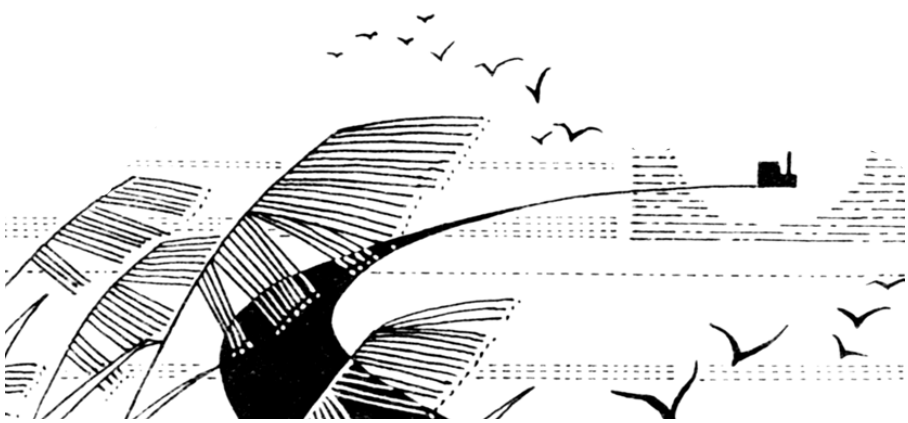
汉字编码设计（二）

马　敏　主编



目　录

形码设计的比较研究（二）

3.4 形码典型方案分析

比较是为了区分优劣。既然输入速度、重码率、信息含量等都不能作为编码优劣的衡量标准，那么衡量的标准是什么呢？我们认为标准就在于看这些编码在何种程度上反映了汉字和编码设计的结构规律。尽管各种形码都在不同的侧面接近了这些规律，但对汉字规律的探索毕竟是一个长期的过程。我们拟从众多的形码方案中，选取部分典型方案：仓颉码、五笔字型、郑码、见字识码、大众码、宏观码、许毕码、王码象形码、表形码等，加以剖析，扬其长，弃其短，以逼近对编码设计客观规律的进一步认识。

一、仓颉字母

二、五笔字型

三、郑码“ZN”电脑汉字 26 键拆根编码方案

四、见字识码

五、大众码

六、汉字宏观编码

七、许毕码和王码象形码

八、字根码的突破与表形码

3.4 形码典型方案分析

一、仓颉字母

仓颉字母把汉字部件分为 24 类，24 个字母分为 4 个大类：

甲哲理类日月金木水火土

乙笔画类丿丶十乂丨一亅

丙人身类人心手口

丁字形类尸艹山女田卜

甲、哲理类

字母	产生原则	举例
日	①日的变形为 ⺆	巴
月	①取月的外形成冂，⺆ ②月的变形为 ，	巾，冗 祭，胖，采
金	①取金的分形丷 ② 变化而成八，儿	弟 谷，四
木	①取木的主形 ⺍ ②由 ⺍ 变化成	寸 皮
水	①取水之左右并重合为又 ②水做偏旁时为氵	支 沿
火	①取火字的上半形为小丬 ②小变为丬 ③灬做字尾时的字形	肖 紧 热
土	①土的变形为士	任

乙、笔画类：

字母	定义	举例
竹 （斜）	①斜的定义为丿 ②丿的变形成为厂	舌 反
戈 （点）	①点的定义为丶 ②点的变形成为厶、广	之 公、广
十 （交）	①交的定义为十，而其变形成为亠	宋
大 （叉）	①叉的定义为  ②大的部分形成	刈 友
中 （纵）	①纵的定义为丨即由上而下 ②变形为丿 ③丰亦为纵向意义相近	巾、川
一 （横）	①横的定义为左右向，故工属于辅助字形 ②横下加一撇为变形	江 原
弓 （钩）	①钩的定义为亅 ②变形成为マ ③再加变化成为 	了 甬 久

丙、人身类：

字母	产生原则	举例
人	①人的变形为 ②取部分形为 ③偏旁为亻 ④变形为	兆 孑 仁 乞
心	①匕七 皆为心字形变化 ②偏旁作忄 ③字尾作	尼、虎、民、句 快 恭
手	①手之主要字形为 ②变形为、 ③偏旁作扌	青、夫 年 拍
口	无	

丁、字形类：

字母	产生原则	举例
尸（侧）	①侧之定义为向左右开口，所以为 、匚 ② 、 为变形	尹、巨 刀
廿（并）	①并之定义为两形相并所以为廾 ②变形为 、 。	草 黄、羊
山（仰）	①仰之定义为向上开口，所以为凵 ②变形为 、 、屮	凶 巴、民
女	①纽的定义为曲	巢

（纽）	纽，故为𠂇 ②变形为𠃌、	亡、衣
田 （方）	①方之定义为一方框，内有其它形所以为口 ②变形为	国 母
卜 （卜）	①卜的变形为𠂆、冫、二、卜 ②礻为卜的类似形	言、冰 斗、被、非

据上所述，仓颉码的主要问题是：

一是分类零乱主观。仓颉码的分数系统是：几万个汉字，由十三个简单的汉字、四个部首、两个笔画结构（十乂）、五个笔画组成。这个系统，是怎样组成的？发明人没有加以任何有依据的说明，使人感到它在设计上是主观的，在结构上是零乱的。例如“十、乂”，并不是笔画，但设计者却把它们摆在笔画类里，这显然是不恰当的；“女”，摆在字形类里也不合适。

七个笔画类，不是笔画的将近三分一，这个类名，即类的定义就难以成立。一个完整的系统，分类和命名上的客观性和科学性是很重要的。

二是取形混乱，例如哲理类有七个小类，其中有五个类的取形就有五样：

取月的外形成冂、冖；取金的分形丷，丷的变化而成八、儿；取木的主形𠂇；取水之左右并重合为又；取火的上半形为小。月取外形；金取分形；木取主形；水取左右并重合；火取上半形。每取一个形，即多了一个记忆单元，取与不取，没有什么不同，不如干脆规定几个就作为这个类方便。有了一个“取”字，设计者的意

思是使它们之间有联系，但这个联系很勉强，不是唯一的。因为，规定取外形，就要撇开内形；规定取上形，就要撇开下形。这就无形中增加了记忆负担。

三是字母定义的随意变形。仓颉码依据二十四个字母联系形和义，又产生了如个小类。我们可以从上面的字母表中看出它的随意性来：

斜的定义为丿丿的变形为厂

点的定义为丶丶的变形为厶、广

交的定义的十而其变形为宀

纵的定义为丨即由上而下变形为丿丰亦为纵向意义相近

横的定义为左右向，故工属于辅助字形，横下加一撇为变形。

钩的定义为亅，变形成为マ，再加变化成为  。

从以上例子可以看出：由“丿”变为“厂”，即加上一横，如果这个“丿”加出头了，就会成为“  ”，但偏偏又不出头，与第五类“横下加一撇的‘原’”字头并没有区别。

“丶”变为“广”，由“十”变为“宀”，实在说不通，根本无法解释。“丶”是任何一个笔形的始点，变为某个笔形，或许可以解释，设计者却使它变为“广”。“十”，如果说变为“  ”，也会讲得通；说“冂”变为“宀”，可以说得过去，因为，两者的命名，一为“秃宝盖”，一为“宝盖”，“盖子”上多了一点。说是由“丶”变过来的，却讲不通了。

“丨”（纵）的变形为“丿”，与撇笔的形象一致。一个相同的形象分为两类，在分类学上是难以接受的。

这些“变形”，都是十分随意的，不合情理的。每

个类有个定义是必要的，定义准确就毋须解释。但如何变形，是需要解释的。例如说“丶（点）”的变形为“㇏（捺）”，如“木”的末笔在字中或字右，均为捺笔，若作为左边旁，则变形为点笔。必须这样解释，不能没有依据。尽管有个别的字母其“变形”是可以说得过去的,例如“日”的变形为“冂”;人的变形为“亻、 ”，但这是极少数。卜的变形为“𠂉”，还可以说得过去，但再变为“冫、二”，就无法解释了。所谓“变形”的本意，也是为了增加相互联系，减轻记忆负担，但需要讲得通。现在讲不通，不能帮助理解，也说明这种解释是主观的，牵强的，不科学的，不仅不能减轻记忆负担，还起了反作用。

四是取码主观随意。分类系统的主观性、随意性，必然导致拆分取码上的主观性和随意性。仓颉码有许多字要一个一个规定，例如：

正确取码错误取码

王一土一十一

容十金人口戈月金人口

羊艹手金一手

九大山大弓山

乍竹尸人卜卜

屯心山十山山

言卜一一口戈一一一口

非中一卜卜卜中一中一一

这样一个字一个字记忆，需要多少记忆量？

五是无理映射记忆困难。仓颉码二十四个类代表通过“变形”或某种随意的解释，再产生 1~4 个字母。由于字母数量太少，结构简单，造成取码（即分解）的困

难。例如：

妻——十中女

商——卜金月口

雨——一中月卜

函——弓山水

求——戈十水

乘——竹木中心

舟——竹月卜戈

焉——一卜中火

这些规定教人难以理解，而且每一个字都要如此规定，记忆实在太困难。

仓颉码 24 小类安排在键符上，这种映射方法是任意的，无理的，记忆难度太大。

台湾的电脑科技和生产发展较快，一起始就进入字根码的使用，人们一般很难发现汉字编码的发展轨迹。也许设计者朱邦复先生已经知道一个设计方案，只不过是纸上文章，并不那么值钱，他干脆让人家无偿使用。后来即使有人设计了好一点的方案，想从中赚钱，用户自然不愿意拿钱买，也就无法替换它。汉字编码作为一种基础设计，是要加上软硬件技术才可作为商品走上市场。软硬件成为编码方案的主要支柱后，产生很大的稳固性。如果另一种设计方案在技术上、经济上、主观能力上不具有压倒性优势，就无法立足。编码方案仅是纸上文章，与已经占领市场的方案比高低，就好比秀才造反，光靠嘴巴，永远不能成功。仓颉码在台湾能稳稳保住主导地位，其原因就在于此。

3.4 形码典型方案分析

二、五笔字型

五笔字型首先通过对字根的使用频度与组字频度的统计，挑选了 199 个（设计者宣布为 125 个，故意压缩数字，引人上钩，是一种商业手段，表现它的非科学性）基本字根作为建立方案的基本材料。所用的字根是汉字中独立的结构“ 块 ”，不是通过主观任意的“ 变形 ”来产生的。虽然数量比仓颉码多，但是学习者分解汉字要比仓颉码方便。字根的产生是通过使用频度和组字频度统计得到的,(虽然是“ 优选 ”的,总比乱“ 变形 ”的好）也比仓颉字母有依据。

五笔字型利用 1-5 个数目字对应横、竖、撇、捺、折五个笔形，再把键盘上的英文字母键符，划为五区五位，选用一些使用频度高的部件（称基本字根），根据这些部件的第一、二笔的对应数字代号，分配到相应的键符上。如下：　一丨丿 

尽管五笔字型的设计者事先规定用字根的第一个笔画的笔形代码数字作为该字根应分到的“ 区 ”, 第二个笔画笔形代码数字作为该字根的“ 位 ”, 但具体的安排差异很大，只有不到 50%的字根符合设计方法。即是说，有50%以上的字根是无理安排的。原因是按字根的起首两个笔形为区、位的安排方法并不符合实用要求。因为，字根的起笔以横竖为多，点（捺）折为少，这样的安排使字根的键位分配极不均匀。例如以点（或捺）为区，以竖、撇、折为位的字根，就很难找得到。这些键位就成为空位。许多字根以横为区，横或竖为位的很多。这些

地方就会“根满为患”。为了使每个键位上都能分配到一些字根，有的不能死板地按区、位安排，只能无理地安排在这些空位上。这种以无理化作为代价是为了使每个键位得到“合理分配”，以换取降低重码率。可以认为，由于把键盘上的 26 个英文字母键符，分为五区五位的设计安排在实践中碰到了困难，设计者只得服从现实，不得不进行修改。实践证明了这种设计方法的不成熟性。

五笔字型利用数字的序（1.2.3.4.5）与笔形的序（横、竖、撇、点、折）对应，又把键盘上的 26 个字母，划为五个区五个位方式（即有理化），设计考虑仍然离不开“笔形——数字”这样一种转换方式，在部件与键符的转换关系上，比笔形码又多了一道“手续”。这道“手续”在人的脑子里使字根与键符的对应，多拐了一道弯。笔形码是直接采用数字键的，五笔字型则是用数字换区位，区位再换拉丁字母键，间接加间接的转换。已经学会英文打字的人，还不能直接打字，得重新记忆区位。示意式如下：

笔形码字根-笔画-数字（键符）

五笔字型字根-笔画-数字-区位-键符字母

五笔字型方案除了间接加间接的转换毛病外，还有几个主要缺陷：

（一）在设计思想上忽视汉字的整体性，没有认识到选用少数基本字根涵盖全部汉字是很困难的。因为，它的设计根本在键盘上，只考虑键盘安排，不考虑设计原则。所以，进入实用时觉得不顺手，就随意修改事先的规定。但为了使人觉得它的易学和“科学”，还是“照本宣科”。这种实用主义手段，作为商业行为为可以的，但作为科学行为却是不应该的。在中文电脑应用的专业

时期，人们考虑的是技术性能，因此，商业手段可以行得通。因为，学的人是一些寻找职业的年轻人，记忆力好，多花一些时间学习也不要紧。但中老年就大不相同。由于文字使用和软硬件在社会上的稳固性，在中文电脑全面普及的时期，社会需要一种易学、符合汉字自身规律的编码，而五笔字型的市场占领就会起很大的反作用。因为，它的社会使用惯性影响计算机应用向全社会展开，阻碍计算机全面普及，阻碍我国信息化的进程。

（二）对部件的客观性缺乏认识，认为汉字的字根（部件）没有客观标准，只能靠设计人“精心筛选”。王永民认为：“什么算构件（即字根），什么不算，构件选多大合适，都因人而异，这里的根据是设计的需要。”“字根并不像汉字那样，有公认的标准和一定的数量。”这样的理论，必然导致编码方案的无限制的花样翻新。字根没有“公认的标准和一定的数量”，设计者要多大就多大，要多少就多少，实质是说，设计编码不需要给它的基本材料定性定量。这样的主张，实质是背离设计学的基本原理和原则。

（三）“精心筛选”的字根不能套拆全部汉字，有的汉字便只能一个一个地主观决定，增加记忆负担。

（四）由于以上原因，基本字根就不可能排成一个体系严密的、完整的、统一的系统。例如王码有一句助记词：“王旁青头兼五一”，即是说，把“王、⺹、戋、五、一”都安排在 1 区 1 位。让我们想想，这“王、⺹、戋、五、一”五个字根之间，存在着什么样的联系呢？设计者也许会说：“它们的起首都是两个横笔。”其一、“五”的第二个笔画不是横笔；其二、还有很多字根起首两个笔画都是横笔的却没有安排到这个位置上。后来

的版本对一些键位上的字根进行了调整，例如 1 区 1 位安排了“ 干、戋、王、夫、千 ”。“ 千 ”的头上是一撇，第二笔是横，本来应放到 3 区 1 位才对。如果我们认真检查它的八个版本，就会发觉，无论如何调整，都没有办法增加它的有理成分。

（五）由于拆分的主观决定，必然背离汉字规范，损害识字教育。例如下面一些字的拆法：

那拆分为刀二阝 面拆分为厂冂丨三

曲拆分为冂艹身拆分为丿三丿

击拆分为二丨山其拆分为艹三八

乐拆分为 小拜拆分为手三十

不要说这些拆分方法在记忆上的难度，它在中小学语文识字教育上的不规范性使人无法接受。

（六）由于它对汉字字根的规范性认识不足，以致对许多汉字的拆分，采用基本字根代替的方法。如：

东七小派氵 厂 丑 ㄱ土不一小

母 一 ㇀凸丨一几一聚耳又丿水凹几冂一

段亻 三几又舆亻 二车 ㄱ二一八

这样的字很多，如果中小学生们都这样去学习汉字，那么，头脑里就会经常打架。

（七）为了减少重码，五笔字型采用了“ 末笔识别码 ”。但是，这种末笔有许多是设计者自己规定的，不是国家规定的。就是说，这种末笔是损害国家书写统一规范的。例如根据设计者规定：

“ 国 ”的末笔为“ 丶 ”

“ 远 ”的末笔为“ 乚 ”

“ 刀 ”的末笔为“ ”

“ 力 ”的末笔为“ ”

“九”的末笔为“乁”

“匕”的末笔为“乚”

“必”的末笔为“丿”

“乘”的末笔为“乚”

“戈”的末笔为“丿”

这些末笔规定都违反了国家的标准规范。

（八）除末笔识别码外，五笔字型的汉字拆分仍然有许多地方不符合国家的语言文字规范，例如：

“冒”应拆为“冂二”，但五笔字型却拆为“曰目”，“帽、瑁、冕”等均错。

“象”，中间是“  ”，五笔字型拆为“  ”（称横日），“像、蟓、橡”等均错。

“麸”，“  ”下面是“夂”，五笔字型却作“夕”。

“北”的左旁是竖起笔，五笔字型拆为“  ”，成了点起笔，所以，“背、褙、邶、乖、乘、剩、嵊、冀、骥、燕”等都拆错了。

“非”的首笔是中间一竖，再写左边三横，按国家规定应拆为“丨丨三三”，五笔字型却拆为“三丨丨三”。

“燕”按规定应拆为“廿口北灬”，“口”先于“北”，五笔字型却是“北”先于“口”。

总之，笔顺、形象错拆的很多，例如“母、毋、幽、丑、凹、凸、官、敝、及、貌、里”等都拆错了。而且，由于这些字又可作为偏旁，拼出另一些字，因此，一个拆错，累及一批。“本节资料，得自云南大学张在云副教授的研究）。

五笔字型是数码字根码的典型代表，尽管已有上述八大缺陷，但比之仓颉字母，它所用的基本材料增加了，分解汉字更方便。仓颉码虽然只用了 83 个字母，但它的

“变形”“取”字母，都是很随意的，没有任何依据。五笔字型用的 199 个基本字根，毕竟还有个使用频度和组字频度作为依据。尽管这种依据并不十分科学，有总比没有好。因此，记忆量相对减少了。说明它在设计上比仓颉码有一定的进展。

3.4 形码典型方案分析

三、郑码“ZN”电脑汉字 26 键拆根编码方案（以下简称郑码）

实践经验证明，利用数目字代表字根的笔形或某些简单的结构，再将其转换为键盘符号的做法，并没有给一般人的记忆带来很多的好处，反而增加了拆分上的麻烦，是得不偿失的工作。但是有的设计者还是觉得笔形多少总有一点用处，不愿意全部丢掉，郑码的设计思想基础可以作为这方面的代表：

郑码选用了 182 个基本字根，外加 21 个笔画作为编码方案的基本单元。把 182 个基本字根分为 50 个主根和 132 个副根。这 182 个主根和副根又根据起笔笔形分为横、竖、撇、点、折五个区，每个区中含有 3——8 个类不等。每个类中根据安排的顺序取得一个英文键符为代码。每类中有 1——2 个主根，其他的为副根。

50 个筛选出来的字根，按笔形分为 5 类，根据起笔分别安排如下：

横起笔类：

根区代码	主根	副根	基根笔形特征
A（一）	一	丁	一横的基根