蒙古文编码转换软件的设计与实现

图格木勒

（内蒙古大学 蒙古学学院 蒙古语文研究所 呼和浩特市 010021）

摘 要： 虽然蒙古文编码国际标准(ISO10646)发布几年了，但是显现字符码位不统一。各种蒙古文系统之间互相不兼容。使用不同系统输入的蒙古文文档不能共享，成了制约蒙古文信息化发展的一个颈瓶。公众急切需要蒙古文编码转换软件。本文以设计和实现蒙古文编码转换软件的角度分析和分类了现有的蒙古文显现编码及蒙古文转写方案。将现有的蒙古文显现编码方案和蒙古文转写方案分为名义字符方案、准名义字符方案和显现字符方案等三个方案。对每个方案进行了分析后，将转换算法分为基于规则转换和综合方法转换的两个形式。详细分析了两种转换算法的流程，并实现了一些使用基于规则编码转换的蒙古文编码转换。

关键词： 蒙古文；编码方案；编码转换

Design and the Implementation of Mongolian Encode Conversion Software

Algoi Tugemel

(The Institute of Mongolian Studies, Inner Mongolia University, Huhhot 010021)

Abstract: Although international standard of Mongolian codes (ISO10646) were promulgated in quite a few year, but present Mongolian presentation codes are not unified. The Mongolian presentation coding systems were not compatible each other. The codes of documents that be used the dissimilarity Mongolian input systems can't shared each other. Mongolian codes became a neck bottle of the information-based development of Mongolian. Public anxiously demanded the conversion software of Mongolian codes. In this paper, for to design and carry out conversion software Mongolian codes analyzed and classified the present presentation Mongolian codes and Mongolian transfer letter systems. We divide presentation Mongolian coding projects and Mongolian transfer letter systems into the three major types that the nominal characters project, the likeness nominal project and the presentation characters project. We analyzed the three types of characters project, confirm that the conversional algorithms should be divided into two type that the rule based conversion and the conversion of synthesized method. We deeply analyzed the two types of conversional algorithms and carried out some coding conversions that be used the rule based conversional algorithm.

key words: Mongolian; coding project; code conversion

1 引言

在蒙古文编码国际标准制定以前，就存在着很多蒙古文编码方案。为了将蒙古文显示在 Windows 操作系统

作者简介：图格木勒（1979—），男，内蒙古呼伦贝尔人，研究生，E-mail：algtgml@sohu.com

或其他操作系统上，很多软件开发者纷纷设计了各种显现字符编码方案。由于各种蒙古文显现字符编码方案之间互相不兼容。使用不同输入系统输入的蒙古文文档互相不能转换。如此混乱的蒙古文显现字符编码方案和蒙古文转写方案对蒙古文信息化进程带来了很大的麻烦，所以迫切需要开发蒙古文编码转换软件来实现各种蒙古文显现字符编码方案和蒙古文转写方案统一到蒙古文编码国际标准上，并且实现各种蒙古文显现字符编码方案和蒙古文转写方案的互转换。

2 蒙古文编码方案的分类

我们对存储蒙古文的文档进行编码分析后，以编码转换实现角度将现有的蒙古文编码方案和蒙古文转写方案可分“名义字符方案”、“准名义字符方案”和“显现字符方案”。

2.1 名义字符方案

这类编码中包含蒙古文编码国际标准及其等价的拉丁转写方案。例如，蒙古文编码国际标准的拉丁转写方案。此方案的优点是符合国际标准，是蒙古文编码统一的大方向。但目前 Windows 通用控件中无法正确显现蒙古文；即蒙古字母在词首、词中、词尾等位置用一个形式表示，不能以其位置不同而自动变换形式，这样不符合蒙古文的书写方式。OpenType 字体技术可实现蒙古文自动变换形式。在 Window Vista 版本中将会支持蒙古文自动变换形式。如今的 Linux 系列操作系统中也能实现蒙古文自动变换形式。

2.2 准名义字符方案

“准名义字符方案”是以蒙古文编码国际标准的蒙古文字母序列排列显现字符的编码方案或与其近似的转写方案。“准名义字符方案”中不包括强制性合体字和非强制性合体字，区分形同音不同字母。现有以下几种方案可归类于准名义字符方案：

（1）表音显现字符方案

表音显现字符方案是以蒙古文编码国际标准的蒙古文字母序列排列显现字符的、显现字符占码位的编码方案。此编码方案中不包括强制性合体字和非强制性合体字，区分形同音不同字母。例如，内蒙古明安途互联网技术开发有限公司的显现字符编码方案、内蒙古蒙科立软件有限责任公司的显现字符编码方案等。此方案的优点是在 Windows 中正确显现蒙古文。其缺点是编码占用保留区域。用这显现字符编码只可以显示蒙古文，如用这显现字符编码表示、传输、交换、处理、存储、输入蒙古文的话不符合蒙古文编码国际标准。不过很多蒙古文处理软件就是用显现字符编码处理蒙古文的。

（2）拉丁转写方案

此类拉丁编码方案指的是根据蒙古文读音的拉丁转写方案。包括各种记录蒙古语读音的拉丁转写方案其中不包括蒙古文编码国际标准的拉丁转写方案。有些拉丁转写方案中有音形兼顾的特点。如，内蒙古大学蒙古文语料库专用拉丁方案（下面简称内大拉丁）。

拉丁编码方案的优点是在计算机上不依靠任何软件都能显示和输入。其缺点是容易跟拉丁文字混淆，对蒙古文进行检索、排序比较麻烦。

2.3 显现字符方案

“显现字符方案”包括强制性合体字或非强制性合体字，不区分形同音不同字母。例如，方正编码等。此方案的优点是在 Windows 中正确显现蒙古文。用这显现字符编码只能显示蒙古文，用这显现字符编码表示、传输、交换、处理、存储、输入蒙古文的话不符合蒙古文编码国际标准。

3 编码转换方案

设计编码转换软件时，首先根据以上编码方案的分类和具体转换需求，将编码转换分为以下四种形式。对四种形式进行分析后，确定了将转换算法分为基于规则转换和综合方法转换两种形式。

3.1 转换形式

根据以上编码方案的分类，将编码转换分为以下四种形式：

①“准名义字符方案”之间的转换

实现“准名义字符方案”之间互转换时，先将源“准名义字符方案”都转换到蒙古文编码国际标准上，再将转换到蒙古文编码国际标准的编码转换到目标“准名义字符方案”。转换时可用基于规则的方法。

②“准名义字符方案”向“显现字符方案”的转换

“准名义字符方案”向“显现字符方案”的转换也可用基于规则的方法进行转换。

③“显现字符方案”向“名义字符方案”的转换

此类转换比较复杂。“显现字符方案”向“名义字符方案”转换时有一对多，多对多转换的情况。具体转换需要用词典或者使用概率统计模型。

④“显现字符方案”之间的转换

“显现字符方案”之间转换也可用基于规则的方法进行转换。

3.2 转换算法

编码转换软件需要实现上面四种转换形式。根据程序实现方法，可以把编码转换算法分为基于规则转换方法和综合转换方法两种：

① 基于规则转换方法

基于规则的方法适用于“准名义字符方案”之间的转换、“准名义字符方案”向“显现字符方案”的转换和“显现字符方案”之间的转换。

② 综合方法转换

综合方法针对的是“显现字符方案”向“名义字符方案”的转换。综合方法中包括基于词典转换和基于概率统计模型的转换。

4 软件设计

编码转换软件在转换文档时需要经过文档机构分析、“词”分切、编码转换和存储结果等四个过程。

4.1 文档结构分析

文档结构分析过程中对需要转换的文件进行分析，提取需要转换部分。

4.2 “词”切分

“词”切分的过程中对需要转换部分进行“词”识别。这里涉及到的“词”是根据形式上的词，不是表意层次上的词。比如，蒙古语中的  （红的比较级，合起来才是真正意义上的词）的  也算作一个“词”（形式上的一个单位）来看待。一般情况下蒙古文以空格以及标点符号作“词”分界标准。此过程中根据蒙古文的词分界将需要转换部分的蒙古文切分成蒙古文“词”表，为下一步转换作准备。

4.3 编码转换

编码转换过程中要实现对蒙古文“词”表中的“词”进行转换。

编码转换过程分为前面提到的基于规则的方法和综合方法等两种方法。下面详细讨论其实现的方法。

4.3.1 基于规则的方法

蒙古文编码的规则转换的实现算法流程图：

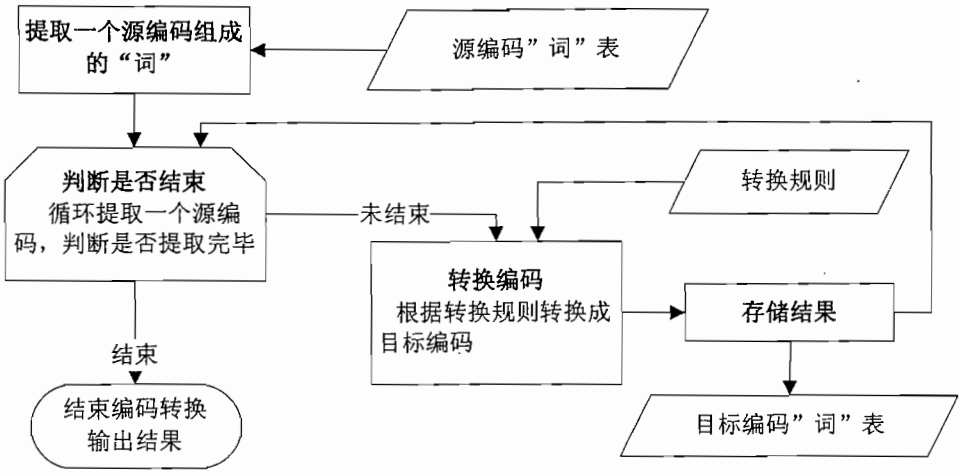

图 1 基于规则的蒙古文编码转换流程图

Fig.1 the flow chart of conversional algorithm that rule based

编码转换时，根据蒙古文正字法制定了一套规则；即特殊词的转换规则、词首转换规则、词中转换规则、词尾转换规则。并在每个规则中描述了前后字符制约关系。

具体编码转换流程为：

4．1 判断“词”中的源编码的位置选择不同的转换规则。

5．1 再用转换规则中的前后字符制约关系确定目标编码。

转换规则可以实现一对一的转换、无条件的多对一的转换、根据正字法条件的多对一的转换或者根据编码的变形选择符的多对一的转换，但无法实现一对多和多对多的转换。基于规则的方法无法实现“显现字符方案”向“名义字符方案”的转换。

4.3.2 综合方法

“显现字符方案”向“名义字符方案”的转换时有一对多，多对多转换的情况。综合方法中包括基于蒙古语词典转换和基于概率统计模型的转换。这里主要讨论基于词典转换。

① 基于词典转换

基于词典转换需要转换规则、蒙古语词干库和构形附加成分库。蒙古语词干库和构形附加成分库以“显现字符方案”描述的“词”干或构形附加成做索引并有“名义字符方案”描述的转换结果。具体流程图如图 2 所示：

基于词典转换方式是基于规则转换方法的一种拓展方法，对用规则转换不出正确结果的“词”进行特殊处理；首先对“词”进行切分附加成分处理，然后切出来的“词”干与“显现字符方案”描述的词干库中进行匹配。如匹配成功提取词干库中的“名义字符方案”描述的转换结果。然后词干与附加成分合并输出转换结果。

② 基于概率统计的方法

基于词典的转换的缺点是覆盖面不够大。在词典中没有的词的转换就没什么限制条件。基于概率统计的方法可以克服覆盖面窄的缺点而且可利用需转换编码的上下文的搭配模型对同形异码字符进行消歧。我们尚未对基于概率统计的方法进行分析。

4.4 存储文档

存储文档过程中将转换的编码以所需的文件格式进行存储。

5 目前成果和存在的问题

目前本蒙古文编码转换软件中实现了以下几种蒙古文编码的互相转换：

① 蒙古文编码国际标准和内大拉丁的互转换

② 内大拉丁和明安途显现字符编码的互转换

③ 内大拉丁到方正显现编码的转换

在蒙古文编码转换过程中遇到的几个问题：

① 显现字符向名义字符转换时歧义现象比较多，实现消歧很困难。

② 转换使用拉丁编码方案的文档时将文本中的英文文字或数字也进行了转换。

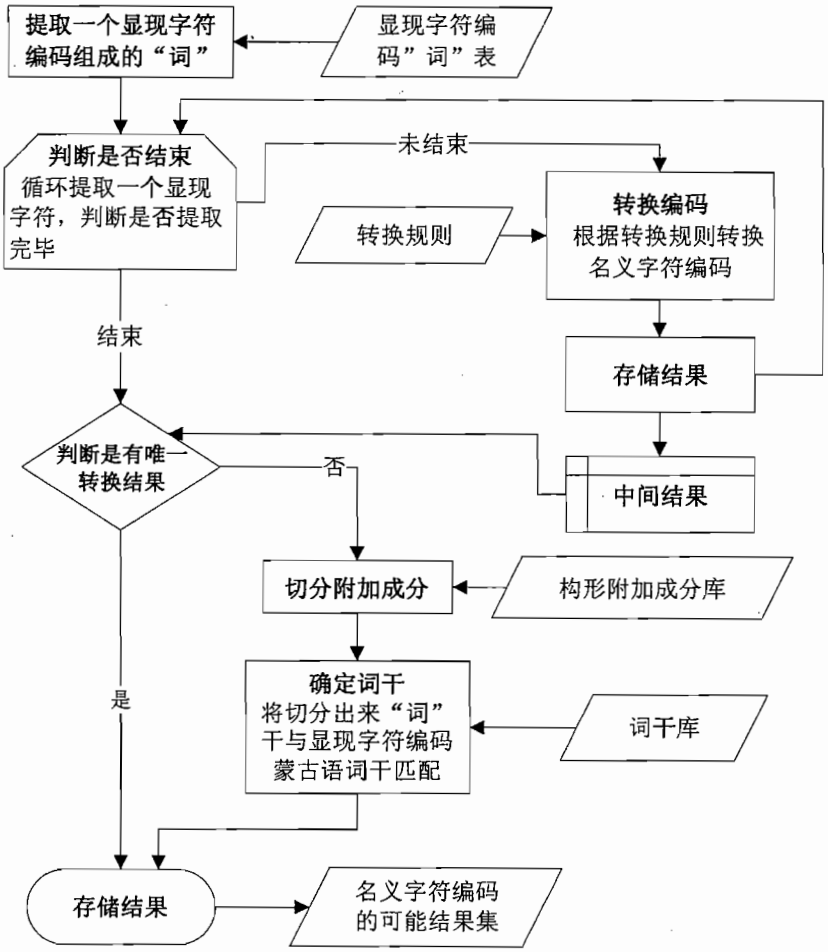

图 2 基于词典转换算法流程图

Fig.2 The flow chart of conversional algorithm that dictionary based

6 总结

我们的蒙古文编码转换软件现如今未实现所有蒙古文编码的互转换，但实现了好几个准名义字符编码方案的互转。目前努力研究实现基于词典的转换方法。

参考文献：

[1] 确精扎布.蒙古文编码[M].呼和浩特：内蒙古大学出版社.2000 年：P.128-268

[5] 那顺乌日图.蒙古文信息处理[M].内蒙古科学技术出版社.1998 年：P.113-121.

[6] 华沙宝. 从方正蒙古文码到 ASCII 码的转换软件——MTOA[A].论文与纪念文集[C]. 呼和浩特：内蒙古大学出版社，1997 年

[7] 确精扎布.蒙古文编码国际标准通过以后研制的几种蒙古文输入系统比较[A].第十届全国少数民族语言文字信息处理学术研讨会[C].青海 西宁：中文信息学会等， 1997 年. P.132-139