心 理 学 报
第4期 ACTA PSYCHOLOGICA SINICA 1982年

汉字结构的统计分析[1)]

彭 瑞 祥
中国科学院心理研究所

摘 要

以小学语文课本的三千个印刷体汉字作统计材料，每个字划分为四个象限，把在左上角象限、右下角象限的笔划构成的形状分别统计，形状相同或类似的归为一类并称之为子模式。统计结果表明，左上角的子模式组字能力比右下角的强。但左上角子模式的形状较复杂，除构成子模式的部首的笔划外，总带有其它笔划。右下角子模式的形状较简单，除它本身的笔划外，无其它笔划。两者比较，在设计汉字多步自动识别的系统时，以右下角的子模式作初分类的依据，比较有利。

一、问 题

作者在前一实验中[1]曾证明，被试者对左上角象限被掩盖的汉字较难再认，而对右下角象限被掩盖的字则较易再认。这结果可能与汉字在这两象限上的笔划结构的特点有关。为了探明这些特点并联系机器识别汉字的问题，本文就汉字在这两个象限上的笔划结构作初步的统计分析。

有关机器识别汉字的问题，国外特别是日本已有较多的研究。桑原启治(1977)[2](1981)[3]曾就日本研究机器识别印刷体汉字和手书汉字的情况分别作了综合的报道，并从中归纳了多种识别方案。由于汉字数量巨大(据估计约数万个，报纸常用的也近4000个)，结构复杂(笔划一般约为5划至20划，常用汉字平均为10.4划，部首约200个，它们的位置和大小不固定)，相互类似的字又多，这就对机器识别汉字带来很大的困难。Yajima, S. (1981)[5]曾提到，在汉字自动识别的问题上，输入这个环节还存在许多问题。为此，对这个问题从各方面进行研究，看来是有必要的。

二、统计用的材料

统计用的材料是小学语文课本(全日制十年制学校用)一至十册中的三千个印刷体汉字。以这些字作统计材料乃是假定它们是较常用的。统计的方法是，每个字划分为4个

1) 本文于1982年5月8日收到。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

象限，把在左上角象限、右下角象限中的笔划所构成的形状分别作分类统计。凡在这象限中，形状相同或类似的，不管它是否字的部首，都归为一类并称之为子模式。例如：委、秀，油、法这两对字，左上角的子模式前者为相同的，后者为类似的。散、数，约、讨这两对字，右下角的子模式前者为相同的，后者为类似的。又按照这些子模式最外边笔划名称（左上角从上往下从左到右最先遇到的笔划，右下角从下往上从右到左最先遇到的笔划），把它们分为点笔划型的，横笔划型的，竖笔划型和斜笔划型的。

三、统计结果

1. 左上角子模式统计结果如表 1 。

表 1 左上角子模式统计表

类 型	子模式数目	字 数*	占总字数%
横笔划型	46	735	24.53
竖笔划型	36	799	26.62
斜笔划型	41	688	22.92
点笔划型	23	582	19.38
合 计	146	2804	93.45

* 字数这列表示由该类型子模式构成的字。
占总字数%这列表示由该类型子模式构成的字占统计用的三千字的%。下同。

表 1 说明，由左上角的146个子模式所组成的字不能包括全部统计用的字，约有200个字，例如：贯、畅、君、巢、幽等，它们的左上角没有共同的子模式。由此可见，汉字结构的复杂程度。在这146个子模式中，它们各自出现的频率相差很大。频率愈高的子模式，组字的能力愈强。表 2 是左上角的子模式出现频率的情况（频率 0.5%以下的未予列出）。

从表 2 看到，由频率 1 %以上的

表 2 左上角子模式出现频率的统计

频 率	5%以上	2.0%—4.99%	1%—1.99%	0.5%—0.99%
子模式数目	2	7(9)*	13(22)	25(47)
字 数	357	722(1079)	504(1583)	551(2134)
占总字数%	11.9	24.07(35.97)	16.78(52.75)	18.37(71.12)

* 表中( )内为累积数值。

这22个子模式所组成的字占统计用的总字数的一半以上，而且它们全部都是汉字的部首。这点与李毅民等(1981)[4]的统计结果颇为接近。他们的结果表明，在二百余个属于形旁结构的汉字部件中，由五十余个大部首组成的字占汉字总字数的80%以上。上述这22个子模式可作为重点子模式，在设计机器识别汉字的系统时，应予以优先考虑。

必须指出，在左上角这一象限中，除构成子模式的部首的笔划外，绝大多数情况下还带有其它的笔划。例如：清字除部首的笔划外还带三横，许字除部首的笔划外还带一撇。表 3 是在该象限中其它笔划出现频率的统计（统计时只计算从上到下从左到右除部首外最先遇到的笔划。例如：请、讲、谁、详这 4 个字最先遇到的笔划依次是横、竖、撇、点，让字

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3　左上角象限其它笔划出现频率的统计

笔划名称	横	撇	竖	点	〔 〕*
出现次数	1067	463	258	191	155

* 〔 〕表示在该象限中除构成子模式的部首的笔划外，无其它笔划。

无其它笔划)。

表 3 表明，在由频率 0.5%以上的47个子模式所组成的2134个字中，只有 155 个字是无其它笔划的。绝大多数情况下，在这象限中除构成子模式的部首的笔划外还带有其它的笔划，其中出现次数最多的横笔划占50%。这点与Sakai, K. 等 (1973)[6] 的结果大致相近。他们的统计结果表明，汉字主要是由横笔划和竖笔划组成的，前者约占40%，后者约占20%。看来横笔划在汉字中占重要的成份，无论就整个字来说或就左上角这个象限来说都是如此。

2. 右下角子模式统计结果如表 4 。

表 4　右下角子模式统计表

类　　型	子模式数目	字　数	占总字数%
横笔划型	59	827	27.56
竖笔划型	52	818	27.26
斜笔划型	26	631	21.03
点笔划型	15	164	5.47
空(无笔划)	1	18	0.60
合　计	153	2458	81.92

与表 1 比较，右下角子模式的数目比左上角的多，而且由这153个子模式组成的字数只占统计用的总字数的81.92%，约有 600 个字没有共同的子模式。这又进一步说明汉字结构的复杂性和多样性。表 5 是右下角子模式出现频率的情况(频率 0.5%以下的未予列出)。

与表 2 相比有很大的不同，在表 5 中由频率 1 %以上的22个子模式所构成的字数只占统计用的总字数的 33.92%，而在表 2 中由频率 1 %以上的22个子模式所构成的字数却占总字数的 52.75%。这表明汉字右下角的子模式就组字能力而言，远低于左上角的子模式。这点与李毅民等的结果也颇相符。他们的结果表明，汉字按声旁结构(即按汉字右半边)组成的字数分布比较匀整，差距不甚悬殊。

表 5　右下角子模式出现频率的统计

频　　率	2%—4.99%	1%—1.99%	0.5%—0.99%
子模式数目	3	19(22)*	36(58)
字　　数	262	756(1018)	748(1766)
占总字数%	8.73	25.19(33.92)	24.93(58.85)

* 表中( )内为累积数值。

但在这里必须指出，在右下角象限中除子模式本身的笔划外，绝大多数情况下无其它笔划。这给机器识别汉字带来很大的方便。

为了便于比较，把这两个角的子模式的形状依次列表如下(见表 6)。

左上角子模式的形状在结构上看来比右下角的复杂些，前者的22个子模式平均约为3.3笔划，后者的22个子模式平均约为2.7笔划(这里所说的笔划是指黑的成分)。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 6 汉字左上角、右下角子模式的形状(频率1%以下的不列出)

左上角

序号	1	2	3	4	5	6	7	8	9	10	11
子模式	氵	扌	冂	讠	木	艹	亻	日	纟	宀	饣
字数	183	174	125	117	112	112	110	78	68	49	46
序号	12	13	14	15	16	17	18	19	20	21	22
子模式	忄	十	灬	冫	厂	力	阝	屮	钅	口	工
字数	44	42	41	40	37	37	36	34	34	33	31

右下角

序号	1	2	3	4	5	6	7	8	9	10	11
子模式	又	乙*	十	冫	口	土*	刂	刀*	乚	又	刂*
字数	110	91	61	57	54	53	50	49	45	38	38
序号	12	13	14	15	16	17	18	19	20	21	22
子模式	彐	丨*	人	卜*	刂	卜*	乂*	匕	儿	刁	彐* 匕* 凵* 乂*
字数	38	37	36	34	34	33	33	33	32	31	31

注：带有标记(*)的子模式表示频率最高的汉字右下角的子模式。

从本文的统计分析可以合理地说明作者前一实验的结果。由于汉字左上角频率较高的子模式全都是字的部首，而且由它们所组成的字绝大多数都是左右分隔型的，同时在左上角象限中除构成子模式的部首的笔划外还带有其它的笔划。当这象限被掩盖后，构成汉字的某些重要的特征或信息已被丢失，从而给被试者的再认带来了困难。在右下角象限中除子模式本身外不带有其它笔划，也就是说，该象限所包含的特征或信息是不多的，而且这些子模式的形状也较简单，这都有利于被试者从记忆中填补缺失的部分。所以掩盖右下角象限对被试者的再认影响不大。看来被试者对汉字再认的特点很大程度上取决于汉字结构的特点。

四、几点建议

把上述的统计结果联系机器识别汉字的问题可能是有意义的。在机器识别汉字的研究中，多步匹配处理技术已被人们所采用。Casey, R. 等(1966)[7] 首先使用两步匹配法，他们把盲模分为组盲模共计58个(即汉字的部首)和个体盲模共计1000个(每个相应于一个汉字)。第一步把输入的字与存贮的58个组盲模逐一进行匹配，一俟输入的字与某个组盲模相匹配后，下一步把输入的字与属于该组盲模的个体盲模进行匹配。Ankeney, L.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(1971)[8] 使用空间滤波技术提出三步的形状/位置匹配的识别方案，他把汉字左上角确定出36个几何形状，右下角确定出32个几何形状。第一步把待识别的字与左上角的36个形状中的一个相匹配，第二步则与右下角的32个形状中的一个相匹配。输入的字经过这两步的处理后，第三步的处理技术可大为简化。Wang, P. (1973)[9] 根据汉字结构的性质提出了三步的识别方案。第一步首先把输入的字区分出在结构上是属于分隔型的还是属于内外型的，第二步把已确认为属于分隔型的待识别的字与参照的分隔型的字进行匹配，借以确定它是属于哪个部首的字。经过这两步处理后，第三步的匹配也可大为简化。Sakai, K. 等 (1976)[10] 采用两步匹配法，第一步使用周边区域编码法把输入的字定出号码，第二步把经过初分类的输入字与该号码所属的字逐一进行匹配。这方法能识别一定数量的印刷体汉字并已取得明显的效果。Widrow, B. (1973)[11]提出所谓"橡皮"盲模匹配技术，首先对输入样本的个别部分(子图像)寻找最佳的匹配，然后把这些部分组合起来再与待识别的整个图像寻找最佳的匹配。

本文的结果应用到汉字多步匹配的设计，下面三点是值得考虑的：

（1） 以汉字左上角的子模式作为第一步匹配的依据，优点是：在这角上子模式数目较少；频率较高的(1%以上的)22个子模式组字能力强，由它们组成的字数占统计用总字数的一半，由其中头9个子模式组成的字数占总字数的1/3。缺点是：在这象限中除构成子模式的部首的笔划外还带有其它的笔划；子模式的形状较复杂；在这象限中四种类型的子模式(横、竖、斜、点笔划型)各自组成的字数彼此相差不大。

（2） 以汉字右下角的子模式作为第一步匹配的依据，优点是：在这象限中除子模式本身的笔划外无其它笔划；子模式的形状较简单，图形中的交点折点和歧点都较少，在匹配过程中如规定扫描或搜索方式为从下到上从右到左，则对输入字的右下角的子模式能迅速作出决定；在右下角中的五种类型的子模式各自组成的字数彼此相差较大。缺点是：右下角的子模式数目较多；子模式组字能力差，由频率0.5%以上的58个子模式组成的字数只占统计用的总字数的60%弱。当然，如果在右下角中除子模式本身外容许其它笔划存在，则由这象限的子模式组成的字数会大为增加。而且这些笔划的分布是有规则的，在子模式左边出现其它笔划，这些字大多数是由垂直分隔的三个部件构成的，在子模式的上方出现其它笔划，这些字大多数是笔划较多的。在第一步初分类时运用这些规则，可以避免盲目的搜索。基于上述的分析，权衡两者的利弊，在设计汉字多步匹配的系统时，第一步匹配以右下角的子模式作依据，似乎较为有利，可以节省匹配的时间。

（3） 每个汉字出现的频率相差很大。据一份统计材料报道[12]，在所统计的六十多万字中频度在一百五十分之一以上的计有二十个字。其中如：的、是、不、和、在、中、人、这、战、们、了、有、个、我这些字，可能是更常用的。在进行第一步初分类匹配时，首先把输入的字与存贮的这些字的右下角的子模式(即表6右下角的带有标记的子模式)进行匹配，然后再与表6右下角的其它子模式依次进行匹配。这样有可能进一步压缩匹配的时间，提高识别系统的效率。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参 考 文 献

〔1〕 彭瑞祥、喻柏林，中国心理学会普通心理与实验心理学第一届学术年会资料，1980年。
〔2〕 桑原启治，日经エレクトロニクス，154，42—59，1977.
〔3〕 桑原启治，日经エレクトロニクス，158，148—167，1981.
〔4〕 李毅民，王建新，李金铠，电子计算机动态，第1期，50—54，1981年。
〔5〕 Yajima, S. et al., IEEE Trans on Pattern Analysis and Machine Intelligence, Vol. PAMI-3, No.2, 221—230,1981.
〔6〕 同〔5〕
〔7〕 Casey, R. Nagy, G., IEEE Tran. Elect. Comput., 15,91—101,1966.
〔8〕 Ankeney, L., Pattern Recognition, 8,87—98,1976.
〔9〕 Wang, P., Shau, R. C., Pattern Recognition, 5,303—321,1973.
〔10〕 Sakai, K. et al., Proc. 3rd Int. Conf. On Pattern Recognition, 122—126,1976.
〔11〕 Widrow, B., Pattern Recognition, 5,175—211,1973.
〔12〕 《毛泽东选集》用字统计资料，1974年。

A PRELIMINARY REPORT ON STATISTICAL ANALYSIS OF THE STRUCTURE OF CHINESE CHARACTERS

Peng Rui-xiang
(Institute of Psychology, Academia Sinica)

Abstract

The structure of 3000 printed Chinese characters which were selected from the Chinese textbook for elementary school was analysed. Each character is divided into four quadrants i. e. upper-left (UL), upper-right (UR), bottom-left (BL) and bottom-right (BR). When the shapes composed of the strokes in UL and BR were the same or similar respectively they were attributed to one group and were called subpatterns.

The result showed that the subpatterns of BR were more simple than that of UL and the number of junctions of the former was less than that of the later. The author argued that the subpatterns of BR are used as group mask in the early stage of the processing of multi-stage matching the computation time may be reduced.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net