《香港语文建设通讯》，2006 年 8 月，第 84 期。

从人名用字谈汉语规范化

冯志伟（教育部语言文字应用研究所，北京）

近来，全国开始换发第二代身份证,今后的身份证需要通过网络认证。如果没有统一的汉字编码，生僻字将无法认证。据说，中国十三亿人口中，仅名字中带有生僻字的就多达六千多万人。全国各地报到公安部的生僻字（GB3212-80 中没有的汉字）达 3 万个之多，经公安部验证，起码有 4600 个字是目前方正的典码的 56000 字（目前最大的汉字字库）中没有的。这种情况说明，汉字不够用了！！这是对于国家汉字标准的严重挑战！！

下面是一些人名中的生僻字：“旻 珺 玥 弢 劼 璟 琤 赟 喆 堃 翀 璠 甦 彧 祎 韡 芃 濛 昇 暐 鈃 禤 垚 瑀 媺 嫚 孋 瑢 昉 犇 钰 銶 汭 蒯 忞 燮 镱 瑄 奊 骎 桊 玙 瓅 晞 颋 璘 燚 竸 鹍 嫶 珮”。

发音为 kun 的生僻字：“鲲 锟 髡 琨 醌 阃 悃 堃 堒 壸 壼 婫 猑 崐 崑 髡 梱 涃 潉 焜 熴 猑 瓂 硱 裍 稇 稛 綑 罤 臗 菎 蜫 裈 裍 褌 褌 貇 貇 錕 閫 閸 餛 馄 騉 髠 髨 鯤 鰥 鰥 鵾 鶤 鹍”。

各位文字专家看看，您认识多少个？

多年前，瑞典也发生过人名问题，因为瑞典人喜欢使用 Anderson 这个名字，结果，全国很多人都叫做 Anderson，彼此难以区分，出现了姓名管理的危机。在这种情况下，瑞典政府花了将近一年的时间让重名的人改名，终于解决了这个问题。这意味着，公民的姓名权是相对的。

但是，我国有关法律彼此之间还不协调，为了解决这个问题，我国的有关法律还需要协调一下。例如，《中华人民共和国民法通则》第九十九条规定： “公民享有姓名权，有权决定、使用和依照规定改变自己的姓名，禁止他人干涉、盗用、假冒。……” 第一百二十条规定： “公民的姓名权、肖像权、名誉权、荣誉权受到侵害的，有权要求停止侵害，恢复名誉，消除影响，赔礼道歉，并可以要求赔偿损失。”这些规定太严格，把“姓名权”看得至高无上，需要改一下。按照《中华人民共和国国家通用语言文字法》的要求，可以保留或使用繁体字、异体字的只有五种情况，其中虽然包括了“姓氏中的异体字”，却并没有包含名字当中的冷僻字。这样的规定为限制人名的冷僻字提供了法律保障，但是却挡住了限制姓氏异体字的大门。按照《中华人民共和国身份证法》第四条的要求，“居民身份证应使用规范汉字和符合国家标准的数字符号填写。”简单来讲，居民在取名时应该符合规范汉字的要求。这样的规定似乎又可以限制姓名中的冷僻字。

可见，我国的这些法规彼此还不十分协调，需要在更高的层次上进行调整。

我的意见是：对于命名实体的汉字问题以及化学名、生物名的命名的汉字问题，最好不要影响到原有的国家标准，如 GB4312-80,GB13000, GB18030。今后我们还可以向国际标准化组织争取更大的汉字字符集空间，但是，上限不要超过 6 万个左右。这个空间不是越大越好。对于各种特殊领域的汉字，分别制定特定的汉字用字表，例如，人名用字表，姓氏用字表，地名用字表，古籍用字表，化学用字表，昆虫用字表，方言用字表，...，等等。总的汉字字数不要超过 6 万个左右。6 万个左右的汉字肯定足够了，不能再超过这个上限了。

最近，方正公司制定了“人口信息生僻字方案”。该方案包括“人口信息字库”和“人口信息输入法”两部分。其中，“人口信息字库”共收录三万多汉字，已经涵盖了各地整理的全部地名、姓氏和人名用字。目前这个“人口信息生僻字方案”的“新典码输入法”技术已经通过了有关部门的检验。该项技术类似于在计算机上查字典，遇到生僻字，可以采取“偏旁部首查询”等方式，让计算机把生僻字翻出来。 但是，我觉得，人口信息的汉字数量似乎

太多了，而且有些是个人临时制造的只使用一次的怪字或者使用频度非常低的字，应该适当地缩减一些，并且把它变成一个封闭集。

我的理想是：从汉字字符集的总体来考虑，把汉字从特大的开放的字符集改变成一个特大的封闭的字符集，严格限制新造汉字，一定堵住新造汉字的各种源头和渠道。

我们应当从信息处理的角度来比较一下世界文字的各种字符集的情况：

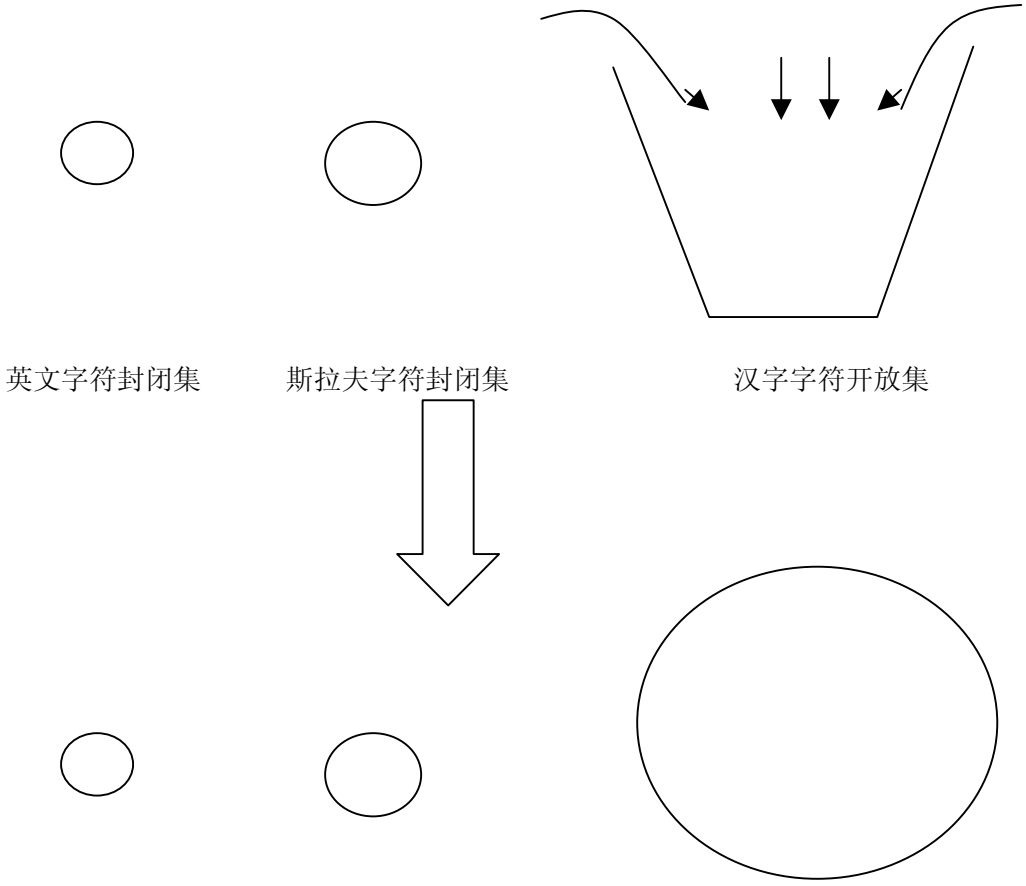
目前世界上的表音文字，其字符数目都很有限。英文字母 26 个，斯拉夫字母 33 个，阿尔明尼亚字母 38 个，塔米尔字母 36 个，缅甸字母 52 个，泰文字母 44 个，老挝字母 27 个，藏文字母 35 个，韩国文字母 24 个，日文假名 48 个。

英文字母只有 26 个字符，而且是一个封闭的集合，任何人，包括英国女皇、莎士比亚、爱因斯坦都不能随便给这个字符集增加新的字符。

斯拉夫字符有 33 个字符，也是一个封闭的集合，任何人，包括列宁、斯大林、托尔斯泰、普希金都不能随便给这个字符集增加新的字符。

在计算机上可以使用的汉字字符现在已经达到 60600 字符，而且，这个字符集是一个开放的集合，方正典码原有汉字 56000 个，加上典码中没有的人名汉字 4600 个，估计方正典码的总字数一共为 60600 个。目前即使我们承认方正典码的 60600 汉字是合法的，汉字还有进一步增加的可能性。例如，最近由北京时代瀚堂科技有限公司自主开发四字节汉字编码技术中的汉字就增加到 71500 个。如果没有严格的法律加以控制，任何人都还有可能继续给这个特大的字符集合增加新的字符，从而使得这个字符集合日益庞大，不堪重负。

对于这样大的开放字符集，实现全面的信息管理有很大的难度，我们应该清醒地认识到这个问题，搞好汉字的规范化和标准化，限制汉字的数量，改变汉字字符集的开放性质，并进一步探索有效的技术手段和方法，搞好中文信息处理工作，迎击信息时代对中华民族的挑战。目前首要的问题，就是把汉字这个特大的开放字符集改变为一个封闭字符集，并且有必要在语言文字政策上保证汉字字符集的封闭性质。我的这个意见可以图释如下：



英文字符封闭集　　斯拉夫字符封闭集　　汉字字符封闭集

（26 个字符）　　（33 个字符）　　（60600 个字符）

图　从汉字开放集到汉字封闭集

这个问题，实际上只是字符集这个平面上的规范标准问题，这只是初级标准和二级标准中的问题。

我认为，从国家语言文字数据管理的角度来看，语言文字的标准可以分为 5 级：

初级标准：包括汉字基本构成成分（如笔划、部件）以及汉语拼音字母的标准，这样的初级标准，在英语中就是字母标准（如 ASCII 字符集）早已完成并且非常成熟了，而在我国有的还没有做（例如，关于汉字笔划的标准）；有的还不完善（例如，关于部件的标准，《汉字统一部首表（草案）》有 201 部，但没有成为标准；国家语言文字工作委员会公布的《信息处理用 GB13000.1 字符集汉字部件规范》，在学术界还有争议，也没有成为正式的国家标准）。现在，面对信息化的要求，越南开始利用异拼定型，使越南文的各种标调符号与附加符号减少 50%，再利用约十五年到二十年的时间，达到文字的完全线型化，以提高本国信息化处理能力。汉语拼音的标调法不是线型化的，汉语拼音的有的字母的写法与国际标准的拉丁字母写法不一致，这些将给计算机处理带来麻烦，我们是否应当学习越南的办法，进行适当的改革？及早地考虑汉语拼音线型化的问题，以避免走越南拼音文字的弯路。我认为，初级标准的问题还没有完全解决。

二级标准：二级标准就是汉字字符集的标准。目前，国家标准 GB 18030-2000 《出版用汉字字符集》最多只包含 27484 汉字，而汉字的字符集总数估计有 6 万字左右（方正的典码有 56000 字），这样，在信息处理时，必定会出现计算机无法处理的汉字，在古籍整理、户口管理、科技术语、人名地名管理、化学物质命名、新的生物物种的命名中，在汉语方言的调查中，经常出现汉字不够用的问题。目前公安部采用方正公司的典码的 56000 超大字库再加上 4600 个新的生僻字来解决第二代身份证中的人名用字问题。随着人口的增加，还会出现新的人名生僻字，我们是否还要继续造新字？这说明，二级标准还不完善。

解决的办法是：

第一，　制定“全汉字字符集”的国家标准，使得不再出现计算机不能处理的汉字，如果我们承认方正“典码”以及他们提供的 4600 个人名用字都是合法的，把它们一股脑都纳入国家标准，那么，这个字符集可能要达到 6 万零 6 百个汉字（56000 字+4600 字=60600 字），一旦这个字符集成为国家标准，便原则上不允许新造汉字，使得全汉字字符集基本稳定，成为一个封闭的集合。

第二，　规定“基本汉字字符集”的字数，我过去曾经提出过“汉字容量极限定理”，认为可以把基本汉字的规模控制在 1 万 2 千个左右，这样的基本汉字规模，估计我国大约在晋朝时就形成了。汉赋中使用了大约 9100 多个汉字，在晋代李忱的《字林》中已经收集了汉字 12824 个。从信息科学的角度来看，在晋代时期（公元 265-420 年），汉字已经是能够完满地实现记录汉语功能的符号系统了，汉字已经不仅是在单个的文字上而且也在整个的系统上完全成熟的文字了。我曾经指出，在这 1 万 2 千多个汉字中，大约有 4000 个常用字，4000 个次常用字，4000 个罕用字；其他的汉字，在汉字发展的历史长河中曾经发挥过一定的作用，不过，现在已经成为“死字”了，基本上已经不再使用了。我觉得，在一般的汉字使用中，把基本汉字的规模限制在 1 万 2 千个左右是比较合适的。这个基本汉字字符集应当成为我国正在研制中的《规范汉字表》的基础。

第三， 在基本汉字字符集的 1 万 2 千个汉字中，常用字和次常用字为 8000 个，它们是最有用的汉字，也就是通用的汉字。台湾中央研究院郑锦全教授在统计了古代中国各个朝代的用字情况之后，提出了“字涯七千”的推测。所以，把通用汉字控制在 7000 - 8000 汉字是符合科学道理的。这些汉字应组织专门的班子认真研制和筛选，研制出初步的草案，初步草案应当在群众中进行试验，确实得到群众的认可之后再作为国家标准公布。

第四， 在基本汉字字符集当中，选出一些特殊用途的汉字制订特殊用途汉字集，如“人名用字集”“地名用字集”“方言用字集”“化学物质用字集”“动物新物种用字集”“植物新物种用字集”“方言用字集”等，严格规定特殊用途汉字集的数量，这些特殊用途的汉字原则上不要超过基本汉字字符集的范围。

第五， 改进命名的方式。对于姓名来说，可以改进姓名的命名方式。例如，提倡采用父母双姓命名（如：“郑张尚芳”），提倡使用双名（如：“冯志伟”），尽量减少单名（如：“李军”），提倡除名之外使用“字”（如：“白居易：白乐天”），尽量增加姓名结构的复杂性和区别特征。

第六， 对于方正提出的 4600 个人名用字重新进行审查，如果使用频度还比较高，可以选为人名用字，纳入“人名用字集”；如果是只使用一次或者使用频度很低的生造字，可以劝告当事人改名，不再作为合法的人名用字，不能纳入“人名用字集”。

第七， 把方正的“典码”以及他们提供的 4,600 个人名用字一股脑纳入国家标准的办法可能会带来的副作用，因为“典码”中的 56000 字实际上就是《汉语大字典》中的全部汉字加上 GBK 中的韩语汉字和日语汉字，《汉语大字典》是有字就录，其中有不少的“死字”或者是死去多年的“僵尸字”，这样的汉字只有记录的价值，而完全没有使用的价值。韩语汉字和日语汉字实际上并不是汉语中的汉字，而方正提供的 4600 个人名用字，其中许多是只用一次或者使用频度非常小的怪字或者私人临时造的特殊字，如果把这样的汉字也纳入国家标准，为了牵就少数人造的特殊汉字而给整个中华民族的汉字背上沉重的包袱，使我们整个中华民族都为这么沉重的大字符集的包袱而受累，似乎有些得不偿失，这样的策略是否明智？这样做是否会给我们的子孙后代带来麻烦和不利？这是应该慎重考虑的。另外，我们还应当从信息时代的要求来考虑，如果将来在所有的 Windows 上都装上这样大的字符集，是否会影响工作效率？这也是值得认真考虑的。

因此，我的另外一个建议是：以 GB18030 的 27484 汉字为基础，认真甄别方正“典码”以及他们提供的 4600 个人名用字中的汉字，如果确实是使用频度比较高的字，从中选择 500 个左右的汉字用来进一步补充 GB18030，使 GB18030 的汉字保持在 28000 个左右。我们可以把这个字符集叫做“扩充汉字字符集”。这个“扩充汉字字符集”中的汉字，原则上也应当封口，如果确实需要增加，可以从“全汉字字符集”中遴选，但是，不能新造汉字。

这样一来，我们可以建立四个级别不同的汉字字符集：

通用汉字字符集：8000 个左右（比 GB2312-80 的字数多一些）
基本汉字字符集：12000 个左右（在此基础上制订《规范汉字表》）
扩充汉字字符集：28000 个左右（大致相当于 GB18030 中的字数）
全汉字字符集：60600 个左右

全汉字字符集完全封口，基本汉字字符集以及扩充汉字字符集中的汉字都要从全汉字字符集中遴选，而特殊用途汉字集（如“人名用字集”“地名用字集”“方言用字集”“化学物质用字集”“动物新物种用字集”“植物新物种用字集”“方言用字集”）当中的汉字，原则上要从“基本汉字字符集”中遴选，少数可以从“扩充汉字字符集”中遴选，“全汉字字符集”减去“扩充汉字字符集”之后剩下的 32600 个汉字，实际上已经没有任何使用和流

通的价值，它们唯一的作用就是在字典中作为记录的条目，以显示它们在汉字的发展历史上曾经出现过的事实，它们只是“字典汉字”，而不是“流通汉字”，已经不能承担语言文字的交际功能了。这样做，也许有利于二级标准的稳定。是否恰当？请大家研究。

这个全汉字字符集应当由国家来管理，至于在全汉字字符集之外的其他汉字，例如，日本国字、韩国汉字、特殊情况下生造的汉字，可以由对它们有兴趣的民间机构来管理，因此这些字也就不能成为我国汉字标准化的对象了。

三级标准：三级标准就是词汇标准。我国除了 GB13725《信息处理用现代汉语分词规范》、GB/T 16159－1996《汉语拼音正词法基本规则》、CF1001-2001《第一批异形词整理表》等标准涉及到汉语的词汇之外，基本上还没有做三级标准。而在英语中，词汇标准已经完成了。在英语的书面文本中，单词与单词之间是有空白的，单词的界限基本上是清楚的（当然，英文中也有词例还原的问题，如 I’m 还原成 I am）。在汉语书面文本中，汉字是前后相续连写的，单词与单词之间没有空白，如何在汉语书面文本中发现单词成为了中文信息处理一个特殊的问题。由于汉语研究中“词” 的概念不清楚，给词界判断带来巨大的困难，造成了完全不必要的巨大消耗，使检索、排序和自动翻译等数据处理的歧义错误百出，大大降低了工作效率，甚至使本来可以从容完成的管理变得极其艰难，而且，即便可以容忍歧义错误，其成本代价之高也远远超出一般的想象。

词是自然语言信息处理的基本单位，词法分析、句法分析、语义分析都是建立在词的基础上的，我们应当明确地界定汉语中的“词”的数量和形式，使用“规则+词典”的方法来确定究竟什么是一个“词”。

因此，制订词汇标准的工作似乎也应当提到日程上来了。

四级标准：四级标准就是句法标准，这是信息处理中的高级标准。目前除了《信息处理用现代汉语词类标记集》的标准之外，我国基本上还没有做，而在英语和其他使用拉丁文字的语言中已经基本完成了。例如，英语的词类标记集就有 Penn Treebank 标记集（45 个标记）、C5（61 个标记）和 C7（146 个标记）等多个，用户可以根据不同的要求来选择不同的标记集。

五级标准：五级标准就是语义标准，这也是信息处理中的高级标准。目前我国已经研制了“知网”（HowNet）这样的语义系统，在英语和其他使用拉丁文字的语言中也研究“词网”（WordNet）或“框架网络”（FrameNet）这样的语义系统，可是，这些系统基本上是学者个人或单独的机构研究出来的，还没有进行标准化，因此，难以在语言文字数据的管理中普遍推广。看来，语义标准的建立，在全世界范围内都是一个亟待解决的问题。

所以，我认为，面对信息化的要求，我国的语言文字标准化工作还应当做很多的事情。为了实现中文数据的全面管理，我们还有很长的路要走，我们不能懈怠，我们还要努力！

参考文献

[1] 郑锦全，从计量理解汉语认知，《汉语计量与计算研究》，香港城市大学语言资讯科学研究中心，1998 年。

[2] 孙茂松，语言计算：信息科学技术中长期发展的战略制高点，《语言文字应用》，2005 年，第 3 期。

[3] D. Jurafsky, J. Martin,（冯志伟 孙乐 译），自然语言处理综论，电子工业出版社，2005 年。

附录：

讨　论　　姚德怀

1. 冯志伟先生的文章似只从内地人名用字问题出发，似未涉及台港澳以及海外华人用字问题。谈这个问题，又不能不涉及繁体字。冯文第 2 段所举的“生僻字”不少是繁体字。而其中一些字，在境外并不少见，有些甚至是重要的，如“崑曲”中的“崑”（昆），“畢昇”中的“昇”（毕升），“鰥寡孤独”中的“鰥”，“淳于髡”的“髡”，……
2. 我常常想起与冯文有关的一些实际问题。例如台港澳居民进入大陆，证件上的姓名都是繁体字，但出入境似无困难。实际情况如何，我还未深入了解。
3. 持外国护照入中国境的华裔人士，护照上姓名项只用罗马拼音（各种各样的罗马拼音，不一定是汉语拼音）。罗马拼音 + 护照号码 + 出生日期大概便能确定一个人的身份。
4. 据说内地发身份证 或 护照时，遇到生僻字，便替当事人改用一个常用字。（是否如此？）因此该人便有一个“证件姓名”和一个“本来姓名”。
5. 一人多名现象早已有之。如以前的文人多有“名”+“号”；作家的真名 +“笔名”；艺人的真名 +“艺名”；本来不太好的名字 + 自己喜欢的新名；真名 +“化名”等等。
6. 展望将来，可能有些人会有至少两个姓名：本来姓名及“证件姓名”。换言之，有“真名”和“别名”，英语用 alias 来表示。
7. 再谈谈汉字的“封闭集”及“开放集”。

   封闭集 便是冯文中提到的“全汉字字符集”的国家标准，是计算机能处理的汉字。

   开放集 可包括 (1) 封闭集之外的“汉字”，例如“死字”，(2) 新字及所谓“生造字”，(3) “汉字型字符”或 (4)“方块字型字符”，等等。

   例：107 号元素 bohrium 中文名称现定为“𨨏 + 波”，当时另一候选名“𨨏 + 玻”则可归入“开放集”。111 号元素 roentgenium 现定名为“𨨏 + 仑”，当时另一候选名“𨨏 + 伦”则可归入“开放集”。“𨨏 + 玻”、“𨨏 + 伦”将来都是“死字”，但是将来讨论化学元素命名经过时又不能不提到它们。换言之，“死字”可随时还魂。这种生死不定的矛盾难以解决。

   [参考《科技术语研究》2006 年第 1 期。姚注：bohrium 纪念丹麦物理学家玻尔（Niels Bohr），“玻尔”现在是否应改称“波尔”？roentgenium 纪念德国物理学家 X 光发现者伦琴（Wilhelm Roentgen，也曾译为欒琴），伦琴现在是否应改称“仑琴”？]
8. “字”与“字符”之间没有明显的界限。因此“开放集”里有些是“字”，有些只是“字符”。“开放集”里既有手写体，也可有“印刷体”。“开放集”里的印刷体也会在某些范围内流通。
9. “封闭集”由国家管理。
10. “开放集”可由国家管理（如果它有兴趣），也可由个人或专门机构管理。“开放集”将来可能成为一种 hobby，由对此有特殊爱好的人士（发烧友）组成业余的或专业的社团或 club 来管理。正如地球上的植物标本由设在 London 的 Kew Garden 来管理；天上的星星由国际天文组织来管理。这个组织可能设在北京，也可能设在与北京遥遥相对的智利的 Santiago 或任何其它地方。
11. “开放集”也可包括甲骨文，金文……。开放集里的字或字符要有一个公认的分类命名原则。正如植物品种有一个公认的命名原则。

    发烧友 club 例 1：香港有些朋友为粤语中有音无字的音节创造方块字。这些字，加上已有的粤语字，便构成“粤语字符集”。

    发烧友 club 例 2：内地有朋友为日本汉字（或称日本“国字”）拟汉音，这便构成“汉读日本汉字字符集”。（参考：费锦昌、松冈荣志，“日本‘国字’的汉语读音”，上海《语言文字周报》第 1155-6 号，2006 年 4 月 12 日、4 月 19 日）

发烧友 club 例 3：上述化学元素单一汉字命名组。

12.“开放集”总集可命名为“方块字型字符 集”。以上诸例均属“开放集”的子集。