编辑：左越
E-mail: zuoyue@cesi.cn

# 现行锡伯文满文编码存在的问题与解决方案

Existing Problems and Solutions for Current Sibe and Manchu Coding

新疆大学[1]　乌鲁木齐市索贝特数码科技有限公司[2]
付　勇[1,2]

**摘　要**　从文本编辑、字符排序和应用效率三个方面分析了现行锡伯文、满文编码存在的文字顺序和控制符等方面的问题与弊端，并给出了解决这些问题的可行方案，包括对相应文种的规范字符给出对应的编码等。

**关键词**　锡伯文　满文　编码　排序

**Abstract:** From three aspects - text editing, characters sorting, efficiently using, the thesis analyzes the problems of current coding on the Sibe and Manchu characters in word order, control character and etc., and gives out the feasible solutions for these problems, including giving corresponding coding for standardized characters in specified languages.

**Keywords:** Sibe language; Manchu language; coding; sorting

## 1 引言

现行锡伯文、满文编码是融入蒙古文编码之中的。2011 年 1 月发布的国家推荐性标准 GB/T 26226-2010《信息技术 蒙古文变形显现字符集和控制字符使用规则》规定了蒙古文 ( 标准所指蒙古文为传统蒙古文、托忒文、锡伯文、满文以及传统蒙古文阿礼嘎礼、托忒文阿礼嘎礼、满文阿礼嘎礼字母的集合 ) 变形显现字符的最小集合，还包括传统蒙古文、托忒文、锡伯文、满文的每个名义字符的变形显现形式的图形以及控制字符的使用规则。

该标准规定的蒙古文名义字符如图 1(GB/T 26226-2010 中为表 2) 所示。

图 1 所示表中，传统蒙古文是按照元音在前辅音在后这种传统蒙古文的字母顺序排列的，其他文种则以此为基础，附加了传统蒙古文中没有的托忒文、锡伯文、满文和阿礼嘎礼字母。图 1 中用蓝色框框住的是补充的锡伯文字母，红色框中的是补充的满文字母。该表中 180 列和 181 列给出了 10 个蒙古文符号、3 个自由变体选择符、1 个元音间隔符和


图 1　GB/T 26226-2010 中的蒙古文名义字符编码表

10 个蒙古文数字。从代码 1820(182 列 0 行 ) 所代表的元音字母 a ( 用独立式ᠠ表示 ) 开始则用 131 个编码位置表示蒙古文、托忒文、锡伯文、满文以及传统蒙古文阿礼嘎礼、托忒文阿礼嘎礼、满文阿礼嘎

**项目来源：**新疆维吾尔自治区科技型中小企业技术创新基金无偿资助项目“锡伯文、满文输入法的研发”，项目编号：2014531069。

礼字母。其中多数情况是：若字母有独立式则代码是字母独立式的代码，否则是字母词首式的代码，如果字母没有独立形式和词首形式则是字母词尾式的代码，这些具有代码的字符称之为名义形式或名义字符。字母的其余形式则没有代码，被称为变形显现形式也称作变体显现字符，是在特定上下文中使用的可选形式，依赖于该字符相对于其他字符的位置，也就是在一个单词中才会出现。然而，对于锡伯文、满文而言，这种编码方式则存在一些问题。

## 2 锡伯文、满文编码存在的问题

从应用的角度来看，GB/T 26226-2010 中给出的锡伯文、满文编码存在诸多问题。下面仅就三个典型问题加以讨论。

### 2.1 文本编辑中所表现出来的问题

在锡伯文、满文文本编辑 ( 如 PPT) 中，当输入一个锡伯文单词，例如：alin( 山 )，在小学语文教学中往往需要进行字母分解组合的解说，即将单词的每一个字母用空格分开，然后进行单词的组合讲解，如图 2 所示。

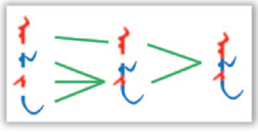

图 2　锡伯文单词的分解与组合

另外，在语言文字方面的文章和论文中也往往需要某个字母的某种字符形式单独出现，如双引号中需要有某个字母的词中形式或词尾形式等。但是，采用这种编码方案的输入法输入锡伯文的实际状况如图 3 所示。

图 3　采用蒙古文名义字符表的规定编码的锡伯文、满文单词插入空格后的变化示意图

采用蒙古文名义字符表规定编码方案的锡伯文输入法输入的单词 alin，在用空格将其字母分开后，结果成为图 3 右边所示的状态。

产生这种情况的原因是：采用 GB/T 26226-2010 规定的编码方案，输入的是其字母对应的名义字符的编码，当字母在单词中时，具有位置关系，可以按照字库设计的变形显现脚本语言，通过微软的变形显现处理程序 (usp10.dll) 将字母按位置关系转换成合适的显示形式。例如上面单词中的 a, 因为处在词首位置，因此用 a 的词首变形显现字符显示，同理，字母 l、i、n 分别显示为词中形式和词尾形式，出现的是一个正确的单词。然而，如果中间用空格分开，字母失去了在单词中的位置关系，则只能用名义字符的形式出现，结果就形成了图 3 右边的显示状态，这显然是我们不希望的情况。

### 2.2 文字顺序问题

GB/T 26226-2010 的表 3 给出了表 2 名义字符的名称，从 185D 开始给出的名称依次为蒙古文字母 . 锡伯 E、I、IY、UE、U、ANG、KA、GA、HA、PA、SHA、TA、DA、JA、FA、GAA、HAA、TSA、ZA、RAA、CHA、ZHA，满文 I、KA、RA、FA、ZHA( 上述的大写字母是 GB/T 26226-2010 对字母的命名 )；没有明确给出锡伯文的 (a)、 (o)、 (n)、 (b)、 (m)、 (l)、 (s)、 (ch)、 (zh)、 (r)、 (w)、 (k) 以及满文的第 6 元音 (v)，从具体使用情况来看，这些没有明确给出的锡伯文、满文字母的代码分别使用的是蒙古文 1820、1823、1828、182A、182E、182F、1830、1834、1835、1837、1838、183A 以及锡伯文 1861。其中的锡伯文按代码顺序可以列出如图 4 所示。

通常，文字信息在许多应用中是需要排序的，例如词典、数据库应用以及互联网的应用等。排序往往有以下方式：(1) 按英文字母排序，也就是用文本的拉丁字母转写方式排序；(2) 按汉语拼音方案的顺序排序；(3) 按传统排序方式排序，对于锡伯文、满文而言，就是采用“阿字头”字母出现的顺序排序。现代通常使用的是第一种或第二种方式，在一

| 名义字符 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 汉语拼音 | a | o | n | b | m | l | s | ch | zh | r | w | k | e | i | I | u | u |
| 标准代码 | 1820 | 1823 | 1828 | 182A | 182E | 182F | 1830 | 1834 | 1835 | 1837 | 1838 | 183A | 185D | 185E | 185F | 1860 | 1861 |
| 名义字符 | | | | | | | | | | | | | | | | | |
| 汉语拼音 | ng | K | G | H | p | sh | t | d | y | f | g | h | c | z | R | ch | zh |
| 标准代码 | 1862 | 1863 | 1864 | 1865 | 1866 | 1867 | 1868 | 1869 | 186A | 186B | 186C | 186D | 186E | 186F | 1870 | 1071 | 1872 |

**图 4　锡伯文字母在 GB/T 26226-2010 中的编码顺序**

些锡伯文、满文辞书中则常用第三种方式排序。然而图 4 中字母的顺序与上述三种方式均不相同，其顺序是极其混乱的，我们从图 4 表格中的汉语拼音就能看出。这在以后的应用中会造成极大的不便和困惑。

产生这种情况的原因是：这套编码系统字符的顺序首先是传统蒙古文，其次是托忒文、锡伯文和满文则仅作为该系统的补充附加在蒙古文、托忒文之后。

还有一个问题是其中的字符 (1861) 在锡伯文中应该是元音 (u)(1860) 在与辅音、、相拼时的字符形式，若作为锡伯文，按名义字符的编码规则 u 是不应该给出这个字符编码的，若作为满文第 6 元音，就不应该命名为“蒙古文字母 . 锡伯 U”（见 GB/T 26226-2010 第 6 页 )。另外，还有一个奇怪的现象是编码为 185D、185E、1863 的字符都有相应的独立形式或词首形式，却没有采用，似乎是在回避蒙古文字母 1821、1822 和 182C。再就是编号 1868 的字符没有采用 t 的词首第一形式，似乎也是在回避蒙古文字母 1832。

## 2.3　控制符的困惑与资源的浪费

由于采用了只给名义字母编码的形式，名义字符所对应的其他形式的字符如词首形式、词中形式、词尾形式等除了需要根据特定位置关系来确定外，经常还需要使用一些变形控制符来加以控制。这些控制符有：自由变体选择符 1 FVS1、自由变体选择符 2 FVS2、自由变体选择符 3 FVS3、元音间隔符 MVS，此外还有窄无间断空格符 NNBSP、零宽连接符 ZWJ、零宽禁连接符 ZWNJ 等。

下面通过两个典型示例看它们是如何使用的。

示例一：GB/T 26226-2010 第 41 页给出了输入字母的示例如图 5 所示。

| 名义字符 | | | 单个显现字符 | | | 认同 | | | | 字符序列 |
|---|---|---|---|---|---|---|---|---|---|---|
| 代码 | 字符 | 中、英文名称 | 序号 | 图形 | 中、英文名称或作用 | 蒙. | 托. | 锡. | 满. | |
| 1828 | | 蒙古文字母 NA<br>MONGOLIAN LETTER NA | —— | | *蒙字母. na 词首第一形式<br>ml. na first initial form | na | na | na | na | ZWJ |
| | | | 0097 | | 蒙字母. na 词首第二形式<br>ml. na second initial form | na | | | | FVS1 ZWJ |
| | | | 0005 | | *蒙字母. na 词中第一形式<br>ml. na first medial form | na | na | na | na | ZWJ ZWJ |
| | | | 0014 | | 蒙字母. na 词中第二形式<br>ml. na second medial form | na | na | na | na | ZWJ FVS1 ZWJ |
| | | | 0015 | | 蒙字母. na 词中第三形式<br>ml. na third medial form | na | na | na | na | ZWJ FVS2 ZWJ |
| | | | 0008 | | *蒙字母. na 词末形式<br>ml. na final form | na | na | na | na | ZWJ |

**图 5　GB/T 26226-2010 中的控制符的使用示例**

其中，第 5 列“图形”是要显示的字母 n 的不同形式，最后一列是输入的字符序列。从中可以看出输出 n 的带点的词中形式和带点的词尾形式竟然需要敲击四次不同的按键。

示例二：GB/T 26226-2010 第 23 页给出的锡伯文和满文使用窄无间断空格符的举例如图 6 所示。

显然，采用这种控制符的输入方式有两个非常明显的弊端：

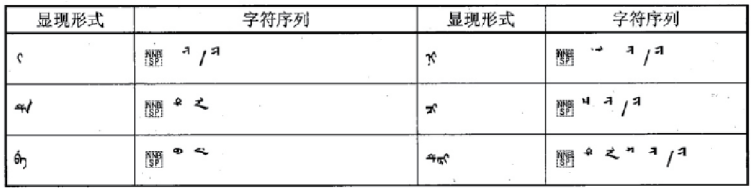


图 6　GB/T 26226-2010 中的锡伯文和满文窄无间断空格使用方法

(1) 这种采用控制符进行字符转换的方式对于普通民众和学生而言，使用上是非常困难的，即使对于专业人员而言也是难学易忘。

(2) 整篇文章中除了文字所对应的字符编码外，必然会有许多这样的不可见的控制符。一篇两篇尚且不明显，但社会发展必然会有大量的文章在计算机中存储和传输，日积月累，这些控制符所消耗的资源无论是存储空间还是传输成本都是不可估量的。从长远来看，会给国家和人民造成巨大的经济浪费。

## 3 解决方案

对于 2.1 中的问题，解决方案其实非常简单，就是对相应文种的规范字符 ( 指字母表中给出的各种独立形式、词首形式、词中形式和词尾形式，以下称规范字符形式 ) 都给出对应的编码，但不包括特定情况下的变形形式 ( 以下称特殊形式 )。例如：对锡伯文、满文字母 b 规范的词首形式、词中形式和词尾形式都进行编码，但不包括与 o、u 相拼时变形显现的特殊形式。这样，无论是在单词中插入空格或者对部分字母或单词进行复制粘贴操作等，字符都会以合适的形式显示，解决了字符形变所产生的令人困惑的问题。当然，采用这一方案需要对字符编码区间进行一定程度的扩展和调整。

对于 2.2 中的问题，应建立在第一个问题的解决方案的基础之上，有两种解决方案：

(1) 不同的文种按各自希望的顺序进行字符的编码。建立相应文种的编码字符表和字库文件，编码区间也可以重叠共用。这样就比较容易地满足各文种字符编码顺序要求和变形显现要求，实现起来简单方便。但这一方案可能有悖于国际标准的统一要求。

(2) 摒弃牺牲其他文种以一个文种为主的方式，将几种文字字符混合在一起，采用一种均认同的字母顺序进行编码。需要编码的字符由各文种按认同的字母顺序给出，字形相同且不影响字母顺序的字符可归并为一个共享的字符形式和编码。这种编码字符表按不同文种分离，都可形成按认同字符顺序满足各自文种要求的一个完备的字符编码表。这一方案需要相关部门统筹，各文种达成共识。这一方案的字库文件可以按文种或部分文种合并加以实现。如果采用统一字库，则由于变形显现可能会有不同方式而存在一致性问题，特别是手写体。

对于 2.3 中的问题，若建立在第一个问题解决方案的基础之上，对于锡伯文、满文而言，则完全不需要任何控制符就能实现，这已经被实践所证明。对于传统蒙古文和托忒文，也会极大地减少控制符的数量和提高应用效率。

## 4 结语

现行编码标准至少对于锡伯文、满文而言存在一定的问题，这些问题事实上已经影响了锡伯文、满文信息化建设的健康发展。解决这些问题所给出的方案是可行的，也在具体实践中得到了证明。乌鲁木齐市索贝特数码科技有限公司开发完成的“锡伯文满文输入法的研发”解决了第 2 节中提到的采用现行编码方式所产生的在单词中插入空格以及局部文本复制粘贴时字形发生变化的困惑问题，也实现了锡伯文和满文字符编码的汉语拼音序的对等有序性，并且没有使用任何控制符，简化了用户实际操作的复杂度。这一编码方案的实施极大地改善了锡伯文、满文的应用环境，也为锡伯文、满文信息化建设提供了更好的应用基础。

（收稿日期：2014-11-19）