Xiandai Hanyu Yongzi Xinxi Fenxi

陈原
主编

现代汉语用字信息分析

上海教育出版社

Xiandai Hanyu

现代汉语用字信息分析

Yongzi Xinxi Fenxi

陈 原 主编

● 上海教育出版社

(沪)新登字107号

现代汉语用字信息分析

陈　原 主编

上海教育出版社出版发行

(上海永福路123号)

各地新华书店经销　上海市印刷十二厂印刷

开本850×1156　1/32　印张6.5　插页5　字数148,000

1993年12月第1版　1993年12月第1次印刷

印数 1—1,000本

ISBN 7-5320-3138-1/G·3067　定价:(精)11.50元

作者简介

陈　原　研究员(语言文字应用研究所)
刘连元　研究员(国家语言文字工作委员会文字管理司)
常宝儒　教　授(北京语言学院语言教学研究所)
康加深　助理研究员(语言文字应用研究所)
李　燕　助理研究员(语言文字应用研究所)
张一清　助理研究员(语言文字应用研究所)
佟乐泉　副研究员(语言文字应用研究所)
傅永和　研究员(语言文字应用研究所)
张书岩　助理研究员(语言文字应用研究所)

目　录

没有今日的基础研究就没有明日的开拓和应用〔导论〕

陈　原

1　汉字的基础研究——

没有今日的基础研究，就没有明日的开拓和应用。我确信这样的论断。正是从这样的认识出发，我和我可敬的同道，在研究语言文字的应用过程中，从来没有忘记进行语言文字基础问题的探索。这部集体著作《现代汉语用字信息分析》，正如1989年出版的第一部集体著作《现代汉语定量分析》①一样，在很大程度上几乎可以说，也是一种对现代汉语的基础研究结果。

六年前我在语言文字应用研究所主办的汉字问题学术讨论

会②上曾经说过，对汉字的研究分析已经越出了传统文字学范畴，很多学科对汉字的研究作出了新的贡献，其中包括心理学、教育学、人类学、社会学、社会语言学、神经生理学和神经心理学、信息论、控制论、系统论、电子学、机器人学(人工智能学)、群众媒介学、音声学以及文字改革等领域的许多专家在考察和研究汉字和汉字书写系统中，不断获得新的成果。

这些新成果部分地记录在上面提到过的《现代汉语定量分析》一书中。这些成果主要是八十年代对现代汉语若干要素进行的数量测定，即在这期间进行的多次规模巨大的语言工程。毫无疑义对现代汉语书写系统进行的基础研究对此后的语言工作是很有意义的。

此书出版后三年间，对现代汉语用字——汉字——的研究继续深化，进一步取得了一些基础数据和作了有效应用。目前这部取名为《现代汉语用字信息分析》记录了这些基础研究的若干成果。简言之，它力图从信息科学的角度出发，分析汉字和汉字书写系统一些基础特征。自然，这只不过是小小的局部成果，但尽管如此，公之于世还是有益处的。这部小书可以认为是我在上一部集体著作《序论》中所提到的"应用(实用)语言学讲座"中的第二种。我当时使用的术语"应用(实用)语言学"是不确切的，前年我在海外作研究工作时发现，不如使用"应用社会语言学"③这个早已存在的术语更为确切。

应用社会语言学的内涵，简单地说，就是对一些语言现象或语言文字要素进行社会语言学的考察和分析，从而将分析的结果应用到语言规划上来——也就是对某一特定社会群体使用的语言文字进行一番整理，使它更加有效适应社会信息交际的需要。这里说的"语言规划"是广义的，当然也包括我们中国读者几十年来日日接触的"文字改革"在内。西方一个著名的社会

语言学家费希曼(J. Fishman)④带着感情说过以下的一段话：

"语言规划作为一个理性的和技术的进程，事先有符合实际的数据，事后在进行中又有反馈，这当然至今还是一个梦想，但是无论如何，这个梦现在已不是十年前那样可望而不可即的了。"

他说得真好。因为我国有多少可敬的专业和业余语言文字工作者在长达半个世纪甚至一个世纪的不疲倦的奋斗中，进行过很多基础研究和实际应用，为的就是实现这个"梦"。换句话说，就是要使我们民族习用的传播媒介(其中特别指文字)，能够更加有效地适应信息社会的挑战和需要。

2 拓扑学与汉字研究——

把拓扑学理论导入汉字研究(特别是对字形的研究)中去，这是本书《汉字拓扑结构分析》一文的主题。

拓扑学虽然是上个世纪创立的一个数学分支，但是随着近年信息科学的发展，它被赋予了新的生命，打开了新的前景⑤。拓扑学研究几何图形在连续变形(continuous transformation，例如在弯曲、伸缩，却又不致破裂或粘合的变形)中保持不变的性质(invariant properties)。几何图形这种不变性质，称为拓扑性质；使几何图形保持拓扑性质的种种变形，称为拓扑变换；在拓扑变换下的种种变形，称为同胚变形 (homeomorphism)。社会日常生活中容易看到很有趣味的拓扑变换——最简单的例子是在一块擦字橡皮上，随便画一个几何图形，然后用手将这块橡皮扭曲，橡皮上的几何图形随之而变形——如果这个几何图形是圆形，扭曲之后它会变成椭圆或不规则的封闭线圈，但是无论变得多利害，这个图像尽管已变成不是通常的圆圈，但它却仍然是封闭的，它的线条决不会因变形而开口。这

种不变性质就是拓扑性质。人在哈哈镜前变形，长了，胖了，矮了，瘦了，但是人的两只耳朵决不会粘合在一起，而人的一个鼻子总不会撕成两个。

汉字也有这样的拓扑性质。当一个汉字变成长仿宋或变成扁楷体时，它总保持着某些不变性质，比方说洁字的"氵"旁永远是在"吉"的另一边，它不会跳到"吉"的"口"字当中去。这就是最简单的拓扑性质的例子。由这种不变性质所规定的结构，就是拓扑结构——应用这种理论来分析、整理汉字，汉字的这种不变性质就是汉字的拓扑结构。作者在这篇简明的入门论文中，介绍了汉字概念的三个层次，即图形，字形和字符这样三个层次，由是研究分析各个层次的同胚或不同胚，这样就可以在认读、书写、理解(认知)汉字过程中采取有效的方法来达到预期的目的甚至加强效果。随着计算机文字技术的发展，产生了所谓字形信息处理的许多理论和实际问题，拓扑学应用在汉字研究上已经被提上日程了。它将越来越显得重要。

论文没有使用艰深的数学分析，语言学工作者都能很快就熟悉它所讨论的内容。

3 汉字属性/形声字研究——

汉字属性问题在八十年代初曾经困扰过我们的读书界，人们带着迷惘的眼光问道：汉字属性是什么？

从本质上说，汉字属性问题就是对汉字本身所蕴藏的信息以及汉字对它周边延伸引导的信息运动进行质的和量的分析问题——正因为这样，汉字属性问题是语言信息学或信息论语言学所要考察的一个重要问题。过去十年间围绕这个问题所得到的数据，引起许多语言文字工作者和信息科学工作者的广泛注意；八十年代末期先后出版了两部篇幅浩繁的汉字信息（汉字

属性)词典⑥，就是一个证明。

《现代汉语定量分析》收录的《汉字属性字典的编制》一文，对这个问题作过初步的探讨；收在本卷中的《论汉字属性》则把这个问题引导到深入的境界，从而补足了前一篇所缺少的理论概括。这篇论文第二节和第三节，对汉字字形因素(笔画)和字音因素(音节和声调)进行的定量分析，是饶有兴味的，而且是很有启发意义的。这里给出的数据是基础数据，值得注意。前人对此曾经做过一些分析，这里公布的调查结果也许可以说是前人研究的继续。由于这些研究主要以八十年代几次大规模的语言工程所得数据为依据，因而具有切合实际的科学价值。

论文作者对三种高频字表作了调查统计，得出的结论是：最常用的高频汉字有一半以上都分布在六画(六笔)到九画(九笔)这四个笔画范围内。而在字数更多的字表中——例如在1988年公布的《现代汉语通用字表》7000字中，有2355个是在四至九笔的区域内，换言之，即四至九笔的汉字在七千字的大范围内占到33.64%。这两个数据说明了，六到九画的汉字在常用(高频)汉字中占一半强，在通用(一般)汉字中占三分之一。这个结论同齐普夫对现代拼音文字所做的论断——即最常用的词是由很少几个字母组成的单音词——本质上是近似的。这个数据对实现汉字教学、认知、记忆和重现过程，有重要的启发作用，因此对于语文教育学、社会语言学和心理语言学的研究都很有用。

对汉字音节和声调的调查研究结果表明了一个饶有兴味的事实。对七十年代以来大陆销行最广、几乎达到家家户户都存有的《现代汉语词典》进行分析的结果，在一万个汉字(10567)中读第四声的汉字有3453个，占32.67%，约为三分之一——依次递减为第一声、第三声、第二声。

对3500字的《现代汉语常用字表》作的调查结果，也得出

了与上面数据相近似的结果。3500个汉字中有469个是同形异音字。因此从语音角度出发，这里共有3969(而不是3500)个“汉字”，分布在405个不带调音节和1183个带调音节中——其中读第四声的有1339个(包括324个音节)，约占29.8%，不到三分之一。

仅仅举出上面的数据，就可以明了：作这样的基础研究，对于许多有关学科和语文教育实践都具有可资利用的价值（而这篇论文当然还不止这一项数据）。

从这里出发，论文对通用汉字的发展趋势作了推断——这实际上是现代汉语用字趋繁还是趋简的推断。在众多的论述中，作者引用了当代外国语言学家提出的语言“经济原则”和“省力性原则”，（可惜文中没有展开理论探讨）认为用字简化符合社会生活的发展。时至今日，也许已经很少人抗拒这个论点，但是确实还有少数人坚持不同的意见，这不要紧，学术问题能(而且只能)通过心平气和的争辩来解决——或者永远得不到解决。但现实生活却在前进，这是不以个人的主观意志为转移的。正如我最近为外国一家语言学杂志所写的论文所说：⑦

“今年(按：1992)七月一日，中国大陆最有影响的传媒《人民日报》(海外版)改排简化字，这意味着繁体字在传媒中最后的“堡垒”终于悄然消失了，从而结束了长达七年传媒繁简并存并由此每年引起激烈争辩的奇妙局面。这个现象可以理解为现代汉语某些演变(变异)趋势是不可抗拒的。”

我把这称为“不可抗拒的趋势”——这是本书不曾阐明的，而在客观世界中确实存在的“规律性”的东西。⑧

在对汉字进行信息分析的过程中，最能吸引注意的语言现象——在某一种角度上说——是汉字的形声字。形声字是汉字所具有的独特性质(甚至可以称为一种很特别的拓扑性质，虽然这里牵涉到的并不完全是两维的平面几何图形)。本书所收关

于形声字信息分析的两篇论文，重新确定了“形声字”的范围和确认原则，据此，在7000个现代汉语通用字当中，属于形声结构的有5631个，约占通用字总数的80.5%。这里导入“形声结构”的概念，同时导入“声符”，“形符”的概念。根据这些概念统计的结果，在7000个通用汉字中总共有246个形符，而其中54个构字力很强的形符构成了4898个形声结构，约占形声结构总数的87%。在5631个形声结构中共包含了1325个不同的声符。从这些基础数据出发，论文分析了声符表音度和形符表义度，这正是我在《现代汉语定量分析》序论中所论述的，由定量分析回到定性分析的过程。这样，我们到达了前人（例如《现代汉字形声字字汇》一书⑨）所未曾详尽探索过的领域。

几年前我在奥地利的维特根斯坦国际学术讨论会以及其他会议上⑩，曾就汉字的这种形符加声符的语言现象作过分析——欧美一些语义哲学家听了这个分析，顿时对汉字发生浓厚的兴趣。我把某些形符称为“类别标态”（我用了可以释为“指示器”的indicator一词），我把声符称为“音声标志”，并认为“类别标志”（形符）带有某种语义信息，而“音声标志”（声符）则蕴藏着更多的语义信息，当这两部分语义信息密切融合到一起时，就显示出这个汉字的语义。

不过我的概括分析并没有达到本卷收录的两篇论文（《现代汉语形声字形符研究》和《现代汉语形声字声符研究》）的深度——看来，这项研究还可以进一步深入的。

4 一个实验报告——

《视觉因素在儿童书写汉字中的作用〔实验报告〕》一文把我们引入另外一个领域，这个领域是前人研究汉字时常常接触到而又没有“登堂入室”的地方。视觉，动觉对于接受信息的作

用，是近年来信息科学家所确认的——这关系到心理学、实验心理学、神经生理学以及其他周边学科，看来是一个很复杂的研究对象。

这是一个朴实无华的实验报告——这个实验设计了非常简单却又很有趣的五种条件，在北京城区一所小学的中年级随机挑选了70名在学儿童作为这项实验的被试，取得了很有启发性的数据，这些数据以及实验方法本身，都给语言文字工作者，特别是语言教育学研究者提示了富有教益的猜想、设想和联想。实验结果表明，在特定条件下，当视觉信息传入大脑发生障碍时——这障碍不是由生理因素而是由于环境因素产生的——，大脑接收到的信息不全面，于是产生了这样的后果，即大脑神经中枢对臂、腕、指发出的指令也就遇到一些困难。实验表明，在汉字书写过程中对视觉的剥夺程度每加深一步，书写的准确性就随之下降一步。因此，可以得出这样的推断：在书写汉字过程中视觉的参与非常重要，训练儿童书写汉字，不仅要让他们学会对臂、腕、指大小肌肉的控制，还要注意训练手眼的协调和配合。

晚近信息科学的进展，已经注意到视觉神经和听觉神经在接受语言信息中所起的作用，耳眼并用能收受最大信息量，已经得到理论上和实验上的证明。“video-tape”(录像带，包含着视觉信息和听觉信息的工具)远比录音带的效果好，早已进入人们的日常生活；新近的CD-ROM CDV或CD-I⑩一类的传媒就向传统的书报(即印在平面上的单凭视觉而不能给出听觉信息和动觉信息的传播工具)提出最有力的挑战。从电视新闻所能接受到的信息量比之从广播新闻所能接受的信息量要多得多。多年前控制论创始人维纳早已说过，在大脑的感觉皮层中，视觉与听觉面积之间的比例约为100:1。换句话说，如果将听觉皮层全部用于视觉，则信息的接收量约“相当于眼睛得到的信息量

的百分之一”⑫。

从信息角度对儿童用字，写字和识字的研究，是一个大可拓展的领域，所得结果必定大大有助于语文教育；同时也必定很有益于计算机的人工智能研究。

5 汉字的结构和构造成分——

汉字构造形态的研究，是用计算机进行汉字信息处理的一项重要的基础研究。没有对这方面认真细致的基础研究，就不可能满足高技术的需要。

根据信息学和拓扑学原理，现在可以认为：汉字是由一个一个部件构成的，不同的部件构成不同的汉字，而相同的部件出现在方框内不同的部位，也形成不同的汉字。部件是由笔画形成的，构成部件的笔画，按照约定俗成的惯例，在书写时有先后之分，这就是通常所说的笔顺——笔顺的约定俗成不是绝对随意的，它服从书写时的心理状态和书写时的技术方便。因此，对汉字的结构，部件，笔画，笔顺这几个要素的研究，是适应新技术需要的基础研究。

本书收载的一组研究成果，总名为《汉字结构和构造成分的基础研究》，是名实相符的。“汉字的结构”一节给出了迄今最详尽的合体字构造成分（结构成分）组合方式，绘制了最简明的合体字构造框图（结构框图）——其中包括由两个部件构成的九种结构方式，由三个部件构成的二十一种结构方式，由四个部件构成的二十种结构方式，由五个部件构成的二十种结构方式，由六个部件构成的十种结构方式，以及由七、八、九个部件构成的三，一，一种结构方式。这一节还采用层次分析法概括出十三种结构方式。这里给出的成果对于生成汉字有重大的理论意义和实用价值。第二节和第三节研究部件和部件的结构部位（即

在结构框图中出现的部位),应当认为这里提出的论点和给出的数据都是很有用的,其中论述部件的名称和对部件名称规范化的论点很值得注意,这无疑对语言文字教学和汉字信息处理、语音移入和口语(oral)通讯方面都有特殊的实践意义,第四节和第五节是对汉字笔画和笔顺最详尽的基础研究,也是综合了前人研究成果得出的概括,例如第四节提出汉字笔画的排序规则和第五节提出的笔顺规则,都是有创见的概括。这些概括可能在不同学者间产生不同意见,但即使有若干分歧意见,这里给出的数据和论点至少是有启发性的——不只对语文教学方面,而且对计算机文字处理方面。

6 字形和人名用字的规范化——

作为本书最后的部分,是两篇关于规范化的论述和设想——即关于汉字的字形规范的综合论述和关于人名用字规范的设想(创制"人名用字表"的建议),这里的论述都是建立在严谨认真的调查统计基础上的,对于进行汉字信息分析不啻是两项基础工程。

《汉字的字形规范》是近四十年来汉字规范化过程的历史概括,它的范围包括汉字历史形态变异的规范化,地理名词用字的规范化;计量用字的历史演变和规范;汉字本身演变过程中的规范化活动(包含自然演变的调节和人工干预即汉字简化活动)。人们常说,没有规范化就不能有现代化;或者换句话说,社会生活包括科学技术进步越来越要求规范化,没有规范化即达不到高速度和高效率的要求。在这个意思上对汉字在中国大陆近四十年的规范化过程作一次综合考察,是完全必要和急需的。

在社会用字规范化过程中,似乎人名用字比地名用字更复

杂些⑬——也许因为人的命名比地的命名更富有社会意义，更富有情感信息的原故。为了避免与别人雷同，或者说，为了突出自己特有的个性，人在命名时往往喜欢选用一些生僻字。这是可以理解的。我记得二次大战后我在英国遇见一对夫妇，男的原是英国人，战争时期空降到法国去同地下抵抗运动联系，女的原是法国人，是地下抵抗运动的活跃分子，他俩因此结识并结成夫妇，他们两人各取了一个其"怪"无比却又无比"普通"的"姓名"——男的叫"英国人"(Englishman)，女的叫"法国人"(Frenchman)，他们现在就使用这个地下时期所用的"隐名"或"假名"。这个极端的例子使我深深感到命名的社会性。在《现代汉语定量分析》一书中提到过在那"史无前例"的"文革"十年中，人名用字大量采用了"东"，"彪"，"向"，"卫"这样的字眼，而不愿或不敢采用花花草草那些历来认为美丽的字眼，也就是命名社会性的生动例子。

《人名用字调查和规范化设想》是以第三次人口普查的抽样作为研究基础的。两次抽样调查得到人名用字共计4542个(自然只限于汉族的人名用字)，这4542个汉字在《现代汉语通用字表》中只见3913个，其余629个字不见于通用字表——即七千通用字以外的汉字(也可以称为生僻字)，而通用字表中有3087字没有机会使用，可见即使以通用字表为依据，也还有很大的利用潜力。(当然，这个研究结果有若干局限性，局限性在于抽取的样本太小，比起整个汉族人口来只占很小的百分比，因此只能看出命名用字的倾向性，而不能认为是全面的科学分析。)这篇研究论文提出了制定人名用字表的建议，并且提出一些设想，这些设想一旦实现了，对于很多社会部门的信息存储和提取将会是很有益的。

7 汉字研究展望——

写了以上的六节，传来了朱德熙教授⑭今年七月十九日在美国加州斯坦福大学医院辞世的消息，引起我无限的惆怅。在1989年我和他曾约好年底去新加坡参加一个汉语学术讨论会。他早去了美国，便由美国直接去目的地了——而我却没有成行。没有成行的原因是众所周知的。但我为此始终感到遗憾，失掉向他请教的最后一次机会。幸而1986年底朱德熙同吕老(叔湘)一起参加那年在北京举办的第一届汉字问题学术讨论会，应我们的请求在开幕式上发表了有关汉字问题的长篇精辟讲话——现在留下来的这篇讲话，提及汉字问题的好几个方面的状况和前景。他在那里讲过，汉字可以说是一种语素文字——当然有极少数汉字不代表语素，只代表音节，但绝大部分汉字都是代表语素的；就汉字本身的构造看，汉字是由表意、表音的偏旁(形旁、声旁)和既不表意也不表音的记号组成的文字体系。他还指出，如果字形本身既不表音，也不表义，变成了抽象的记号时，那汉字可以说是一种纯记号文字；不过事实上并非如此，只有独体字才是纯粹的记号文字；合体字是由独体字组合而成的，组成合体字的独体字本身虽然也是记号，可是当它作为合体字的组成成分时，它是以有音有义的"字"的身份参加的。——这些精辟的意见，对于从事汉字信息分析的后人来说，是很有启发的。也是在那次讲话中，他指出过去研究文字学的人只讲字形，讲六书，对语言不感兴趣，这是传统文字学很大的弱点。他说我们研究汉字学，要突破这个框框。字形当然要研究，但尤其要研究汉字和汉语的关系。说得好极了。此刻，我想补充说，即使研究字形，也要突破传统的框框，例如需要从信息学、传播学和实验心理学等角度来探究汉字蕴藏的信息，这不只对语言学研究有益，而且对整个社会现代化有益的。