中华人民共和国国家标准

信息处理用现代汉语分词规范

GB/T 13715—92

Contemporary Chinese language word segmentation specification for information processing

1 主题内容与适用范围

1.1 主题内容

本规范规定了现代汉语的分词原则，以满足信息处理的需要。它对汉语信息处理的规范化，对各种汉语信息处理系统之间的兼容性有重要的作用。

1.2 适用范围

本规范适用于汉语信息处理各领域，其他行业和有关学科可以参考使用。

汉语信息处理各领域可以根据其专门需求，进一步补充和细化本规范的规定。

2 引用标准

GB 12200 汉语信息处理词汇

3 术语

以下术语引自 GB 12200。

3.1 汉语信息处理

用计算机对汉语的音、形、义等信息进行的处理。

3.2 词

最小的能独立运用的语言单位。

3.3 词组

由两个或两个以上的词，按一定的语法规则组成，表达一定意义的语言单位。

3.4 分词单位

汉语信息处理使用的、具有确定的语义或语法功能的基本单位。它包括本规范的规则限定的词和词组。

3.5 汉语分词

从信息处理需要出发，按照特定的规范，对汉语按分词单位进行划分的过程。

4 概述

本规范以信息处理应用为目的，根据现代汉语的特点及规律，规定现代汉语的分词原则。

本规范用下划线“_____”作为分词单位标记。

4.1 空格或标点符号是计算机中分词单位的分隔标记。作为分隔标记的标点符号有：句号、逗号、顿号、分号、冒号、问号、叹号、引号、括号、破折号、省略号、书名号、间隔号、连接号及符号“/”等。

4.2 二字或三字词，以及结合紧密、使用稳定的二定或三字词组，一律为分词单位。例如：

国家技术监督局 1992-10-04 批准　　1993-06-01 实施

发展　可爱　红旗

对不起　自行车　青霉素

4.3　四字成语一律为分词单位。例如：

胸有成竹　欣欣向荣

四字词或结合紧密、使用稳定的四字词组，一律为分词单位。例如：

社会主义　春夏秋冬　由此可见

4.4　五字或五字以上的谚语、格言等，分开后如不违背原有组合的意义，应予切分。例如：

时间 就 是 生命

失败 是 成功 之 母

人 心 齐，泰山 移

结合紧密、使用稳定的词组，分开后如违背原有组合的意义，或影响进一步的处理，则不予切分。例如：

不管三七二十一

4.5　惯用语和有转义的词或词组，在转义的语言环境下，一律为分词单位。例如：

妇女能顶半边天

他真小气，象个铁公鸡

4.6　略语一律为分词单位。例如：

科技　奥运会　工农业

4.7　分词单位加形成儿化音的“儿”，一律为分词单位。例如：

花儿　悄悄儿　玩儿

4.8　在现化汉语中出现的非汉字符号，例如其他语言的字符串、数学符号、化学符号、阿拉伯数字等，仍保留原有形式。例如：

CAD CO ：= cm 1 247 1 298 576 3.14

4.9　现代汉语中其他语言的汉字音译外来词，不予切分。例如：

巧克力　　吉普

4.10　不同的语言环境中的同形异构现象，按照具体语言环境的语义，根据本规范的规定进行切分。例如：

把 手抬起来

这个把手是木制的

5　具体说明

为叙述方便，本规范沿用了把词分为名词、动词、形容词、代词、数词、量词、副词、介词、连词、助词、语气词、叹词、象声词等十三类的方法。

5.1　名词

5.1.1　普通名词

5.1.1.1　二字的名词或结合紧密的二字名词词组，一律为分词单位。例如：

火车　牛肉　钢铁

5.1.1.2　结合紧密，分开后如违背原有组合的意义的名词性词组，一律为分词单位。例如：

有功功率　被子植物

5.1.1.3　由形容词加名词组成的词组，应予切分。例如：

绿 叶　　小 床

形容词加名词组成的有转义的词组，一律为分词单位。例如：

小媳妇　　戴高帽儿

5.1.1.4 前加成分加名词性分词单位应为分词单位。例如：

阿哥　　老鹰　　非金属　　超声波

5.1.1.5 名词性分词单位加如下类型的后加成分：

家　手　性　员　子　化　长　头　者

应为分词单位。例如：

科学家　拖拉机手　革命性

理发员　椅子　　标准化

科长　　木头　　学者

名词性分词单位后如有多个后加成分，则它们是一个分词单位。例如：

物理学家

5.1.1.6 名词性分词单位前后如有前加成分和后加成分，则它们是一个分词单位。例如：

非党员　　超导性

5.1.1.7 各类专业的基本术语为分词单位。例如：

加速度　　中央处理器

5.1.1.8 方位词应予单独切分。例如：

桌子上　　长江以北

5.1.1.9 除“人们”之外，仅表示前一个名词性分词单位复数的“们”单独切分。例如：

朋友 们　学生 们

但是“哥儿们　爷们儿”等是分词单位。

5.1.1.10 时间名词或词组的分词规则如下：

a. 一年的十二个月份以及每周的七天，一律为分词单位。例如：

五月　　元月　　3 月

星期日　　礼拜三

b. “年、日、时、分、秒”分别为分词单位。例如：

1988 年 3 月15 日

11 时 42 分 8 秒

c. “前、后、上、下、大前、大后”等直接与时间名词或量词组合时，它们为一个分词单位。例如：

前天　　后年　　上星期

下月　　大前天　　大后年

d. “初”加十以内的数字一律为分词单位。例如：

初一　　初八

5.1.2 专有名词

5.1.2.1 人名、称谓等处理如下：

a. 汉族人名的姓和名分别单独切分。例如：

张 胜利　　欧阳 海

b. 其他国家、其他民族的人名按其习惯形式切分。例如：

卡尔·马克思　　牛顿　　小林 多喜二　　才旦卓玛

c. 带职务、职称的称呼一律切分。例如：

张 教授　王 部长　李 师傅

d. 简称、尊称等为分词单位。例如：

老张　小李　郭老　陈总

e. 带排行的亲属称谓一律切分。例如：

三叔　大 女儿

5.1.2.2 民族名、地名中的“族、省、市、州、县、乡、区、江、河、山”等应单独切分。但包括“族、省、市、州、县、乡、区、江、河、山”等只有两个字的民族名、地名，则不予切分。例如：

汉族　哈萨克 族　北京 市　浙江　省　正定 县　长江　忻县

专名部分不能单独存在而保持原有意义的地名，不予切分。例如：

牡丹江　横断山

街、路、村镇名称，各大洋和各大海一律为分词单位。例如：

长安街　学院路　周口店　刘家村　大西洋　地中海

5.1.2.3 国家全名一律为分词单位。例如：

中华人民共和国　大不列颠及北爱尔兰联合王国

5.1.2.4 组织、机构、单位的全名按组成其全名的分词单位切分。例如：

联合国　教科文　组织

中国 共产党

5.1.2.5 商品牌号、品种、产品系列名称中的专有名词与普通名词一律分别切分。例如：

永久 牌　中华 烟　牡丹 Ⅱ 型

5.2 动词

5.2.1 动词的重叠形式较多，具体规定如下：

a. 单字动词重叠使用为一个分词单位。例如：

看看　动动

b. 二字动词性分词单位的重叠方式“AABB”为一个分词单位。例如：

来来往往　拉拉扯扯

c. “AAB、ABAB”重叠形式的动词词组应予切分。例如：

说说 看　研究 研究

d. “A 一 A、A 了 A、A 了一 A”重叠形式的动词词组应予切分。例如：

谈 一 谈　想 一 想

读 一 读　想 了 想

想 了 一 想

5.2.2 动词前的否定副词一律单独切分。例如：

不 写　不 能　没 研究　未 完成

5.2.3 用肯定加否定的形式表示疑问的动词词组一律切分，不完整的则不予切分。例如：

说 没 说　看 不 看　相信 不 相信

相不相信

5.2.4 动宾结构的词或结合紧密、使用稳定的二字动宾词组，不予切分。例如：

开会　跳舞

解决吃饭问题

孩子该念书了

结合不紧密或有众多与之相同结构词组的动宾词组一律切分。例如：

吃 鱼　学 滑冰

写 信(写 文章；写 论文；写 书；……)

动宾结构的词或词组如中间插入其他成分，则应予切分。例如：

吃 两 顿 饭　跳 新疆 舞

5.2.5 动补结构的二字词或结合紧密、使用稳定的二字动补词组，不予切分。例如：

打倒　提高　加长　做好

“2＋1”或“1＋2”结构的动补词组一律切分，三字以上的动补结构词组也一律切分。例如：

整理 好　说 清楚　解释 清楚

动补结构的词或词组如中间插入“得、不”，应予切分。例如：

打 得 倒　提 不 高

5.2.6　偏正结构的词，以及结合紧密、使用稳定的偏正结构的词组，不予切分。否则应予切分。例如：

胡闹　瞎说　死记

早 来　晚 走　重 说

5.2.7　复合趋向动词一律为分词单位。例如：

出去　进来

当插入“得、不”时应予切分。例如：

出 得 去　进 不 来

5.2.8　动词与趋向动词结合的词组一律切分。例如：

寄 来　　跑 出 去

5.2.9　单字动词无连词并列，并且均保持各自独立动词意义的词组，一律切分。例如：

苫 盖　　听 说 读 写

多字动词无连词并列，一律切分。例如：

调查 研究　　宣传 鼓动

5.3　形容词

5.3.1　形容词的重叠形式“AA、AABB、ABB、AAB、A 里 AB”一律为分词单位。例如：

大大　　高高

高高兴兴　　匆匆忙忙

绿油油　　红通通

蒙蒙亮　　马里马虎

“ABAB”重叠形式的形容词应予切分。例如：

雪白 雪白　滚圆 滚圆

5.3.2　“一 A 一 B、一 A 二 B、半 A 半 B、半 A 不 B、有 A 有 B”等类型的形容词性词组，不予切分。例如：

一心一意　　一清二楚

半明半暗　　半生不熟

有条有理

5.3.3　形容词的并列形式按以下规则切分：

a.　两个单字形容词并列且改变词性的，一律不予切分。例如：

长短　深浅　大小

b.　形容词并列且各自保持原有形容词语义的词组，应予切分。例如：

大 小尺寸　　光荣 伟大

5.3.4　有关颜色的形容词或词组不予切分。例如：

浅黄　橄榄绿

5.3.5　用肯定加否定的形式表示疑问的形容词词组一律切分，不完整的则不切分。例如：

容易 不 容易

容不容易

5.4　代词

5.4.1　单字代词加“们”为分词单位。例如：

我们　你们　它们　他们

5.4.2　“这、那、哪”加量词“个”或“些、样、么、里、边”等为一个分词单位。例如：

这个　这么　这边
那些　那样　那里
哪个　哪里　哪些

5.4.3 “这、那、哪”加数、量、名词性分词单位一律切分。例如：

这 十 天　　那 人　那 种

5.4.4 疑问代词或词组为分词单位。例如：

多少　　怎样
为什么　　什么

5.4.5 “各、每、某、本、该、此、全”等代词与后面的量词或名词一律切分。例如：

各 国　　每 种
某 工厂　　本 部门
该 单位　　此 人
全 校

5.5 数词

5.5.1 数词与量词一律切分。例如：

三 个　一 种

5.5.2 汉语数位词分别为分词单位。例如：

一亿八千零四万七百二十三

5.5.3 表示序数的“第”与后面的数词一律切分。例如：

第 一　第 四　第 五 十 三

5.5.4 分数中的“分之”为一个分词单位。例如：

五 分之三　百分之二　万分之五

5.5.5 数字并列表示概数时，表示概数的数字为分词单位。例如：

八九 公斤　十 七八 岁

5.5.6 表示概数的“多、来、几”等在数词或量词之后时，一律为分词单位。例如：

两 点 多　一 千 多 人　十 来 家　十 几 个

5.5.7 “些、一些、点儿、一点儿”等表示概数的词在形容词或动词之后时，一律切分。例如：

大 些　　懂 一些
快 点儿　快 一点儿

5.5.8 “近、约、数”等在数词或数位词前，与之连用表示概数时，应予切分。例如：

近 千 人　约 三 百　数 万

“成、上”在数位词前，与之连用表示概数时，不予切分。例如：

成百　　上千

5.6 量词

5.6.1 量词重叠使用不予切分。例如：

年年　　天天　　个个　　家家户户

5.6.2 复合量词或词组为分词单位。例如：

人年　　人次　　架次　　吨公里

5.7 副词

5.7.1 副词一律为分词单位。例如：

很好　　都来了
刚走　　互相协助

5.7.2 以下经常使用，起副词作用的词组为分词单位：

越来越　　不得不　　不能不

起关联作用的“越…越…、又…又…”等应予切分。例如：

越 走 越 远　　又 香 又 甜

5.8 介词

介词一律为分词单位。例如：

生于　　走向胜利　　按照规定

5.9 连词

连词一律为分词单位。例如：

工人和农民　　光荣而伟大

5.10 助词

5.10.1 结构助词“的、地、得、之”一律为分词单位。例如：

他的书　　慢慢地走　　说得快

美丽的城市　　中国的大熊猫　　成功之路

5.10.2 时态助词“着、了、过”一律为分词单位。例如：

看着　　看了　　看过

5.10.3 助词“所”与其后的动词一律切分。例如：

所 想　　所 认识

5.11 语气词

语气词一律为分词单位。例如：

你好吗？

快去吧！

5.12 叹词

叹词一律为分词单位。例如：

啊，真美！

唉呀，他走了！

5.13 象声词

象声词一律为分词单位。例如：

嘟　　当当　　轰隆隆

附 录 A
分 词 举 例
（参考件）

A1 略语

离退休　零部件
石化　火电
四化　农副业
亚运会　联大
教委　奥委会
环保

A2 惯用语及有转义的分词单位

喝西北风　闲人免进
好家伙　对台戏
进一步　吃香
吃醋　批复
这件事真 扎手
进 一步 说

A3 动宾结构

“2＋1”或“1＋2”结构的动宾词组一律切分。

开窍　上班
讲课　洗澡
开学　开 锁
进兵　进 村
生病　生 火
生 炉子

A4 动补结构

毁坏　耗尽
认清　了不起
来得及　搞好
搞活　搞 脏
打倒　打 坏
看透　看 清楚

A5 偏正结构

火热　冰冷
滚烫　感冒 药
象牙　兔 牙
农药　兽 药

创建　　新建
原油　　原书

A6　主谓结构

眼红　　性急
人造　　民办
头痛

A7　“于”的处理

介词：
对于　　关于
由于
“于”作为后加成分：
属于　　在于
敢于　　善于
“于”作为助词：
生于　　应用于
出现于　　逝世于

A8　“不”作为前加成分的几种情况

不好　　不送
不禁　　不能(能愿动词切分开)
不论(连词)

A9　趋向动词

提出　　发起
指出　　引起(已是词)
售出　　拿起
购进　　走进
寄来　　跑出去

A10　前后加成分

完全虚化的前后加成分：

阿哥	阿妹	阿爸
初一	初五	初十
老鹰	老头	老张
儿子	桌子	鞋子
花儿	悄悄儿	玩儿
石头	枕头	木头
党员	运动员	演员
学者	作者	压迫者
党性	规律性	酸性
现代化	深化	蜕化

科学家　　作家　　发明家
枪手　　拖拉机手　　爆破手

部分虚化的前后加成分：

超导体　　超 时代
多边形　　多 功能
泛神论　　泛 希腊
可爱　　可 采纳
泥巴

接头词接尾词：

被 打倒　　代 军长　　所 称赞
侦察 班　　进度 表　　工具 厂
航空 馆　　棉花 库　　工程师 室
展销 楼　　副 部长　　计算机 处

附加说明：

本标准由中华人民共和国机械电子工业部提出。

本标准由北京航空航天大学、燕山公司系统部、北京师范大学、中国标准技术咨询服务中心、机电部计算机与微电子中心、北京语言学院、水电科学院计算所、中国软件技术公司、机电部第四研究所负责起草。