中 文 信 息 学 报
第9卷 第3期 JOURNAL OF CHINESE INFIOMATION PROCESSING Vol.9 No.3

部件组合--潜在的汉字结构层次

韩布新

(中国科学院心理研究所)

【摘要】本文提出了汉字结构中的一个潜在层次--部件组合，并对其在汉字编码字符集（基本集）中的分布特征进行了统计分析，发现绝大多数组合的组字次数和频率都很低，高频组合很少。文中列出了低频组合中的60个高频部件，以供汉字编码输入参考。最后讨论了部件组合在汉语教学及认知心理学研究等方面的应用意义。

一、问题的提出

在已公布的500多种汉字编码输入方案中，形码方案占绝大多数。"形码"之形，指笔画、部件等字形特征，但有一个重要的结构层次尚未被注意，那就是介于整字和部件之间的一个层次--部件组合(Combination of Chinese character constituents,CCCC)。"部件组合"是指汉字中同时存在的一组部件，比如"部"字有三个部件，可形成三个部件组合，即"立--口"、"立--阝"和"口--阝"。提出"部件组合"这一概念是基于以下三点理由：①它是不同于部件和整字的一个中间结构层次；②它可能是一组汉字的共同特征，具有区别价值，因此在人类汉字识别中可能有重要作用；③它在汉字计算机输入编码、汉字学和汉字教学等方面有应用意义。

英文字词识别研究表明熟悉的字母串是字词知觉的基本单位之一，识别字母串的概念是其使用频率的单调递增函数[1]。按照Feigenbaum & Simon(1984)的初级知觉和记忆模型，熟悉单元是最大的刺激成分，而频率则是熟悉性的一个良好的指标[2]。汉字中与字母串相当的结构层次是部件组合，它的作用又如何呢？部件和部件组合作为一个整体反复认读的经验积累，使得它们在整字识别时表现出一种整体性，可以成为信息加工时的基本表征。由于汉字独特的二维平面特征，其整体和局部的结构联系可能要比英文更加紧密。

因此，对于汉字集合中部件组合的统计分析，将可为汉语信息的计算机处理（编码和自动识别）提供一些有益的启示[3]，并为开展有关人类汉语言认知机制的心理学实验研究准备必要的数据资料。本文报告了对"信息交换用汉字编码字符集--基本集"(GB2312-80)中6724个汉字部件组合特征的统计研究结果，并讨论了该结果的应用意

①本文1994年8月22日收到

义.

二、统计方法

根据《汉字信息字典》对部件的界定，GB2312-80 的 6724 个汉字中可拆分出 567 个部件，包括独体字、部首和最小不可切分单位，比如"木"、"氵"、"⺾"等[4]。整字频率数据引用 748 工程的统计结果[5]。"部件组合组字次数"是指 GB2312-80 中含有某部件组合的汉字个数；"部件组合频率"是指在 748 工程所统计的各类书面材料中含有某部件的汉字个数与总字数（2165 万）的比率，实际是 GB2312-80 中所有含该部件组合汉字的字频之和。如 GB2312-80 中含有"隹"和"页"的只有一个"颧"字，所以该部件组合的组字次数为 1，组合频率就是"颧"的字频 0.0039‰。

以上海交通大学计算机科学系建立的"汉字属性信息数据库"为基础[6]，用 FoxBASE 的有关命令编程运算。从部件集合中随机选出两个部件分别作为部件 1 和部件 2，再在"汉字属性库"中检索出含有部件 1 的所有字形成一个新的数据库，新库的平均容量是前者的百分之一左右，再用部件 2 在新库中与各记录进行字符串比较运算。两个部件分别判定的算法较之同时判定某字中是否存在两个部件的算法大大提高了运算速度。另外利用微机的扩展内存（虚拟盘）工作以避免硬盘读写操作和提高运算的速度，有效地提高了统计效率。

三、统计结果

1、部件组合在 GB2312-80 中的分布特征

本项统计生成了"汉字部件组合信息数据库"。该库含有每个部件组合的"组字次数"和"部件组合频率"信息。统计结果表明，在 GB2312-80 的 6763 个汉字中实际存在的两部件组合有 7583 个，部件组合组字次数和组合频率有显著的正相关，相关系数为 0.44,$P<0.001$。表 1 列出了出现次数最多和最少的部件组合各 15 个，以及它们的有关各项数据。

表 1　部件组合有关参数举例

部件1	部件2	组字次数	组合频率(‰)	部件1	部件2	组字次数	组合频率(‰)
口	一	101	13.10	隹	食	1	0.00
口	冖	80	9.21	隹	矢	1	0.00
口	十	78	8.19	礻	末	1	0.01
口	艹	62	1.99	隹	牙	1	0.04
口	冂	50	7.19	隹	页	1	0.00
口	木	50	2.24	礻	今	1	0.00
口	丁	49	5.30	礻	犬	1	0.00
曰	一	48	7.10	隹	止	1	0.08
口	土	46	2.76	隹	匕	1	0.08
口	月	43	0.95	隹	冂	1	0.01
口	夂	39	5.20	隹	氵	1	0.49
口	扌	39	1.82	隹	亻	1	0.00
一	人	39	7.31	隹	犭	1	0.00
口	亻	38	2.81	隹	彐	1	0.01
艹	曰	38	1.68	礻	手	1	0.00

两部件组合的频率分布情况如图 1 和图 2 所示。从图 1 可以看出，GB2312-80 中组字次数越高的部件组合越少，而组字次数仅为一次的部件组合却有 4868 个。部件组合在书面统计语料中的分布情况与此类似，如图 2 所示，即高频组合很少，绝大多数为低频组合。

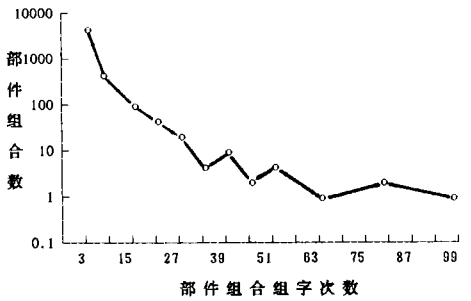

图1 部件组合组字次数分布图

说明："组字次数"是指GB2312—80中含有某部件组合的汉字个数 .

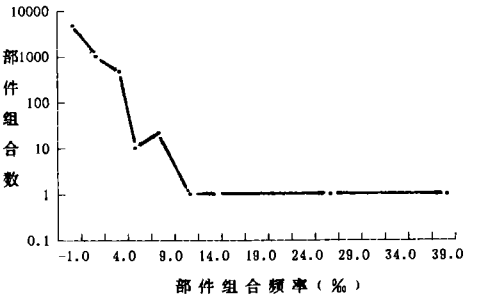

图2 部件组合频率分布图

说明："频率"是指在748工程所统计的各类书面材料中含有某部件的汉字个数与总字数的比率.

2、低频部件组合中的高频部件

在 7583 个由两个部件形成的组合中有 4868 部件组合的组字次数为 1，也就是说 6763 个汉字中只有一个字含有这种组合。用这样的部件来编码当然就不会有重码。如果将这些低频部件组合中的高频部件挑选出来，则一方面可以保证单个部件码元的覆盖面尽量大，一方面又可以有效地减小部件编码的重码率。表 2 列出了前 60 个这种部件的组字次数。

表 2 低频部件组合中的前 60 个高频部件的组字次数

部件	组字1	组字2	部件	组字1	组字2	部件	组字1	组字2
氵	397	161	足	87	71	礻	51	36
扌	296	144	山	130	68	广	91	34
艹	419	134	疒	102	68	酉	66	34
饣	220	134	目	132	66	夂	112	32
木	421	133	曰	414	61	牛	40	31
虫	182	117	米	97	56	人	196	31
亻	284	116	贝	139	56	尸	109	29
口	1056	116	鱼	80	54	门	68	29
女	207	109	讠	198	50	又	182	28
讠	158	107	车	85	48	十	246	27
纟	149	100	心	130	47	攵	108	27
竹	114	96	犭	78	46	冖	97	27
石	114	89	一	285	46	亠	160	27
月	297	89	八	224	46	口	93	27
土	255	88	刂	110	45	力	112	25
阝	159	83	马	65	43	隹	110	25
忄	129	77	禾	107	43	匕	88	24
火	129	77	鸟	63	39	钅	39	24
辶	136	76	大	148	38	穴	43	24
王	132	72	田	132	37	兀	46	24

说明：① "组字1" 为该部件在GB2312—80中的组字次数；
"组字2" 为该部件在4868个低频组合中的组字次数；
② 本表部件根据组字2降序排列。

四、讨论

汉字部件组合的频度特征可以在下列三个方面得到应用。

1、人类识别和阅读汉字

部件组合作为一种新的变量，在人类汉字识别中有一定的作用。特别是在研究整体和部分的关系时，部件组合有承上启下的重要作用。因为同一个部件组合在不同的汉字中既可能是整体特征（比如在两部件合体汉字"另"中的"口-力"），也可能是局部特征（比如在三部件合体汉字"别"中的"口-力"）。认知心理学的研究表明，部件组合频率在汉字识别中有易化和干扰两方面的作用，位于汉字左边或上边的部件组合频率越高，则汉字识别的反应时越短；而位于右边或下边的部件组合频率越高，则识别反应时越长[7]。这从结构层次认知方面解释了彭瑞祥等（1983）发现的汉字的左上象限对识别最重要的现象[8]。

在识别汉字中，出现次数不同的部件组合显然具有不同的信息量，其区辨价值也有所不同。这里以本文开头所提到的"部"字中的三个部件组合为例做一简单说明。三个部件组合中，"立-口"的组字次数为 20（含有该组合的有'赔、倍、剖'等 20 个汉字），组合频率为 4.31‰；"立-阝"的组字次数为 2（含有该组合的只有'部、陪'二字），组合频率为 3.12‰；"口-阝"的组字次数为 24（含有该组合的有'邵、郡、阿、陨'等 24 字），组合频率为 3.85‰。显然识别时以"立-阝"为单位最不容易混淆，而以"立-口、口-阝"为识别单位则需要做出许多排除工作才能找到目标汉字。如此说来，似乎是部件组字次数越高，则识别时越困难。

这个推论同前面所叙述人类识别汉字的实验结果并不矛盾。因为首先部件的"组字次数"和"部件频率"是不同的概念，前者是一个静态的指标，和单个汉字的使用情况无关；而后者则是一个动态的指标，直接受含有该组合的各个汉字之使用频度的影响。显然这两种指标在人类和计算机识别汉字中的作用有所不同。虽然都是要达到唯一地确认某个汉字，但机器识别时往往只是根据形态特征的区别来实现这一目的，部件组字次数可以满足这一要求；人类识别时则常常利用语境（在上下文中的字义特征）、字的使用频度（对字的熟悉度）等等特征，并结合字型特征进行综合判断。那么人类识别汉字时的这种"自上而下"和"自下而上"相结合的高速判断能力，是否可以应用到机器识别之中？象人类那样利用语境信息是不是消除机器识别时那 1%以下的错误率的有效办法？这些都需要进一步研究。

2、汉字信息的计算机输入

在选择部件作为小键盘上配置的码元时，以词为单位的汉字计算机输入方法不仅可以使输入速度大为提高，同时也可以降低文本录入的重码率。部件组合次数可作为简码字或简码词输入的定量测量指标，作为优化编码方案的依据。这么说有两方面的理由：一是部件组合次数表示了该组合在多少个汉字中存在，而简码字或简码词输入是将频度最高的汉字或词用一个字母键加空格键或一个选字数字键来选择上屏。部件组合次数低可以保证较低的重码率，而含有这种低频部件组合的高频汉字就完全可以考虑用简码输入了；二是根据部件组合次数可以合理地安排汉字输入时小键盘上的码元配置，使连续输入编码时双手合理分担敲键工作量，避免同一个字的码元一直用同一侧手输入。比如在汉字集合中存在

的部件组合，其两个组成部件均作为码元时便不宜安排在同一侧键位上。已有研究表明[9]，单字编码码元用双手轮流输入的速度和效率均比只用单侧手输入要高。

另一方面，部件组合频率和次数作为汉字使用频度的相关指标，可为高频先见的智能处理提供参考。同时，如前文表 2 所示，将低频部件组合中的高频部件选出作为形码输入的首选码元，既可以扩大编码的覆盖面，因为高频部件在汉字集合中和汉语书面材料中使用会有较多；也可以在一定程度上减少汉字输入的重码率，因为这些部件组合只在很少一些字中存在。

从表 2 可以看出，在低频组合中抽取的 60 个高频部件基本上都在《新华字典》的 189 部首之内。这说明了两方面的问题：一是以部首作为汉字输入的码元有很强的可行性；一是还有一些部首是不宜作为码元使用的，因为它们与其它部件所形成的组合在汉字集合的很多字中都有，如果作为码元使用，将不可避免地导致输入文本重码率高。比如象"寸、刀、户、巾、小、皿、厂、冂、氵、子、礻、页、立、耳、欠"等部件，就不宜选作部件。

就是在表 2 列出的 60 个部件中，也有一些值得考虑的问题。以部件"口"为例。它在 GB2312-80 中的组字次数为 1056，利用它确可对基本汉字集合中将近 1／6 汉字进行编码。但是，这个数字同时也表明它可以与大量其它部件形成两部件组合，其中有不少是高频组合。因此，"口"这一部件作为编码键元是有一定局限性的[7]。类似这样的部件在汉字部件集合中有多少？每个部件可能与其它部件形成组合的规律如何？这是需要进一步分析的问题。

3、汉字学研究、应用和教学

"计量汉字学"、"工程汉字学"等新的边缘学科正在逐渐形成[10]，并在中文计算机、汉字信息处理、自然语言的阅读和理解等研究领域发挥越来越重要的作用。对汉字部件组合规律的探讨，必将在丰富汉字学研究成果的同时，促进这种作用的进一步发挥。

汉字教学中如加入有关高频部件组合的知识，结合汉字的结构组合特征（如上下、左右和包围等）进行教学，并与小学计算机普及教育活动相结合，则既可以加强学生对汉字形体结构的理解，促进汉字学习和记忆[11]，又可以促进青少年儿童对中文信息处理技术的掌握和应用，具有深远的意义。

参考文献

[1] Solso R L, King J F. Frequency and versatility of letters in the English language. Behavoiral Research Methods and Instrumentation, 1976,8(3): 283-286.

[2] 石绍华，国外文字认知研究回顾，心理学动态，1991，(1)：46-54.

[3] 韩布新，陈一凡，汉字认知心理研究对机器自动识别汉字的启示，中文信息学报，1993,7(4): 60-66.

[4] 上海交大计算机科学系编制，《汉字属性信息库》使用手册，上海科技文献出版社，1988.

[5] 上海交大等单位编．汉字信息字典，北京：科学出版社，1988:999-1060.

[6] 郑林曦，高景成，汉字频度表（按字音查），北京新华印刷厂，1980.

[7] 韩布新，汉字识别中部件和部件组合的频率效应，中科院心理所博士学位论文，1993 年 7 月.

[8] 彭瑞祥，喻柏林，不同结构汉字的再认研究．见：普通心理学与实验心理学论文集，甘肃人民出版社，

1983:182–195.

[9] 陈一凡等．键盘相关速度当量的研究．中文信息学报．1990,4(4):12–18.

[10] 林川等．汉字的工程技术和工程汉字学刍议．中国印刷．1993,39:35–40

[11] 刘鸣．汉字字形心理学研究述评．心理学动态．1993,1(1):1–9.

Combination of Chinese character constituents–– A latent structural unit

Han Buxin

(Institute of Psychology, Chinese Academy of Sciences)

Abstract

A latent structural unit of Chinese character––combination of Chinese character constituents (CCCC)–– was introduced. It was found that CCCC had an uneven distribution in the Basic Chinese Character Set (GB2312–80). It were also discussed about the cognitive effects of CCCC on the human cognition psychology, and their application in Chinese information processing and education.

(本文是作者在荆其诚教授和林仲贤教授指导下完成的博士学位论文的一部分．承蒙何厚存、李公宜两位教授允许使用《汉字属性信息库》，并得到张侃研究员、陈一凡教授和张普教授的帮助．陈一凡教授阅读全文并提宝贵意见．谨此一并致以谢忱！)