ICS 35.240.01
L 70

中华人民共和国国家标准

GB/T 36452—2018

信息处理用藏文分词规范

Specification on Tibetan segmentation for information processing

2018-06-07 发布　　2019-01-01 实施

国家市场监督管理总局
中国国家标准化管理委员会　发布

目　　次

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本标准起草单位:中国电子技术标准化研究院、西藏大学、西北民族大学、西藏自治区藏语文工作委员会办公室、青海师范大学、青海民族大学、中国科学院软件研究所、西藏自治区工业和信息化厅。

本标准主要起草人:扎西加、欧珠、尼玛扎西、熊涛、格桑多吉、多拉、拉巴泽仁、大罗桑朗杰、高定国、拉琼、仁青诺布、索南尖措、旺堆、小尼玛扎西、普次仁、顿珠次仁、赵栋材、边巴嘉措。

引　言

本标准以现代藏语的词类和分词研究成果为基础，根据藏文词汇特点与构词规律，并参考汉语分词及词类标记相关标准（见参考文献）的部分内容，规定了信息处理用藏文分词规范。

信息处理用藏文分词规范

1 范围

本标准规定了信息处理用藏文分词规范。

本标准适用于藏文信息处理各领域，其他行业和有关学科可参照使用。

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/T 36337—2018 信息处理用藏语词类标记集

3 术语和定义

下列术语和定义适用于本文件。

3.1

词 word

区别事物意义的最小的语法单位。

3.2

词组 phrase

两个或更多词组合成的语言单位。

注：词组可以是实词与实词的组合，也可以是实词和虚词的组合。

3.3

藏文信息处理 Tibetan information processing；TIP

用计算机对藏文的音、形、义等信息进行处理。

3.4

分词单位 segment unit

在分词过程中出现的词。

注：分词单位不仅限于语法词，其中也包含了信息处理所需的一部分结合紧密、使用稳定的词组。

3.5

藏文分词 Tibetan segmentation

将连续的藏文音节序列按照一定的规范重新组合词序列的过程。

4 分词规范

4.1 藏文分词单位和词类的标记

本标准以“/”作为藏文分词单位的标记，藏语词类标记依据 GB/T 36337—2018 的规定。

4.2 一般名词<སྤྱིར་བཏང་གི་མིང་།>(nn)

4.2.1 由单音节名词和单音节形容词组成的词为一个分词单位。