·检索工具·

《汉字属性信息数据库》在汉字识别研究中的应用*

韩布新

中国科学院心理研究所（北京100012）

关键词：汉字属性　汉字识别　部件频率　数据库

〔提要〕本文简要介绍了《汉字属性信息数据库》，并结合汉字识别研究，初步分析了其中的笔画数、部件数、结构类型、字频等字段数据，重点就有关部件属性特征进一步分析了结构字段的数据。文中还举例说明了如何在汉字识别研究中运用《汉字属性信息数据库》设计实验、确定分组水平、找刺激字、匹配控制条件等，最后探讨了对《汉字属性信息数据库》的可能完善途径。

1980年国家标准总局会同有关部门颁布了《信息交换用汉字编码字符集·基本集》，标准号为GB2312-80[1]，规定了6763个信息交换用汉字图形字符的编码原则和通信、排序标准。1988年上海交通大学计算机科学与工程系结合有关这6763个汉字的属性研究，建立了《汉字属性信息数据库》[2]。

《汉字属性信息数据库》是对汉字属性研究成果的总结性汇集。6763个汉字对应6763个记录，每个记录包括30个字段，比如拼音、音调、部首、笔顺、组字结构、部件数、结构类型、构词能力、出现次数、频级、频序、频率、国标码、笔画数等。这些丰富的内容，加上数据库强大的数据处理、实时检索和比较功能[3]，为继续研究汉字属性及其对汉字识别的影响，提供了极大的方便，同时具有一定的指导意义。GB2312-80以国家标准的形式界定了信息交换用字范围，事实上同时也界定了心理学汉字识别研究中选择刺激材料的最佳范围。本文报告对《汉字属性信息数据库》的几个主要字段进行的初步分析结果，并谈谈在汉字识别研究中应用此数据库的一些经验和体会。

1　对《汉字属性信息数据库》有关文字段的初步分析

1.1　部件数、笔画数、结构类型、字频等属性不同水平下的字数分布

GB2312-80共有6763个字，它们在部件数、笔画数、结构类型、字频等几项属性方面各有其自己的特点。了解在这几个常用属性的不同水平下各自分布有多少个汉字，对于汉字识别研究中设计变量水平、寻找刺激字都是很有帮助的。使用dBASE-Ⅲ的count命令，可以很快地从《汉字属性信息数据库》中算出字数分布情况如表1.所示。

*本文为作者在导师荆其诚教授、林仲贤教授指导下撰写的博士学位论文的一部分。在使用《汉字属性信息数据库》过程中，曾得到张侃副教授、陈一凡教授、张普教授、何厚存教授、李公宣教授的帮助。在此，谨向各位老师致以衷心感谢！

表1　几种汉字属性不同水平下的字数(个)分布表

属性	分组水平				
	1	2	3	4	5
结构类型*	318	1456	4266	636	44
部件数**	318	2373	2702	1035	298
笔画数***	423	3038	2572	623	67
频级****	560	807	1037	1864	2456

说明：*结构类型的5个水平分别代表独体字、上下字、左右字、半包围字、嵌套字；

**部件数为5一栏中的数字表示部件数为5—8的字数为298个；

***笔画数的5个水平为1—5、6—10、11—15、16—20、21—；

****频级的5个水平指的按字频高、低顺序所划分的最常用字、常用字、次常用字、不常用字、罕用字[4]

从表1.可以看出，左右字的字数最多，占整个GB2312-80一半以上；3个部件的字最多，2部件、4部件的字也不少，都在1000个以上；5/6以上的汉字笔画数在6—15之间。以上3个属性不同水平字数分布近似于正态分布，而字频属性不同水平下的字数分布就比较特殊。随着频率的降低（频级由1向5增加），字数越来越多。这在一定程度上也说明了随着科学技术的发展，新学科不断增加，新造字也不断增加。这些字在某个领域中是常用的，但总体来讲，在汉字集中却是罕用字。这个趋势给汉字信息处理带来了一定的困难，也是研究汉字识别时应该注意的。图1.直观地表示了汉字各种属性的不同水平下字数分布状态。

表1.所含的信息，在进行汉字识别研究中具有很重要的意义。因为每一个实验设计者，首先必须决定要考虑哪些变量（自变量、因变量、控制变量），各个变量要分成哪几个水平，其次是在什么样的条件范围下找实验所用的刺激字，其研究结论的代表性及推论范围如何等等问题。比如，要研究笔画数在汉字识别中的作用，由于6—15画的字数最多，那么在这个范围中找出刺激字进行实验，其结果的代表性就比较好。因此在设计中可以将6—10、11—15定为笔画数的两个分组水平。

1.2　部件的位置分布特征

GB2312-80中可以拆分出的500多个部件，它们在汉字中有不同的分布特征。通过分析《汉字属性信息数据库》中的“组字结构”字段，我们可以大致得出一些常用部件在汉字中的位置分布特征。表2.给出了一些例子。

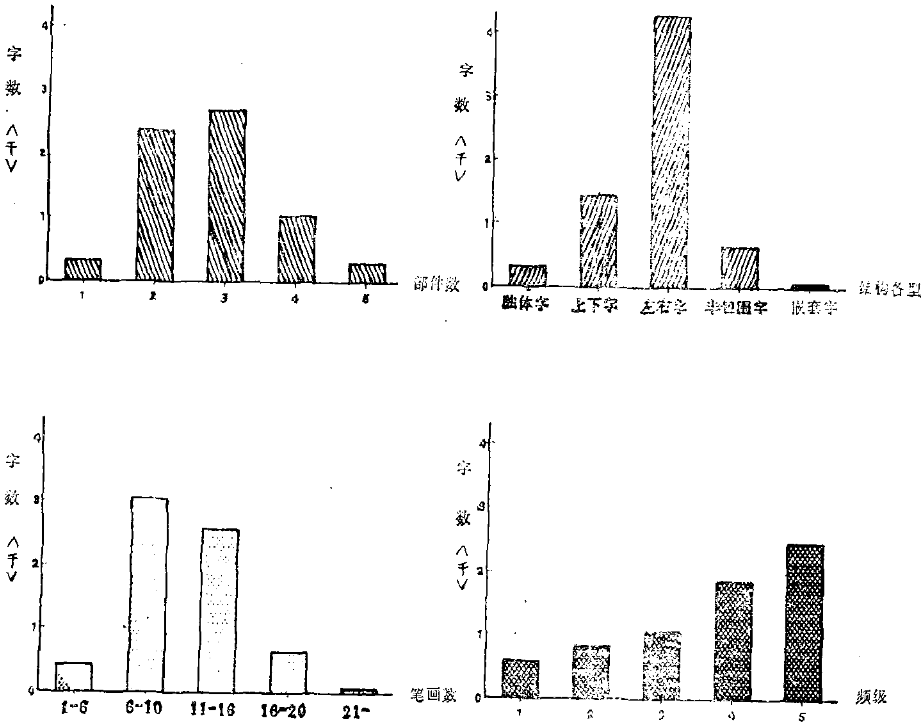

图1 汉字四种属性水平下字数分布示意图

表2 常用部件的位置分布特征示例

位置	常用部件
左	丬犭氵扌亻讠彳饣忄纟衤
右	刂勺攵
上或左上	宀艹冖亠雨广尸疒穴
右上	弋夂巛气
下或左下	皿心廴巾辶
不定	阝口一丨丿大山彡十子女

这种关于特定部件在汉字中可能的位置的知识，对于我们识别汉字应该是有帮助的，它可能是部位作用的一部分。

1.3 部件的固定搭配

分析结构表达式，我们可以发现有些部件在汉字中总是同时出现的，即形成固定的搭

配，它们实际上是高频的部件组合。如“匚——乂”、“几——又”、“一——口”等。从广义上讲，所有的两部件或三部件合体字都可以看成是固定的部件搭配，比如“各、尧、彭”等，而且许多包含这几个字的合体字中也同样地含有这几个部件。

2　对《汉字属性信息数据库》中“结构”字段的进一步分析

《汉字属性信息数据库》有30个字段，包含了巨大的信息量。这些信息有的可以给我们直接的启示，有的在稍作分析后也表现出了重要的意义，比如上面所谈的几点。对于这些字段作进一步的分析，有可能从中发掘出更多的重要信息。这里简要谈谈针对其中的“结构”字段所做的一些分析研究思路。

英文字词识别中的频率指标有字母、字母组合、单词三个层次。这三个层次单位在字词识别中的作用被很多研究所证实[3]。与此相应，在汉字识别中也有三个对应的层次——部件、部件组合、整字、它们各自有自己的频率指标。

2.1　部件频率

一个部件在特定字集中的出现次数，与我们对此部件的熟悉程度有关，因此在汉字识别中可能有一定的影响。部件的频率指标可以体现在以下两个方面：

部件组字频率——GB2312-80中含有此部件的字数与总字数（6763）之比；

部件使用频率——在特定语料范围内，此部件的出现次数与总字数之比；

2.2　部件组合频率

前面我们已经谈到汉字集中有一些固定的部件搭配，它们或者组成字，或者作为成字的一部分。那么在GB2312-80中有多少这样的搭配，它们各自出现了多少次？换句话说，在GB2312-80的部件集中随机抽出2个、3个甚至4个部件形成组合，有多少是实际存在的？它们各自的频率是多少？这就是部件组合频率的问题。

部件组合频率也可以分为组字频率和使用频率两种指标。同字母串或部件一样，这种特定的部件组合对识别含有这种组合的字可能也有一定作用。部件组合作为一种特征，其作用可能表现在频率指标上。具体而言，部件组合频率指标也有一种熟悉性的作用。比如“士＋口→吉”、“夂＋口→各”等。

2.3 部件的位置频数

部件位置频数就是在特定字集中，在各个汉字部件序列的特定位置上，某部件的出现次数。英文字词识别中字母位置频率(positional frequency)是影响字母和单词识别的一个重要变量。比如，T在4字母单词的第一个位置出现30738次。汉字部件也有同样的问题。GB2312-80中汉字的最高部件数为8，那么每个部件在这1—8的顺序位置上分别出现多少次呢？利用dBASE-Ⅲ的有关命令编程运算，从《汉字属性信息数据库》中计算出了GB2312-80的部件集的每一个部件的部件位置频率。部件位置频率在考察汉字的空间位置的作用时是非常重要的。

《汉字属性信息数据库》的6763个汉字可以拆分出567个部件。比如“十”这个部件的3

类频率指标分别是：在GB2312-80中出现246次，故其部件组字频率为246/6763＝36.37‰；GB2312-80 中含有“十”这个部件的246个字频率之和为23.48(‰)，所以其部件使用频率为23.48(‰)；而在GB2312-80中“十”的部件位置频数则从第一位至第五位分别为33，110，67，25，11；类似地，“十”与“口”的组合在GB2312-80中出现了78次，这就是该部件组合的组字频率；含有此组合的78字之频率总和为6.19（‰），这就是该组合的使用频率。其它部件上述三类属性量值的统计分析结果，限于篇幅，将另文介绍。

3 《汉字属性信息数据库》的应用体会

（1）《汉字属性信息数据库》充分体现了数据库语言易学、易用的优势[3]，信息量大而且全面，因此在使用中我们可以在控制条件范围内，最大限度地找出刺激字，同时严格匹配各种控制条件，因而可尽量减少研究者在选择刺激字时有意无意的主观偏向。同时由于采用了数据库，每个字的笔画数、结构类型、部件数、字频等变量一次考虑进去，可以逐步删、取每个刺激字，并迅速查看控制每一种条件后的目标字匹配情况。这种实时检验、迅速反馈的选字过程，给研究者带来了极大的方便。

比如要研究部件组合的组字次数（GB2312-80中汉语该部件组合的字数）在汉字识别中的作用，在GB2312-80中2部件组合的最高次数为87次，最低为1次。因此以16为组距分成6组，则各组下限分别为80、64、43、32、16、1。

将7639个部件组合所在的字找出并归入各组，用dBASE-Ⅲ的replace和update命令去除组内和组间的重复字（由于有的字含有多个部件组合，因此被重复选中；同时保证各字按其所含部件组合次数最高者归组）。建立与分组对应的6个数据库，每个库的每个记录代表满足分组条件的字，这6个库分别含有87、163、143、218、552、2198个字。

匹配控制条件：虽然每组找出这么多字，但要在实验中作为刺激还必须考虑部件数、笔画数、结构类型、字频等变量，并在组内或组间匹配这些变量，使组间控制条件相等，以使组间差异确实可以归之于组合频率之差异。首先考虑结构类型和部件数，用使用dBASE-Ⅲ命令编写的程序xz.prg检查这6组字，其分布情况如下：

表3 部件组合出现次数各分组水平下的字数分布

结构类型	部件数	分组水平 1	2	3	4	5	6
左右字	4	36	19	43	79	107	298
	5	21	15	23	3	23	44
上下字	3	2	3	16	17	85	369
	4	13	8	3	26	33	77

由于其它水平组合下缺字，故此处便不再考虑。从上表可以看出，要保证结构类型、部件数组间条件相等，每组最多可以找出19＋3＋2＋3＝27个字。再考虑笔画数和字频。由于每组字数都多于27个，因此可能在组内匹配笔画数和字频这两个重要变量。鉴于笔画数（7-10、11-15）、字频（高频——1、2级字；低频——3、4、5级字）均分2个水平，每组找字数定为26个。经过各组内不断取舍，最后基本匹配了笔画数和字频两个变量。以第四组为例，26个字及其有关参数如字头、部件数、结构类型、字频等如表4。所示：

表4　第四组刺激字的各种属性参数

汉字	结构表达式	部件数	结构类型	字频	笔画数	频级
查	木／（日／一）	3	2	0.5230	9	1
奇	大／（丁＞口）	3	2	0.2242	8	2
荷	艹／（亻//（丁／口））	4	2	0.1013	10	2
宣	宀／（一／（日／一））	4	2	0.3531	9	2
哥	（丁＞口）／（丁＞口）	4	2	0.1299	10	2
榇	木//（亻//（口／木））	4	3	0.0000	14	5
畸	田//（大／（丁＞口））	4	3	0.0223	13	3
垣	土//（一／（日／一））	4	3	0.0052	9	4
渣	氵//（木／日／一）	4	3	0.0372	12	3
刮	（大／（丁＞口））//刂	4	3	0.0000	10	5
倘	亻//（B／（冂＞口））	4	3	0.0027	10	4
淌	氵//（B／（冂＞口））	4	3	0.0037	11	4
躺	身//（B／（冂＞口））	4	3	0.0167	15	3
喳	口//（木／日／一）	4	3	0.0033	12	4
啊	口//（阝//（丁＞口））	4	3	0.0917	10	2
倚	亻//（大／（丁＞口））	4	3	0.0023	10	4
骑	马//（大／（丁＞口））	4	3	0.0214	11	3
桓	木//（一／（日／一））	4	3	0.0024	10	4
碍	石//（（日／一）／寸）	4	3	0.0672	13	3
洹	氵//（一／（日／一））	4	3	0.0000	9	5
馄	饣//（木／（日／一））	4	3	0.0000	12	5
恒	忄//（一／（日／一））	4	3	0.0452	9	3
得	彳//（（日／一）／寸）	4	3	2.1847	11	1
操	扌//（（口／（口//口））／木）	5	3	0.1876	16	2
歌	（（丁＞口）／（丁＞口））//欠	5	3	0.1303	14	2
灌	氵//（艹／（口//口）／隹）	5	3	0.1697	20	2

注：结构表达式中的(1)表示次一级结构；//、／、＞分别表示左右、上下、半包围结构

（2）　经过对《汉字属性信息数据库》现有资料的深入分析，还可以得出更深层次的属性

资料，如本文第二部分介绍的部件频率等。这一方面丰富了《汉字属性信息数据库》，另一方面可根据我们研究课题的需要，总结新的统计数据为汉字识别研究服务。再比如，关于笔画信息在GB2312-80中的分布资料，请参见文献〔6〕第二章。另外，如笔画组合等数据，尚有待进一步研究补充。

4 结语

《汉字属性信息数据库》是对1988年以前有关汉字属性研究资料的总结，它含有的丰富信息与结构化数据库的强大数据处理能力结合，使其在汉字识别研究中表现出了广泛的应用前景。一方面我们可以利用它的现有资料来设计实验、选择材料、分析研究结果，另一方面我们也可以根据课题的需要，对它进行深入研究和分析，得出更深层次的属性数据来，以使其得到不断的完善，更方便于应用。

5 参考文献

〔1〕国家标准总局，信息交换用汉字编码字符集—基本集，1981
〔2〕上海交通大学计算机科学与工程系编制，《汉字属性信息数据库》使用手册，上海科技文献出版社，1988
〔3〕李良材主编，汉字dBASE-Ⅲ结构化程序设计，电子工业出版社，1991
〔4〕郑林曦，高景成，按字音查汉字频度表，北京新华印刷厂，1980
〔5〕Solso R.L. & King, J.F. Frequency and Versatality of letters in the English language Behavior Research Methods & Instrumentation, 283-286, 1976, 8(3)
〔6〕张忻中，汉字识别技术，清华大学出版社，1992

（上接第41页）

以上为实例。

（7）研究中统计方法的使用　系统经过运算将显示不做统计、描述统计、推论统计、多元统计和其它统计方法的数量和各种统计方法随年度变化情况。形式同研究方法的分析。

（8）　引文文种情况　系统经过运算将显示引文各文种数量和引用特种文献的数量。

（9）　课题来源情况　系统经过运算将显示自选课题、国家下达任务、基金课题、横向联系、中外合作、研究生论文等各种课题来源的数量和各种课题来源随年度变化情况。形式同研究方法的分析。

（10）　研究性质的分析　系统经过运算将显示基础研究、应用研究、开发研究等各种性质研究的数量和各种性质研究随年度变化情况。形式同研究方法的分析。

每当CCDEP系统给您列出一个表，屏幕都会提问：是否作图？回答“Y”，系统将提示如何制图（略）。

想进一步了解CCDEP系统的读者请来函索要《中国儿童发展教育心理学文献数据库》（《CCDEP》）手册。