ICS 01.040.01
A 22

中华人民共和国国家标准

GB/T 20532—2006

信息处理用现代汉语词类标记规范

Standard of POS tag of contemporary Chinese for CIP

2006-09-18 发布　　2007-03-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

目　　次

前　言

本标准由教育部语言文字信息管理司提出。

本标准由教育部语言文字信息管理司归口。

本标准起草单位：教育部语言文字应用研究所。

本标准主要起草人：靳光瑾、肖航、郭曙伦、富丽、章云帆、于桂英、陈玉泉、王立。

信息处理用现代汉语词类标记规范

1 范围

本标准规定了信息处理中现代汉语词类及其他切分单位的标记代码。

本标准适用于汉语信息处理，也可供现代汉语教学与研究参考。

2 术语和定义

下列术语和定义适用于本标准。

2.1

汉语信息处理 Chinese information processing；CIP

用计算机对汉语形、音、义等信息进行输入、排序、存储、输出、统计、提取等。

2.2

切分单位 segment unit

汉语信息处理使用的、具有确定语法功能的基本单位。它包括本标准的规则所限定的词、短语及其他单位。

2.3

词类 parts of speech；POS

词的语法分类，主要是根据语法功能划分出来的类。

2.4

标记 tag

对文本中切分单位的类别进行标注的代码。

3 总则

3.1 切分单位的范围

本标准的切分单位包括词、短语和其他切分单位，如习用语、缩略语、前接成分、后接成分、语素字、非语素字、标点符号、非汉字符号等。

3.2 词类划分的原则

本标准的词类分类体系参考了吕叔湘、朱德熙、胡裕树等先生的语法体系和《中学教学语法系统提要》。

本标准根据汉语信息处理的特点和要求，主要依据语法功能原则划分词类。

3.3 标记代码的制定原则

依据国际通常做法，标记代码主要采用英文术语的字母。例如，“名词”，采用英文术语“noun”的首字母“n”作为标记代码；“数词”，采用英文术语“numeral”的第三个字母“m”作为标记代码。

汉语独有的，或使用英文术语字母不便的，依据国内通常做法，标记代码采用汉语拼音字母。如，“缩略语”，采用汉字“简”汉语拼音的首字母“j”作为标记代码；“语素字”，采用汉字“根”汉语拼音的首字母“g”作为标记代码。

4 词类及其他切分单位分类

本标准将词类划分为 13 个一级类，16 个二级类；其他切分单位划分为 7 个一级类，13 个二级类。用户可根据需要自行增补。