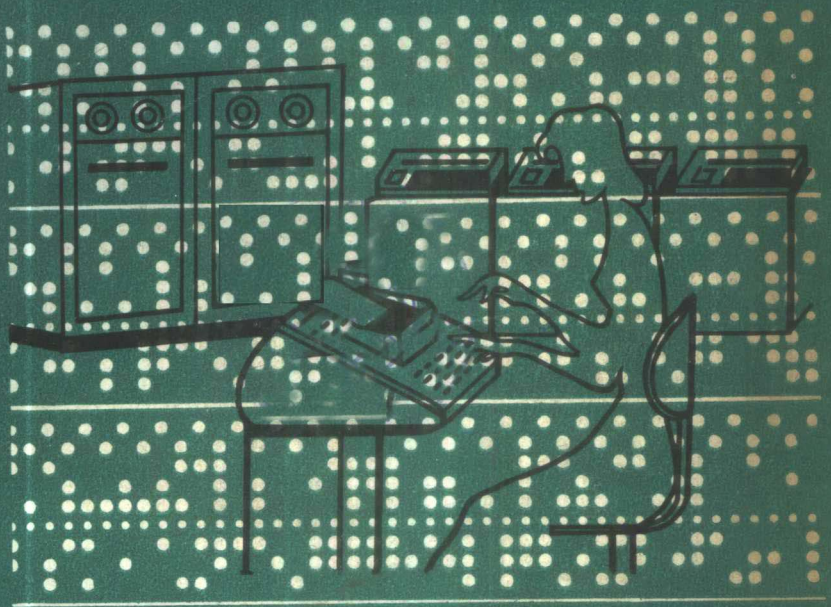
汉字编码方案汇编

中国汉字编码研究会编



科学技术文献出版社

汉字编码方案汇编

（限国内发行）

编 辑 者：中国科学技术情报研究所

出 版 者：科学技术文献出版社

印 刷 者：中国科学技术情报研究所印刷厂

新华书店北京发行所发行　　各地新华书店经售

☆

开本：787×1092 1/16　印张：12 1/2　字数：320千字

1980年8月北京第一版第一次印刷

印数：1—8,200册

科技新书目：154—30

统一书号：17176·234　　**定价**：1.30元

前　言

一九七八年十二月在青岛召开了全国汉字编码学术交流会，这是我国汉字编码研究工作者的第一次盛会。出席会议的有来自全国十七个省、市、自治区的八十多名代表，其中有自然科学工作者，也有文字研究工作者，有老教授老专家，也有人数众多的业余汉字编码研究人员，显示出这项研究工作在我国有广泛的群众基础。会议共收到各种方案四十多个，基本上反映了我国编码的主要设计思想和已取得的成绩。这次交流会也是对我国汉字编码研究工作的一次检阅。为了加强交流，相互促进，减少重复劳动，推动研究工作的深入开展，在这次交流会上成立了中国汉字编码研究会，并决定由研究会负责编辑出版会议的学术交流资料。本汇编就是根据这个决定产生的；内中共收报告和方案四十篇，大体上反映出了交流会的概况和水平。

计算机科学技术已列为我国科学技术发展规划的八大重点项目之一。要在我国广泛应用计算机技术，首先必须解决汉字信息处理的有关问题，也就是要让计算机能够输入、存贮、处理和输出汉字。汉字字数多，结构复杂，如何解决好汉字输入计算机的问题，是汉字信息处理中的一个十分困难而又紧迫的重要课题。

围绕着这个难题，国内外都在奋力开展研究试验工作，但迄今尚未形成受到普遍赞许的解决方法。我国许多专家和研究工作者，经过多年艰苦努力，提出了不少具有一定

水平的汉字编码方案，为解决这个难题打下了比较厚实的基础。粉碎“四人帮”以后，在新长征号角的激励下，这项研究工作十分活跃，不同特点的新方案不断涌现，呈现出百花齐放的兴旺景象。可以相信，今后通过加强基础性研究工作，开展上机试验和逐步试用，不断进行改进和提高，并对一些重大难点发挥集体力量，组织攻关，我们一定能够以较快的速度解决好这个关系到我国加快实现四个现代化的重要课题。

最后，需要说明的是，这里收入的各种编码方案没有包括大键盘整字输入类型的方案；未参加青岛交流会的同志可能有一些水平较高的方案，这里也没有收入。本汇编中对方案的排列，粗分为按“形”和以“形”为主、按“音”和以“音”为主两大类。本着贯彻“双百”方针，发扬学术民主的精神，我们对作者所持的不同学术观点，所用的不同术语，都没有进行统一。由于人力和时间的限制，对文中的数据也没有全都进行校核。按照惯例，文责自负。

本汇编所收的各种汉字编码方案，欢迎有需要的单位与作者取得联系，创造条件，进行检验和试用，以期尽快使之完善和提高。

对编辑工作中的缺点和错误，欢迎批评指正。

中国汉字编码研究会

《汇编》编辑组

目　　录

有关汉字信息处理研究的几个问题

刘涌泉

（中国社会科学院语言研究所）

摘　　要

本文对汉字信息处理研究中的几个问题，如汉字编码在汉字信息处理系统中的地位、汉字编码和拼音文字、词汇码和有关汉字编码的基础理论研究等，进行了论述，并提出了一些看法。

世界已进入电子计算机时代，计算机不仅用来进行数值运算，而且越来越多地应用于非数值运算。科学技术发达的国家都在利用电子计算机实现文字工作的机械化自动化。

在我国，要实现文字工作现代化，必须解决汉字信息处理问题。不解决这个问题，什么情报工作自动化，什么印刷排版现代化，什么管理科学化，都将落入空谈，四个现代化的进程也必将受到影响。问题如此重大，这次学术交流会的意义也就不言而喻了。

下面就汉字信息处理问题的研究谈几点看法。

一、汉字编码和汉字信息处理

汉字编码是汉字信息处理的一个组成部分。汉字信息处理一般包括这样几项内容：编码、输入、存储、处理、输出和传输。汉字编码是最关键的部分，也是最难解决的部分。这些都是大家熟悉的。我要强调说明的一点就是：我们不能孤立地谈论编码，我们必须从整个系统的角度来论证这个问题，即考虑编码时，还必须考虑输入是否方便、处理是否容易、存储是否节省，等等。否则，编出的码，一旦在机器上加以实现，就会出现这样或那样一些问题。同样，评定一个编码方案时，绝不能单单以某种编码法“规则很少、容易掌握”为依据，也绝不能以没有同码字为依据，必须综合各项指标（字码无二义性、操作方便易学、输入和处理效率高、存储节省、传输可靠、设备经济实用等等）加以全面比较才能得出正确结论。

这里，还必须指出一点：汉字编码是属于半自动化输入方式的关键问题，一旦全自动化输入方式（即电子扫描自动识别）研究成功，汉字编码问题也就大大改观了。那时虽也有汉字编码，但要自由得多，因为机器内部编码，可以忽视人的因素。

二、汉字编码和拼音文字

汉字编码是给汉字编代码，拼音文字是汉语未来的书面形式。这二者之间既有联系，又有差别。考虑到文字改革要走拼音方向而主张以音为主来解决汉字编码问题有它的道理，因为这样，一可培养人们的拼音习惯，二可为文字改革摸索一些经验，三可使编码与拼音文字趋向一致。但是，要知道未来的拼音文字是代表汉语，而不是代表汉字的，换句话说，是代

替汉字表达汉语的。而目前的编码却是为了代表汉字的，否则也就不叫做汉字编码了。由于有这样的差别，所以在一些问题（甚至是原则问题）的处理上，二者不能强求一致，更不能混为一谈。代表的对象不同，要求不同（例如，在码的长短、有无同码等方面，编码比拼音文字要求高），处理的方法也应允许不同。例如，有的主张以字音编码的人，为了解决同码字问题，一般求助于汉字的部首或其他形体信息。这从编码上说是完全合理的，但是从拼音文字角度来说，就是画蛇添足了，给未来的拼音文字添上“部首尾巴”，显然是行不通的，因为后代人（绝大多数）不学汉字又怎么知道汉字的部首呢？又如，有的人考虑声调不能完全区别同音，又考虑方言区的人不易掌握正确声调，主张在有限的范围内为了区别同码使用汉字形体信息而不利用声调，这也是一种可行的办法。但是，如果以此来推论将来的拼音文字可以不标声调，又未免太武断了。看来，这样一种论点也许是可以肯定的：代表汉字的编码在一定条件下可以忽视声调，而依靠汉字形体信息，但是，代表汉语的拼音文字却不应忽视声调，因为声调是汉语固有的基本要素。

三、词汇码问题

词汇码的应用不外两个目的：一是提高效率，二是区别同码。这种方法如果使用得当，的确不失为一种可行的办法。但是，我们还应充分估计到一些矛盾问题。

先看看提高效率方面的问题：人们常愿把最长的词，即音节最多的词（有时其实是词组）作为规定的词汇码，这种情况更常发生在部件式方案中。显然，这可以大大减少击键次数，提高输入速度。例如，有的方案把“人民共和国”作为规定的词汇码，这样，输入该词（组）时，只需击完“人”字键，再按一下“词汇键”就行了。但是，由此却产生了一个问题：“最长”和“通用”的矛盾。词汇码被“人民共和国”占去，而“人民”这样常用的词便排斥在外了；象“人民公社”、“人民解放军”、“人民英雄”等常用词组也不可能利用“人民”这个当之无愧的词汇码了。由此可见，在规定词汇码时，必须多多考虑词的概率性。当然，这种情况可以利用数标来处理。如“人”＋1＋词汇键＝“人民”，“人”＋2＋词汇键＝“人民性”，“人”＋3＋词汇键＝“人民公社”，“人”＋4＋词汇键＝“人民共和国”。不过，这给操作人员的负担未免太重了。

再谈谈区别同码方面：的确，“词儿连写”是解决同音（即同码）的一个方法。但是，也不能认为它是万能的，尤其在其他条件不足时（例如不标声调；又没有其他可依靠的特征）更不能完全依靠它。《汉语拼音词汇初稿》一书中收词20133词，其中2125个同音词（同音同调词），如不标调，同音词还要增两倍左右。汉语多音词中双音节词占绝大多数，据统计，同音现象也相当严重（占20%左右）。为了说明问题，让我们拿shi字的同音情况同shi字所组成的双音节词的同音情况作一个比较：shi字是同音字比较多的一个，一共有72个。其中阴平14个，占19.4%，阳平14个，占19.4%，上声7个，占9.7%，去声35个，占48.6%，轻声2个，占2.8%。〔以上数字根据《新华字典》得出〕

shi字组成的双音节词共537个（这里只计算了作词的第一部分的shi，如“食油”、“事由”等词，没有计算“分式”、“分时”等词），其中同音同调词155个，占28.8%，同音异调词295个，占54.9%，不同音的词只有87个，占16.2%。

从上面可以看出，如果不标声调，同音的双音节词占83%，这对区别同码没有多少补益。〔以上数字根据《汉语拼音词汇（增订稿）》得出〕

词汇是不断发展的。随着科学技术的迅猛发展，术语大量涌现（化学名词已发展到二百多万）。在这种情况下，同音现象只能是有增无减。例如，shishi 这个双音节词（失实、失时、诗史、失事、失势、施事、实施、时时、时事、时势、时世、时式、史诗、史实、试试、誓师、事实、适时、事势、逝世、世事、视事）的同音行列中近年来又增加了一个新伙伴：实时（实时系统）。fenshi（分式、粉饰、偾事、愤世）的同音行列中又增加了“分时”（分时系统）。

另外，还必须估计到，许多双音节，分开写就成了两个单音节词。如“改变”——“改”（这个数一定要改）和“变”（这个数不能变）。这些单音节词的同音问题（改：该、盖；变：边、编、煸、匾、便、遍、辩、贬），反正总是要单独解决的。因此，我们应该首先着眼于汉字编码（一般是几千，最多是几万），而不应首先着眼于汉词编码，因为给汉词编码是编不完的。

顺便指出，词汇码还不同于拼音文字的词儿连写，词儿连写后剩余的同音词可以保留其形式由人根据上下文加以判别。而词汇码产生的同码词则必须加以处理，因为要输出不同汉字。

总之，单字码总得先解决好。不能把区别同码问题完全推到词汇码上，也不能把提高效率问题完全推到操作员身上。最好的办法是多在机器身上打主意，即作为词汇码输入也行，作为单个字输入也行，二者都可得到正确结论。例如，字音式方案真要想利用词汇来区别一定数量的同码的话，可以在机器内部增加一些处理程序，让机器通过查词典来进行辨别。甚至还可不输入词，而输入字，让机器通过“最大匹配法”自己组成词，再去查词典。当然，还可通过一些语法规则来区别同音词，如 feizhǐ 前有“一张”是“废纸”，前有“曾”字或后有“了”字是“废止”。不过，这样做，对于汉字信息处理似乎太繁琐了。

四、应该加强基础理论研究

汉字编码不仅是一个技术问题，而且也是一个理论问题，不弄清它的理论方面，有时对实际工作当中出现的问题得不到正确解释，也有时会头痛治头，脚痛治脚，不能从根本上解决问题。

编码方案（这里主要谈“字形”式）为什么这样多，它们的异同在什么地方？如果我们把文字看成是一个多层的符号系统，这类问题不难得到答案。

语言是一个符号系统，文字更是一个符号系统。正如语言因其类型不同而各有不同符号系统一样，文字也因类型不同而各有自己不同的符号系统。英文的基符是字母，只有26个，由它们直接组成了一个庞大的符号系统。俄文的基符，也只有33个，整个文字符号系统也是由它们直接组成的。汉文的符号系统是怎样构成的呢？这是研究汉字的人必须解答的一个问题。

我们所敖镜浩同志在这方面进行了一定研究，最近得到了某些成果。他的研究可以证明汉文是一个客观存在着的符号网络系统。这个系统是一个多层结构，点是最初始的，它是一切笔划、笔形的根基，下笔即为点，向不同方向延伸，即构成五种最简单的单名笔划：横、竖、撇、捺、挑。再以它们作为起笔，又构成十二种双名笔划（㇆、㇇、㇈、㇋、㇌、㇍、㇎、㇏、㇐、㇑、㇒、㇓）。这十八种基本笔划就是汉字构形的基本单位。

在汉字构形过程中，下一步骤便是这些笔划的进一步组合，其组合方式不是任意的，而

是按一定规则进行的。笔划一般采用“离”、“交”、“接”三种基本形式组合。例如：

离（旦、八）　　离（贝）　　交（力、右）　　交（内）

接（且、人）　　接（刀、石）

“离”、“交”又各分甲乙两型，“接”又分甲乙丙丁四型。

笔划通过上述组合方式结合在一起，便构成一个个彼此不同的形，从这些不同的图形中寻找出共通点，比较归纳，又可得出不同的结构格局。结构格局分单体结构和复体结构两种：单体结构的基本图式有四：(1) ⧅，用“离”方式构形，例丿、一、川。(2) ⊤，主要用“接”方式构形，例卩、亻、丁。(3) □，主要用“接”方式构成的框形，例匚、勹、口。(4) ⊠，主要用“交”方式构形，例十、廿、才。复体结构的基本图式有三种：(1)左右形，又可细分四小类，例：梆、鸿、粥、乘。(2)上下形，又可细分六小类，例：全、鼻、息、哀、事、品。(3)框形，又可细分六小类，例：凶、冈、医、处、匈、右。

以上的简单介绍不过是对汉字符号系统研究的一种尝试。其中可能还有不少地方值得商榷。但是，可以肯定地说，作为一种多层符号系统来考察和研究汉字，这是一个十分可取的、富有成效的途径。

五、加强统计研究

音位、音节、声调等语音特征的统计研究，无疑对于以字音为主的编码方案的设计（如是否标调，如何标调），对于文字和代码中字母的配置，以及对于打字机键盘的设计都有重要意义。我所高玉振同志曾对普通话语音频率做过精密统计（共统计10万字）。例如，其中对元音频率的统计，百分比如下：

a：17.396	ə：11.665	ɿ：3.332	ã：0.113
o：7.037	e：8.175	i：22.012	ə̃：0.052
ɔ：3.569	ɛ：3.445	u：17.213	
ɤ：2.398	ɿ：1.052	y：2.521	

辅音频率的统计，百分比如下：

p：4.291	t'：2.933	tʂ：5.661	ɕ：3.640
p'：0.898	n：16.650	tʂ'：2.092	k：4.244
m：3.892	l：5.250	ʂ：6.788	k'：1.702
f：1.593	ts：3.593	ʐ：1.600	ŋ：11.520
v：0.268	ts'：0.898	tɕ：5.274	ŋ'：0.024
t：9.603	s：1.123	tɕ'：2.569	x：3.894

声调频率统计，百分比如下：

阴平：15.106	上声：1.692	半上：13.426	轻声：22.748
阳平：20.044	去声：23.760	半去：3.224	

字元和字的频率统计，无疑对于以字形为主的编码方案的设计，对于键盘设计都有重要参考价值。词的频率统计对于词汇码的确定，也有很大参考价值。但是，到目前为止，除去字汇方面有些资料可供参考外，其他方面几乎等于零。

总起来说，无论是字音还是字形方面的统计工作都还差得很多。为了更好地开展编码研究，有必要加强各种参数的统计。可喜的是，目前已可利用计算机来代替人做这种工作了。

最后还想就加强情报工作谈两句。汉字信息处理研究是为科技情报工作现代化服务的。它本身应该带头搞好情报工作。美、日等国、海外侨胞在汉字信息处理方面也已研究多年。他们的经验教训，应该加以借鉴。"他山之石，可以攻玉"，这是郭老早就指出的。

拉杂地谈了这样一些。不难看出，这是一个急就章。不妥之处，请指正。

中文编码原理研究

陈明远　　管汀鹭

（中国科学院声学所）　（中国科学院生物物理所）

摘　　要

本文扼要介绍作者们近十几年来在中文编码原理方面所作的一些探索性工作，尝试从数理语言学、工程心理学和计算机科学的各种角度，分析中文的规律：方块字的使用概率分布及信息量，字形的拓扑分类，字音的音位分析，等等。文中给出了中文打字的数理模型——符合响应系统，讨论了人—机器系统的操作特征。还比较了各种可能的编码方式，并提出了对于中文输入方式的几点原则性要求（鉴定标准）。

1. 中文（方块字）的特点

1.1 中文在目前世界上的地位

中文是联合国法定的六种工作语文之一。使用中文的人数世界上最多，除我国（包括台湾、香港）以外，还有新加坡及许多海外华侨，占世界总人口四分之一。日本和南朝鲜使用的“汉字”与我国读音不同，形体和意义也有差别，不是“中文”。中文打字、电报、排版的速度，不到英文的十分之一，是世界上效率最低的。国际上以“中文”为第二语文的人数的比例，不到英文的万分之一，是世界十大语文中通用率最低的。外国人学中文都愿意首先使用汉语拼音，很少有几个外国人能得心应手地掌握方块字。到七十年代中叶，世界上有10万种定期的科技学报，每年在学报上刊登150万篇科技文献，26000种定期性科技索引，每年出版25万种科技书，用60种文字印刷，其中英文占50.5%，而中文少于0.5%。[1] 我国七十年代上半叶平均每年出版物约30万万字，但译成英、法、德等外国文字并出版的，不到这个数量的1%。[2]

我国的文字机械化、自动化，离开现代世界水平差距很大，必须急起直追，迎头赶上。

1.2 中文的历史演变（简述）

中文的历史悠久，仅次于古埃及的圣书字和古巴比伦的丁头字，从公元前十四世纪的殷商甲古文到现在，共有三千多年。在这漫长的历史过程里，中文的字数不断增加，字形几度变化，字音逐渐更改。鲁迅说过：中文早已成了“不象形的象形字”。周有光指出[3]：中文的“形声字”声旁大半已不再具备表音的功能。

殷代的甲骨文和殷周金文只有2000字左右。东汉末年许慎（公元121年）的《说文解字》收入9353字；清朝的《康熙字典》（1715年）增加到47043字[4]。历代所增加的大都是“形声字”。郭沫若[5]指出：“看来汉字的数目大体上有五万字的光景。这五万字中，有绝大多数的字已经不使用了，目前一般知识分子所日常使用的大概有五、六千”。

关于我国语言、文字的历史发展，著作很多，兹不详引。我们在这里提出三个观点：（一）

必须把现代汉语、现代汉字与古代汉语、古代汉字严格区别开来，在电子计算机上作为两种语文进行处理（类似于古拉丁文和现代西方语文的区别）；(二)必须尽早制订《现代标准汉字表》，在字数、形体与读音上都标准化（规范化），不再翻来复去地改变；(三)必须同时考虑方块字与汉语拼音两种"中文"的电子计算机处理。

下面，我们着重对于现代中文的使用概率分布、字形的算法表示和拓扑分类、读音的音位分析，分别进行讨论。

1.3 中文的使用概率分布

1971年修订版《新华字典》收入7262字头，1964年制定的《印刷通用汉字字形表》收入6196字头，但其中有一千多字是几乎不用的（在每千字中平均用不到一次）。那么，现代中文用字的情况究竟如何？中文编码究竟应当处理多少字头？这个问题只有通过数理统计和工程心理学的科学方法，才能合理地作出答案。

从三十年代以来，我国先后有许多研究者对于总共约三千万字的书面材料进行了统计（包括老教育家陈鹤琴等人，及教育部、文改会、语言所、声学所、心理所、新华印刷厂……等。又，对于毛选四卷，许多单位作了重复性的用字统计）。我们应用统计数学和概率论的原理，对于中文统计的抽样方法、可靠性、误差范围进行了研究，并使用电子计算机作了大量运算。得到了从三十年代到七十年代的中文使用率统计分布（另有专文）。

分析的结果表明，中文的使用率确实有一定的规律。例如，最常用的一个"的"字，无论在三十年代或六、七一年代，它的出现率总是稳定在4%左右，即平均每25个字中出现一次。又如，在大量书刊中，3000常用字的出现率，总占99%左右。《毛泽东选集》前四卷共计66万字，总共使用的字头为2981个。还有一个规律，就是从三十年代到七十年代，常用字的出现率越来越集中。

对于中文使用率分布的另一种研究方法是工程心理学方法。这就是通过心理学实验，调查分析各种人的"符合响应系统"的内存储字库、音—字响应的概率分布、联想字的条件概型等等。

按不同文化水平分组，每组可有数人到数十人。在一定条件下进行"音—字响应"实验，由发音者唸出普通话字音表（包括1200个音节的随机排列），让听音人对每个音写下他首先反应的字。把大量的结果汇总，经过整理分析，就得到现代汉语的"基本字表"。如果再让听音人对每个音依次写出他所能联想的同音字，依照所写出的同音字的次序，也可以推算出"音字响应"分布。

这两种科学方法（数理语言学的统计方法和工程心理学的实验方法）所得的结果基本是一致的。前一方法处理的是书面文字材料，后一方法处理的是口语—心理材料，两种结论的一致证实了人的"符合响应系统"所具有的客观规律。

1.4 汉字形体的算法表示和拓扑分类

我们曾将汉字用"写字程序"加以表示[6]，就是把二维结构的汉字写成由基本笔划和算符按照笔顺排列成的线性表达式。或者说，所有汉字可以表达为一个"结构树"。基本笔划的某些组合，形成"字根"。字根的结合形式是各种拓扑关系。

首先对于中文形体进行拓扑分类的，是杜定友[7]和郑易里[8]。在他们之后，美国的Rankin，Siegel，McClelland和James Tan等[9]，日本的坂井（T. Sakai），长尾（M.Nagao）等[10]，以及W. Stallings.[11] Paul.P. Wang[12]等研究者也开始从拓扑结构方面分析汉字的形体。由我国学者丁西林、杜定友、郑易里等人所提出的概念，如：基本笔划（包括单笔、

复笔），单体与合体字，字根，汉字结构的拓扑分类，等等，已在不同程度上被各种汉字编码方案所采用。

字根组合形式可归纳为如下几种拓扑结构：

(1)单体字：□ 例字：王。

(2)合体字：A．左右字：  例字：明。

B．上下字：  例字：星。

C．包孕字：  例字：国。

左右结构还有：  湘，  绮，  邵，  韶；

上下结构还有： 善， 碧， 品， 琵；

包孕结构还有：  同，  凶，  匡，  反，  司，  近。

我们采纳丁西林等人[13]的一种意见，把汉字最基本笔划分为6种：

名　称	横	竖	撇	点(及捺)	勾	弯
笔形特征	一	丨	丿	丶	亅	乚

对于9000汉字作字形的起笔与末笔统计，结果如下：

笔　划	起笔		末笔	
	字　数	占百分比	字　数	占百分比
点（丶）	2240	24.9 %	1104	12.3 %
捺（㇏）	0		1942	21.6 %
横（一）	2753	30.58 %	1961	21.8 %
直（丨）	1290	14.33 %	871	9.6 %
撇（丿）	2087	23.19 %	99	1.1 %
勾（亅）	194	2.15 %	1962	21.8 %
弯（乚）	436	4.85 %	1061	11.8 %

方块字的笔划数目，未简化前是平均每字11.2划，简化后是9.8划。由于基本笔划数过多，所以长期以来许多人从事“复合笔划”或称为“字根”的研究，但是迄今没有得到统一的结论。

尽管如此，不论何种“字根分类法”的字根，都可以写成基本笔划的某种算法表达式，再加上概率统计分析，就可以把字根问题从数学角度得以解决。

1.5　中文读音的音位分析

如果说汉字的字形结构的规律性比较混乱，不便于单纯从字形的角度进行计算机处理，那么我们却看到汉语语音结构的规律性十分整齐，很便于进行计算机处理。

当我们从数理语言学特别是数理逻辑的角度来考察现代汉语的结构时，会发现它本身就符合进位制的体系。由二十几个声母、三十几个韵母、四个声调组成了语音结构的三维空间。从信息论观点来说，它本身就具备一种“组合编码”的形式。

我们使用电子计算机经过大规模的数据处理，算得声、韵、调的概率分布及信息量，和现代汉语（书面语）字音的概率分布及标准差总表（另行发表）。

从信息论的观点来看，现代方块字的独立信息量为9.5bit，普通话单音节的独立信息量为7.5bit。因此一个方块字的信息可分为两部分：字音信息和字形信息。我们认为可取前一

部分（字音）为主要信息，后一部分（字形）为附加信息。

从文字学的字音、字形关系来看，二十几个声母与三十几个韵母共可有600多个组合，其中400个组合是载有消息的，利用率为75%；400个无调音节与四个声调共可有1600个组合，其中1200个组合是载有消息的，利用率也是75%。

在普通话中，平均每个标调音节有5—7个同音字（按6000或8000通用字计算），但它们的分布是很不均匀的。同音字最多的音节（超过30字以上者）是BIH，FUL，JI，JIH，LIH，QIL，SHIH，XI，YIH，YUL，YUH，ZHIH等12个，仅占标调音节总数的1%（-L表示升调，即阳平，-H表示降调，即去声）。

2．中文打字的数理模型——符合响应系统

“符合响应系统”是我们用数理语言学及工程心理学方法研究“人—机”关系所提出的一个新的基本理论。由于数学性比较强，并且尝试应用现代正在发展的数学工具（FUZZY集合论、非标准分析、突变理论等），所以在这里不准备详述，只介绍一些基本概念。

2.1 语音集合、文字集合及层次概念

“符合响应系统”就是刺激——符合——响应——反应过程的机制。

人的内存储有两个空间：声音语空间和书面语空间。声音语空间包括音素平面、音节平面、语词平面、语句平面；书面语（文字）空间包括字符（字母或笔划）平面、文词平面、文句平面。每个平面是一些元素的集合，各元素有一定的概率分布，元素之间各有一定的距离，这些距离依它们的“区别性特征”而定。各元素可以按“区别性特征”进行内部编码，内部码之间的“差”就是它们的“距离”。例如汉语音节平面是汉语所有音节的集合，各音节有相应的概率及条件概率（在一个音节后面出现另一音节的概率）分布……而且，这些元素（音节）的集合是“模糊”集合。就是说，它们不是确定不变的“点”，而是一些相对模糊的“域”。人耳听到的口语音节，是相当“模糊”的。特别是方言的发音有许多差异。例如南方人说普通话，往往Z——ZH、C——CH、S——SH不分，EN——ENG、IN——ING不分，甚至有些方言N——L不分，H——F不分。此外，同是标准的普通话，个人差异也是存在的。例如男、女、老、幼的语言频谱差异也相当大。但是人们相互交谈，怎么会听得懂呢？这主要是由于各平面之间“综合”、“分析”与“对比”、“联想”等等的作用。

实际上，无论是口授打字（听讲话记录）或复写打字（看原稿誊清），都是人对于语言及文字的输入输出过程。打字者听到口授的一连串音节，或看到已写好的一连串字迹，在他的神经系统中，就引起一连串的、分层次的“符合”和“响应”，也就是对输入信号进行加工，在他的存储空间（包括声音语和书面语——文字空间）中加以对比、联想、分析、综合、转换等等分层处理。

2.2 识别与联想

第一个处理过程是“输入符合”或者叫“识别”过程。单独听一个音节时，我们往往不能确定这是什么字，而只作可能性的猜测。但如果听一个词或词组（双音节、三音节），多半就能判断出这两个（或三个）音对应着什么字了。如果听一个语句，那几乎百分之百地能判断这语句写出来是一串什么字了。这在声音语空间里，就是从音节平面向上一层语词平面的“综合”过程，以及从语词平面向文字语空间的“响应”过程，等等。有时候（在多数情况下）人听到的语言不是清晰、准确的，而是模糊的（如带有口音、方言音），甚至含有错误的

成分。要通过符合响应过程使模糊的输入信号清晰化，使包含错误的输入信号正确化（纠错）。在音节平面中相应元素的组合，经过“综合”对应于上一层次——语词平面的元素，进一步综合对应着更上一层次平面——语句平面中的模式。然后由语句平面“分析”到语词平面，再分析到音节平面。原来的“模糊”的输入信号，这时就成为“确定”的元素组合了。当然，这过程取决于㈠输入信号的模糊程度、错误程度；㈡人本身的声音语空间和文字语空间的内存储情况（认识方块字的多少，听方言发音的能力等等）。最后，达到输入的信号串与内存储记忆的某组元素串完全符合，也就是完成了“识别”过程。但是，如果输入的信号与内存储的相应元素距离过大或造成错位、混淆，就会造成“误解”或“拒绝”（听错或听不懂）的情况。

还有一个处理过程是“联想”。听到某一个音节，可以产生许多“同音字”的联想，这是“符合联想”。看到某一个字，可以产生与该字相连组成双音词或词组的许多字的联想，这是“响应联想”。在与字母打字机有关的联想过程中，最重要的是字母的响应联想，就是打出一个字母后紧跟着第二个字母可能是什么。这就存在着某种条件概率分布。对于中文编码输入说来，操作员的记忆库里主要的联想过程是“方块字——字音和字形”的符合联想。

“符合”关系大致是“一一”对应或“多一”对应的，“联想”关系大致是“一多”对应的。

最后，在打字过程中，由操作系统——手指的动作进行输出。我们应用数理逻辑、数理统计及信息论控制论的方式，把这个全过程化为一种定性、定量化的模型，由此进行合理化的运算，作出中文编码的设计方案。

2.3 打字过程的工程心理学研究

打字操作是一种主要的人—机联系方式。手指和键盘之间的配合问题，是提高打字效率的最关键问题。

众所周知，法文键盘（AZERTY）是与英文键盘（QWERT）排列不同的。必须根据汉字编码的特点和不同的需要，设计若干种专用的中文输入键盘。

键盘的键数以多少为最佳呢？这是一个工程心理学问题。中文打字，可以有几千个键，也可以少到二十几个键（小键盘）甚至十个键（四码电报用）。日本实务用字研究协会（S.T.B）发表过日本打字员平均每分钟打字速度如下：[15]

拉丁字母打字机（26键）	450击
五十假名打字机（50键）	250击
汉字假名混合打字机（2000键）	50击

本世纪，打字机的效率问题愈来愈成为一个突出的问题。国外许多研究者总结出“触打法”。触打法规定两手的八个手指分管三排字母键，以中排为“导键”，即各手指常在位置，大拇指专管空位（间隔）。打字时，不用眼睛看键，只有手指按一定的规则操作；当一个手指击键时，另外七个手指处于准备状态。这类似于弹钢琴——看谱不看键，以提高效率。使用这种“触打法”，英文字母打字速度平均为每秒8—10击，最高可达14击。

从“触打法”的理论与实践看来，三排、不超过30键的小键盘是最可取的。

还有一个问题就是各键位的排列问题。

根据国外的工程心理学实验数据，[16] 同一手指敲击运动间隔平均为0.09秒（同手连击间隔）；同一手的不同手指之间敲击运动的间隔为0.03秒（同手轮击间隔）；不同手的手指之间敲

击运动的间隔为0.02秒(左右轮击间隔)。

根据我们初步实验结果，人的各个手指每分钟最多连续敲击次数如下表：

手　指	左　手	右　手
食　指	400	420
中　指	360	380
四　指	330	360
小　指	280	300

不仅是各手指敲击速度不同，并且敲击力及疲劳程度更为不同。这是很容易理解的：小手指最弱，最易疲劳，四指、中指依次增强，而食指灵活性最强。对于绝大多数人（除了左撇子）右手比左手更方便。

在汉字编码输入问题上，如果片面地只强调减少击键次数，而不注意键盘排列问题，那并不能保证高速度。提高汉字输入效率的关键之一是设计最佳键盘。

2.4　最佳键盘的设计问题

国外许多专家一致认为：目前使用的英文打字键盘QWERT是不合理的，许多人发出了要求改进键盘的呼声。[17] 看来，英文键盘（特别在排字工业上）必将全盘改动，并且已经出现了新的设计和产品，如PALANTYPE。因此，我们完全不必沿用不合理的英文字母键盘。

对于字母键盘的合理化要求是：

(1)中排字键应安排最常用的字母码，上排负担其次，下排负担应最少；

(2)同一手指越排连击的次数尽量少；

(3)食指、中指、四指、小指所负担的工作量应当依次降低；

(4)右手总负担应略大于左手总负担；

(5)打字时应尽量做到左右手交替。

对于键盘的这些要求，可以化为数理逻辑和数理统计的运算。我们经过十几年的研究，已经设计出汉语拼音的DUIHGE打字机和中文音形码输入键盘。

还有一种新的高速打字法，叫做“并时打字”，所使用的是“速记键盘”。它的主要设计思想是：在速记键盘上由几个手指同时动作，一下子打出几个符号的组合，代表一个音节，印在专门的纸带（或穿孔带）上；记录完了以后进行整理，翻译成通常的文字。这种速记符号的翻译改写也可以由电子计算机进行。速记键盘的速度可比打字机快三、四倍，完全能跟上人的讲话速度。我们也设计了汉语的速记键盘SUJIPAN。

键盘设计是中文编码问题的一个重要组成部分。不论哪一种字码，都可以而且应该设计相应的键盘，以便利使用，提高效率。

关于这方面的问题请参看我们的另一组研究报告《电子计算机的汉语信息处理》。

诚谢：我们从事本项研究十几年来，得到华罗庚教授、王力教授、周有光教授的指导和帮助，在此向这三位前辈表示衷心感谢。已故陈越同志热心提供了宝贵资料，已故王宗桥同志在使用电子计算机的过程中给予了方便，在此向他们两位致以深切的悼念。

参　考　文　献

【1】Bauer, C.K., International Information System for Physical Scientists, Lockheed-Georgia Com-