中文信息处理丛书

信息处理用现代汉语分词规范及自动分词方法

刘　源　谭　强　沈旭昆　著
杨铁鹰　梁南元　王德进　王虹青　审校

清华大学出版社
广西科学技术出版社

内 容 简 介

本书着重介绍了国家“七五”科技攻关项目“信息处理用现代汉语分词规范”的制订的研究成果。内容包括《信息处理用现代汉语分词规范》及制订过程、汉语自动分词方法、《信息处理用现代汉语常用词词表》及制订过程等。《信息处理用现代汉语常用词词表》配有软盘。

《信息处理用现代汉语分词规范》为汉语信息处理提出了一整套实用、科学、系统的分词规则。该规范已被批准为国家标准。

本书可供汉语信息处理有关研究人员参考。

(京)新登字 158 号
(桂)新登字 06 号

信息处理用现代汉语分词规范及自动分词方法
刘 源 谭 强 沈旭昆 著
杨铁鹰 梁南元 王德进 王虹青 审校
☆
清华大学出版社出版
北京 清华园
广西科学技术出版社出版
南宁市河堤路 14 号
×××× 印刷厂印刷
新华书店总店科技发行所发行
☆
开本: 787× 1092 1/16 印张: 字数: 千字
1994 年 1 月第 1 版 1994 年 1 月第 1 次印刷
印数: 00001—00000
ISBN 7-302-01430-2/TP · 556
定价: 00.00 元

清华大学出版社　广西科学技术出版社
计算机学术著作出版基金

评 审 委 员 会

主 任 委 员　张效祥

副主任委员　周远清　汪成为

委　　员　(按姓氏笔画排列)
王鼎兴　杨芙清
李三立　施伯乐
徐家福　夏培肃
董韫美　张兴强
徐培忠

出 版 说 明

近年来,随着微电子和计算机技术渗透到各个技术领域,人类正在步入一个技术迅猛发展的新时期。这个新时期的主要标志是计算机和信息处理的广泛应用。计算机在改造传统产业,实现管理自动化,促进新兴产业的发展等方面都起着重要作用,它在现代化建设中的战略地位愈来愈明显。计算机科学与其它学科的交叉又产生了许多新学科,推动着科学技术向更广阔的领域发展,正在对人类社会产生深远的影响。

科学技术是第一生产力。计算机科学技术是我国高科技领域的一个重要方面。为了推动我国计算机科学及产业的发展,促进学术交流,使科研成果尽快转化为生产力,清华大学出版社与广西科学技术出版社联合设立了"计算机学术著作基金",旨在支持和鼓励科技人员,撰写高水平的学术著作,以反映和推广我国在这一领域的最新成果。

计算机学术著作出版基金资助出版的著作范围包括:有重要理论价值或重要应用价值的学术专著;计算机学科前沿探索的论著;推动计算机技术及产业发展的专著;与计算机有关的交叉学科的论著;有较大应用价值的工具书;世界名著的优秀翻译作品。凡经作者本人申请,计算机学术著作出版基金评审委员会评审通过的著作,将由该基金资助出版,出版社将努力做好出版工作。

基金还支持两社列选的国家高科技重点图书和国家教委重点图书规划中计算机学科领域的学术著作的出版。为了做好选题工作,出版社特邀请"中国计算机学会"、"中国中文信息学会"帮助做好组织有关学术著作丛书的列选工作。

热诚希望得到广大计算机界同仁的支持和帮助。

清华大学出版社
广西科学技术出版社
计算机学术著作出版基金办公室

1992 年 4 月

中文信息处理丛书

序　　言

中文信息处理技术在我国现代化及信息化建设中,越来越起着重要的作用,作为一个高新技术的重点,它已经列入国务院批准的“国家中长期科学技术发展纲领”。十几年来,我国的中文信息处理领域里,在技术的研究、产品的开发以及产业的建立等方面都取得了显著的成绩。现在很需要把这些方面的成果加以综合并且提炼出来,以便推广应用,并且作为一个起点,再上一个新台阶。这就是我们组织编写并出版这套中文信息处理丛书的目的。

在这套丛书即将开始出版之际,我愿向读者介绍以下两点:

第一　为什么我们要把中文信息处理技术作为高新技术的一个重点来发展呢?

我们日常工作中的信息,绝大部分是以语言文字作为媒介,传播交换和记载的。因此随着计算机的推广应用,由数据处理、信息处理发展到知识处理,对语言文字的处理的要求的深度和广度越来越高。这个问题在西方国家并不突出。因为计算机从诞生之日开始,就是以处理西方语言为基础的。换言之,他们无须经过呼吁和宣传,随着计算机的推广应用的发展,很自然地都会主动地研究和解决自己国家使用计算机如何不断地适应自己国家的语言文字问题。可惜,我们的汉语与西方语言的差别很大。能够处理西方语言的计算机,面对汉语,却显得无能为力。例如:

· 西方语言为拼音文字,而汉语是表意文字。西文字符只有 20 余个,而汉语文字仅常用的就有六、七千个,总数超过五万。这是一个根本性的问题。仅这一个差异就引起了处理汉语的计算机与处理西方语言的计算机一系列的差异,需要我们自己去解决。包括键盘输入、汉字打印与显示、内部代码、汉字识别、程序语言的数据类型、数据库的检索和排序等等。

· 西方的书面语言,词与词之间有空格。而汉语的词与词之间无空格。于是词的切分问题就成了计算机处理汉语的首要问题。

· 西方语言的同音词很少,而汉语的同音词很多。例如,JI 音汉字就有一百多个。辨析同音词就成了汉语语音处理的关键。

· 西方语言多有形态变化(例如:多数、少数,过去、现在,男、女等等),而汉语缺少形态变化。计算机对汉语的处理(例如,机器翻译、人机接口等)无法利用形态,只能在语法、语义上找出路。

· 汉语的语法尚未形成规范化,而且人们习惯于非规范化的语法。于是语义的研究的重要性比西方语言重要得多。例如,“吃饭”“吃大碗”和“吃食堂”的理解只能靠语义来解决。

· 汉语的自动(计算机)处理是多学科和跨学科的研究工作,特别需要计算机科学与

语言学的密切结合,而且要依靠长期积累的语言学的研究成果。但我国语言学界多着重汉语教学,对象是人,而不是机器,因此对其丰硕的研究成果要经过改造、深化、量化,甚至要从头开始。要清醒地认识到它的艰巨性,要持续不懈地抓下去。

以上只是几个突出的问题。还有很多其它问题,不再赘述。这些语言上的特点造成了计算机处理汉语的很多障碍,每前进一步都会遇到新问题,使我们不得不花费自己很多力量去解决。

再就计算机的发展趋势而言,计算机产业面临转型期,多媒体和笔记本式计算机将成为热门产品。这些产品的核心技术无不与中文信息处理技术有关。因此,加强中文信息处理的研究更为必要。

第二　中文信息处理技术包括哪些科目呢?

大体上包括下列一些科目:

· 词的切分和频率统计
· 汉语句型和短语的研究及频率统计
· 汉语语义的研究
· 键盘和非键盘汉字输入技术及处理系统
· 汉语语料库的开发及应用
· 汉字的机器代码,程序设计语言的数据类型
· 汉语开放系统的接口规范
· 语声输入与合成
· 汉字识别
· 字形生成
· 汉语分析及理解
· 汉语生成
· 人机接口
· 机器翻译
· 情报检索
· 自动标引和抽词,自动文摘
· 全文检索
· 电子印刷出版系统
· 汉语辅助教学
· 电子词典

以上这些科目,有些是基础研究,有些是技术研究,也有些可以直接转化为产品。这些科目的分类并非学科分类,不过是按照编者本人日常接触的项目,把它们罗列出来而已。其分类的科学性、正确性和完整性尚待商榷。必须指出,有些基础性研究虽然看不到直接的经济效益,但它的研究成果则是其它研究工作所必需,而且要先行。

到目前为止,在上述这些项目中,有些已经产业化,例如电子印刷出版和少数几个汉字输入系统;有些项目已经商品化,正向产业化迈进;很多项目已经实用化。但每个领域都有很多问题等待我们去解决。今后的工作只能加强,不能削弱,使我们中文信息处理的每

个领域，每个项目都沿着实用化、商品化和产业化的道路奋勇前进。我相信我们这套丛书必将在促进中文信息处理技术的发展方面发挥它应有的作用。这套丛书大约十册左右，将在“八五”期间陆续出版。

最后，感谢“计算机学术著作出版基金评审委员会”把出版中文信息处理丛书列入了“八五”出版计划。感谢清华大学出版社和广西科学技术出版社给予出版基金的支持。

中国中文信息学会理事长　陈力为

1992 年 5 月　于北京

中文信息处理丛书编委会

前　言

《信息处理用现代汉语分词规范及自动分词方法》一书实际上是“七五”期间集体攻关的成果。其汉语分词规范部分，已于1992年被国家技术监督局批准为国家标准(GB13715)，并于1993年5月1日在全国正式实行。

正如序言中所述，汉语不同于西方语言，词与词之间无空格、无明显切分标记，于是词的切分问题就成了计算机处理汉语的首要问题。

早在1987年，按照国家有关部门下达的“七五”攻关和国标制订任务，北京航空航天大学、燕山公司系统部、北京师范大学、中国标准技术开发公司、北京语言学院等13个单位组成了刘源、梁南元为组长的国标研制工作组，并聘请了中国计算机与信息处理标准化技术委员会主任陈力为教授、国家语言文字工作委员会王均教授、中国社会科学院语言研究所刘涌泉教授等三位著名专家为该项任务的顾问。工作组成立后，历经三年多的研究、探索和实践，先后易稿十余次，最后由梁南元副教授执笔，于1990年底形成了报批稿。该稿规定了用于信息处理的现代汉语分词规则，将对汉语信息处理的规范化起到积极作用。

本书的内容共分五章：第一章是现已成为国家标准(GB13715)的信息处理用现代汉语分词规范。第二章是分词规范的编制说明，其目的是使人们能够了解分词规范中每一条内容的制订过程和制订的原因。第三章汉语的自动分词方法也是本书的重点之一，简要说明了近十多年来国内在汉语自动分词方面所做的研究工作。第四章对第五章信息处理用现代汉语常用词词表的制订过程和制订词表的原则进行了说明。第五章为信息处理用现代汉语常用词词表，总计近四万条词。这些词经过许多计算机专家、信息处理专家、语言专家和制订组的多次审订，都是满足分词规范的。

本书的著者和审校者都是分词规范及常用词表制订组的成员，所有工作都是由大家共同完成的。

在分词规范的研制和出版过程中，陈力为教授不仅对本书的具体内容给予了热情的指导和帮助，而且对本书的出版也给予了大力支持。

我们从信息处理角度推出本书，供国内外研究汉语信息处理的广大读者使用，但由于我们的水平有限，无论在分词规范方面、分词方法方面、常用词表的收词方面，一定会存在不少缺点和错误，欢迎广大读者批评指正。

著者：刘源　谭强　沈旭昆

1993年3月于北京

目　　录

第一章　信息处理用现代汉语分词规范 (中华人民共和国国家标准 GB13715) …… 1
附录 A　分词举例(参考件) …… 8
第二章　信息处理用现代汉语分词规范的编制说明 …… 11
2.1 制订《信息处理用现代汉语分词规范》的目的和意义 …… 11
2.2 制订过程 …… 11
2.3 指导思想和制订原则 …… 13
2.4 《分词规范》综述 …… 14
2.5 详细说明 …… 16
2.6 《分词规范》的验证 …… 33
参考文献 …… 34
第三章　汉语分词方法、分词模型与歧义字段 …… 36
3.1 研究现状 …… 36
3.2 自动分词方法与技术 …… 36
3.3 已实现的分词系统 …… 43
3.4 自动分词模型 CWSM …… 45
3.5 歧义字段的分析 …… 46
3.6 分词评测方法 …… 49
3.7 结论 …… 49
参考文献 …… 50
第四章　信息处理用现代汉语常用词词表编制说明 …… 52
4.1 目的和意义 …… 52
4.2 制定过程 …… 52
4.3 指导思想 …… 53
4.4 词条来源 …… 55
4.5 词表的构成 …… 56
4.6 词表验证 …… 62
参考文献 …… 63
第五章　信息处理用现代汉语常用词词表 …… 65
5.1 词表编排格式说明 …… 65
5.2 信息处理用现代汉语常用词词表 …… 65

第一章　信息处理用现代汉语分词规范

(中华人民共和国国家标准 GB13715)

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,其它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

汉语信息处理词汇 GB 12200

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、使用稳定的二字或三字词组,一律为分词单位。例如:

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 五字和五字以上的谚语、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 在现代汉语中出现的非汉字符号,例如其它语言的字符串、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 1247 1,298,576 3.14

4.9 现代汉语中其它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　爷儿们”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. 其它国家、其它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,名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 Ⅲ 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它成分,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,三字和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　　红彤彤

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　　二 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和动词之后时,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: