文章编号: 1003－0077(2003)03－0059－07

正易全：一个动态结构笔组汉字编码输入法*

张小衡

(香港理工大学中文及双语学系 香港)

摘要:“正易全”是一个以“正”、“易”和“全”为基本指导思想的笔组型汉字编码输入法。在“正”方面，采用国际标准汉字集 ISO10646 CJK，并以《GB13000.1 字符集汉字字序(笔画序)规范》和《信息处理用 GB13000.1 字符集汉字部件规范》指导编码；在“易”方面，以单双笔笔组和十来个常用部件为码元，按笔顺和音托等简单原则映射到 26 个英文字母建元上，从而避免了传统的繁复字根—键元对应表；在“全”方面，支持 CJK 中的所有 20902 字符，包括简体字、繁体字、日韩字和偏旁部首等，而且可以在不改变编码方案的前提下进一步扩充字集。正易全的单字最大码长为 5 个字母，平均码长 4.315，键选率 16.4%。该输入法的笔组—键元设计和取码模式是在对整个CJK 字集作了全字编码以后多次试验、统计和优化后确定下来的。

关键词: 计算机应用；中文信息处理；动态结构笔组；字形码；汉字输入

中图分类号: TP391 文献标识码: A

Towards Correctness, Easiness and Completeness: Building a Chinese Character Coding Input Method Based on Dynamic Structured Stroke Groups

ZHANG Xiao-heng

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

Abstract: In Chinese character input, the form-based coding method is an indispensable complement to the Pinyin-based method. The former is preferable in the cases where high-speed input is needed, where a large character set is required, where words of single characters or words missing in normal dictionaries are abundant, and where unfamiliar or rarely-used characters are more frequently used. The present paper introduces ZYQ, a stroke-group-based Chinese character input method whose development has been kept under the guidance of being Correct (in respect to the norms of language education and language application), Easy (in respect to user friendliness and convenience) and Complete (in respect to the Chinese character set available). ZYQ has a key-selection rate of 16.4%, while the maximum and average code lengths are 5 and 4.315 respectively.

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

一、导言

对于普通用户来说，一般汉字输入宜用拼音输入法[1]。但目前在港澳地区能熟练运用拼音的人比较少，普通话和拼音的推广工作还需要相当一段时间。即使熟练的拼音输入法用户

* 收稿日期: 2002－07－15
基金项目: 香港理工大学研究资金资助项目(1－9827).
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.

也时常会遇到需要形码输入法来辅助的情况。例如有的字无规范读音或不知其正确读音。即使是一个知其读音的字，当出现在古文、律诗、人名地名、单字表等难于以词定字的地方时也是不便用拼音输入的。还有不成字的偏旁、首部、部件和日韩特有的汉字等等。加之，社会上有一定数量的先天聋哑患者，要求他们掌握拼音恐怕是强其所难。因此，形码输入法对于一般用户仍有不可忽视的实用价值。

形码输入法成百上千，但普遍存在规范性差和难学易忘的缺点，不适合一般用户使用[2]。现有形码输入法的另一个缺点是或面向繁体字（如仓颉）或面向简体字（如五笔），很少有在开始设计时就两者兼顾或兼顾得较好的，这样不利于在港澳地区和国际上推广。

本文介绍面向一般用户的“正易全”汉字输入法的研制，其基本指导思想是力求做到：

正：正确、合理，遵循语言规范和汉字的结构规律，同语文知识保持一致；

易：易学难忘，使用方便，具有小学语文水平的人稍加提示就能上机打字；

全：支持大而全的字符集，包括繁体字、简体字、日韩字和偏旁部首等。

此外，还要做到汉字输入速度一般可达到和超过传统的纸笔书写速度。正易全的研制过程可分为三个阶段：（1）选择汉字集、笔组集和键元集，并确定键元与笔组的对应关系；（2）为整个汉字集作全字编码，并通过统计分析，优化全字编码的策略；（3）在全字编码的基础上试验和选择最优取码模式，建立总体编码方案。下面逐一介绍与讨论。

二、汉字集、笔组集与键元集

2.1 汉字集的选择

正易全输入法选用的汉字字符集是 ISO 10646 中的汉字部分，即其 CJK 子集。这个国际标准字符集同时又是中国国家标准 GB13000.1 和 Unicode 的汉字集，含中、日、韩汉字字符 20902 个，其中的中国汉字涵括《现代汉语通用字表》的全部，同时又是汉字输入法普遍采用的 GB2312 简体字集和 Big5 繁体字集的母集。可见，CJK 汉字集是很“全”的。同时由于该字集既是国际标准又是国家标准，所以自然“正”。此外，由于标准往往会带来使用上的方便，因而也利于“易”。

2.2 笔组、键元及其对应关系

字集选好后，下一步是确定作为基本编码单位的汉字码元。传统上形码码元一般采用部件（即字根）或笔画，两者各有利弊。

部件编码颗粒度较大且能较好地保护字形结构，故能以较短的码长赢得较低的键选率。但是这种输入法往往难学易忘。其主要原因在于需要用少量的键元来直接表示大量的部件。《信息处理用 GB13000.1 字符集汉字部件规范》[3] 列出的汉字基础部件多达 560 个，而且规定基础部件不得拆分成更小的部件，只能有条件地合并。这样很难减少所需部件的总数。要用五百多个部件码元通过 ASCII 键盘上的二三十个键元来为 CJK 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。

笔画编码虽然简单易学，但由于一码（指一个键元字符）只代表一笔画，编码颗粒度小，为了保证可接受的键选率，单字平均码长必然很大，严重影响汉字输入速度。

张普教授等以单笔和双笔笔组为码元的简便方法[4]给了我们很好的启示。二画笔组不仅可以缩短逐笔编码的码长，又可以方便地映射到字母键上。笔组和键元的对应关系如图 1 所示。冒号左边是键元，右边是它所代表的笔组码元。例如“a：〈一，一〉”指键元 a 表示笔组

〈一, 一〉(即〈横, 横〉)。

单笔笔组:

h: 一(包括提笔),

s: 丨(包括㇙),

p: 丿,

d: 丶(包括㇏),

z: 乛(ㄥ乙 [illegible]…).

双笔笔组:

a:〈一, 一〉	b:〈一, 丨〉	c:〈一, 丿〉	d:〈一, 丶〉	e:〈一, 乛〉
f:〈丨, 一〉	g:〈丨, 丨〉	h:〈丨, 丿〉	i:〈丨, 丶〉	j:〈丨, 乛〉
k:〈丿, 一〉	l:〈丿, 丨〉	m:〈丿, 丿〉	n:〈丿, 丶〉	o:〈丿, 乛〉
p:〈丶, 一〉	q:〈丶, 丨〉	r:〈丶, 丿〉	s:〈丶, 丶〉	t:〈丶, 乛〉
u:〈乛, 一〉	v:〈乛, 丨〉	w:〈乛, 丿〉	x:〈乛, 丶〉	y:〈乛, 乛〉

图 1　键元— 笔组对应关系

在张普教授的方法的基础上我们作了三点改良:

(1) 将横、竖、撇、点、折 五种标准笔画分别按音托映射到其拼音首字母上, 即 h(横)、s(竖)、p(撇)、d(点)、z(折)。这样既方便掌握又有利于将键元集限制在 26 个英文字母之内。

(2) 25 种双笔序列笔组(简称"笔对")和键元字母都是按其自身的先后顺序相互对应的, 即笔对〈一, 一〉、〈一, 丨〉、〈一, 丿〉、…〈乛, 丶〉、〈乛, 乛〉依次对应字母 a, b, c, …, x, y。一个笔对在字母表中的大体位置一目了然, 例如给出笔对〈一, 丿〉和〈乛, 丿〉, 我们马上就可断定与之对应的键元分别位于字母表的前部和后部。用户还能从一个笔对的键元联想到邻近笔对的键元, 例如, 如果知道字母 m 表示笔对〈丿, 丿〉, 就可推出〈丿, 丶〉对应 n。

(3)引入结构信息, 只有结构较为紧密的笔对才可按双笔笔组编码, 例如, 汉字"纠"按笔顺应该拆成笔组串"〈乛乛〉, 一, 〈乛, 丨〉", 而不是"〈乛, 乛〉, 〈一, 乛〉, 丨"等, 前者蕴含了汉字的结构信息, 较为自然规范。把结构信息纳入笔组中还有利于降低键选率。一个笔对的编码合法性是根据它在字中的结构环境动态决定的, 详细情况将在下一节说明。

三、全字编码及其优化

为了取得最佳的取码模式, 我们先给汉字集的每个字符作全字编码, 并在此基础上进行分析和优化。

3.1　全字编码

全字编码指汉字的每一部分都为代码所表示。为了保护汉字的固有结构, 维持部件编码的一致性, 我们在编码规则中使用了"基本范围部件"和"双笔结构紧密度"两个新概念。

3.1.1　基本范围部件

基本范围部件包括以下几类: a. GB 13000.1 部件规范中的基础部件, 例如: 立氵土戈犬; b. 相交或相接在一起的笔画, 例如: 日十土生车; c. 单笔与其它笔画相配相从所形成的结构, 例如: 三八戈太犬; d. 基本范围部件的连接体, 例如: 辛韭早。

基本范围部件的选用遵循"大者优先"的原则。

上述定义的 a 项属国家标准, 是核心部分。然而 GB 基础部件多达 560 个, 且规律性不强, 因而不易掌握。这是增加其它三项定义的原因之一。b 和 c 两项定义直接引用教育部和国家语委等单位联合推荐的《笔形查字法》中关于"单结构"的定义[5], 其优点是内容明确合理, 可操作性强。d 则是上面几项定义的自然延伸。bcd 三项除了本身比较容易掌握之外, 还覆盖了 a 类部件的绝大多数, 大大降低了使用 a 的难度。定义中还规定基本范围部件的选用遵循大者优先的原则, 这就保证了使用的部件是 GB 基础部件或它们的组合, 较好地满足了 GB 部件规范的要求。

基本范围部件是一种非码元部件, 用于限制一个笔组码元的笔画范围, 以保护字形结构。由于基本范围部件不是直接编码单位, 所以不需要分组映射到键元上, 因此部件的数量多寡对

编码的难度影响不大，主要是应该做到规范和易于识别。

3.1.2 双笔结构紧密度

两个笔画的结构紧密度定义为：相交=2＞ 相接=1＞ 相离=0。

3.1.3 全字编码规则

全字编码严格按照国标笔顺[9]进行，首先以待编码汉字的首笔为起点确定首个笔组码元并按照图1的笔组—键元对应关系取得相应的键元代码，接着再以紧接该笔组的下一笔为起点确定下一笔组并取其代码，以此类推，直至字末。一个笔组码元不得跨越其首笔所在的基本范围部件。在这个范围内单个笔组的处理方法是：如果首笔同第二笔的结构紧密度大于或等于第二笔同第三笔的紧密度，则取前两笔的笔组代码；否则，取首笔的单笔代码。

3.2 测试与优化

根据上述方法，我们对CJK字集的所有20902个汉字字符作了全字编码。各键元字母在全字码中及其首码部位上出现的字数如表1所示(按全字码字数降序排列)。例如，在CJK字集中有15509个汉字的全字码含有建元字母h，其中2589个以h为首码。

表1 CJK字集全字码键元使用字数统计

键元	全字码	字首码		键元	字码	字首码
h	15509	2589		w	2594	500
b	14879	4107		t	2488	470
n	10295	2042		x	2413	147
j	9934	1609		z	2380	63
p	9418	3233		u	2310	274
s	9334	1775		v	2292	309
a	7746	842		k	2090	882
d	7688	1774		l	1980	774
o	4965	1337		y	1347	652
r	4130	714		m	701	61
c	4059	698		g	637	28
f	3278	252		i	214	3
e	2665	121		q	187	55

从表1可以看到键元字母在汉字集中的分布很不平衡，全字码字数标准偏差高达4370.46。这不仅妨碍击键操作，还会增加键选率。

字母y、m、g、i、q是用得较少的键元，如果让它们各自再管辖一个笔组，基本上不会产生新的重码，却能有效地平衡键元负担。因此，我们选择了GB13000.1部件规范基础部件表中使用频度最高的五个三画以上的部件让这几个键元附带表示，其对应是q：口，i：日，g：土，m：木，y：草，其中较高频的部件对应较低频的键元，以便进一步降低产生新重码的可能性。

按上述设计修改全字码表后再统计，发现作为字首码出现次数在100以下的字母键元有z和a，键元a的首码出现次数由原来的842降到82低位的主要原因是大量的“日”部件编码由原来的ja改成了i。为了进一步缩短码长，我们让a和z这两个键元分别兼管一个常用的多笔字首部件，即a：扌，z：氵。其中a在全字码中出现的次数较高，对应于字首出现几率最高的扌，以免冲突。接着，我们统计了各个键元作为末码的使用情况。由于三笔以上的字末部件除“灬”之外，其它的出现次数都在500以下。所以我们只选用这个部件，对应k（k的字末码使用次数为2，在全字码中的使用也较少）。

经上述调整优化后，键元的使用均衡度有了可观的改善，全字码字数分布区间由原来的15509(h)~187(q)缩短为10428(h)~1970(l)，标准偏差从4370.46降到1892.63(由于文章篇幅所限，与表一对应的"优化后键元使用情况统计表"从略)。另外由于新增的少量笔组都是三画以上的常用部件，使得平均码长也得到进一步缩短。

四、取码模式的选择

尽管全字码的键选率低达3.68%，但码长仍相当大，例如笔画最多的"龘"字(48画)的全字码"prhieeahprhieeahprhieeah"长达24个字母。所以应该在全字码的基础上寻找一个适当的取码模式，使得按此模式产生的输入码在键选率和码长两方面达到最佳平衡。

4.1 合体字的二部切分

与传统的字根编码输入法相比，笔组取码颗粒度较小，为了保证可接受的码长、键选率以及取码的方便度，在全字码中取码不宜集中在字的头尾部位，而是应该有较离散的分布。汉字可分为独体字和合体字，其中合体字占绝大多数，而且可以切分为部件。我们根据汉字结构把合体字的全字码用"-"分为字前部代码和字后部代码。例如：香：pm-i，港：z-ycduz。

合体字的二部切分以形为主要依据，参考字源理据，力求简便。具体做法是：

◇ 不得切断基本范围部件。例如，章：prh-ib(不切断联体结构"早(ib)")。

◇ 可按横向或纵向划分(不破坏基本范围部件，且逐块书写不影响整字笔顺)的字按"上(中)下"或"左(中)右"结构处理。例如：碧(hgpi-cq)、赢(pz/q/oajnod)。

◇ 参考字源，或按字源类推，例如：幕(莫/巾)、鹕(胡/鸟)。

◇ 上中下结构和左中右结构的字如果上述原则没有规定其切分线，则分别按(上(中下))和(左(中右))结构处理。例如：赢(pz-qoajnod)、修(l-spxmp)。

◇ 嵌套式的字以"框"和"心"的第一个交界处为切分点。如：国(j-hgdh)、乘(pb-fhon)。

4.2 取码模式的试验与选择

将码表中的所有合体字代码作二部切分标注后，我们利用软件作了多种自动取码试验，表2所示是在独体字统一取头二码尾一码时，合体字采用不同取码模式时的性能数据统计。

表2 CJK 汉字集笔组码不同取码模式性能比较

取码模式	2-2	3-2	2-3	2-21	11-21	3-3	2-2-2	11-22	全笔顺	全笔组
最大码长	4	5	5	5	5	6	6	6	48	24
平均码长	3.589	3.978	4.315	4.315	4.315	4.704	4.836	4.836	12.844	6.321
不同码数	11373	12578	16198	17474	17968	17198	18622	19096	19710	20133
键 选 率	45.6%	39.8%	22.5%	16.4%	14.0%	17.7%	10.9%	8.6%	5.7%	3.68%

取码模式表达式中，"-"左边的一组数字表示字前部的取码方式，右边一组数字表示字后部的取码方式。每组数字的第一位表示在相应取码范围内的头部(连续)取几码，第二位(如果存在的话)则表示在相应取码范围内的尾部(连续)取几码。如取码模式"2-21"表示在字前部的部头取二码，接着在字后部部头取二码，部尾取一码。如"衡"字的全字笔顺码(数字式)是3323525121134112，全字笔组码是pl-ojgcdhb，"2-21"型笔组码则是pl-ojb。键选率的计算公式是：(20903－不同字码的个数)/20903。(注：在CJK原20902字中增加了"〇"(零)字。)

由表2最后两列的数据可以算出，全字笔组码的平均码长是全字笔顺码的49.2%，键选率为64.6%。这样的性能是纯笔对编码无法达到的。较低的码长是因为我们选用了8个高频

多笔部件作为笔组码元，更高的字形分辨力来自于笔组中携带的结构信息。

根据实验结果，我们最后选用综合效果较好的 2-21 取码模式：单字最大码长为 5 个键元符号（与香港普遍使用的“仓颉输入法”一样），平均码长 4.315，键选率为 16.4%，三个取码点较好掌握，同时避免了在字末反笔顺取多个代码所带来的不便。独体字取码仍采用 21 方式，即头二码加尾一码，与合体字字后部的取码方法是一样的。

至此，“正易全”输入法的整体编码方案已经形成。

五、软件实现与测试

经过一年多来的努力，我们建立起了一个按笔画数和笔顺排列的 CJK 字集码表，并在全衡汉字输入系统[7] 上成功实现了正义全笔组型汉字输入法。在初期试用中我们逐字输入了《现代汉语通用字表》的全部 7000 汉字，另外还测试了许多繁体字。总的来说，该输入法的工作性能令人满意，一般不需要选字。在软件测试过程中，还发现和更正了少量的代码错误。

我们还把全字笔顺和全字笔组码当作辅助输入法实现，并且同 2—21 输入法相联系，使得通过这三种输入法中的任何一种检索到一个汉字后可以方便地“横向”查找其它两种输入法的代码。例如用 2—21 型输入码 preeh 找到“龙”字后（无重码），转至另外两个输入法就可以看到该字的全笔顺数码是 4143125111515111，全笔组码是 prhieeah。用户还可以在输入码中使用方便击键的统配符“;”（等价于通常用的“*”）和“.”（相当于“?”）。

六、结束语

“正易全”是一个以“正”、“易”和“全”为基本指导思想建立起来的笔组型汉字输入法。在“正”方面，正易全遵循语言规范，采用作为国家和国际标准的 ISO10646 CJK 汉字集，并以《GB13000.1 字符集汉字字序（笔画序）规范》和《信息处理用 GB13000.1 字符集汉字部件规范》来规范编码设计。在“易”方面，正易全主要以单双笔笔组为编码元素，并将它们按音托和笔顺的规律与建元对应，避免了传统的繁复字根表。在“全”方面，正易全支持 CJK 中的全部 20902 个字符，包括繁体字、简体字、日韩字和偏旁部首等，而且可以在不改变编码方案的前提下进一步扩大字符集，直至全汉字集。正易全最大码长为 5 个键元字符，CJK 字符集平均每字码长 4.315，键选率 16.4%。这些性能是在未使用词语编码、字词频排序、简码处理和智能消歧等技术的前提下实现的。

正易全使用的基本范围部件是一种易于识别的非码元部件，用于规范一个笔组码元的笔画范围，以保护字形结构。这样既利用了传统部件的结构特征，又摆脱了键位的束缚，可以用最合理最自然的方式选用部件，数量不限。

该输入法的另一个特点是所使用的取码模式是在对整个 CJK 大字集作了全字编码以后多次测试、统计、择优而产生的。同时，由于有全字码为依托，可以根据需要灵活产生相应的输入法，例如可以用 11-22 取码模式甚至全字码来满足更大字集和更低键选率的要求。

然而，由于汉字是一个历史悠久，内容丰富，字形复杂多变的文字系统。要达到理想的正易全，还有许多问题要解决。在理论支持方面，首先是目前还没有一个关于合体字结构切分的国家规范，例如“赢”字是按字源分为“贝”和余下部分，还是依字形简单处理为上中下结构？此外《信息处理用 GB13000.1 字符集汉字部件规范》同一般语文教育字典存在一些分歧。关于使用该规范遇到的困难和体会比较多，请容另文讨论。

两年多前发布的 GB18030—2000 是 GB13000.1 汉字集的母集，因此正易全的输入法码表

也应作相应的扩充。等与新标准字集相配套的笔顺和部件规范公布后我们即着手码表的扩充工作。此外,今后除了进一步优化和规范单字输入法之外,还要发展词输入和智能处理。

参 考 文 献:

[1] 张普. 一般人使用什么输入法. 汉字编码键盘输入文集. 北京: 中国标准出版社, 1997. 12—14.

[2] 许嘉璐. 语言文字学及其应用研究. 广州: 广东教育出版社, 1999, 132.

[3] 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. 北京: 国家语委, 1997.

[4] 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, 广西科学技术出版社, 1994.(见 3.6 节: PJS/TLS 中文汉字输入系统)

[5] 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, 164.

[6] 国家语委. GB13000.1 字符集汉字字序(笔画序)规范. 上海: 上海教育出版社, 2000.

[7] Zhang, X. AllBalanced: A Web—Based Chinese Character Input System to Meet Hong Kong's Needs. Proceedings of ICCPOL 2001. Seoul, Korea, 2001, 333—338.

(上接第 **20** 页)

五、实验结果和结论

本系统采用的基于格助词和接续特征的藏文分词方法是一种利用藏文字、词、句各类接续特征的分词方法,其主要目的在于提高自动分词的精度。通过对各类语料的对比测试表明,系统切分精度均达到了 96%以上。同时,实验结果还表明,本系统对不同领域的藏文语料表现出较强的适应性,因而说明了系统具有较强的通用性。

藏文自动分词是藏语信息处理中的基础性课题,由于自然语言固有的复杂性,加之本系统主要又是基于规则的系统,对藏语句子中的诸如堆块、词截断等错误的处理能力方面还有待改进。在此基础上,积极进一步引入统计技术,尝试规则与统计相结合的分词技术研究进而开展藏文分词、词性标注和句子分析的一体化分析技术研究,是我们下一步的努力方向。

参 考 文 献:

[1] 陈玉忠, 李保利, 俞士汶, 兰措吉. 基于格助词和接续特征的书面藏文分词方案. 语言文字应用, 2003(1).

[2] 才旦夏茸, 藏文文法详解, 西宁: 青海民族出版社, 1988.

[3] 朱德熙. 语法讲义. 北京: 商务印书馆, 1999.

[4] 刘挺, 吴岩, 王开铸. 串频统计和词形匹配相结合的汉语自动分词系统. 中文信息学报, 1998(1).

[5] 陈小荷. 自动分词中未登录词问题的一揽子解决方案. 语言文字应用, 1999(3).