

蒙古语言文字数字资源建设与共享工程  
信息处理用蒙古文相关标准  
MGC/01-11

# 信息技术 传统蒙古文字词数统计

Information technology — Specification for the counting of  
traditional Mongolian characters and words

内蒙古自治区民族事务委员会

内蒙古大学

2016年11月

# 目 次

目次 .....	I
前言 .....	II
1 范围 .....	1
2 规范性引用文件 .....	1
3 术语和定义 .....	1
4 计入传统蒙古文字数统计的名义字符 .....	2
5 有条件计入传统蒙古文字数统计的通用标点符号 .....	3
6 有条件计入传统蒙古文字数统计的通用控制符 .....	4
7 传统蒙古文字词数统计原则 .....	4

# 前 言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准由内蒙古自治区民族事务委员会提出。

本标准起草单位：内蒙古大学、内蒙古民族事务委员会。

本标准主要起草人：斯·劳格劳，欧日乐克

# 信息技术 传统蒙古文字词数统计规范

## 1 范围

本标准规定了传统蒙古文字（字符）数统计规范。

本标准规定了传统蒙古文词数统计规范。

本标准适用于具有传统蒙古文字词数统计功能的电子信息产品。

## 2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB 13000—2010 信息技术 通用多八位编码字符集（UCS）

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

## 3 术语和定义

下列术语和定义适用于本文件。

### 3.1

**字符** character

共组织、控制或表示数据用的元素集合中的一个元素。

### 3.2

**编码字符** coded character

字符及其编码表示。

### 3.3

**编码字符集** coded character set

一组无歧义的规则，用于建立一个字符集和该字符集的字符及其编码表示之间的一一对应关系。

### 3.4

**控制功能** control function

影响数据记录、处理、传输或解释的动作。

### 3.5

**控制字符** control character

一种具有控制功能的字符。将它插在别的字符之间，是为了启动、修改或停止某种控制功能的执行。控制字符由机器解释，它不是图形符号。

### 3.6

**蒙古文控制字符** Mongolian control character

一种具有控制功能的字符。将它插入在名义字符之前或之后，完成变体选择或位置表示。

### 3.7

#### 名义字符 nominal form

蒙古文字母的主要形式。它适用于蒙古语的书面形式以及附加符号的表示、传输、交换、处理、存储、输入及显示。

### 3.8

#### 文本 text

由编码字符构成的数据。

### 3.9

#### 传统蒙古文单词 traditional Mongolian word

仅由传统蒙古文字母、控制符以及词中连接符（U+180A）构成的编码字符序列。

### 3.10

#### 字数统计 character counting

统计文本中的编码字符个数。

### 3.11

#### 传统蒙古文字数统计 character counting for traditional Mongolian

统计文本中传统蒙古文编码字符以及蒙古文引用的控制符及标点符号的个数。

### 3.12

#### 词数统计 word counting

统计文本中的单词个数。

### 3.13

#### 传统蒙古文词数统计

统计文本中传统蒙古文单词个数。

## 4 计入传统蒙古文字数统计的名义字符

表 1. 计入传统蒙古文字数统计的名义字符

	180	181	182	183	184	185	186	188	189	18A
0	᠎	᠎	᠎	᠎	᠎			᠎	᠎	
1	᠎	᠎	᠎	᠎	᠎			᠎	᠎	
2	᠎	᠎	᠎	᠎	᠎			᠎	᠎	
3	᠎	᠎	᠎	᠎		᠎		᠎	᠎	
4	᠎	᠎	᠎	᠎				᠎	᠎	

5	ᠰ	ᠰ	ᠰ	ᠰ				ᠰ	ᠰ	
6		ᠰ	ᠰ	ᠰ				ᠰᠰᠰ	ᠰ	ᠰ
7		ᠰ	ᠰ	ᠰ				ᠰ	ᠰ	ᠰ
8		ᠰ	ᠰ	ᠰ		ᠰ		ᠰ		
9		ᠰ	ᠰ	ᠰ				ᠰ		ᠰ
A	.		ᠰ	ᠰ				ᠰ		
B	[FV S1]		ᠰ	ᠰ		ᠰ		ᠰ		
C	[FV S2]		ᠰ	ᠰ		ᠰ		ᠰ		
D	[FV S3]		ᠰ	ᠰ				ᠰ		
E	[M VS]		ᠰ	ᠰ				ᠰ		
F			ᠰ	ᠰ				ᠰ		

5 有条件计入传统蒙古文字数统计的通用标点符号

表 2. 有条件计入传统蒙古文字数统计的通用标点符号

标点符号		标点符号	
字形	编码	字形	编码
!!	U+203C	ᠰ	U+FE36
?!	U+2048	ᠰ	U+FE3D
!?	U+2049	ᠰ	U+FE3E
;	U+FE14	ᠰ	U+FE3F

!	U+FE15	∨	U+FE40
?	U+FE16	┌	U+FE47
	U+FE31	└	U+FE48
∩	U+FE35	•	U+00B7

## 6 有条件计入传统蒙古文字数统计的通用控制符

- 1) 窄宽度无间断空格 NARROW NO-BREAK SPACE (U+202F)、
- 2) 零宽连接符 ZERO WIDTH JOINER (U+200D)、
- 3) 零宽禁连符 ZERO WIDTH NON-JOINER (U+200C)

## 7 传统蒙古文字词数统计原则

### 7.1

#### 传统蒙古文字数统计原则

传统蒙古文字数统计结果为文本中出现的传统蒙古文名义字符、传统蒙古文引用的通用控制符以及传统蒙古文引用的通用标点符号的总数。

#### 7.1.1 名义字符统计原则

表 1 所列的字符在文本中出现时，不带任何条件计为一个传统蒙古文字符。

#### 7.1.2 通用控制符统计规则

文本中出现的满足下列条件的通用控制符计为一个传统蒙古文字符。

- 1) 如果零宽禁连符 (U+200C) 的前或后邻接字符为传统蒙古文字母，则将该控制符计为传统蒙古文字符；
- 2) 如果零宽连接符 (U+200D) 的前或后邻接字符为传统蒙古文字母，则将该控制符计为传统蒙古文字符；
- 3) 如果窄宽度无间断空格 (U+202F) 的后邻接字符为传统蒙古文字母，则将该控制符计为传统蒙古文字符。

#### 7.1.3 通用标点符号统计规则

方案一：

文本中传统蒙古文名义字符及引用的通用控制符的总比例超过 50% 时包含在表 2 中的“!、?、;、!、?、∩、∪、∩、≅、∩、∪、┌、└、|、•”等标点符号计为一个传统蒙古文字符。

方案二：

1) 表 2 中的“!、?!、!?、 ;、 !、 ?、 ~、 ≡、 ~、 一、 .”等标点符号在文本中出现时，如果其前邻接字符为传统蒙古文或控制符，则将该标点符号计为一个传统蒙古文或控制符；

2) 表 2 中的“^、 ^、 ^、 一、 |”等标点符号在文本中出现时，如果其后邻接字符为传统蒙古文或控制符，则将该标点符号计为传统蒙古文或控制符。

## 7.2


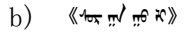
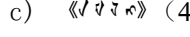
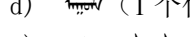
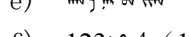
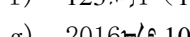
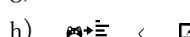
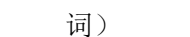
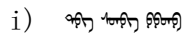
### 传统蒙古文词数统计原则

#### 7.2.1 不包含标点符号的传统蒙古文词数统计原则

满足下列条件的字符串计为一个传统蒙古文单词：

- 1) 由传统蒙古文或控制符、蒙古文自由变体选择符 FVS1 (U+180B) FVS2 (U+180C) FVS3 (U+180D)、蒙古文元音间隔符 MVS (180E)、窄宽度无间断空格 NARROW NO-BREAK SPACE (U+202F)、零宽连接符 ZERO WIDTH JOINER (U+200D)、零宽禁连符 ZERO WIDTH NON-JOINER (U+200C) 以及词中连接符 NIRUGU (U+180A) 构成；
- 2) 包含一个或一个以上蒙古文或控制符；
- 3) 其前邻接字符和后邻接字符均为非 1) 中所列字符；
- 4) 不满足条件 1)、2) 和 3) 的其他任何字符串不计入传统蒙古文词数统计。

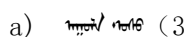
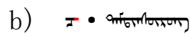
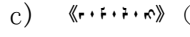
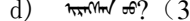
词数统计例子：

- a)  (2 个传统蒙古文单词)
- b)  (4 个传统蒙古文单词)
- c)  (4 个传统蒙古文单词)
- d)  (1 个传统蒙古文单词)
- e)  (1 个传统蒙古文单词)
- f)  (1 个传统蒙古文单词)
- g)  (5 个传统蒙古文单词)
- h)  (10 个传统蒙古文单词)
- i)  (3 个传统蒙古文单词)

#### 7.2.2 包含标点符号的传统蒙古文词数统计原则

- 1) 满足 7.2.1 所列条件的字符串计为一个传统蒙古文单词；
- 2) 传统蒙古文特有的标点符号 BIRGA (U+1800)、省略号 (U+1801)、逗号 (U+1802)、句号 (U+1803)、冒号 (U+1804)、四点 (U+1805) 以及满足 7.1.3 所列条件的通用标点符号 (如表 2) 计为一个传统蒙古文单词。

词数统计例子：

- a)  (3 个单词)
- b)  (3 个单词)
- c)  (9 个单词)
- d)  (3 个单词)