张时钊：书同文、文同码与无字库

一、书同文

［日期：2007-10-23］　　来源：　作者：张时钊　　［字体：大 中 小］

春秋战国五百多年，诸侯割据，各自为政，造成文字相异，迫使秦始皇要搞书同文。此后汉字字形几经变化，好像再没有提出这种任务。两千年后的今天，不到五十年工夫，大陆成功推广了简化字，海峡两岸文字就不一样了，又需要搞第二次书同文。文字本身虽没有阶级性，但也会受政治政策的影响。不过政策不符合文字发展的本性，恐怕也难起作用，如武则天造的字和第二简化字表，就都以失败告终。第一批简化字既已被大陆和世界最大多数华人所接受，决无倒退回去的道理。至于由于简化和归并造成的某些问题，在使用中完全可以理智地避免。实际上任何文字，包括繁体字，都可以找到这种会造成歧义的特例。同样不能把繁体字定为非法。我认为可以搞三五千字的常用字表，但不应该搞什么规范字表，限定汉字数。规范文字（包括规范字音、字义）通常应该由字典去承担。写文章总要求能让读者看得懂，采用读者熟识的字，不会用大量的生僻字，除非不想给别人看。为做到这点，作者可能翻字典而不会去查字表。我想现在最影响中文纯洁性的网络词语中的字母数字，不会纳入规范字表吧，但不能禁止许多人用它。如果不计偏旁替代产生的简化字，简繁不同的字就不多了，大陆青年即使没学过繁体字，现在接触多了也认得了，第一次常常是猜得的。因为大多数简体字来自草书或手写俗字，港澳台也是熟识的。简繁转换有时不一定是必要的，尤其对不对称的简繁字，不同时期、不同地区、不同的人用不相同的字，可能有不同含义，转换之后会丢掉一些信息。总之，我主张兼容并包，两岸交流多了，自然会书同文了。

限定汉字数的做法更要不得。现在的电脑只能使用字库里有的汉字，还要依靠种种输入法，已经扼杀了汉字的发展，逼出奇形怪状的网络词语。我们应该让电脑也能自由使用任何字形，由实践来选择、形成新时代的字集。比起上百万的英文字来，九、十万汉字并不算多。虽然绝大部分是死字，但不能抹杀它的存在，说不定某时某刻要用到它，甚至复活或获得新含义。我们搞文字的，都应该首先致力于研究汉字的科学排序法，通过自动组字软件，把电脑的输入码、内码统一为同一个，编出易用易查的字典，那么什么问题都解决了。

二、文同码

［日期：2007-10-23］　　来源：　作者：张时钊　　［字体：大 中 小］

我说的不是输入码，而是汉字内码。二十年前，两岸文字不只是简繁不同，而是更严重的内码不同，软件不配套时，会显示一堆乱码，要经过内码转换才能阅读。现在中日韩所有汉字都统一在统一码 Unicode 中，没有乱码问题了，而且可以简繁转换。但是问题并没有完全解决。按 Unicode 的计划，要把世界上所有文字都纳进来，任何一个不同的字形，都有一个唯一的内码，完整的内码要 4 个字节，可容纳 20 亿个码位。因为世界上到底有多少种文字，复杂的如汉字到底有多少个字，都不能一次确定，妥善安排，只能由各个国家或地区分次申请注册。汉字已申请到 7-8 万个码位，但不是连成一片，而是割裂成许多段。字太多了，容易出错，有网友发现所谓的“电脑错字”，也有重复的。更难的是输入法，如何从近十万个汉字中选取您需要的。最后，还是有缺字，尤其缺一些人名用字，使不少人办不了第二代身份证。有一本电子书“国学备览”，就要用一千个图片，用来显示字库里没有的汉字。这说明用扩大字库的方法是不能解决问题的。于是 Tom Bishop 和 Richard Cook 提出汉字描述语言 CDL，据说他们就在 Unicode 工作的。使用 CDL，任何汉字都可以用比它简单的汉字或部件，最后都可以用笔画组出。

比较一下中文和英文，英文字（词）虽然上百万，电脑里也永远不会缺字。原因在于：他们不是对字而是对字母编码的。字是开放的，可以任意造新字，且有自然的字典序。我们也改为对笔画编码，行吗？英文字母是线性排列的，汉字笔画是平面排列的，有可能笔画序列相同而汉字不同，怎么办？这个问题以及笔顺等等，都可以加一些约定来解决，难解决的还是汉字笔画数（平均 10-11）比英文字母数（平均 5-6）大一倍，码太长，也不直观。如果改为对部件编码，部件又太多。能不能将部件归并成 100 类，每类定一个高位为 1 的字节作为内码，只有该类内出现频率最高的部件直接用该码，其他部件则另加一个数码来分辨。这样，使两个高频部件的 10000 个组合能够囊括两千左右高频字，每字两字节。其他低频字，码长些，击键次数多一些也没有关系。这需要摸索试验，如果成功，输入码与内码相同，自然排序也有了，而且永不缺字，所有问题都解决了。为此，如果需要对常用字形作少量改变或限制，也是值得的，文字工具的改变引起汉字形态的一些变化，历史上就发生过，是正常现象。

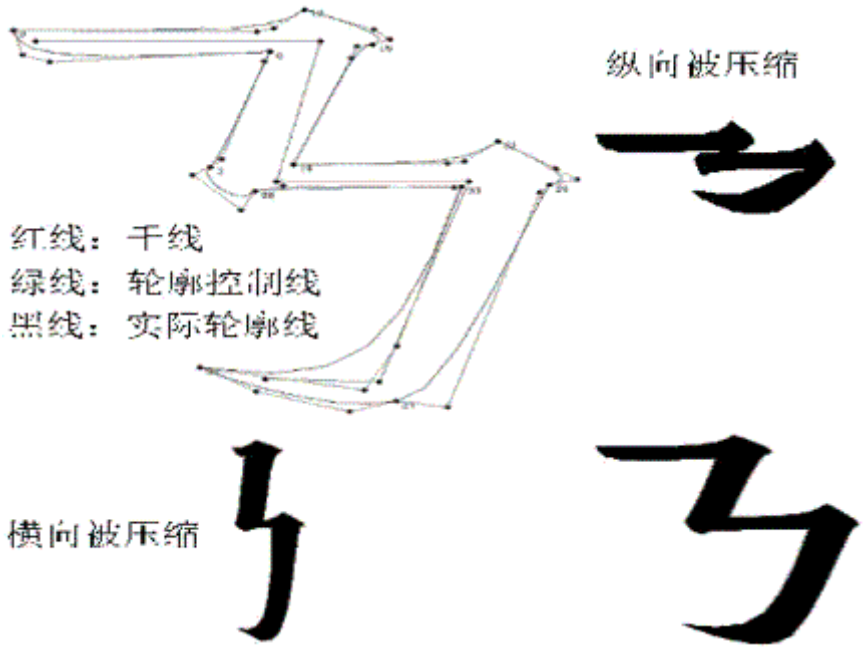
三、无字库

［日期：2007-10-23］　　来源：　　作者：张时钊　　［字体：大 中 小］

在汉字系统中使用组字技术，远不是我开始的。1985 年陕西省气象局领导要求我，把已开发的袖珍机 PC1500 无字库汉字搬到微机上去。因而接触到当时的苹果机时，就知道了

朱邦复先生的汉字系统，猜想他就是使用组字技术的。可惜这个课题很快就被停止了。到了新世纪之交，在网上联系到朱先生，下载了他的技术资料，但最终未能读懂。看来读懂别人的这种资料是很困难的，我写的也可能难以被人看懂。下面我尽量简单地介绍我的笔画组字概要。

2003 年，台湾易符公司叶健欣先生等 6 人来西安看我，他曾在朱先生手下工作过，当时与戚桐欣先生合作，是戚先生推荐了我。他向我介绍了他的无限字库及CDL（汉字描述语言）等。虽然因我的英文太差，未能通读所有资料，但基本搞懂了一般的组字办法是：给每个部件指定在汉字中的位置（左上角和右下角）或在两部件间加一个结构码（Unicode已有这种码）。我早在 1984 年搞的袖珍机无字库汉字，各部件已不必指定位置，其位置由各部件的笔画数按比例自动计算。每个部件都规定一个缺省的组字特性，只有小部分不按缺省特性组字的，才加结构符。进入新世纪，我又提出了由笔画组成部件（独体字）的层积理论。该理论认为，笔画都是按笔顺由上而下，逐步层积才形成汉字的。与横向笔画相交的，按笔顺必排在该笔画下部能到达的那个层，可以认为是它向上伸展了N层而发生的，在该笔画后加一个数N即可。凡左右并列的笔画，放在方括号内，被看作是一个层，可以各自上伸不同的高度，如果第一个笔画后加*N，则都从上一层开始上伸N层。这样一来，笔画序列就可正确地组成字形。所有汉字都可由笔画组出来了。但是，若用通常的图形拷贝方式，笔画和部件经过压缩、拉长等变形，笔画转角的特有形态及首尾笔锋就会变形，笔画粗细也不能保持一致，所以笔画只能用画线方式，即字体只能采用明线体。我循笔画的转折，在每笔中心设一条骨干折线，它可以伸缩变形。真正的笔画轮廓线采用贝塞尔曲线，其控制点则相对于最近的骨干线折点来定位。这样，问题就就解决了，如图所示。



我的软件已经可以组出所有字形了，只是字形还不够美观，显示速度跟不上。最近推出的小字库 WORD，原计划常用字采用字库字，罕用字才用组出的合成字，2.0 版实际上可用所有字库字，且字库字也可用来组字。