第26卷 第2期 心 理 学 报 Vol.26 No.2
1994年5月 ACTA PSYCHOLOGICA SINICA May, 1994

汉字部件信息数据库的建立[1)]

——部件和部件组合频率的统计分析

韩 布 新
(中国科学院心理研究所,北京,100012)

摘 要

用FoxBASE语言统计了6763个基本汉字集合中的部件和部件组合的频度信息,建立了"部件数据库"和"部件组合数据库"。前者包含567个部件;后者包含汉字中实际存在的7583个两部件组合。统计结果表明部件和部件组合均呈偏态分布,绝大多数的频率很低。这两个数据库不仅可应用于研究汉字认知中整体和局部的关系、汉字的学习和记忆,而且也可供汉字学的定量研究、中文信息的计算机处理研究参考。

关键词 部件,部件组合,组字次数.频率,数据库.

1 前 言

认知心理学认为字词频率是反映人类对特定字词熟悉性的指标[1],是字词识别中最重要的变量之一。E. L. Thorndike在1944年开始进行英文字词频率的统计研究,其后陆续有人分别统计了不同年级和不同学科阅读材料的分类字频[2]、字母频率和字母位置频率[3]。这些频率在字词识别中的作用已被大量的心理学实验研究证实[1,4,5]。关于汉字频率的统计研究不少*,但有关汉字部件频率的统计研究却并不多见。大量实验研究业已证实,汉字不同结构层次的频率信息在汉字识别中有重要作用[6]。

部件频率研究始自《汉字信息字典》[7],但该字典选字范围较大,包括7785个正体字,3469个繁体、异体、别体字,因此含有很大一部分不常用的信息。在大规模字频统计基础上确定的"信息交换用汉字编码字符集——基本集(国家标准号为GB2312-80)"[8]包括6763个汉字,其外的字出现机会极少,在汉字识别中的认知意义不大,因此有必要以GB2312-80为基准,重新计算部件频率数据。本文旨在结合汉字认知心理学和汉字信息处理的有关研究,进一步统计分析部件和部件组合的频率分布情况。"汉字属性信息数据库"[9]包含了GB2312-80所有汉字的各种信息,为我们的统计工作提供了良好的基础,但其中没有部件频率数据。关于部件组合频率未见研究报道。

2 部件频率的统计分析

2.1 统计方法和有关概念的界定

1) 本文于1993年11月19日收到。

*作者注:韩布新。汉字识别中部件和部件组合的频率效应。中国科学院心理研究所博士学位论文,1993年7月,3-5。

根据《汉字信息字典》对部件的界定，GB2312-80 中可拆分出 567 个部件，包括独体字、部首和最小可切分单位，比如“木”、“氵”、“艹”、等。“结构表达式”是指将单个汉字按笔顺拆分后形成的一维部件序列，比如“数”的结构表达式为“(米/女)∥攵”。以某一个部件作为字符串 1，汉字的结构表达式作为字符串 2，用 FoxBASE+(汉化版本 2.0) 的子字符串运算符“$”进行两个字符串的包含比较，若结构表达式中包含有目标部件，则运算结果为真，该部件的组字次数赋值为 1(参见后面举例)，再以单字频率加权计算部件频率。整字频率数据引用 748 工程的统计结果[10]。“部件组字次数”是指 GB2312-80 中含有某个部件的汉字个数；“部件频率”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 中所有含该部件汉字的字频之和。比如，部件“丈”的组字次数为 3，部件频率为 0.0771‰，因为 GB2312-80 中含有“丈”的字是“丈、仗、杖”，其字频分别为 0.0151‰、0.0591‰、0.0029‰。

2.2 部件频率的统计结果

本项统计生成了“汉字部件数据库”，包含 567 个部件的“部件组字次数”和“部件频率”数据。统计表明，部件组字次数的范围是 1—1056 次；部件频率的范围是 0-111.50‰。两者的相关系数为 0.84，$p<0.001$。表 1 列出了按部件组字次数高低排列的前 14 个高频部件。

表 1 14个高频部件的统计结果示例

部件	组字次数	部件频率(‰)	部件	组字次数	部件频率(‰)
口	1056	111.50	一	285	60.97
木	421	35.28	亻	284	44.17
艹	419	13.11	土	255	37.65
日	414	45.16	十	246	23.48
氵	397	27.62	八	224	27.55
月	297	28.96	饣	220	5.42
扌	296	25.55	女	207	16.50

GB2312-80 中拆分出的 567 个部件的分布情况如图 1 表示。从图中可以看出，部

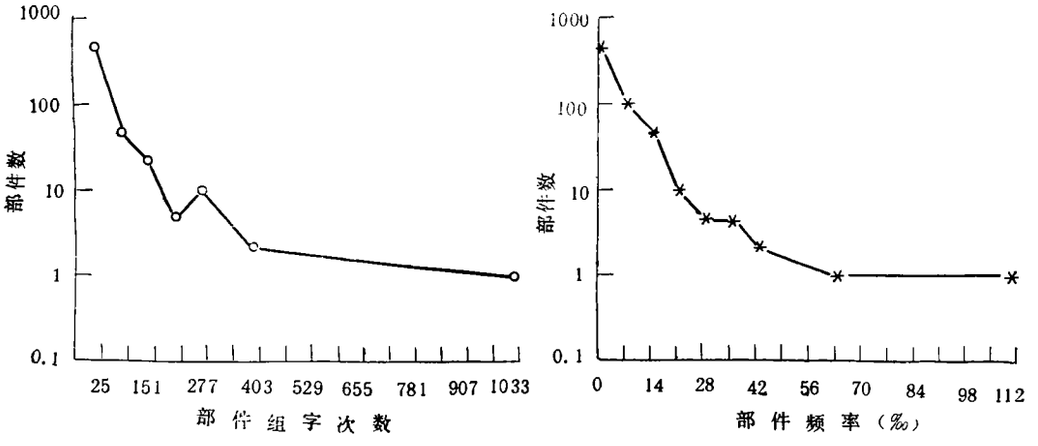

图 1 部件组字次数和部件频率的分布图

件组字次数或部件频率越高，则该水平下的部件数目越少，反之则越多。部件频率不同水平下部件的分布数目相差很大，高频部件很少，绝大多数为低频部件。这是一种不均匀的偏态分布，绝大多数部件都位于曲线的低频端。

图 1 表明部件组字次数和部件频率有类似变化趋势，这与有关英文字母组字次数和字母频率变化趋势的研究结果一致[8]，但某个具体部件在部件组字次数和部件频率高低序列中的前后位置可能有些差别。比如表 1 中部件"钅"的组字次数为 220，比"女"要高，但由于含有"钅"的汉字多为金属名称用字，其整字频率不高，所以"钅"的部件频率反而相对较"女"的为低。

部件频率的偏态分布给实验中控制部件变量带来了一定的困难，因为每个部件所代表的汉字数目相差很大，各等距分组的部件频率水平内可选汉字的多少也大不一样。但低频部件对应某个特定汉字的代表性比较强，也为我们选定汉字计算机输入的码元提供了契机。

3　部件组合频率的统计研究

很多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(词或词的一部分)的概率是其使用频率的单调递增函数[11]。按照 Feigenbaum & Simon 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[12]。

汉字结构中与字母串相当的层次是部件组合，由汉字中同时存在的两个部件组成。多个单词共有的字母组合在单词识别中有促进作用，汉字部件组合的作用又如何呢？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。另外，部件组合作为一组不同汉字的共同特征，有类似部件的熟悉性作用，但它在汉字识别中的作用也可能比部件的大，因为部件组合的结构层次更高，有许多的两部件组合本身就可以形成单字。要定量地研究这一问题，首先必须对部件组合的频率特征进行系统的统计研究，以期为开展有关的实验研究准备必要的数据资料，同时也希望能为汉字信息的计算机处理(编码、自动识别)提供一些有益的启示。

3.1　概念的界定和统计方法

"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"。"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 GB2312-80 中所有含有某部件组合汉字的字频总和。利用 FoxBASE 的有关命令编程，结合"汉字属性信息数据库"和"部件频率数据库"的有关字段进行字符串比较运算来进行统计。如 GB2312-80 中含有"隹"和"页"的只有一个"颥"字，所以该部件组合的组字次数为 1，组合频率就是"颥"的字频 0.0039‰。

3.2　统计结果

本项统计生成了"汉字部件组合信息数据库"，在 GB2312-80 中实际存在的两部件组合有 7583 个。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，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 2

列出了出现次数最多和最少的30个部件组合的有关数据。

表2 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	亠	80	9.21	隹	矢	1	0.00
口	十	78	6.19	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
日	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	讠	1	0.49
口	扌	39	1.82	隹	彳	1	0.00
一	人	39	7.31	隹	犭	1	0.00
口	亻	38	2.81	隹	彐	1	0.01
艹	日	38	1.68	礻	手	1	0.00

两部件组合频率的分布情况如图2所示。图2表明，部件组合在GB2312-80中的分布情况与部件的类似，即高频组合很少，绝大多数为低频组合，有4868个组合的组字次数仅为一次。

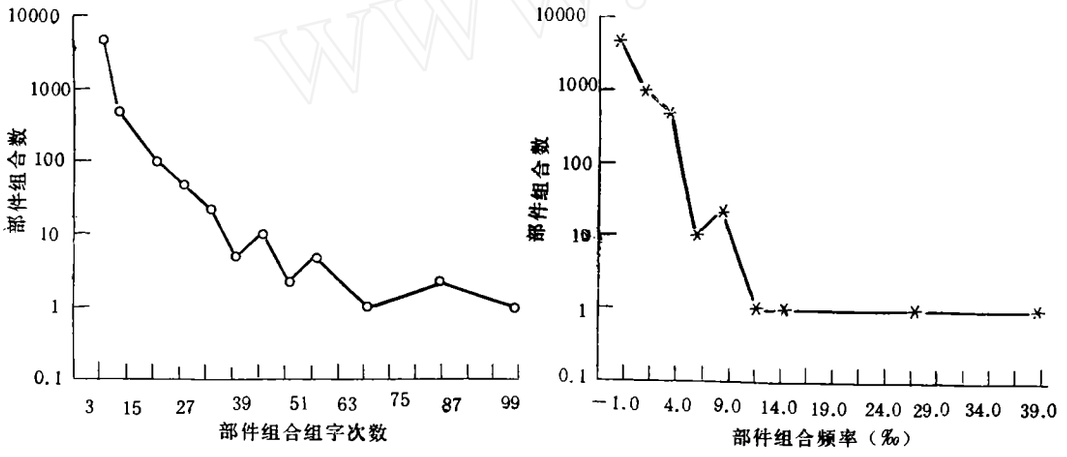

图2 部件组合频率和组字次数分布图

4 讨 论

4.1 优选算法以提高统计效率

统计部件频度时，由于对每个汉字只需进行一次字符串搜索比较运算，速度比较快，在主频10MHz的IBMPC286上运算四小时即可完成。统计部件组合频率时，则要进行多步运算，比较费机时。以两部件组合为例，因不知道在GB2312-80中实际存在多少两部

件组合，因此必须首先判断某个组合的存在与否；其次，两部件组合在同一个汉字中必须同时存在，因此对每一字要完成两次判定——甲、乙两部件各一次。为此可采用“bjA ＄某汉字. and. bjB ＄某汉字”的算法，但实践证明这种算法是很不经济的。原因有两点：①必须将全部 6763 个汉字都核算一遍才能知道某部件是否存在及有多少，而实际上大多数部件只存在于不到 100 个汉字之中。②在单一命令语句中用逻辑与方式同时完成两种判定运算，既使操作复杂化，又增加了机时，所以采用了下面的算法。从“汉字属性信息数据库”中考出含有部件 1 的所有字形成一个新的数据库，新库的容量是前者的百分之一左右，再用部件 2 与新库中各记录比较运算。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)，既避免了频繁的硬盘读写操作，又可以提高 10 倍左右的速度，有效地提高了运算的效率。其它一些编写命令文件的细节不再赘述。

4.2　汉字部件信息数据库的应用前景

数据库技术在汉字识别研究中有广泛的应用前景。本研究建立的“部件频率数据库”和“部件组合频率数据库”不仅为汉字识别的心理学实验提供了良好的刺激字选择工具，也为对汉字形体结构特征的定量研究提供了行之有效的操作手段。在研究整体和局部关系时将部件频率作为变量来考虑是必要的。

计算机汉字输入编码的码元可以是字音、字形，或音形结合。这几类编码方案均取得了巨大的成功。在目前已完成的四百余种编码方案中，形码方案占了大多数*。汉字输入编码若以部件为码元，当然要研究自然状态下部件组合的规律，以使编码方案更合乎实际。从部件层次分析汉字形体结构规律，符合汉字造字法的基本原理。在此基础上形成的汉字部件编码输入方案有很强的科学性，也便于学习、掌握和运用[13]。但以部件或笔画为码元的汉字计算机输入编码方案存在码元多(按不同的拆法或不同的字集范围，部件数目从一百多至一千多不等)、重码率高，为消除重码就必须增加码长(即增加击键次数)等困难。如果将低频组合中的高频部件找出作为首选码元，结合汉字编码输入软件的智能化处理(高频先见、自学习等)技术，将有助于减少计算机汉字编码输入时的重码率高的困难，并缩短码长。汉字的机器自动识别如果以部件组合为搜索匹配的主线索，则可以大量减少搜索范围。另外，研究人类如何利用部件频率或部件组合频率等局部信息来识别整字的认知策略，或可为汉字机器识别提供一些方法学上的启发。

汉字系统的形、音规律是汉字教学的重点[14,15]。“集中识字”教学法就是集中学习一组具有共同结构成分的汉字，以加强区别效果。该方法经过十多年的不断应用、改进、研究，已取得了巨大的社会效益，培养了几代青少年。对于汉字形体结构规律的深入研究，必将推动汉字教学的发展，在汉字学研究、汉字改革等方面，也有一定的理论和应用意义。

致谢　本文是作者在荆其诚教授、林仲贤教授指导下所完成博士学位论文的一部分。谨向两位导师表示衷心的感谢！在建立汉字部件信息数据库的过程中，承蒙何厚存、李公宜两位教授允许使用“汉字属性信息数据库”，并得到张侃副教授、陈一凡教授和张普教授的帮助，在此一并致以谢忱！

*　作者注：周志农，李锋。《自然码》使用手册。北京超想电脑技术开发公司，1992 年 4 月，105。

参 考 文 献

1 Solso R L, et al. Cognitive Psychology. New York: Harcourt Brace Jovanovich, Inc., 1979, 97-102.
2 Kucera H, Francis W N. Computational analysis of present-day American English. Providence: Brown University Press, 1967.
3 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.
4 Mason M. Reading ability and letter search time: effects of orthographic struture defi-nded by single-letter positional frequency. Journal of Experimental Psychology: General, 1975, 104(2): 146-166.
5 Massaro D W. Orthographic regularity, positional frequency, and visual processing of letter strings. Journal of Experimental Psychology: General, 1979, 108(1): 107-124.
6 彭聃龄. 阅读的认知心理学研究. 北京师范大学学报,1989,(5): 75-84.
7 上海交大等单位编. 汉字信息字典. 北京: 科学出版社,1988,999-1060.
8 华北计算技术研究所起草. 信息交换用汉字编码字符集-基本集(GB2312-80). 国家标准总局发布, 1981 年 5 月实施.
9 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.
10 郑林曦,高景成. 汉字频度表(按字音查). 北京新华印刷厂,1980.
11 Morton J. Interaction of information in word recognition. Psychological Review, 1969, 76(2): 165-178.
12 石绍华. 国外文字认知研究回顾. 心理学动态,1991,(1): 46-54.
13 张普. 汉字部件分析的方法和理论. 人大复印资料(语言文字学),1984,(5):51-56.
14 万云英. 儿童学习汉字的心理特点与教学. 见: 高尚仁,杨中芳主编. 中国本土化心理学论文集, 台北: 远流出版公司,1991,403-448.
15 刘鸣. 汉字字形心理学研究述评. 心理学动态,1993,1(1): 1-9.

DEVELOPMENT OF DATABASE OF CHINESE CONSTITUENTS INFORMATION ——STATISTICAL ANALYSIS OF THE FREQUENCY OF THE CONSTITUENTS AND THEIR COMBINATION

Han Buxin

(Institute of Psychology, Chinese Academy of Sciences, Beijing, 100012)

Abstract

Frequency parameter of Chinese characterconstituent and their combinations in GB2312-80 were computed using FoxBASE techenique. "The Database of character constituents" andThe Database of Character"Constituent Combinations" were produced as the result. The former consisted of 576 character constituents, the later consisted of 7583 2 constituents combinations. Every constituent or combination had 2 attributions, one was the number of Chinese characer combining by character constituents or their combination, another was the frequency. The character constituents or their combination, another was the frequency. The character constituents and their combinations had similar pattern of uneven distribution, and most of them had low frequences. These 2 databases could be applied in the experimental research of Chinese cognition, learning and memory, it also could be used in the qualitative analysis of Chinese character and computer processing of Chinese information.

Key words Chinese character constituent, combination of Chinese character constituents, number of Chinese characters combining by character constituents or their combinations, frequency, database