

《蒙古语语法信息词典字符分库》的建立及意义

艳花

(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: 《蒙古语语法信息词典》是为蒙古语语句的自动分析与自动生成而研制的一部电子词典。它是由总库及各分库所组成的。其各分库是《蒙古语语法信息词典》的有机组成部分,《蒙古语语法信息词典字符分库》(以下简称“字符分库”)也包括在内。本文主要通过“字符分库”与“总库”的关系和“字符分库”与印刷体蒙古文识别之间的关系等两个方面来阐述了建立“字符分库”的意义,并以传统语法研究和传统蒙古文字符自身的特点为基础,设计出了利于机器处理的各种属性字段及其取值规格。

关键词: 蒙古语语法信息词典; 字符分库; 印刷体蒙古文识别

Development and Meaning of “Character Bank of the Mongolian Grammatical Information Dictionary”

Yanhua

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

Abstract: “Mongolian Grammatical Information Dictionary” is an electronic dictionary developed for automatic analysis and automatic generation of Mongolian sentences. It is consisting of general bank and the sub-banks of various word classes. Every sub-bank is organic part of “Mongolian Grammatical Information Dictionary”, and “the Character Bank of the Mongolian Grammatical Information Dictionary” (for short “Character Bank”) also be included. The article mainly discussed the meaning of developing “Character Bank” from the two aspects that “Character Bank” against general bank relationship and “Character Bank” together with the relationship between the recognition of printed Mongolian characters. Based on the traditional Mongolian grammar and traditional Mongolian characters feature itself, the article designed various kinds attribute field which is easy for machine processing, and filled the relevant values in it.

Keywords: “Mongolian Grammatical Information Dictionary”; “Character bank”; the recognition of printed Mongolian characters

1 建立“字符分库”的意义

1.1 “字符分库”与“总库”的关系

蒙古文信息处理研究工作自 20 世纪 80 年代开始至今,进行了一些基础工程的建设 and 基础理论的研究,也开发过一系列应用系统。如,各种语料库的建设、研制蒙古文编码国际标准、进行蒙古语语法属性、语义属性的研究、建立蒙古语音参数数据库、研制蒙古语语法信息词典、开发汉蒙机器翻译系统等等。在这二十几年中,蒙

基金资助: 获教育部、国家语委(项目号: MZ115-005), 国家自然科学基金(项目号: 36963005)资助。

作者简介: 艳花(1979—), 女, 内蒙古通辽人, 研究生, E-mail: svrgvgceceg@126.com

古文信息处理研究工作取得了丰硕成果。其中《蒙古语语法信息词典》的研制具有深远的理论意义和实践意义。以往编撰的蒙古语词典从类型上来讲,大致可分为语义解释词典、正音正字法词典、双语或多语对照词典及专用词典等,这些词典主要是面向人的、印刷形式的词典。

《蒙古语语法信息词典》是面向计算机的、为实现自动分析和自动生成蒙古语语句而研制的一部电子词典。它囊括了蒙古语词法形态、句法功能、搭配特征以及正字法等方面的知识,是蒙古文信息处理工作的重要环节。

《蒙古语语法信息词典》由不同层次构成。第一层是包括该词典所有词条的总库;第二层是各类词的分库。如,名词分库、动词分库、形容词分库、副词分库等。第三层是由一些子类组成的若干个分库。

目前,我们已建立了有38893个词条、19个属性字段的“总库”;有13896个词条、33个属性字段的“动词分库”;有297个词条、42个属性字段的“构形附加成分分库”;现在,我们正在建设有34个词条、18个属性字段的“标点符号分库”;有7426个词条、45个属性字段的“形容词分库”;有14117个词条、54个属性字段的“名词分库”;有1006个词条、19个属性字段的“副词分库”。

《蒙古语语法信息词典》的各“分库”是《蒙古语语法信息词典》的有机组成部分。其“总库”是建立各“分库”的重要的词语来源。各“分库”与“总库”相辅相成形成一个知识体系。

蒙古文是世界上从左到右、由上而下连写的一种特殊的文字。它有着自己特有的标点符号和数字系统。如,蒙古文句号是“ᠰ”、蒙古文省略号“ᠰᠢ”、蒙古文数字一“ᠠ”、蒙古文数字五“ᠤ”。并且每一个蒙古文字符(除标点符号、控制符、数字以外)都有独立、词首、词中、词尾等四种形式。如,蒙古文“ᠠ” (A)的独立形式是“ᠠ、ᠡ”,词首形式是“ᠠ᠊”,词中形式是“ᠠ᠊、ᠡ᠊、ᠢ᠊”,词尾形式是“ᠠ᠊、ᠡ᠊”。

作为蒙古文信息处理工作的重要基础工程之一的《蒙古语语法信息词典》。理应尽收蒙古文字符以及它独有的一些特点,使它成为《蒙古语语法信息词典》的有机组成部分。为蒙古文信息处理研究的进一步深入打良好的基础。所以建立“字符分库”是蒙古文信息处理工作难以避开的基础性工作。

1.2 “字符分库”与印刷体蒙古文识别之间的关系

蒙古文是一个历史悠久的文字,至少有800年的历史。蒙古文在漫长的历史发展过程中给我们留下了许多宝贵的历史文献资料和优美的文学作品。把这些文献资料全部用键盘输入计算机进行保存,使这些优美的作品代代相传将是我们这一代人应尽的职责。

自蒙古文信息处理工作开展至今,研制蒙古文计算机输入法工作就已经开始。目前蒙古文计算机输入法主要有读音输入法、拉丁输入法、整词输入法、词组输入法、读音选择输入法等,但这些输入法都是键盘输入法。其他的输入法,像语音输入、扫描输入等比较先进的输入法,在蒙古文信息处理领域中才刚刚开始研究。

其中扫描输入法涉及到文字识别技术(OCR)。文字识别是模式识别的一个重要分支,是将人工智能与图像处理技术相结合的新技术,也是新一代计算机智能接口的一个重要组成部分。通过各国科学家的尽心努力,文字识别技术发展得相当快。现在,汉字识别技术已逐步走向成熟,有汉王、清华紫光等商业化文字识别软件产品不断出现。

跟其他文种的识别一样,印刷体蒙古文识别的原理概括起来是:将原始图像输入到计算机中,以二值图像表示出来。预处理包括对原始图像的去噪、倾斜校正或各种滤波处理以及轮廓边缘平滑等。识别阶段包括行、字的切分、特征提取以及字符的识别等。识别后处理则包括利用词义、词频、语法规则或语料库等语言先验知识对识别结果进行校正过程。对印刷体蒙古文识别原理用流程图表示如下:

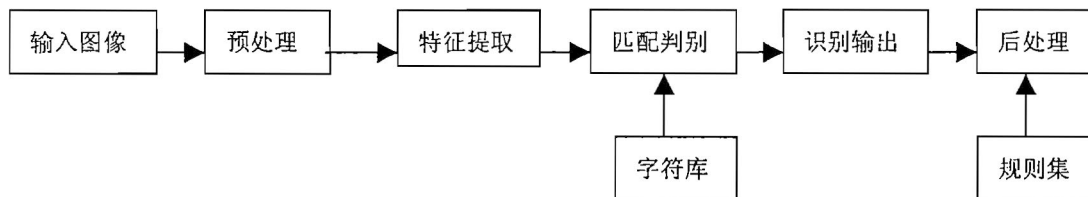


图1 印刷体蒙古文识别基本流程图

Fig.1 Flow chart in the recognition of printed Mongolian characters

从上面的流程图我们不难看出，在蒙古文识别研究中字符库充当着举足轻重的作用。在对字切分结果进行匹配判别以及识别后处理工作中都将是必不可少的。蒙古文识别研究工作才刚刚开始，所以“字符分库”的建设将是不可缺少的准备工作。蒙古文识别研究中同形异码字识别问题是整个识别过程中的最难的一个环节。能否准确地识别出同形异码字将会影响识别率的提高。解决同形异码字的识别问题时，我们可以利用一些规则，也可以以某一字符的词中位置为上下文信息来解决同形异码字识别问题。我们在“字符分库”中所设置的一些属性字段的属性值将会对解决同形异码字识别问题提供判别条件。如，“形式”、“阴阳性”、“前面出现的字符”、“后面出现的字符”等属性字段。

2 “字符分库”中所设置的语法属性字段及该分库中所使用的一些符号的说明

“字符分库”收入传统蒙古文字符共 249 个，其中蒙古文特有的标点符号、数字、控制符 28 个，名义字符、变形显现字符 151 个，强制性合体字 70 个，并设置了 13 个语法属性字段。



2.1 语法属性字段及其取值范围的说明

(1) 总序号

填写该传统蒙古文字符在本分库中的记录序号。

(2) Unicode



如果该传统蒙古文字符是名义字符，直接填写该字符的 Unicode 编码。

在研制蒙古文编码国际标准时，一个蒙古文名义字符及它的所有变形显现字符共同享有同一个 Unicode 编码。所以一个蒙古文变形显现字符的 Unicode 编码跟它相应的名义字符的 Unicode 编码是一样的。如，传统蒙古文名义字符“” (A) 的 Unicode 编码是 1820，它有“”等八个变形显现形式。这些变形显现形式的 Unicode 编码都是 1820。在填写这些变形显现形式的 Unicode 字段时，我们依据传统蒙古文字符有独立、词首、词中、词尾形式的特点，按 1820-0 独立、1820-1 首写、1820-2 中写、1820-3 尾写、1820-3-1 第一尾写、1820-A (Alternative) 选择等方式来填写相应字符的 Unicode 编码。



在制定蒙古文编码国际标准时，我们规定不对蒙古文强制性合体字进行切分。所以对蒙古文强制性合体字的“Unicode”字段无法填写其属性值。因此蒙古文强制性合体字的“Unicode”字段值为空。

(3) 方正编码 (A)


有两种方正编码。一种是方正样本编码 (称为 A)，一种是方正显现字符编码 (称为 B)。其方正编码 (A) 与 (B) 之间的差别是：高字节为 b0，如 b040 等，A=B；高字节为 b1，如 b144、b145 等，A-2=B；高字节为 b2，如 b274、b275 等，A-3=B。

在该字段中填写该传统蒙古文字符的相应方正编码 (A)。例如，“” (A 的第二个词中形式) 的方正编码 (A) 是 b148；“” (BA 的词首、词中形式) 的方正编码 (A) 是 b06a。

(4) 方正编码 (B)

在该字段中填写该传统蒙古文字符的相应方正编码 (B)。例如，“” (A 的第二个词中形式) 的方正编码 (B) 是 b146；“” (BA 的词首、词中形式) 的方正编码 (B) 是 b06a。


(5) 编号

我们专门为“字符分库”设置了一套编号系统，第一个数字如果是“0”代表该字符是传统蒙古文特有的标点符号、数字、控制符；第一个数字如果是“1”代表该字符是名义字符、变形显现字符；第一个数字如果是“2”代表该字符是强制性合体字。其他的三个数字表示的是该字符在字符分库中的记录编号，其中有必要说明的是同形的字符有相同的编号。我们可以用“编号”字段排序，很容易就得到同形的传统蒙古文字符。如，元音 A 的第一个词中形式、元音 E 的词中形式、辅音 NA 的第一个词中形式都是“”，所以这三个记录的编号都是 1028。


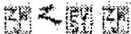

(6) 字符

传统蒙古文特有的标点符号、数字、控制符、名义字符、变形显现字符、强制性合体字的字形。


(7) 与其他文种的交叉

如果对该字符在其他文种（托忒、锡伯、满文、阿礼嘎礼）中也认同，我们就填写那个交叉文种的名称。如，“”（CHA）在托忒、锡伯、满文中也有。所以我们就在该字段中填写“托忒、锡伯、满文”。


(8) 录入方法

传统蒙古文字符有两种录入方法。一种是在字符串（或词）中的录入，一种是单独出现时的录入方法。我们在“录入方法”这一属性字段中填写该字符的这两种录入方法。如，元音 A 的第二个词中形式“”的单独出现时的录入法是。书写在字符串（或词）中时把控制符去掉。

(9) 字符说明

给出该字符的名称及简单的说明。例如，“”字符是用于传统蒙古文和托忒文的文章或段落末尾。


(10) 形式

填写该字符是独立、词首、词中、词尾形式等特点。如，“”是元音 OE 和 UE 的词首形式，我们就在该字符的“形式”字段中填写“词首”即可。


(11) 阴阳性

在蒙古文正字法中有元音和谐律，阳性元音后必须跟阳性元音或中性元音，阴性元音后必须跟阴性元音或中性元音。在“阴阳性”字段中我们就填写该传统蒙古文字符的这一特征。传统蒙古文强制性合体字可以有两种或四种读音，在具体的单词中才能知道它的阴阳性，所以我们在传统蒙古文强制性合体字的“阴阳性”字段中填写“阴阳性”这一属性值。

(12) 前面出现的字符

填写在传统蒙古文文本中该字符前面可能出现的字符条件。如，元音 A 的第一个独立体形式“”，在传统蒙古文文本中这一字符的前面出现的字符是空格，我们就在该字段中填写“space”。

(13) 后面出现的字符








填写在传统蒙古文文本中该字符后面可能出现的字符条件。如，元音 A 的第二个词中体形式“”的后面出现的字符可以是任一个传统蒙古文字符。我们就在该字段中填写“*”。

2.2 “字符分库”中所使用的字符的说明

“space”表示空格；“mspace”表示蒙古文空格；“*”表示任一个传统蒙古文字符；“c1”表示强辅音，“c2”表示弱辅音；“v1”表示阳性元音，“v2”表示阴性元音；“cs (control symbol)”表示控制符。

3 建立“字符分库”时遇到的一些问题

3.1 对特殊问题的解决

在蒙古文字符中有一个“”（MONGOLIAN BIRGA，用于传统蒙古文和托忒文的文章或段落首）字符，它也有五种（、、、、）变形显现形式。但是这四种变形显现形式不分独立、词首、词中、词尾等。所以不能按上述对显现字符的“Unicode”字段值的填写方法来填写这些字符的 Unicode 编码。我们采用了其他的方法解决了这一特殊问题。“”的 Unicode 是 1800，我们对它的五种变形显现形式用以下方法来表示。即，1800-0，1800-1，1800-2，1800-3，1800-4。

3.2 有待于探讨的问题

“字符分库”中是否包括不是传统蒙古文特有的一些通用的标点符号呢？如，“!、{}、[]”等标点符号。从理论上讲，“字符分库”应该囊括传统蒙古文中的所有字符才能保证其完整性。但是我们在“标点符号分库”中已经翔实地描述了这些通用的标点符号，所以我觉得在“字符分库”中就不收入这些通用的标点符号。

4 结束语

目前，建立“字符分库”的主要目的是为印刷体蒙古文识别研究做准备。因为印刷体蒙古文识别工作还处于起步阶段，所以在今后的工作实践中，可能会遇到很多实际问题，届时我们将及时对“字符分库”的内容做相应的调整。

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. 5-107
- [2] 那顺乌日图. “蒙古语语法信息词典”框架设计[D]. 内蒙古大学蒙古学学院资料室：内蒙古大学，2000年
- [3] 那顺乌日图. 蒙古文信息处理[M]. 赤峰：内蒙古科学技术出版社. 1998年. P. 185-200.
- [4] 丁晓青, 郭繁夏. 汉字识别研究和技术的发展与现状[J]. 电子与电脑, 1995, 2: P. 109-110
- [5] 李振宏, 高光来. 印刷体蒙古文文字识别中常用特征的获取[J]. 微机发展, 2003, 11: P. 117-118
- [6] 李振宏, 高光来, 侯宏旭等. 印刷体蒙古文文字识别的研究[J]. 内蒙古大学学报, 2003, 7: P. 454-455
- [7] 李伟, 高光来, 侯宏旭等. 印刷体蒙古文识别技术中切分方法的设计与实现[J]. 内蒙古大学学报, 2003, 5: P. 357-358

附件：“字符分库”样本

总序号	Unicode	方正 GBK 编码(A)	方正 GBK 编码(B)	编号	字符	叉 与其它文种的交叉	字符说明	录入方法	形式	阴阳性	前面出现的字符	后面出现的字符
023	1818	b058	b058	0023	᠈	托忒	蒙古文特有的数字：8	᠈				
024	1819	b059	b059	0024	᠉	托忒	蒙古文特有的数字：9	᠉				
025	1820-0-1	b060	b060	1025	ᠠ	托忒, 锡伯文, 满文	元音 [a] 的第一个独立体形式	ᠠ	独立	阳性	space	space
026	1820-0-2	b061	b061	1026	ᠡ		元音 [a] 的第二个独立体形式	ᠡ ᠡ	独立	阳性	space	space
027	1820-1	b146	b144	1027	ᠠ	托忒, 锡伯文, 满文	元音 [a] 的词首体形式	ᠠ ᠠ	词首	阳性	space	*
028	1820-2-1	b147	b145	1028	ᠠ	托忒, 锡伯文, 满文	元音 [a] 的第一个词中体形式	ᠠ ᠠ	词中	阳性	*	*
029	1820-2-2	b148	b146	1029	ᠡ	托忒	元音 [a] 的第二个词中体形式	ᠡ ᠡ	词中	阳性	*(cs)	*
030	1820-2-3	b149	b147	1030	ᠢ		元音 [a] 的第三个词中体形式	ᠢ ᠢ	词中	阳性	mSPACE	space