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。

4.1 词类划分及标记代码

4.1.1 名词(n),表示人或事物的名称,在句子中主要充当主语和宾语。

4.1.1.1 普通名词(ng),表示事物的名称。如:

人 马 书 教师 飞机 电冰箱 阿姨 桌子 木头

道德 理论 历史 思想 文化 因素 作风 哲学

4.1.1.2 时间名词(nt),包括一般所说的时量词。如:

年 月 日 分 秒

现在 过去 昨天 去年 将来 宋朝 星期一

4.1.1.3 方位名词(nd),表示位置的相对方向。如:

上 下 左 右 前 后 里 外 中 东 西 南 北

前边 左面 里头 中间 外部

4.1.1.4 处所名词(nl),表示处所。如:

空中 高处 隔壁 门口 附近 边疆 一旁 野外

4.1.1.5 人名(nh),表示人的名称的专有名词。

华罗庚 阿凡提 诸葛亮 司马相如 松赞干布 卡尔·马克思

4.1.1.6 地名(ns),表示地理区域名称的专有名词。如:

亚洲 大西洋 地中海 阿尔卑斯山 加拿大

中国 北京 浙江 景德镇 呼和浩特 中关村

4.1.1.7 族名(nn),表示民族或部落名称的专有名词。如:

回族 藏族 壮族 蒙古族 维吾尔族 哈萨克族

4.1.1.8 机构名(ni),表示团体、组织、机构名称的专有名词。如:

联合国 教育部 北京大学 中国科学院

4.1.1.9 其他专有名词(nz)。如:

五粮液 宫爆鸡丁 桑塔纳

4.1.2 动词(v),表示动作、行为,心理活动、生理状态及事物的存现、变化等,在句子中主要充当谓语。

4.1.2.1 及物动词(vt),能够带宾语。如:

吃 打 擦 洗 喂 借 送 买 捧 提 填

喜欢 告诉 接受 羡慕 考虑 调查 同意 发动

4.1.2.2 不及物动词(vi),不能够带宾语。如:

病 休息 咳嗽 瘫痪 游泳 睡觉

4.1.2.3 联系动词(vl),表示关系的判断。如:

是

4.1.2.4 能愿动词(vu),表示可能、意愿。如:

能够 能 应该 可以 可能 情愿 愿意 要

4.1.2.5 趋向动词(vd),表示趋向。如:

(走)上 (趴)下 (进)来 (回)去

(跑)上来 (掉)下去 (提)起来 (扔)过去

4.1.3 形容词(a),表示性质、状态,在句中主要充当谓语、定语、状语和补语。

4.1.3.1 性质形容词(aq),表示性质。如:

好 高 美 大 勇敢 危险 漂亮 干净 伟大

4.1.3.2 状态形容词(as),表示状态。如:

雪白 黢黑 通红 冰凉 绿油油 亮堂堂 白花花 冷冰冰

4.1.4 区别词(f),表示事物的区别性特征,在句子中只能做定语修饰名词或跟助词“的”组成“的”字结

构。如：

男　女　公　母　雌　雄　微型　国产　军用

4.1.5　数词(m)，表示数目和次序。如：

零　一　半　百　千　百万　一百零八

第一　第十八

4.1.6　量词(q)，表示人、事物或动作的单位。如：

个　条　片　匹　辆　尺　斤　两　吨　支　回　次　遍　千瓦时

4.1.7　代词(r)，起替代和复指作用。如：

我　你　他　这　那　谁　我们　你们　他们

这个　那个　大家　自己　什么　哪里　怎么　怎么样

4.1.8　副词(d)，修饰或限制动词和形容词，表示范围、程度等。在句子中做状语。如：

都　只　就　仅　很　再三　屡次　将　不　却

总共　正在　常常　重新　曾经　竟然　居然

4.1.9　介词(p)，引介名词性成分，不单独充当句子成分。如：

把　被　从　向　对　凭

按照　对于　为了　自从　关于

4.1.10　连词(c)，连接词、短语或句子，表示两者之间所具有的某种关系。如：

和　同　与　及　并　或

并且　而且　或者　因为　所以

4.1.11　助词(u)，附着在词、短语、句子后面表示某种附加意义。如：

的　地　得　了　着　过　等等　似的　一样

4.1.12　叹词(e)，表示感叹、呼唤或应答，可独立成句或在句中充当独立成分。如：

啊　嗯　唉　哎　哼　哦　哎哟　哎呀

4.1.13　拟声词(o)，模拟自然界事物的某种声音，不能单独成句。如：

砰　滴答　扑通　咕咚　丁丁当当

4.2　其他切分单位划分及标记代码

4.2.1　习用语(i)，一种相沿习用的定型短语。

4.2.1.1　名词性习用语(in)。如：

海市蜃楼　井底之蛙　蛛丝马迹

4.2.1.2　动词性习用语(iv)。如：

跑龙套　打官腔　吃老本　与时俱进　励精图治

4.2.1.3　形容词性习用语(ia)。如：

丰富多彩　艰苦朴素　光明正大

4.2.1.4　连词性习用语(ic)。如：

总而言之　由此可见　综上所述

4.2.2　缩略语(j)，专有名词或常用语的简缩形式。

4.2.2.1　名词性缩略语(jn)。如：

人大　五四　奥运

4.2.2.2　动词性缩略语(jv)。如：

调研　离退休

4.2.2.3　形容词性缩略语(ja)。如：

短平快　高精尖

4.2.3　前接成分(h)，词根前面的附加构词成分。如：

阿　老　初　第

4.2.4　后接成分(k),词根后面的附加构词成分。如:

子　儿　头　化　们　式　性　者

4.2.5　语素字(g),汉字字符集中一般不单独使用的汉字。

4.2.5.1　名词性语素字(gn)。如:

民　农　材

4.2.5.2　动词性语素字(gv)。如:

抒　究　涤

4.2.5.3　形容词性语素字(ga)。如:

殊　遥　伟

4.2.6　非语素字(x),汉字字符集中单独使用时不具有意义的汉字,如:

垃　琵　蜘　蜍　鸯　蜻

4.2.7　其他(w)

4.2.7.1　标点符号(wp),如:

，　。　、　；　？　！　：　“”　……

4.2.7.2　非汉字字符串(ws),如:

office　windows

4.2.7.3　其他未知的符号(wu)。

5　词类及其他切分单位标记代码表

词类及其他切分单位标记代码表见表1。

表1　词类及其他切分单位标记代码表

(按标记代码的字母顺序排列)

序号	标记代码		类别名称	代码说明
	一级类	二级类		
1	a		形容词	adjective
2		aq	性质形容词	adjective-quality
3		as	状态形容词	adjective-state
4	c		连词	conjunction
5	d		副词	adverb
6	e		叹词	exclamation
7	f		区别词	difference
8	g		语素字	“根”的汉语拼音首字母
9		ga	形容词性语素字	“根”的汉语拼音首字母-adjective
10		gn	名词性语素字	“根”的汉语拼音首字母-noun
11		gv	动词性语素字	“根”的汉语拼音首字母-verb
12	h		前接成分	head
13	i		习用语	idiom
14		ia	形容词性习用语	idiom-adjective
15		ic	连词性习用语	idiom-conjunction

表 1（续）

序号	标记代码		类别名称	代码说明
	一级类	二级类		
16		in	名词性习用语	idiom-noun
17		iv	动词性习用语	idiom-verb
18	j		缩略语	“简”的汉语拼音首字母
19		ja	形容词性缩略语	“简”的汉语拼音首字母-adjective
20		jn	名词性缩略语	“简”的汉语拼音首字母-noun
21		jv	动词性缩略语	“简”的汉语拼音首字母-verb
22	k		后接成分	依据通常做法
23	m		数词	numeral
24	n		名词	noun
25		nd	方位名词	noun-direction
26		ng	普通名词	noun-general
27		nh	人名	noun-human
28		ni	机构名	noun-institution
29		nl	处所名词	noun-location
30		nn	族名	noun-nation
31		ns	地名	noun-space
32		nt	时间名词	noun-time
33		nz	其他专有名词	noun-“专”的汉语拼音首字母
34	o		拟声词	onomatopoeia
35	p		介词	preposition
36	q		量词	quantity
37	r		代词	pronoun
38	u		助词	auxiliary
39	v		动词	verb
40		vd	趋向动词	verb-direction
41		vi	不及物动词	verb-intransitive
42		vl	联系动词	verb-linking
43		vt	及物动词	verb-transitive
44		vu	能愿动词	verb-auxiliary
45	w		其他	依据通常做法
46		wp	标点符号	依据通常做法
47		ws	非汉字字符串	“w”-string
48		wu	其他未知符号	“w”-unknown
49	x		非语素字	依据通常做法

中 华 人 民 共 和 国
国 家 标 准
信息处理用现代汉语词类标记规范
GB/T 20532—2006
*
中国标准出版社出版发行
北京复兴门外三里河北街16号
邮政编码:100045
网址 www.spc.net.cn
电话:68523946 68517548
中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销
*
开本 880×1230 1/16 印张 0.75 字数 12 千字
2007年3月第一版 2007年3月第一次印刷
*
书号:155066·1-28954 定价 14.00 元

GB/T 20532-2006