利用 OpenType 字库特征建立蒙文字库

巩政[1]，敖其尔[2]，吉日木图[3]，乌达巴拉[4]
（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

摘要：本文介绍了利用 OpenType 字库特性技术建立蒙文字库，解决在 Unicode 标准编码中，蒙古文字符集一个名义字符对应多个变形显现字符的问题。

关键字：蒙古文；Unicode 编码；OpenType 字库
中图分类号：TP391.1 **文献标识码：**A

一、引言

现行蒙古文是世界上公认的一种最为复杂的文本语言之一。它属于拼音文字，但是由于诸多的历史原因，蒙古文字在发展过程中又借鉴了其它文字的一些特性，因此它又有别于一般的拼音文字。蒙古文文字处理技术的发展已经经历了近 30 年的时间，在国外一般是按照西方文字的特点，将字母拼接组合的方式进行处理，在国内受汉字处理方法的影响，一般采用按字形连接或字形与发音字母相结合的方式处理。从文字处理的角度讲，现在的蒙古文文字处理技术还是比较成熟的。但是蒙文信息处理不仅仅局限于文字方面的处理，还包括许多方面，如检索查询、排序、信息交换等。特别是随着蒙文国际标准编码的出台，传统的蒙文字体技术如 TrueType 字体技术已经难以实现在显现出规范的书写形式的同时，又支持 Unicode 标准编码。要坚持走蒙文国际化的道路，必须寻求新的技术支持。按照目前的发展趋势，利用 OpenType 字体技术建立蒙文字库，应该是一个比较理想的方案。

二、蒙古文字符集的特点

《蒙古文字符集》是包括传统蒙古文、托忒文、锡伯文、满文以及蒙、托、满三种文字用于转写藏文和梵文的阿礼嘎礼字母、标点符号、数字和控制符的蒙古文字符的编码集。托忒文、锡伯文、满文均用蒙古文字形。

传统蒙古文、托忒文、锡伯文、满文的大部分字母根据它们在词里的位置（词首、词中、词尾）等有不同的变体形式，有时一个字母能有十种以上的变体形式。

根据 ISO/IEC10646 的有关规则，只对这些变体形式中的一个进行编码，称为“基本字符”或“名义字符”。对于元音，原则上采用它们的独立形式。对辅音，原则上采用出现在元音“A”前面的词首形式。所有其它形式称为“变形显现形式”，不对其进行编码。每个“基本字符”的“变形显现字符”在绝大部分情况下能够用位置的不同等条件区别清楚，但也有很少几个只用位置的不同等词内条件无法区别，而通过设置控制符区别。

在 ISO/IEC10646 编码中，Unicode 平面 1800-18AF 区确定了 155 个码位用于蒙古文基本字符集。传统蒙古文字符只占 35 个编码位置，而由此产生的变形显现字符实际上可达 200 个左右。除了“名义字符”的“变形显现字符”外，蒙古文文字还存在强制性合体字的情况。一般合体字的字形与由单个字符组成的合体字形不太一样，这样可能还要增加一些变形显现字符。

在 ISO/IEC10646 中规定的《蒙古文字符集》，从发音上区分开了蒙古文字符，但按照现有的制作蒙古文字库技术，在不增加编码的情况下，很难区分“形同音不同，音同形不同”的字母。

三、OpenType 字库特征简介

OpenType 是 Microsoft 和 Adobe 联合开发的一个新的字体文件格式。

OpenType 字库格式的主要特点是：

（一）跨平台兼容性：OpenType 字体中所用的轮廓、metric 和位图数据使用一个字体文件进行文件管理，在 Windows 和 Macintosh 操作平台中均可以使用。

（二）支持 Unicode 编码字符集：OpenType 字体格式是 TrueType sfnt 格式的扩展，同时支持 PostScript 字体数据和新的显示字形特性。OpenType 字体包括一个大字符集和布局属性(feature)，提供多语言支持和更精确的高级显示字形控制。

（三）OpenType 的 feature 支持：OpenType 的布局 feature 支持字符的自行替换和定位。它不但允许字符到字形的映射，还可以存在字形到字形的映射，而且这种映射可以是一对多、多对一、多对多。OpenType 技术的核心就是字形到字形的映射，通过这种映射它支持替换(substitution)、重定位(repositioning)、组合字符(ligature)等。

OpenType 字体根据字体轮廓类别分别使用.OTF 或.TTF 扩展名。包含 CFF（压缩字体格式 Compact Tont Format）数据的字体（非 TrueType 轮廓数据）使用.OTF 扩展名；包含 TrueType 轮廓数据字体使用.TTF 扩展名。

OpenType 字体描述表

对于使用 TrueType 或 Postscipt 轮廓信息的字体，下表是实现字体功能所必需的。

标志位	名　称	描　述
Cmap	Character to glyph mapping	字符代码到文字轮廓序号的映射表
Head	Font header	文件头表
Hhea	Horizontal header	水平度量头信息
Hmtx	Horizontal metrics	水平度量信息
Maxp	Maximum profile	最大值描述表
Name	Naming table	名字表
Os/2	Os/2 and Windows specific metrics	OS/2 和 Windows 度量信息
Post	PostScript information	PostScript 信息

cmap 表给出了字符代码到文字序号的映射。head 表给出字体的制造商、制造日期、所有文字的上下左右定位位置、字体风格和字体方向等重要标志。hhea 表提供字体的水平刻度信息，如字符的基线位置、线宽、字宽、左右边界。hmtx 表为每一个字形提供水平空间度量信息。Maxp 表给出相关数据的最大值，例如简单文字或复合文字的最大轮廓数和轮廓点数，解释器堆栈的最大容量，复合文字包含的最大简单文字个数等信息。name 表给出字体的操作系统、压缩方法、所用语言、字体名称、商标版本信息等。OS/2 表给出字体的平均宽度、平均字高、宽高比、笔划粗细、最大最小编码、基线位置、字型种类等。post 表给出字体的倾斜程序、底线位置和线宽等。

OpenType 布局表（Layout Table）

标志位	名　称	描　述
BASE	Baseline data	基线数据
GDEF	Glyph definition data	字形定义数据
GPOS	Glyph positioning data	字形位置数据
GSUB	Glyph substitution data	字形替换数据
JSTF	Justification data	字形调整数据

BASE 表给出每一种脚本的基线和扩展数据范围。通过 JSTF 表文字处理端可以在调整字符时打开或关闭字形替换和位置特性。GDEF 表包含文字处理时的有效信息，如光标位置以及和字形相关联的点。GSUB 表和 GPOS 表定义了字形替换和位置特性，处理复杂文本主要运用这两张表。GSUB 表包含了 5 种类型的替换以支持不同类型的字符到字形的映射，例如多到一的字形替换和一到多的字形分解。GPOS 表定义了 7 类位置特性为字形的调整提供

二维位置数据。

GPOS 和 GSUB 的结构完全一致，包括四种数据：Script、Language System、Feature、LookUp，这四种数据也是以表的形式存在的。

通过访问 Script 表中的文字种类确定相应的 Language System，根据 Language System 中的信息找到相应的 Feature，从 Feature 中读出 LookUp 偏移量，从而读出 LookUp 存放的替换和定位条件。在文本处理过程中，应用程序会将一个 LookUp 应用于输入字符串中的每一个显现字形之后才转向下一个 LookUp。在确定某个字符的显现字形之前应用程序必须先扫描所有与 LookUp 操作相关的字符，确定是否能够构成替换操作的 OpenType Layout。

四、OpenType 特性应用于蒙文字库

蒙古文字是一种复杂的文字，每个名义字符可能要对应多个变形显现字符。一个单词在输入过程中，依据上下文相关，在输入每个字符时其对应变形字符随时有可能调整。目前在不增加变形显现字符编码的情况下，利用 OpenType 字符映射和替换等特性应该是解决蒙古文字符变形显示的较好方案。

下面是在制作蒙古文 OpenType 字库时需要用到的一些 OpenType 特性值：

Feature	Feature function	Layout operation
fina	这个特征用来对字型的最后形式到 unicode 字符值的映射。	GSUB
medi	这个特征用来对字型的中间形式到 unicode 字符值的映射。	GSUB
init	这个特征用来对字型的词首形式到 unicode 字符值的映射。	GSUB
rlig	这个特性用来将几个字型合在一块儿形成另一个字型。造字开发者将这些所有的合体字型放到一个表格里，以便可以随时直接找到他们。	GSUB
calt	在具体环境下，根据上下文选择一个字型来代替默认的值，这样能更好的进行连接。<calt> 表中详细说明着每个字形式到代替字型的映射。	GSUB
[GSUB = glyph substitution]		

这些特征值可以在蒙古文 OpenType 字库中作如下应用：

1. <fina>

“fina” 特性用于 Unicode 字符到它的词末形式的映射。(GSUB lookup type 1)

Unicode 字符	映射为词末形式
ᠨ	ᠩ

2. <medi>

“medi” 特性用于 Unicode 字符到它的中间形式的映射。(GSUB lookup type 1)

Unicode 字符	映射为中间形式
ᠨ	ᠬ

3. <init>

“init” 特性用于 Unicode 字符到它的开头形式的映射。(GSUB lookup type 1)

Unicode 字符	映射为字首形式
ᠨ	ᠨ

4. <rlig> (the required ligatures)

“rlig” 特性可用于蒙古文的合体字处理。

Unicode 字符	映射为合体字形式
[illegible]	[illegible]

5. <calt>

“calt” 特性可用于组成蒙古文单词的中间位置的字母根据其前或后一字符的要求去映射相应的变形字符。

字符	映射（在词中时）
[illegible]	[illegible]

除了这些特性值，还有 isol、cswh 特性也应用于蒙古文字库。

如果对 OpenType 字库的文件结构十分了解，则可以直接通过计算机程序设计语言完成建立字库的过程。但这样做是比较费力的。现在已经出现了几种专门制作 OpenType 字库的工具软件。如 FontLab Ltd.的 FontLab 造字工具、Microsoft 的 VOLT 工具软件。

一般情况下，我们可以用 Fontlab 工具软件先制作一个只有轮廓字形，没有 Script、Language、Feature 等信息的字库。虽然 FontLab 也能够在 GPOS 和 GSUB 表中添加字形变换方面的相关数据，但是不如在 VOLT 中更容易实现。

五、结束语

目前，在 Linux 和 Windows XP 操作系统中都能够支持 OpenType 格式的字库，在 Linux 中需要专门编写一个“文本处理引擎”程序才能够正确解释 OpenType 蒙古文字库。而 Windows XP 也必须有微软提供的 The Unicode Script（USP10.dll）程序。否则即使制作好了 OpenType 蒙古文字库也无法使用。现在微软的 USP10.dll 正在研制中。另外，在蒙古文中有一些特殊的词，如外来词，用 OpenType 特征值处理还是有些问题，需要进一步研究。

参考文献:

[1]. http://www.unicode.org/

[2]. http://www.adobe.com/type/opentype/main.html

[3]. 确精扎布.蒙古文编码[M].呼和浩特: 内蒙古大学出版社，2000.

Using the technology of the OpenType feature to found the Mongolian Fonts

GONG ZHeng[1],Ochir[2],Jirimtu[3],Wudabala[4]

(School of Computer Science, Inner Mongolia University, Hohhot 010021)

Abstract : this paper introduces using the technology of the OpenType feature to found the Mongolian Fonts ,and there it solves to map the basis character to its visible character in the Unicode .

Key word: Mongolian character ; Unicode ; OpenType Fonts

收稿日期: 2005-04-12;

基金项目: *本文得到内蒙古自然科学基金项目“蒙文OpenType字库研究”（200408020810）和教育部人文社会科学研究重大项目“蒙古文信息处理平台（MIPP）的研究”（02JAZJD850003）

作者简介: 1.巩政（1965 -），男，内蒙古呼和浩特人，内蒙古大学副教授，主要研究蒙文信息处理。2.敖其尔（1941 -），男，内蒙古兴安盟人。内蒙古大学教授，硕士生导师，主要研究蒙文信息处理和计算语言学。3.吉日木图（1980 -），男，内蒙古通辽人，内蒙古大学计算机学院 2002 级研究生。主要从事蒙文信息处理和计算语言学方面的研究。4.乌达巴拉（1983 -），女，内蒙古通辽人，内蒙古大学计算机学院 2004 级研究生。主要从事蒙文信息处理和计算语言学方面的研究。