ICS 35.240.01
L 70

中华人民共和国国家标准

GB/T 36337—2018

信息处理用藏语词类标记集

The parts-of-speech tagging set for Tibetan 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、西藏自治区藏语文工作委员会办公室、西北民族大学、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，根据藏文自身的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）的内容，规定了信息处理中藏语词类标记集。

信息处理用藏语词类标记集

1 范围

本标准规定了信息处理中藏语词类标记集。

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 术语和定义

下列术语和定义适用于本文件。

2.1

词类 parts of speech；POS

依据词的句法功能原则所划分的类。

2.2

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。

2.3

词类标注 POS tagging

对分词文本中的每个分词单位标注上词类的过程。

2.4

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 标记符号

本标准中标记符号采用英文字母、汉语拼音字母及藏文拉丁转写字母。

4 词类标记集

4.1 名词〈དངོས་པོའི་མིང་།〉(n)

表示人和事物的名称或时间、处所、方位等，在句中主要充当主语和宾语。

a) 一般名词〈སྤྱིར་བཏང་མིང་།〉(nn)，表示人和事物的名称。

示例：

ཚབ་སྦྱིད། སྤྱི་ཚོགས། ཅོག་ཙེ། གསེར། ཀུ། མེ། ས། འཛུལ་ཚས།

b) 人名〈མིའི་མིང་།〉(nr)，表示人的名称的专有名词。

示例：

རིན་ཆེན་བཟང་པོ། བྱམས་པ་ཐུབ་ཚོགས། ཚེ་རིང་སྒོལབས་རྒྱལ། ཡུང་སྒྲུ་རོལ་བའི་རྡོ། བཀྲ་ཤིས་དར་རྒྱས། བཀྲ་ཤིས་ཚེ་རིང་།