文章编号:1007 - 2985(2013)02 - 0031 - 05

板塘苗文的计算机编码及字库创建*

莫礼平[1],周恺卿[2],蒋效会[1]

(1. 吉首大学信息科学与工程学院,湖南 吉首　416000;2. 马来西亚理工大学
计算机科学与信息系统系,马来西亚 柔佛州新山市　81310)

摘　要:民族文字信息处理研究对于保护民族文化遗产和弘扬民族文化有重要意义. 字符的计算机编码及字库创建是民族文字信息处理研究工作的基础内容. 在分析湘西板塘苗文的字形特点和简单介绍字符编码标准及字体技术的基础上,提出了 Windows 环境下基于 Unicode 标准的板塘苗文在计算机中的编码方案,给出了板塘苗文基于 Photoshop 技术的字模制作方法及基于 Truetype 技术的字库创建步骤.

关键词:板塘苗文;字符编码;Unicode 标准;Truetype 字库

中图分类号:TP317.2　　**文献标志码**:A　　**DOI**:10.3969/j.issn.1007 - 2985.2013.02.007

民族文字是传承民族文化的载体,体现了民族的尊严. 在旧中国漫长的历史中,湘西苗族有自己的民族语言,却没有本民族正式的文字. 有语言却没有正式文字,是湘西苗族地区长期贫穷落后的主要原因之一. 清朝末年以来,一些湘西苗族知识分子为了发展苗族文化教育,创制了不同的民间苗文[1-2]. 湘西花垣县龙潭镇苗族秀才石板塘创制的板塘苗文是使用较为广泛的一种民间苗文,当地人至今还用它创作、记录苗歌. 石板塘使用此种文字进行丰富的文学创作,开创了现代苗语书面文学的先河. 100 多年来,板塘苗文为苗族文化教育及经济事业的发展做出了重要贡献.

在当今信息时代,随着计算机和网络技术的飞速发展,民族文字信息化已经成为促进民族发展、保护族文化遗产、弘扬民族文化、使民族优秀文化走向世界的必要手段. 20 世纪 90 年代以来,以藏文、蒙文、维文为代表的多种少数民族文字紧跟汉字信息处理研究的步伐,在字信息处理层面和语信息处理层面的研究取得了显著成绩,并已广泛应用于生产生活的各个领域,使得民族文字的社会功能和经济功能得到了更好发挥. 然而,国内民间苗文的信息处理研究几乎没有,与苗文信息处理相关的研究也非常稀少,见诸报道的仅有 1988 年余乐教授[3]研制的计算机苗文处理系统和 1994 年吴光州等[3]研制的云南规范苗文计算机处理系统. 这 2 个系统的功能仅限于处理 DOS 系统下基于拉丁字母的云南新苗文,无法应用于 Windows 环境,更不能处理非拉丁字母的民间苗文.

笔者研究 Windows 环境下湘西板塘苗文在计算机内的字符编码表示方法、字模制作及字库创建方法.

* **收稿日期**:2012 - 12 - 03

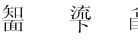
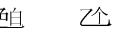
基金项目:湖南省教育厅青年项目资助(10B088)

作者简介:莫礼平(1972 -),女,湖南安化人,吉首大学信息科学与工程学院高级实验师,硕士,主要从事中文信息处理、Petri 网理论及应用、数据挖掘研究.

1 板塘苗文的字形特点

板塘苗文是一种仿汉字结构的方块文字,以假借汉字为主,利用苗语和汉语语音结构基本相同这一特点,根据"六书"的造字规律,借用汉字偏旁,使用汉字 1 个音节 1 个字的方法,运用形声、会意等手段进行创制. 板塘苗文几乎都是合体字,其构件主要来自汉字. 板塘苗文合体字的基本结构包括上下结构、左右结构、内外结构和侧围结构 4 种[1-2]. 表 1 给出了不同结构的板塘苗文示例及其汉义对照.

表 1 板塘苗文结构示例及其汉义对照

	结构类型			
	上下结构	左右结构	内外结构	侧围结构
板塘苗文字例				
汉义	相识 流 打瞌睡	布 砍 瞎	出去	白色 (一)个

2 板塘苗文的字符编码表示

2.1 编码标准

文字在计算机内的编码表示是字库组织的依据,也是文字信息处理的基础. 为了能被计算机处理和显示,每个板塘苗文在计算机内都必须有相应的编码表示. 编码方案和编码标准的选择决定了板塘苗文编码的应用范围.

不同国家和地区的不同语言文字有不同的编码标准. 中国汉字编码标准主要有汉字信息交换用编码标准 GB2312 (全称为"信息交换用汉字编码字符集·基本集")、汉字编码扩展规范 GBK 标准、最新的汉字编码字符集国家标准 GB18030,以及通行于台湾、香港地区的 BIG5 繁体字编码标准等. GBK 向下完全兼容 GB2312,向上支持 ISO 10646 与及 Unicode 国际标准. GB18030 向下完全兼容 GBK 和 GB2312 标准. 此外,中国还制定了藏、蒙、彝、维、朝文等少数民族文字信息处理的字符编码标准.

Unicode 标准是一套适用于全世界所有国家的字符编码标准,2012 年 2 月已推出了 Unicode 6.1 版本[4]. Unicode 标准将全世界所有主要文字和符号统一到一个字符集中,为每个字符提供一个唯一的特定数值,支持世界上所有语言的编码和转换,克服了传统的字符编码方式无法同时处理混合多种语言的局限性. 随着计算机和网络技术的不断发展,基于 Unicode 标准的字符编码已逐渐成为主流.

2.2 板塘苗文编码方案

实现板塘苗文编码表示的最简单方法是,利用已有的汉字 GB2312-80 编码标准或 GBK 交换码标准,占用汉字的位置或保留区来存放板塘苗文的字符. 这样虽然可以显示板塘苗文的文字,但是不能够进行板塘苗文与汉字的混合排版. 而汉字和苗文的混合排版是民间苗文信息处理系统必须具备的基本功能之一. 结合 Unicode 标准兼容不同语言文字且跨平台的特性,考虑到板塘苗文信息处理系统的实用性、开放性和网络化需要,基于 Unicode 的编码方案是实现板塘苗文编码表示的最佳选择.

Unicode 标准中字符的编码方式与 ISO 10646 的通用字符集(Universal Character Set,UCS)概念相对应. ISO10646 对 UCS 的体系结构和基本多文种平面(Basic Multilingual Plane,BMP)做了规定. 在 BMP 内,A 区用于字母文字、音节文字以及各种符号的编码,拉丁文、阿拉伯文、日文的平假名及片假名等都在此区编码;I 区用于中、日、韩(CJK)统一的表意文字编码;O 区保留,供未来标准化使用;R 区作为 BMP 的限制使用区,专用字符、变形显现及兼容字符可在此区编码. BMP 中的 E000—F8FF 是用户自定义区. 中国维文、哈萨克文和柯尔克兹文等少数民族文字的字符与其他阿拉伯文字符一起收入了阿拉伯文字汇. 中国朝鲜族所用的朝文字符也已和南韩的朝文字符一并收入 BMP 中的朝文字汇.

目前,实际应用的 Unicode 版本对应于 UCS-2 和 UCS-4 体系. 基于 UCS-2 的 Unicode 编码字符

集使用 16 位编码空间，每个字符占用 2 个字节，理论上该字符集最多可以表示 2^{16} 个字符，基本能够满足各种语言的使用要求. 实际上，最新版本的 Unicode 并未完全使用这 16 位编码，而是保留了大量空间以供特殊使用或将来扩展. 基于 UCS－4 的 Unicode 编码字符集使用 32 位编码空间，每个字符占用 4 个字节. 其中，首位恒为 0，其余 31 位用于字符集，理论上该字符集最多可以表示 2^{31} 个字符，完全可以涵盖一切语言所用的符号. BMP 中，UCS－4 字符编码形式为 U＋hhhh，其中每个 h 代表一个十六进制数字，与 UCS－2 编码完全相同. 4 字节 UCS－4 编码的前 2 个字节的所有位均为 0，而后 2 个字节与 UCS－2 的 2 字节编码完全一致.

采用基于 Unicode 的编码方案实现板塘苗文编码，需要将板塘苗文的所有字符都放到 Unicode 字符集中，以使板塘苗文的显现形式与其 Unicode 编码一一对应. 在编码设计时，为了使整个编码布局整齐合理，首先，根据笔画数目和构件部首对所有板塘苗文进行了排序；然后，按照 BMP 编码格式分别对独体字、构件字符和合体字进行编排，将编码存放在 Unicode 字符集的用户自定义区 E000—F8FF 范围内.

3　板塘苗文字库的设计和创建

3.1 字体技术

目前，Windows 环境下常见的字体技术主要有以下几种：

(1) 光栅字体(. FON). 该字体是 Windows 系统字体，显示速度快. 它本质上是一种针对特定显示分辨率存储的位图，适用于屏幕显示. 该字体以点阵描述字符，放大以后会出现锯齿.

(2) 矢量字体(. FON). 该字体也是 Windows 系统字体，其扩展名和光栅字体一样，但它由基于矢量的数学模型定义，放大以后不会出现锯齿. 一些 windows 应用程序在较大尺寸的屏幕显示中会自动使用该字体来代替光栅字体的显示.

(3) PostScript 字体(. PFM). 该字体 Windows 不直接支持，常用于 PostScript 打印，采用 Adobe PostScript 矢量语言进行描述. 如果在 Windows 使用它，就需要安装“Adobe Type Manger”(ATM)软件来进行协调.

(4) TrueType 字体(. TTF). 该字体是当前使用最普遍的字体，由 Apple 公司和 Microsoft 公司联合提出，基于二次 B 样条曲线及直线. 该字体用数学模型进行定义，采用数学函数描述字体轮廓外形，与分辨率无关，既可用于作屏幕显示，又可以用于由打印机分辨率决定的打印输出，保证了屏幕显示与打印输出的一致性. 该字体通过字形构造、颜色填充、数字描述、流程条件控制、栅格处理控制、附加提示控制等指令描述字形，比基于矢量的字体更容易处理，且同矢量字体一样可以随意缩放、旋转，而不会出现锯齿.

3.2 板塘苗文字模的制作

笔者以文献[1－2]中收集到的板塘苗文为蓝本设计字稿，对字稿进行扫描处理后，利用 Photoshop 技术把全部字稿扫描图片中的所有板塘苗文字符转换成计算机能识别的图像数据，通过二值化、分割、边缘检测与轮廓提取等处理，形成独立的字模文件.

Photoshop CS5 强大的图像处理功能中融合了高性能的二值化、边缘检测及轮廓提取等算法，为板塘苗文字模制作提供了有利条件. 基于 Photoshop CS5 的板塘苗文字模制作方法如下：

(1) 利用“选择”工具将所有字稿扫描图片文件中的各个板塘苗文分割成独立的字模小图片. 每张图片包含且仅包含 1 个苗文字符的图像数据. 为了保证载入速度，字模图片不宜太大. 根据板塘苗文的字形特点，经计算，以 200×200 像素范围的字模图片比较适合.

(2) 对每张字模小图片进行二值化处理. 在 Photoshop CS5 中，二值化处理通过“阈值”命令完成. 该命令能够将灰度或彩色图像转换为高对比度的黑白图像. 阈值是基于图片亮度的一个黑白分界值，默认值是 128(对应 50％中性灰). 用户可以指定或调整阈值，二值化的效果依赖于阈值的选择. “阈值”命令将所有比指定阈值亮的像素转换为白色，比阈值暗的像素转换为黑色.

(3) 对每张字模图片中的板塘苗文字符进行边缘检测及轮廓提取. 首先，使用“磁性索套”或“魔术棒”工具选取当前字符完整的粗略轮廓. 并在“调整边缘”窗口中将“视图模式”为“黑底模式”. 为了去除字模图片背景的白色，需要利用智能边缘侦测功能，将边缘侦测的“智能半径”勾选，再根据字模图片中背景范围

调节其半径像素值为 50(其值越大,边缘宽展区域越大,可以去除的白边范围越大).通过智能边缘半径调节就可以大致地将字符提取出来了.如果字模图片中背景颜色纯度不高,还可以利用"调整半径"和"抹除调整"等工具进行手动调节.

(4) 对提取到的板塘苗文字符进行边缘净化.当字符边缘有杂色边时,可以通过设置并调整平滑、羽化、对比度及移动边缘等参数进一步调整边缘部分.如果字符边缘存在轻微的半透明颜色,可以在输出窗口勾选"净化颜色"项,设置净化比率(如 80%),以将字符边缘的半透明颜色去除.

最后,将经过以上处理的每张字模小图片保存为独立的 JPG 图形文件.

3.3 板塘苗文 TrueType 字库的创建

基于 TrueType 技术的板塘苗文字库采用 HighLogic 公司推出的一款 TrueType 字体制作软件 Font Creator Program 来创建.该软件能够实现绘制字体、写入字体版权信息、控制字体属性等功能.

使用 Font Creator Program4.1 软件创建板塘苗文的 TrueType 字库时,不必从头绘制各个字符,直接导入已制作好的字模文件进行编辑即可.板塘苗文的 TrueType 字库创建步骤如下:

Step1 以"湘西板塘苗文"作为字库(字体家族)名称,新建一个基于"Unicode"字符集的.TTF 文件.

Step2 打开新建的"湘西板塘苗文.TTF"文件,导入第 1 个板塘苗文字模图片,并在编辑窗口代表字符的网格方框(一个方框对应一个 16 进制 Unicode 编码)对该字模字符进行调整及其字形轮廓的修改.导入的字模字符将对应一个 Unicode 字符,在 Unicode 字符集中具有唯一的位置编码.在编辑字符时,必须将字符控制在一定的范围内.编辑窗口通过 6 条线来控制字符位置.2 条红线的相交处为坐标原点,2 条水平黑线表示行距,2 条垂直黑色虚线表示字距.水平黑线无法调整,制作的字符大小以在 2 条黑线内为宜.垂直黑色虚线可以调整,通过拖拽黑色虚线来决定该字符与其他字符的间距.根据板塘苗文仿汉字的字形特征,经过计算,可以将板塘苗文左右位置坐标范围控制在 0～225 内,上下位置坐标范围控制在 200 ～25 内.字符的编辑修改主要包括以下工作:

(1) 将字模图片居中置于原点位置,通过编辑窗口中的"变换(Transform)"浮动窗口调整字符的位置、大小,通过对话框设置"阈值"、"扩散滤镜值"等参数调整该字形轮廓.

(2) 点击"生成"按钮,将以字形轮廓模式显现的字模转换生成以节点模式显现的字模.字符字形轮廓的细节修改通过调整节点完成.字符上的任意个节点 A 和 B 可以拉直通过它们的线段.当一个不在此线段上的节点 C 作用于该线段时,将形成以 A 为起点、B 为终点的曲线,同时线条会朝 C 点方向弯曲.如果还有其他不在曲线上的点作用于曲线,曲线轮廓也会准确适应这些点的作用变化.字形编辑时,只需在字符相应位置增加节点进行调整即可实现形状的修改.

(3) 使用工具栏"字体"中的"验证"工具对字模字符进行验证,按照验证报告提示的错误信息对字模字符进行相应修改.然后,根据显示效果进一步调整参数值,直到通过验证且字符出现满足要求的精确字形轮廓为此.

(4) 关闭编辑窗口,时代表字符的网格方框左上角显示的字符会由灰色变为绿色,表明当前编码位置的字符已生成,字库中已有一个板塘苗文.

Step3 重复 Step2,在"湘西板塘苗文.TTF"文件中导入余下的苗文字模图片,将所有板塘苗文用同样方法进行编辑调整和验证.直到文件中的所有字符都比较整齐规范为止.

Step4 按"F5"键,对字库文件"湘西板塘苗文.TTF"进行测试.如果发现不满意的字符,就继续按照前述方法重新调整修改.

Step5 完成调整修改后,即可根据字库中字符的字体风格和大小为字体命名,并输入版权等相关信息,然后保存文件.命名板塘苗文字体文件时,必须设置板塘苗文的"版权信息"、"字体系列名称"、"字体完整名称"和"字体子集名称"等相关附加信息;还可以通过"高级"按钮对字体的名称进行进阶设定,通过"平台"设置区选择字体运行平台(更改平台会使命名区中的内容相应改变).通常,在 Windows 环境下,字体的系列名称显示在字体菜单中,字体的子系列名称将作为样式显示;在字处理软件中,当用户选择一种字体时,字体系列名称显示在"字体"中,而字体子系列名称一般显示在字形样式处.注意,若当前字体在风格和大小方面没有特别之处,则此处命名为常规字体"Regular"即可.若当前字体具有汉字楷体或宋体等风

格,则应将字体命名为相应的板塘苗文楷体或宋体等.

4　结语

随着湘西旅游产业的迅速发展,如何借助现代计算机和网络技术手段,将湘西苗族优秀文化推向世界,以更好地促进湘西旅游产业和湘西地区经济的发展,是一个具有重要现实意义的课题. 为了更好地保护湘西苗族文化遗产和弘扬苗族民族文化,为湘西苗族文化研究者提供便利工具,研制和开发一个基于 Windows 平台的湘西民间苗文处理系统势在必行. 板塘苗文编码表示的实现及 Truetype 字库的建立,是湘西民间苗文处理系统的研制与开发的前期工作,为后续湘西民间苗文输入、编辑、翻译等信息处理技术研究奠定了良好的基础.

参考文献:

[1] 赵丽明,刘自齐. 湘西方块苗文 [J]. 民族语文,1990(1):44 - 49.

[2] 杨再彪,罗红源. 湘西苗族民间苗文造字体系 [J]. 吉首大学学报:社会科学版,2008,29(6):130 - 134.

[3] 龙德义. 计算机苗文处理系统研制成功 [J]. 今日民族,1995(5):16.

[4] Unicode Consortium. The Unicode ® Standard: A Technical Introduction [EB/OL]. http://www.unicode.org/standard/principles.html,2012 - 10 - 04.

Computer Encoding and Fonts Creating of Bantang Hmong Language Characters

MO Li-ping[1], ZHOU Kai-qing[2], JIANG Xiao-hui[1]

(1. College of Information Science & Engineering, Jishou University, Jishou 416000, Hunan China; 2. Faculty of Computer Science & Information Systems, Technological University of Malaysia, Johor Bahru, Johor, 80310, Malaysia)

Abstract: National language characters information processing is of great significance for protecting the national cultural heritage and promoting national culture. Both computer coded representation and fonts creating are the basic works of national language characters information processing research. The glyph feature of Xiangxi Bantang Hmong Language Characters is analyzed. On the basis of character encoding standard and font technology being introduced, the encoding scheme of Bantang Hmong Language Characters, which is based on the Unicode standard in Windows environment, is proposed. And then, according to Bantang Hmong Language Characters, the method of making matrix font based on Photoshop technology and the concrete steps of creating fonts based on Truetype technology are given.

Key words: Bantang Hmong Language Characters; character encoding; Unicode standard; Truetype fonts

(责任编辑　向阳洁)