文章编号: 1003－0077(2005)01－0098－07

进一步的“正易全”——三级汉字编码输入法①

张小衡

(香港理工大学 中文及双语学系, 香港)

摘要: 本文报告“正易全”汉字输入法的新进展。从整体上来讲, 正易全已发展成为全字笔顺、全字笔组和2_21笔组三级输入法系列。前两级简单灵活, 键选率极低, 方便大字集查检; 第三级在常用字和通用字中表现极佳, 适合日常快速打字。在编码技术上, 多笔笔组码元的选用、单结构的定义和多结构字的二部划分等方面都作了进一步的简化、系统化和规律化。此外, 码表在GB13000.1字符集的基础上增加了1164个港澳台地区用字或字形。

关键词: 计算机应用; 中文信息处理; 汉字输入; 字形码; 笔组

中图分类号: TP391 文献标识码: A

Further Development of ZYQ: A Three-staged Coding Series for Chinese Character Input

ZHANG Xiao-heng

(Dept. of Chinese & Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, China)

Abstract: The ZYQ Chinese character input method has been developed into a three-staged series including the whole-character stroke order method, the whole-character stroke group method and the 2_21 stroke group method. The first two methods are simple and effective for Chinese character retrieval on large character sets, while the third method is more suitable to normal typing and writing at higher speed. Technically, further simplification, systemization and regularization has been applied to the selection of multi-stroke coding units, the definition of structural components and the bi-segmentation of multi-component Chinese characters. In addition, the coded Chinese character set has been extended to include 1164 characters specific to regions of Hong Kong, Taiwan and Macao.

Key words: computer application; Chinese information processing; Chinese character input; form-based character coding; stroke group

1 引言

“正易全”是一个以“正确”、“易用”和“全面”为指导思想的支持大型简繁体字符集的汉字形码输入法, 其初期成果已在《中文信息学报》2003年第三期上作了汇报[1], 本文重点介绍和讨论一年来的新进展。从总体上来讲, “正易全”的编码方案已发展成一个较完备的三级循序渐进输入法系列, 而且每一级输入法都实现了GB13000.1字符集的码表建设和软件处理。下面将分级介绍。

① 收稿日期: 2004－03－20
基金项目: 香港理工大学研究资金资助(1－9827, G－T766, G－T957)
作者简介: 张小衡(1958－), 男, 博士, 助理教授, 主要研究领域为计算语言学和计算机辅助教学.

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

2 第一级: 全字笔顺输入法

2.1 全字笔顺输入法的码元、键元及其对应关系

码元集: 横(包括提笔)、竖(包括竖提↓)、撇、点(包括捺)、折五种标准笔画。

键元集: 可选用数字键元集{1, 2, 3, 4, 5}或拼音键元集{h, s, p, d, z}。

码元—键元对应关系:

数字式: 五种笔画分别用其序号表示, 即: 横—1, 竖—2, 撇—3, 点—4, 折—5。

拼音式: 五种笔画分别用其普通话拼音首字母表示: 横—h, 竖—s, 撇—p, 点—d, 折—z。

2.2 笔顺输入法的汉字编码

一个汉字的数字笔顺码就是该字笔顺的数字键元表示, 其拼音笔顺码是该字笔顺的拼音键元表示。例如"中国"的数字笔顺码是"2512(中) 25112141(国)", 其拼音笔顺码是"szhs(中) szhhshdh(国)"。两种笔顺码都严格遵循国家笔顺规范[2,3]。用户在我们的软件平台上输入汉字时, 既可以使用全字笔顺码(例如: 龙(dhdphszhhhzhzhhh), 也可在代码中使用方便击键的统配符";"(相当于通常的"*")和"."(相当于"?")。例如, 利用笔顺代码"dh; dh; dh; hhh"就可在GB13000.1字符集中唯一确定48画的"[illegible]"字。

数字笔顺码照搬国家标准, 美中不足是要求用户作"笔画—数码"转换, 略显不便, 这是增加拼音笔顺码的原因。"正易全"笔顺输入法是最简单规范的形码输入法, 据笔者统计, 在GB13000.1字符集中只需取单字笔顺的前7画(不够者取完为止), 在《现代汉语通用字表》中只需取单字笔顺的前6画, 则可将平均同码字数限制在3之内, 显然笔顺输入法比传统的字典排检法要快得多, 而且有助汉语学习。

3 第二级: 全字笔组输入法

全字笔组输入法在第一级的拼音笔顺码的基础上增加双笔和多笔码元, 以降底代码长度。

3.1 全字笔组输入法的码元、键元及其对应关系

码元集: 以笔组为码元, 包括5种单笔笔组(同第一级), 25种双笔笔组(或称"笔对"), 和14种多笔部件, 共44种笔组。

键元集: a~z 26个英文字母。

码元—键元 对应关系:

笔组码元的键元表示请见表1。其中, 单笔笔组"横(包括提笔)、竖(包括竖提↓)、撇、点(包括捺)、折"用各自的汉语拼音第一个字母h、s、p、d、z表示, 与第一级的拼音笔顺码相同。双笔笔组按规范笔顺排列(〈一, 一〉, 〈一, 丨〉, …, 〈乛, 丶〉, 〈乛, 乛〉,)分别对应字母序a, b, …, x, y。多笔笔组是"日月金木水火土, 心口工之大宝草"14组常用国标部件, 分别由各自名称的普通话拼音首字母表示。部件的选用和分组参照有关汉字部件和部首的国家规范[4,5], 使得同组的部件一般都形似或义同, 而且笔顺常常也相同。

3.2 全字笔组输入法的编码

在叙述编码方法之前先定义两个有关的概念, 即"笔画组合紧密度"和"单结构"。

3.2.1 笔画组合紧密度

两个笔画的空间组合分为相交、相接和相离三种, 其紧密度大小关系是: 相交> 相接> 相离。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

表 1　**ZYQ** 笔组—键元对应关系

键元	a:	b:	c:	d:	e:
单笔				丶	
双笔	<一,一>	<一,丨>	<一,丿>	<一,丶>	<一,乛>
多笔		宀（宝盖头）	艹艹□土	大	
键元	f::	g:	h:	i:	j:
单笔			一		
双笔	<丨,一>	<丨,丨>	<丨,丿>	<丨,丶>	<丨,乛>
多笔		扌	火灬		金钅
键元	k:	l:	m:	n:	o:
单笔					
双笔	<丿,一>	<丿,丨>	<丿,丿>	<丿,丶>	<丿,乛>
多笔	口		木朩		
键元	p:	q:	r:	s:	t:
单笔	丿			丨	
双笔	<丶,一>	<丶,丨>	<丶,丿>	<丶,丶>	<丶,乛>
多笔			日曰	水氵 氺	土圭
键元	u:	v:	w:	x:	y:
单笔					
双笔	<乛,一>	<乛,丨>	<乛,丿>	<乛,丶>	<乛,乛>
多笔				心忄 灬	月冃
键元	z:				
单笔	丶				
双笔					
多笔	之辶辶				

3.2.2　单结构

参照国家推荐的《笔形查字法》(草案)中有关“单结构”的说明[6]和傅永和教授的部件定义[7]，单结构包括以下几类：

a. 相交或相接在一起的笔画结构(例如：夫车目王孝)

b. 单笔与其它笔画相配相从所形成的结构(例如：二八氵心戈)

c. 正易全的多笔笔组(例如：金[illegible]水火心)

d. 单结构的连接体，(例如：求，学字头，爱字头)。

个别模棱两可的情形按是否属国标基础部件和方便应用来确定[4]，例如“气”是规范基础部件，所以按单结构处理。此外，单结构的选用遵循大者优先的原则。

3.2.3　编码规则

严格按国家规范笔顺从字首到字尾逐个确定码元笔组并用相应的建元字母表示之。每一个笔组码元都是以其首笔为起点在首笔所属的单结构范围内根据具体字形环境动态确定的，一个汉字的第一个笔组码元的首笔是汉字的第一笔，第 n＋1 笔组的首笔是第 n 笔组的下一笔(n＝1，2，3，…)。

单个码元笔组的确定分两步进行：

第一步：界定两个首尾相交(即共用中间一笔)的笔组：

笔组 1：如果首笔为某多笔笔组码元的起笔，则取该笔组，否则取头两笔；

笔组 2：以笔组 1 的末笔为首笔按同样的方法确定笔组 2。

第二步：分两种情形确定码元的笔画范围(即决定中间公用的一笔是否归前一笔组)。

情形一：笔组 1 和笔组 2 至少有一个是多笔笔组。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.c

处理方法：如果笔组 1 的笔画数≥笔组 2 的笔画数，则取笔组 1 编码（例如，中 ks，羊 chb，夹 hcn，东 hrsn）；否则笔组 1 减末笔后再编码（只可能剩 1 或 2 画）（例如，王 ht，乏 pz，米 rm，束 hjm）。

情形二：笔组 1 和笔组 2 都是二笔笔组。

处理方法：如果笔组 1 两笔之间的组合紧密度大于或等于笔组 2 的二笔紧密度，则按笔组 1 编码（例如：丸 od，山 js），否则笔组 1 减末笔后编码（只剩 1 画）（例如，及 px，巾 sv）。

由于正易全码元中不存在一个多笔笔组完全包含另一个多笔笔组的嵌套关系，所以符合第一步中笔组 1 或笔组 2 条件的多笔笔组一定是唯一的，因此笔组 1 和笔组 2 实际上是未考虑结构环境的最大（即笔画数最多的）可能正易全笔组。而第二步中关于中间笔画的归宿处理则体现了对大部件的照顾（有利于缩短编码长度）和对紧密组合结构的保护。这样处理较合常理，同时还可以将某些笔顺相同的字区分开来。在上一段我们已经看到，同笔顺的“山”和“巾”，“丸”和“及”的代码都是不同的。又如“电”和“号”的笔顺都是“25115”，但是“电”的编码是“rz”，“号”的编码为 “ke”；日、[illegible]、“门”字边/旁的笔顺都是“2511”但是编码分别是“r”、“y”和“ja”。

编码的汉字字形以中文版和国际（英文）版 WinXP 上的 SimSun（宋体）字形为准，该字形同《GB13000.1 字符集汉字字序（笔画序）规范》的标准字形一致。又可任意放大显示汉字的字形结构，例如：“集香李罪变弯显”等字用 72 点尺寸可看出上下结构的分隔沟。

用宋体字形，而不用其它形码输入法常用的楷体字形的原因主要有两方面：（1）国家有关汉字字形、笔顺和部件的规范内容通常（首先）以宋体发布，（2）印刷物和电脑用字也多采用宋体，其实宋体常常作为缺省（default）字形使用，例如，WinXP 上微软拼音（MSPY）输入法就自动配用 SimSun，候选字也常用宋体显示。

由于汉字字形笔画相接/离关系存在着极少数摸棱两可或不便分析的情形，以下笔对一律视为相接：

- 一，例如：立（pc），文（pn），亦（pln）。
- [illegible]、冫、六字底、兆字边，例如：弟（ruvp）、习（zp）、小（sn）、贝（yhn）；脊（pnny）
- 单独平行的两点：头（sd），鼠（lhuhzszsz）.

一个多笔部件同另一个多笔部件的拐角处相接，按相离处理（例如：省着看备）

4 第三级：2 _ 21 笔组输入法

正易全 2 _ 21 笔组输入法在第二级的全字笔组码的基础上套用“2 _ 21”取码模式，以进一步缩小码长，方便日常打字。具体做法是：

单结构字采用“21”模式取码，即字头 2 码加字尾 1 码；多结构字用“2－21”模式取码，即字前部的头 2 码加字后部的头 2 码和尾 1 码，其中字后部的取码方法与单结构字完全相同。在上述取码过程中如果指定的单结构字、字前部或字后部范围内码数不够，则取完为止。例如，单结构字：山（js）、我（pgd）、重（krh）；多结构字：香（pmr）、港（sccz）、龙（pceeh）。单字最大码长为五个英文字母。

多结构字前－后部的切分规则如下：

- 不得切断笔画连续的单结构。（意味着切分处是笔画组合的分隔沟[8]。）
- 有一（或多）条横向或纵向跨越字体的分隔沟，且逐块书写不影响整字笔顺的字统一按

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

上(中)下或左(中)右结构处理,取第一条分隔沟为切分线。例如:立(p _ c)、[illegible](tk _ tk)、器(k k _ ddkk)、鸿(s _ goth)、赢(pz _ kyjnod)。

•嵌套结构(含包围结构和框架结构)的字以"框一心"之间的第一个笔顺交界处为切分点。例如:国(j _ htdh)、区(h _ nz)、或(h _ khwd)、乘(pb _ fhon)、巫(b _ nnh)。

5 性能统计与分析

5.1 三级输入法的性能比较

全字笔顺、全字笔组和 2 _ 21 笔组三级编码方案在 GB13000.1 的 20903 个汉字(原 20902 字加上"〇"字)上的主要性能如表 2 所示。

表 2 三级输入法在 GB13000.1 汉字集上的性能统计

	全字笔顺	全字笔组	2 _ 21 笔组
最大码长	48	21	5
平均码长	12.844	5.965	4.161
键选率	5.70%	6.25%	17.98%

位于第一级的两种笔顺输入法是最为简单易学的输入法,技术性能完全相同,在 GB13000.1 字集上最大单字码长为 48 个键元字符,平均码长 12.844,键选率 5.70%。对于日常计算机打字来说,码长实在太大。因此笔顺输入法适用于拼音输入法用户查找少量的不便用拼音输入的汉字,如果这种字比较多,则可以考虑使用位于第二级或第三级的笔组编码输入法。

第二级的全字笔组输入法应用于 GB13000.1 字符集时,最大单字码长为 21 个英文字母,平均码长为 5.965,键选率 6.25%。与第一级的笔顺输入法相比,码长和键选率的综合效果有了较大的改善,适合于大字集检字。但对于日常的快速汉字输入,码长仍嫌太大。第三级的 2 _ 21 笔组输入法较好地解决了这一问题,其代价是键选率有了明显的回升。因此,我们需要检查 2一21 编码方案在常用字和通用字中的表现。

5.2 正易全 2一21 笔组编码输入法在几个标准字集上的性能表现

正易全 2一21 编码方案在《GB13000.1 字符集》、《现代汉语通用字表》和《现代汉语常用字表》三个标准汉字集上的主要性能如表 3 所示。

表 3 正易全 2一21 编码方案在几个标准字集上的性能统计

	GB1300.1 (20903 字)	通用字表 (7000 字)	常用字表 (3500)
最大码长	5	5	5
平均码长	4.161	3.966	3.776
键选率	17.98%	7.99%	5.80%

可见,2一21 编码方案在通用字和常用字中的表现是令人鼓舞的。考虑到人们日常打字一般只涉及常用字,而在常用字范围内,2一21 编码以平均每字不到 4 个英文字母的码长,就可达到每打 17 个字左右(100/5.8 =17.24)才有一个字需要键选的可喜效果。

5.3 正易全 2一21 编码方案在几个标准字集上的字码分布

2一21 方案的重码分布也较为均匀。三个标准汉字集在 2一21 输入码上的汉字分布情况如表 4 所示。

表 4 三个标准汉字集在 2一21 输入码上的汉字分布

4 a GB13000.1 字符集(20903 字)的"汉字:代码"分布

	16∶1	8∶1	7∶1	6∶1	5∶1	4∶1	3∶1	2∶1	1∶1	合 计
组 数	1	2	7	14	46	137	425	2174	14337	17143
总字数	16	16	49	84	230	548	1275	4348	14337	20903

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

4.b 通用字表(7000字)的"汉字:代码"分布

	5:1	4:1	3:1	2:1	1:1	合计
组 数	6	14	46	401	5974	6441
总字数	30	56	138	802	5974	7000

4.c 常用字表(3500字)的"汉字:代码"分布

	5:1	4:1	3:1	2:1	1:1	合计
组 数	2	5	15	150	3125	3297
总字数	10	20	45	300	3125	3500

在GB13000.1的20903个字上(原20902字加上"〇"字),除有一个输入码对应16个重码字(即 ttox:"悫彀榖穀轂觳殼殼毂毂穀穀轂穀轂"),两个码对应8个汉字(ppv:疔疒冂鄙鄙鄙鄙鄙;pzkyd:赢嬴赢赢赢赢赢赢),其余的重码都在7个字以下。也就是说,实质上,用2-21编码在这样的大字集上检索汉字时,即使需要选字,也不必翻页。在7000通用字中键选率最高的情形是一个码对应五个字,共有六组(hcz:进无元远遣;kb:手午叶牛叮;od:久凡夕丸勺;ppv:疔疒冂鄙鄙;ttox:悫彀榖穀轂;yhlpt:维瞿骓睢睢)。在3500个常用字中,最高字码比也是5:1,但只有两组(kb:手午叶牛叮;od:久凡夕丸勺),4:1的也只有5组(chw:厉勤励勒;cx:友尤芯歹;hcz:进无元远;o:几九儿匕;tpkn:壤谅墩壕)。

6 几点重要改进

与一年前的情况[1]相比,正易全输入法的进展主要包括以下几方面:

(1)从总体上来讲,正易全的编码方案已由原来的单种编码发展成一个循序渐进的三级输入法系列,每一级都已实现对GB13000.1字符集的码表和软件处理。此外,其中的笔顺输入法还提供数字(序号)和拼音(音托)两种代码表示。

(2)多笔码元笔组改用一句简单的口诀:"日月金木水火土,心口工之大宝草。",并采用音托的建元表示,进一步提高编码的系统性和规律性。

(3)单结构(原称"基本范围部件")的定义更贴近国家推荐的《笔形查字法》(草案)中有关"单结构"的说明,而且将原来的"独体字/合体字"直接称为"单结构字/多结构字"。此外,按《GB13000.1字符集汉字字序(笔画序)规范》的术语表达,将"笔画结构紧密度"改称"笔画组合紧密度"。

(4)多结构字的前后部划分以字形为主要依据,重视分割沟的作用,上中下和左中右结构的字统一按第一条分隔沟切分,进一步简化操作。

(5)技术性能方面,正易全2-21编码方案在GB13000.1汉字集上的平均码长4.161,键选率17.98%,与一年前报道的平均码长4.315,键选率16.4%[1]相比,总体上有了明显的改良。此外,考虑到改良后的编码方案对某些等价的繁简体对应部首(如"金-钅"、"魚-鱼")给预相同的编码,由此带来的重码现象(例如,银-銀、铜-銅、锋-鋒;鲤-鯉、鱿-魷、鲍-鮑)在纯繁体字或纯简体字的环境中是不会出现的。

(6)正易全码表在GB13000.1字符集的基础上增加了1164个字和字形。其中476个是从香港政府公布的《香港增補字符集》[9]中精选出来的配有粤语拼音的地方常用字[10],另外的688个是港澳台地区通用的字形。考虑到《GB13000.1字符集汉字字序(笔画序)规范》的20902个标准字形中有一小部分在台湾地区和香港等地电脑上常用的"細明體"和"標楷體"字形集中都没有出现,为顾全大局,我们收录了这些字所对应的"標楷體"字形,例如,反(对应标准字形"反",但第一笔为"横"不是"撇"),骨(对应"骨"但上部包围结构中的"拐"朝右不朝左),选用标楷是因为它比较接近内地的标准宋体字形。

(7)顺便说明,虽然为了方便在大字符集上的操作,我们比以前更重视字形的作用,但是,对于字形方法难以处理的极少数问题仍须借助于"字源理据"和汉字部件规范来解决。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

7 结束语

本文介绍和讨论了正易全的“全字笔顺→全字笔组→221 笔组”三级汉字输入法系列，前两级简单灵活，键选率极低，方便大字集检字；第三级则是原正易全输入法的改良版，在多笔笔组的选用、单结构的定义和多结构字的二部划分等方面都作了进一步的简化、系统化和规范化，平均码长和键选率综合效果也得到改善，使之更适合日常快速打字。

我们在全衡（Allbalanced）汉字输入平台[11]上成功实现了支持GB13000.1 大字集加 1164 个地区用字的正易全三级输入法。通过其中的任何一种输入法检索到一个汉字后可以方便地“横向”查找其它输入法代码，例如用 2－21 输入码 plojb 找到“衡”字后（无重码），转至笔顺输入法就可以看到该字的标准全字笔顺码是 ppspzszhshhpdhhs（或数码式 3323525121134112）。也可以通过（带有统配符的）不完整代码找到汉字后，反查其完整代码。这种代码互查功能还同普通话拼音和广东话拼音连为一体，进一步支持汉语的学习与运用。此外，我们还在 Windows XP 上成功实现了正易全 2－21 型汉字输入法 ime 软件。已经上机打字测试过的数据包括《现代汉语通用字表》（简体字），《简化字总表》（繁体与简体），《商务学生字典》（香港，繁体字）和一些繁体字和简体字文章。我们还准备将码表按字频排序后，测试 221 编码方案在中国内地、香港和台湾地区三地实际语料上的分区与综合，繁体字与简体字等多方面的动态性能。同时统计上述各种情况下的“字－码”分布。

然而，本输入法还存在一些不尽人意的地方，例如，严格按规范笔顺编码，虽然对在校中小学生来说，学习掌握并无困难，但对一般社会人群则可能有一定难度。多字词的编码输入也是应尽早实现的。词库拟以《全衡》词典[12]和《现代汉语词典》（2002 年版）为主要参考资料，收五万词条左右，加上现有的单字部分共七万条左右。由于正易全要照顾到国内和国际不同地区的需要，字/词条目可能会比较多，但是一定要控制在八万条以内，以保证较好的系统工作性能和较低的键选率。此外，词库还应该有简体字和繁体字两个版本。具体发展将在适当的时候报道。

致谢　作者的同事张群显博士，苏咏昌博士，谭世宝博士和本系的一些学生在工作上给了不少帮助和支持。

参　考　文　献：

[1] 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，2003，17(3)：59－65.
[2] 国家语言文字工作委员会. 现代汉语通用字笔顺规范[S]. 北京：语文出版社，1997.
[3] 国家语委. GB13000.1 字符集汉字字序（笔画序）规范[S]. 上海：上海教育出版社，2000.
[4] 国家语委. 信息处理用 GB13000.1 字符集汉字部件规范[S]. 北京：国家语委，2000.
[5] 中国文字改革委员会，国家出版局. 汉字统一部首表（草案）[S]. 北京：1983.
[6] 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.
[7] 傅永和. 中文信息处理[M]，广州：广东教育出版社，1999. 20.
[8] 苏培成. 现代汉字学纲要[M]. 北京：北京大学出版社，2001. 75.
[9] 香港特别行政区政府. 香港增补字符集－2001[S]. http://www.info.gov.hk/digital21/chi/hkscs/introduction.html
[10] 香港语言学学会. 粤语拼音字表[M]. 香港：香港语言学学会，2002. 281.
[11] Zhang, X. AllBalanced: A Web-Based Chinese Character Input System to Meet Hong Kong's Needs[A]. Proceedings of ICCPOL 2001[C]. Seoul, Korea, 2001. 333－338.
[12] 张小衡，张群显.《全衡》词典的设计与建设[J]. 中文信息学报，2002，16(3)：58－62.

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c