价也相当大。首先是要求用户判断一个组合部件是否成字,判断结果会因字集而异。其次,由于编码部件的取用受到限制,整个编码方案所需的码元总数可能会增加(即GB基础部件集+附加成字部件。)

2.5 其它一些问题

《规范》是针对GB13000.1字符集而制定的。这意味着当字符集发生变化时,《规范》也很可能需要修改。例如,作为《规范》主要内容的《汉字基础部件表》是对GB13000.1字符集中的20902个汉字逐个进行拆分、归纳与统计后制定的。那么,当字集扩充为GB18030—2000甚至《中华字符集》[7]时(我们不能忘记形码的一大优点是能处理大字集。),许多新增的字要拆分,从而产生新的基础部件。如果还考虑到字集的变化给构字单位的“构字能力”和“成字与否”的判断所带来的影响,那就得把原来GB13000.1的字符重新处理一次。如此一来,基础部件表的变动可能是相当可观的,这无论对部件规范的制定者还是使用者都会带来很大的不便。可见《规范》的通用性和灵活性有待改善。

前文讲到,严格遵循《规范》的部件编码输入法所需的码元部件应该在560个以上(其中不少是成字部件),如果按照惯例将它们分组与二十多个(英文字母)键元对应的话,则平均每个建元要对应二三十个部件,使得单部件字的键选率相当高。要解决这一问题,我们可给单部件的输入规定特别的编码方法,但这样又会增加输入法的复杂性。

此外《规范》同一般语文教育也存在一些分歧。前文关于部件切分和基础部件的讨论已经给出一些例子,这里再提供一些证据。汉字“熏、黑、非、兆、竹”等字都是《规范》中的基础部件(将它们看成不可分的部件和独体字有悖于字形视觉。),但是由原国家语委主任,全国人大副委员长许嘉璐教授主编的《现代汉语规范字典》却将它们定为合体字。另一方面,“鸟、鸟、马、马、太、良、斗、头、互、页、术、击、夹、夷、来”等都不属于《规范》的基础部件,但是《现代汉语规范字典》却认为它们是独体字。

## 3 造成困难的原因

造成以上困难的原因有几方面:(1)基础部件靠列表的方法来规定,(2)部件强调按理切分,(3)依赖“组字能力”的判别,(4)部件数量的限制过严。

先谈第一个原因。《规范》中的基础部件主要依靠穷尽列表法来确定,要熟练使用《规范》就得首先将基础部件逐个记住(否则就得频繁查表)。因此五百多个部件当然太多了,太难掌握了。况且基础部件表还会随着汉字集的变化而变化。

困难的另一个原因是强调汉字的部件切分应首先考虑字源理据。这对用户来讲又是个很难的“任务”。例如,“太”可以分成“大”和“丶”,“生”可以分为“丿”和余下部分,是因为符合字源理据。而与它们字形特征相似的“犬”和“牛”却属基础部件,不可切分,因为得不到字源的支持。又如,“荤”和“荣”的上部字形一样,《新华字典》和《现代汉语词典》都把它们同归“艹”部。但如按照字源解析则必须采用不同的切分方法:“荤”是“从艹军声”而“荣”则“从木荧省声”[4]。

第三个原因是依赖“组字能力”的判别。例如将“黑”和“熏”定为基础部件,而将“杰”和“鸟”定为组合部件就是因为后面两个字都可以分为具有构成其他字的能力的两个部分,而前面两个字却不行。“兆”和“非”不切分也是因为两个分离对称部分不分开用以构字。

第四个原因是过分重视限制基础部件数。我们来看一组统计数字:《汉字信息字典》从7785个正体字中切分出的异形基础部件有623个[2];傅永和教授的实验结果是,11834字含异形基础部件648个[3]。《规范》的对象字集是GB13000.1,字数达20902个,远远大于前面两个字集,但切分出来的基础部件却只有560个。可见《规范》在控制部件数量方面是下了很大功夫的,这主要是通过上述的切分理据和组字功能来加以限制。结果是损害了基础部件的形态特征规律,又影响与语文教育的一致性。另一方面,560个基础部件对于信息处理来说仍嫌太多。

## 4 克服困难的方法探索

为方便信息处理,汉字基础部件的定义应该通过简明的词语为部件识别提供充分而必要的依据,应该根据汉字的字形特征,总结出一套适用于所有可能汉字的易于操作使用的基础部件判别规则,即使需要列表,也应该做到文字定义能正确覆盖表中的绝大多数部件。《规范》中的《基础部件表》有一个很有用的总体形态特征,就是绝大多数的基础部件都是笔画的连接体。应该进一步强化这一规律,尽量减少"例外部件"。可以考虑将那些非常用的和没有达成语文界共识的分体基础部件分解为连体部件,例如,"非、兆、竹、门"等字可分为左右两个部件。这种想法与傅永和教授的观点是一致的,他认为"从存在形式上看,部件是一个独立的书写单位,凡是笔画串联在一起的,都作为一个部件看待"[3]。这就是一条覆盖面广,适应于任何汉字集的便于操作的部件识别规则,已成功应用于对《辞海》(1979年版)11834个规范汉字的部件切分。此外,教育部和国家语委等单位联合推荐的《笔形查字法》[8]中关于"单结构"的定义* 也是很值得借鉴的。

汉字部件的拆分也应从形出发,订出便捷可靠的规则,使得所有用户,包括汉字初学者,利用这些规则对任何一个汉字进行部件划分时都能得到一致的结果。笔者赞成苏培成教授关于汉字切分应该重视利用分割沟的建议[9]。可在此基础上作进一步规定。

另一方面,我们又得防止部件规范把输入法的研制管得太细太死。例如,在为"非"字编码时,如果将它看成一个不可切分的整体,则只需一个码元部件,但这样做与语文教育和形态感觉有所抵触,而且这个码元需要特别记忆(因为一般来讲码元部件大多是笔画的连接体);如果将"非"字左右两部分开,则需要两个码元部件,但与语文教育和形态视觉都比较一致,容易识别。以上两种处理方法各有利弊,应该让编码设计者自己去选择。因此部件规范应该将"非"的左右两部都定为基础部件,使用户能选择分开使用或者合并使用。又如,在给"意、喜、器、曼"等上中下结构的字做二部切分时如果严格按字源来选择切分线,则比较合理,但难度大;如果统一切分为"(上)/(中下)"则容易操作,但理据性较低。两种切分方法也各有利弊,应该允许用户去选择,为此可规定"上中下"结构的字做二部切分时,切分线应该是在"上—中"之间或者"中—下"之间。总之,部件规范在那些把握不大或仍具有争议的地方宜柔不宜刚,应该给输入法编码方案的设计者留有适当的自由空间,否则就会越俎代庖,把自己弄成一个惟我独尊的"标准"编码方案,而不是一个指导各家编码方案健康发展的良师益友。

部件切分和识别的问题解决好了,基础部件的多寡也就无关紧要了,因此不必去为"构字能力"和"字源理据"的判断而烦恼,同时避免由此带来的与语文教育的不一致。

## 5 结束语

《信息处理用GB13000.1字符集汉字部件规范》对于规范汉字形码输入法具有非常重要的

---

* 该定义把"单结构"规定为两类:(a)相交,相连的笔画;(b)单笔与其他笔画相配相从。

意义，但是在实际运用上难度相当大。上文指出了一些具体困难，分析了困难的原因，并对解决困难的方法作了初步探索。

普通的汉字输入用户大多以拼音输入法为首选，把形码输入法作为拼音输入法的辅助工具，以处理大字集和生僻字等问题。作为辅助工具的形码输入法，除了帮助解决音码输入法所难以解决的问题之外，还应该做到易学和易记。然而，目前部件编码输入法的最大缺点就是难学难记。要走出“难”的困境，不能只重视控制部件数量[10]。治本的更有效的办法是，根据汉字的某些较易掌握的形态结构特征制订出合理简便的汉字部件判别规则和切分规则。这种思路的可行性和有效性已在“正易全”汉字输入系统[11]的实践中得到了证明。

部件规范还需要字形规范工作的支持。现在汉字标准字形的“底板”是1964年编成的《印刷通用汉字字形表》，很少考虑到信息处理的需要。希望正在研制的《规范汉字表》会重视这一点，使汉字的部件结构和构字理据更好地在字形上得到体现，减轻用户的负担。汉字的简化为汉语手写带来了方便，现在很多人都“换笔”了，希望新的字形规范能更多考虑到信息时代用电脑写(打)字的方便。

理想的汉字部件规范应该是面向全体汉字，面向所有应用领域。应该像费锦昌教授[12]所提倡的那样，基础部件定得小一些，容易掌握一些，在此基础上可根据简便合理的规则组成各级合成部件，形成一个统一灵活的部件规范体系，粗细兼供，使用者可以根据自己的需要来“点菜”。这样的部件体系富有前瞻性和通用性，在相当程度上做到一劳永逸。

最后，笔者想顺便说明，尽管本文对《规范》有一定的批评性，但《规范》的成绩远远大于不足，二三十年来的低水平“万码混战”局面说明我们迫切需要统一的指导，有识之士着手研制《规范》的行动本身就是一大创举，况且他们的成果中还有许多可贵的内容。如果说本文的讨论有某些可取之处的话，那也是在实践《规范》的基础上得来的。至于不妥之处，敬请有关专家学者批评指正。祝愿《规范》在信息处理的应用中不断发展完善。

鸣谢　(1)本文初稿承蒙前国家语委副主任傅永和教授垂阅并提出了宝贵意见。但文中仍存在的任何不妥之处由作者负责。(2)感谢审稿专家的宝贵意见。

参　考　文　献：

[1] 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[S]. 北京：国家语委，1997.
[2] 李公宜，刘如水. 汉字信息字典[M]. 北京：科学出版社，1988.
[3] 傅永和. 中文信息处理[M]. 广州：广东教育出版社，1999，26.
[4] 曹先擢，苏培成. 汉字形义分析字典[M]. 北京：北京大学出版社，1999.
[5] 许嘉璐. 现代汉语规范字典[M]. 北京：中国社会科学出版社，2000.
[6] 饶明华，吕游之. 21世纪汉字学习字典[M]. 上海：上海文化出版社，2000.
[7] 李宇明. 搭建中华字符集大平台[J]. 中文信息学报，2003(2)：1－6.
[8] 中国大百科全书出版社. 语言文字百科全书[M]. 北京：中国大百科全书出版社，1994，164.
[9] 苏培成. 现代汉字学纲要[M]. 北京：北京大学出版社，2001，75.
[10] 张小衡. 为何汉字形码输入法难以走出“难”的困境？[A]. 孙茂松，陈群秀. 语言计算与基于内容的文本处理[C]. 北京：清华大学出版社，2003，614－620.
[11] 张小衡，正易全. 一个动态结构笔组汉字编码输入法[J]. 中文信息学报，2003(3)：59－65.
[12] 费锦昌. 现代汉字部件探究[J]. 语言文字应用，1996(2)：20－26.