现代汉语语法信息词典详解

中文信息处理丛书

The Grammatical Knowledge-base of Contemporary Chinese —— A Complete Specification

（第二版）

俞士汶等著

清华大学出版社

http://www.tup.tsinghua.edu.cn

中文信息处理丛书

现代汉语语法信息词典详解

The Grammatical Knowledge-base of Contemporary Chinese —— A Complete Specification（第二版）

俞士汶　朱学锋　王　惠　张化瑞　张芸芸
（北京大学计算语言学研究所）
朱德熙　陆俭明　郭　锐
（北京大学中文系）
共著

清华大学出版社

(京)新登字158号

内 容 简 介

本书（第二版）是对北京大学计算语言学研究所开发的电子版《现代汉语语法信息词典》的详细介绍。内容包括如下两篇：第一篇“《现代汉语语法信息词典》导引”，介绍了语言信息处理与语法研究、《现代汉语语法信息词典》概要、现代汉语词语的语法功能分类、词语的语法属性描述、《现代汉语语法信息词典》的应用与发展；第二篇“《现代汉语语法信息词典》示例”，展现了从电子版词典中挑选的10 000个词语及其语法属性信息。

图书在版编目(CIP)数据

现代汉语语法信息词典详解 / 俞士汶等著. —2版. 北京：清华大学出版社，2002
（中文信息处理丛书 / 陈力为主编）
ISBN 7-302-05911-X

Ⅰ. 现… Ⅱ. 俞… Ⅲ. 汉语—语法—词典—简介 Ⅳ. H146-61

中国版本图书馆CIP数据核字（2002）第073680号

出 版 者: 清华大学出版社(北京清华大学学研大厦，邮编100084)
http://www.tup.tsinghua.edu.cn
责任编辑: 薛 慧
印 刷 者: 中国科学院印刷厂
发 行 者: 新华书店总店北京发行所
开　　本: 880×1230　1/16　**印张**: 61　**插页**: 1　**字数**: 1410千字
版　　次: 2003年2月第2版　　2003年2月第1次印刷
书　　号: ISBN 7-302-05911-X/TP·3508
印　　数: 0001~3000
定　　价: 188.00元

序

第一台电子计算机诞生于20世纪40年代。到目前为止，计算机的发展已远远超出了其创始者的想象。计算机的处理能力越来越强，应用面越来越广，应用领域也从单纯的科学计算渗透到社会生活的方方面面：从工业、国防、医疗、教育、娱乐直至人们的日常生活，计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业，原因在于其高速的计算能力、庞大的存储能力以及友好灵活的用户界面。而这些新技术及其应用有赖研究人员多年不懈的努力。学术研究是应用研究的基础，也是技术发展的动力。

自1992年起，清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展，推动计算机科技著作的出版，设立了“计算机学术著作出版基金”，并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日，本套丛书已出版学术专著近50种，产生了很好的社会影响，有的专著具有很高的学术水平，有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统，继续大力支持本套丛书的出版，鼓励科技工作者写出更多的优秀学术著作，多出好书，多出精品，为提高我国的知识创新和技术创新能力，促进计算机科学技术的发展和进步做出更大的贡献。

中国计算机学会

2002年6月14日

清华大学出版社
计算机学术著作出版基金

评审委员会

名誉主任委员： 张效祥

主任委员： 唐泽圣

副主任委员： 陆汝钤

委员：（以姓氏笔画为序）

王　珊　　李晓明
吕　建　　林惠民
罗军舟　　郑纬民
施伯乐　　焦金生
谭铁牛

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，他们无需经过呼吁和宣传，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。

· 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），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。还有很多其他问题，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他研究工作所必需，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

中文信息处理丛书

第二版前言

《现代汉语语法信息词典详解》一书1998年由清华大学出版社出版。该书介绍的是于1995年底通过技术鉴定的电子版《现代汉语语法信息词典》(以下称“原词典”)。现在,《现代汉语语法信息词典》有了扩充版,规模与内容有很大发展,质量有很大提高。词典的用户迅速增多,研制者自身的认识也更加深入。为了介绍扩充版《现代汉语语法信息词典》及其研究心得,同时也为了满足各界人士对《现代汉语语法信息词典详解》一书的需求,清华大学出版社与作者决定出第二版《现代汉语语法信息词典详解》。经过双方一年多的努力,第二版终于问世了。

第二版保持第一版的基本面貌,但内容更加丰富。更新了第一篇“导引”中的所有章节,几乎每页都有增改,还增加了新的篇幅。对第二篇“示例”中的10 000词语的语法属性信息,也已按扩充版电子词典中的结构与内容进行了修改与调整。

为了详细地讲清第二版与第一版的差别,先概要介绍扩充版对原词典的发展。

1. 增加了词语

扩充版的词语总数达7.3万。原词典收录的5万多词语对一般的中文文本已有很高的覆盖率,但应用单位还要求增加2万词语。为了既能提高覆盖率,又不致引起过多的副作用,研制者对增补词语的原则作了慎重的考虑。

2. 提高了词典的质量

研制者把不断提高词典的质量作为长期的任务。作者已经利用出版《详解》第一版的机会,对书中选用的10 000个词语示例中的信息,极为认真地进行了校对。对于新增加的2万多词语的归类及各项语法属性的描述,研制者坚持按照出书的标准把握质量。近几年来,北大计算语言学研究所及其合作伙伴已在应用《现代汉语语法信息词典》进行大规模语料库加工、汉外机器翻译、中文概念辞书编制等项研究,在应用中我们发现了一些静态检查难以发现的问题与不足,并均已分期分批加

以修正。应该说，第二版中10 000个词语样例的质量又上了一个台阶。

3. 增加了语法属性项目

3.1 全新的语素库

开发原词典时，为了分析少量例句的需要，只在词典中收了1 000多个语素，而语素库的专有属性字段也只有1个，即“类别”。为了研究自动识别未定义词，在扩充版中增加了一个囊括国标GB 2312全部汉字的全新的语素库。国标GB 2312只有6 763个汉字。其中一部分汉字是成词语素，已作为单字词收入了相应的词库，就不再进入语素库。现在的语素库的记录总数超过7 300，这是因为一个汉字往往随着字义的发展变化而分解为多个语素。语素库的专有属性字段也增加到8个：词类(语素类)、组合、姓氏、人名、地名、水名、方古、书面。

3.2 其他词类的语法属性也有所增加和调整

如对名词的“子类”作了调整。名词的“子类”中增加了“过程名词ng”（只能与动量词或时量词搭配）。又如，对动词和形容词都增加了“后接”字段。介词库增加的属性字段最多。同时有些库也删除了一些当初用于调研的属性字段。如状态词库原先设置了“不”、“很”两个字段，是为了调查有没有状态词可以受“很”、“不”修饰。调查结果是任何状态词都不能受“很”、“不”修饰。现在已删除了“很”、“不”这两个字段。同理，也将介词库中的“体词短语”字段删除了，因为任何介词都可以带体词性短语的宾语。

4. 7.3万词的自封闭兼类体系

扩充词典与原词典处理兼类的原则和策略虽然是一样的，但两者的成果还是有差别的。除了由于认识的深化，有些具体的词的兼类信息在扩充词典中有所调整外，最重要的区别在于，原词典对兼类现象不是“封闭”的。例如，在原词典的动词库中的“浮”，其义为“飘”。如果以“不踏实”的意义使用它时，除了可以作谓语，还可以说“很浮”、“浮得很”、“不浮”等，具有形容词特征，因而在动词库的“兼类”字段中填了“a”，不过原词典并不保证形容词库一定也收了“浮”这个词。又如在原词典的动词库的“奉献”记录中，其“兼类”字段填“n”,名词库中也不保证有“奉献”这个记录。扩充词典就不同了，词的兼类信息形成“封闭集”：设形容词库a中有一个“词语”为w的记录,其“兼类”字段有信息“vn”，则在动词库v和名词库n中一定也有“词语”为w的记录，且在动词库v中“兼类”信息为“an”，在名词库n中“兼类”信息为“av”。

第二版《现代汉语语法信息词典详解》全面反映扩充版电子词典的情况，针对电子词典规模、规格说明等的有关章节都重新修订了。像语素库，显然要增加篇幅，以反映新的语素库的面貌。还有，在撰写《现代汉语语法信息词典详解》时，考虑到当时有些语法属性检查得尚不充分，如动词的6个分库和代词的2个分库，故将

10 000个词语样例中的这些信息暂时略去了。因为这个缘故，对印刷版词典的列表格式也作了适应性调整，其中第四章为了说明印刷版同电子版的异同而在某些属性项目前附加了符号★和☆。第二版中的10 000个词语的语法属性信息同扩充版电子词典的信息保持完全一致（包括词典的总体结构和各个库的字段的数目、名称和内容），第四章中这些附加的符号也就不复存在，眉目更清楚了。

在第二版的第一篇“导引”中还补充了一些新内容。如增加了“7万词语归类的实践”作为第三章第4节，原来的3.4节则改为3.5节。新的3.5节还增加了“3.5.4语素的子类”。又如，在5.5节“语料库标注”中，原来的内容归属于小标题“5.5.1预备性工程”之下。新增加了“5.5.2大规模的工程实施”，介绍于1999年4月启动的一项新的语言工程：2 600万字《人民日报》语料的切分与标注。为了便于读者查找《人民日报》标注语料库使用的标记的代码及其含义，在第二篇的后面加了两个关于标记集的附录。参考文献也增加了8篇，新增加的参考文献接在原有参考文献之后编号。

无需讳言，限于条件，无论是原词典还是扩充版，现在的《现代汉语语法信息词典》中的知识都是由研制者在语言学家及其语法理论的指导下，根据自己对语言现象的观测、领悟并参照前人的论著、词典而整理的。虽然也经常从语料中检索例证，但并没有系统地利用大规模的语料库。因此，词典中数以百万计的语法属性信息一定存在与真实语料不尽一致的问题。现在，计算机技术进步了，自然语言处理技术也发展了，北京大学计算语言学研究所拟在大规模的标注语料库上进行词语频度、搭配频度、n元语法参数等各种语法属性的统计与计算，利用这些数据改造《现代汉语语法信息词典》，可将对词语语法知识的定性描述发展为定量描述，进而可建立概率语法体系。科学与技术的进步是没有止境的。

扩充版研制工作由俞士汶、朱学锋、王惠完成。张化瑞参加了出书的校对工作。陆俭明和郭锐也一直关心扩充版的研制工作。那顺乌日图、李峰、赵军、穗志方、詹卫东、赵强、段慧明、亢世勇、郭涛等各位同仁和朋友为扩充版的完成贡献了力量。

在《现代汉语语法信息词典》长达16年的发展历程中，研制者及北大计算语言学研究所得到了诸多部门和单位的支持，得到很多前辈与朋友的帮助，也从与同行的学术交流中获益匪浅。为避免重复，凡在第一版前言中已经列名致谢的，就不再赘述，但作者并没有随时间的流逝而忘记他们。近三年来，《现代汉语语法信息词典》扩充版的研制得到的基金支持有：国家社科基金“九五”重大课题“信息处理用现代汉语词汇研究”的子课题“现代汉语词语语法属性描述研究97@yy001-6”，国家自然科学基金项目“中文信息提取技术研究69973005”，国家重点基础研究973项目“汉英机器翻译G1998030507-4”，北京大学985项目“基础软件研究基地”，国家“九五”重点科技攻关项目“受限汉语处理技术及产品开发”。作者还向下列单位和专家表示衷心的感谢：许嘉璐教授，中国人民大学胡明扬教授，Xerox Research Center Europe（法国）Jean-Pierre Chanod经理，日中韩辞典刊行会（日本）春遍雀来博士，Intel China公司周富秋博士、翁富良博士，松下电器（中国）公司石川敏郎所长、李功俊研究员，Sail Labs公司（德国）Peer van Dristen总经理、Ulrike Bernardi博士，

香港城市大学郑锦全教授、徐烈炯教授、潘海华博士，香港科技大学吴德恺博士，南京大学王启祥教授，清华大学李星教授、丁晓青教授、陈群秀副教授、孙茂松副教授，延边大学金东日教授、北京北佳信息系统有限公司内海晓总经理、姜纪冰副总经理，中国社科院语言所祖漪清研究员，中国科技大学王仁华教授，青海师范大学德盖才郎副教授，烟台师范学院张绍麒教授，上海师范大学范开泰教授，南京师范大学马景仑教授，北京邮电大学王小捷副教授，广西南宁平方软件新技术公司刘连芳经理，中科院声学所杜利民研究员，联想研究院肖航、东芝公司东实常务董事雷海涛部长等。此外，本书的排版由郭涛小姐初步完成，提交清华大学出版社。若不慎遗漏了应该感谢的单位或个人，还望谅解。

步入新世纪，社会信息化的浪潮与知识经济的躁动为语言信息处理技术提供了广阔的发展空间。作者期望《现代汉语语法信息词典详解》第二版和《现代汉语语法信息词典》扩充版的问世能为语言信息处理技术的发展贡献一份力量。作者更期望本书的读者和词典的用户一如既往，继续关爱中文信息处理园地里的这棵幼树，指点疏漏与谬误。作者以“行百里者半九十”的心境看待过去与未来，持之以恒地改进已有的工作，登攀新的高峰。

最后，请允许说明一下第二版署名变化的问题。出第一版时，俞士汶曾就署名问题同陆俭明教授进行了磋商。由于陆俭明教授和郭锐副教授计划基于“七五”攻关成果另撰写一部面向汉语本体研究的有关汉语词类问题的专著，因而陆俭明教授表示就不在这本面向信息处理的书上署名了，只作为审校者。由于客观原因，有关汉语词类问题的专著至今尚未出版。现在出本书第二版，正好提供了一个机会，将朱德熙先生、陆俭明教授、郭锐副教授的名字补上，这样可以历史地、全面地反映这部著作所包含的劳动与成就。我们总是痛惜因朱先生过早仙逝而失去了大师的指导。我们永远怀念朱先生带领我们攻关的那段时光。期望《现代汉语语法信息词典》的传播和《现代汉语语法信息词典详解》第二版的出版为朱先生的英灵送去一份真诚的慰藉。

俞士汶

2001年2月于北京大学

前言（第一版）

《现代汉语语法信息词典详解》终于同读者见面了。本书是根据电子版《现代汉语语法信息词典》编写的，研制这部电子词典的直接目的是为了实现汉语句子的自动分析与自动生成。

开发《现代汉语语法信息词典》的理论基础是朱德熙先生倡导的“词组本位”语法体系。汉语语法体系应以短语(即朱德熙先生所说的词组)为基础，这是合乎汉语实际的。这与以句子为本位的很多外国语的语法体系是不一样的。按“词组本位”的观点来观察汉语，可以更好地认识汉语语法的特点。根据前人的论述和作者的粗浅认识，这些特点可以大致归纳如下：

（1）汉语句子的构造原则与短语的构造原则基本上是一致的，分析短语的结构基本上也就是分析句子的结构。以印欧语的眼光看，语言的语法构造是，由语素构成词，由词构成短语，由短语构成句子，层层是“组成”关系。汉语则跟印欧语不同，汉语是由语素构成词、由词构成短语，这也是“组成”关系；但短语和句子之间不是“组成”关系，而是“实现”关系，即汉语的句子都可以看作是由某个自由短语加上一定的句调而“实现”成的（在汉语中，单独一个自由词也可以加上一定的句调“实现”为句子），那不同的句调以及显示句子终了的停顿，书面上分别用句号、问号或感叹号表示。也可以这样说，汉语的短语和句子，其差别不在构造上，而是在所处的地位上，从构造上说，短语和句子都可以称为“句法结构”。同一个句法结构，当它处于单说的地位时，它是句子；当它处于被包含地位时，它是短语。

（2）汉语的词类与句法成分之间不存在简单的一一对应关系，同一个句法成分可以由属于不同词类的词来充任；而同一个词在句法结构中可以作不同的句法成分，形式上没有任何不同的标志。

（3）各类句法结构的组成成分又可由各种类型的句法结构充任，句法成分可以层层套叠，而且也不需要什么形式标志。其中特别值得一提的是，主谓结构和其他句法结构地位平等，跟其他句法结构一样，既可以独立“实现”为句子，也可以包含在另一个句法结构之中，作某种句法成分；而且在主谓结构中，不仅主语可以由

另一个主谓结构来充任，而且谓语也可以由另一个主谓结构来充任（这就形成了所谓的“主谓谓语句”或“主谓谓语短语”）。

（4）从句子角度观察，汉语的词序是灵活的，“我吃苹果了”、“我吃了苹果”、“苹果我吃了”、“我苹果吃了”等，它们都可以构成合法的句子。但是，汉语的语序是固定的，主谓结构的主语在谓语之前，述宾结构的述语在宾语之前，述补结构的述语在补语之前，偏正结构的修饰语（定语或状语）在中心语之前，如此等等。正是这种灵活的词序和固定的语序使汉语表达灵活多变，丰富细腻。

（5）汉语的虚词虽然有重要的句法功能，但在很多场合下，虚词也并不是必不可少的，常常可以省略。

（6）汉语的语素绝大多数是单音节的。单音节语素在书面上用单个的汉字书写，古汉语中由一个单音节语素构成的词占绝对优势，所以书面上基本上一个汉字也就是一个词（只有极少数连绵词例外）。这就形成了汉字连篇书写的传统。20世纪20年代开始，文章开始分段，并使用新式标点符号，不再连篇书写，基本上改为按句连写，这应当说是一个进步。但由于现代汉语中合成词特别是双音节合成词占优势，从按句连写的书面形式中确认一个个的词仍然有困难。

造成以上这些特点的原因，正如吕叔湘先生、朱德熙先生等所指出的，主要在于“汉语缺乏严格意义上的形态标志和形态变化”。从计算语言学的角度看，这些特点对汉语信息处理有着极为重要的影响，特别是第一、第二两点，对汉语信息处理有全局性的影响；而第六点使词语切分成了汉语信息处理的第一道关口。随着汉语信息处理研究与实践的不断深入，人们会越来越领会到上述汉语语法特点对汉语的计算机自动处理所带来的深刻影响。

正如一个孩子要成长为一个有才干的人需要学习语言、生活、文化、科学、技术等各方面的知识一样，要计算机能理解并处理自然语言，也需要先让计算机“学习”各种知识。以往的文献资料所介绍的语言知识表达形式大多采用规则系统。诚然，每个自然语言处理系统都有一部电子词典，但是早期的词典或者规模较小，或者包含的语法信息量太少，所以远远不能满足自然语言计算机自动理解与处理的需要。若要在自然语言计算机自动理解与处理的研究领域取得突破，必须建设好包括知识库在内的基础设施。通用型《现代汉语语法信息词典》应该是这种基础设施的重要组成部分。北京大学计算语言学研究所与北京大学中文系在承担词典研制任务之初，决定采用当时先进的、又是比较成熟的关系数据库技术作为汉语语法信息词典的支撑技术是恰当的。采用数据库的结构可以在词语分类的基础上详细描述各类词语的语法属性信息。最大量的语法属性信息是各个词类中的每个词语可以同什么样的词类（或者具体的词语）组成合法的句法结构，以及该词语在各种句法结构中能担任什么样的句法成分。显然，如果没有“词组本位”的语法体系作为理论基础，要开发规模如此庞大、内容如此丰富的汉语语法信息词典是难以进行的，容易走上盲目实践的路子。此外，80年代以来，出现了一些新的计算语言学的语法理论，这些新的计算语言学语法理论的共同特点是以复杂特征集和合一运算为基础。用这些语法理论为指导所建立起来的自然语言处理系统中的电子词典所含的语言知识都是很丰富的，比较合乎目前的需要，《现代汉语语法信息词典》的研制也借鉴了这些

理论。

北京大学计算语言学研究所与北京大学中文系合作，根据语法-义项相结合的原则以及词典编纂的其他一些普遍原则，为《现代汉语语法信息词典》选取了5万多个词语；并根据语法功能分布的原则，建立了面向语言信息处理的现代汉语词语分类体系，完成了这5万多词语的归类，即确定了每个词语的词性；由于属于同一类的各个词语的语法属性仍有很多差别，本词典采用关系数据库文件格式描述每个词语及其语法属性的二维关系。词典中共有32个数据库文件，其中包含全部词语的总库1个，各类词库23个。总库设21个属性字段，各类词库又分设若干属性字段，如名词库设27个属性字段，动词库设46个属性字段，等等。另外，某些类词库下又设分库，如动词库下又设立6个分库，代词库下又设立2个分库，分别描述其某个子类的更细微的语法属性。所有的库都可以根据主关键字段（词语+词类+同形）进行连接。这样，这32个库文件构成有上下位继承关系的“树”，子结点可以继承父结点的全部信息，或者说将父结点与子结点连接起来就可以得到关于每个词语的更全面的信息。如果定义每个库所包含的词语数同该库的属性字段数的乘积为该库的信息量，那么现在总库的信息量约为60万，32个库的总信息量达250万。这些信息量所需的存储空间约为16兆字节。

《现代汉语语法信息词典》的研制工作始于1986年，前后大致可以分为两个阶段。第一阶段（1986—1990年）为“七五”计划期间，当时北大计算语言学研究所承担了“现代汉语词语语法信息库”的“七五”科技攻关项目，北大中文系朱德熙、陆俭明和郭锐三位先生承担了“现代汉语词类”的“七五”国家社会科学重点科研项目。为了顺利而有效地完成这两项科研任务，我们双方采取了合作攻关的方式。当时在计算语言学研究所原有的电子词典的基础上，选了25 000多个常用词语，由朱德熙、陆俭明、郭锐三位先生亲自逐个归类，逐个填写语法属性信息；计算语言学研究所则负责软件开发。经过双方的共同努力，“现代汉语词语语法信息库”于1990年底通过技术鉴定。第二阶段（1991—1995年）为“八五”计划期间，北大计算语言学研究所和北大中文系继续合作，联合组成课题组，共同承担了国家“八五”科技攻关项目“现代汉语语法词典”。词语量扩充到5万多，每个词语的语法属性考虑得更为周全、确切。遗憾的是朱德熙先生于1992年7月19日不幸病逝，这对我们的研究工作带来不小的影响。但大家团结合作，如期完成了任务，并于1995年11月30日“现代汉语语法词典”又顺利通过了电子工业部组织的技术鉴定。“八五”成果鉴定之后，为了适应中文信息处理的需要，有利于这方面工作的继续开发和推广应用，作者在征得陆俭明、郭锐两位先生的同意后，综合上述“七五”、“八五”的科研成果及有关的其他科研成果，编写成了《现代汉语语法信息词典详解》一书，并交付出版。在本书出版之际，我们以十分沉痛的心情深切怀念“词组本位”语法体系的创建者、《现代汉语语法信息词典》前期研制工作的规划者与参与者朱德熙先生。

本书是对电子版《现代汉语语法信息词典》的详细介绍。最好的介绍莫过于将词典本身直接呈现在读者面前。只是一本书的篇幅实在容纳不下词典的全部内容。作者从电子版词典中挑选了10 000个词语以及它们的语法属性信息分类列表作为本

书的第二部分。本书的第一部分则是电子版《现代汉语语法信息词典》的导引，共有5章。第一章“语言信息处理与语法研究”是导引的导引，介绍语言信息处理研究的内容及作者对汉语自动分析的困难的认识，从而说明汉语语法信息词典研制的必要性及其在语言信息处理研究中的作用。第二章“《现代汉语语法信息词典》概要”介绍词典的研制过程、设计思想、结构框架、内容梗概，读者若只希望快速地了解词典的主要内容，读这一章也就够了。第三章“现代汉语词语的语法功能分类”介绍面向语言信息处理的现代汉语词语分类体系、词语分类的理论基础以及各类词语在语法功能上的主要特点。词语分类是《现代汉语语法信息词典》研制工作的基础，而词类问题又是汉语语法研究中的老大难问题。显然，这一章的写作对作者来说是相当困难的。作者只是将学习汉语语法理论的心得用自己的语言表述出来，力求简明易懂。如果这一章能为愿意从事语言信息处理研究而对汉语语法研究的最新成果不甚了解的读者提供一点入门材料，作者将感到莫大的欣慰。以语言学家的眼光来评判，这一章大概只能算作是一篇习作。如果能得到语言学界的前辈与专家的指正，作者同样感到莫大的欣慰，并预先致以诚挚的谢意。第四章“词语的语法属性描述”介绍从语言信息处理需要出发所确定的词语语法属性项目的类型和在抽象层次上所划分的语法属性字段值的数据类型。4.3节至4.5节所占篇幅较大，逐一介绍各个数据库文件的各个字段的定义及说明。这3节实际上就是已发表的词典规格说明书的扩充，包括了课题组在词典开发过程中所遵循的填写规范。这3节也包含了一些较深入的汉语语法知识，论述是否准确，剪裁是否得当，也有待专家与读者的评判。由于考虑到篇幅、排版、成熟程度等多种因素，本书第二部分所附的词典与电子版词典有一些差别。这3节详细标明了两者之间的差别。总的说来，印刷版简略些，但电子版也不完全覆盖印刷版。在写书的过程中发现在某些词类中增加某些属性字段是有意义的，便在本书中加上了，为保持电子版的相对稳定性，现在尚未在电子版中加上，这也为听取电子版用户的意见留了余地。第五章“《现代汉语语法信息词典》的应用与发展”采用示例的方法着重介绍《现代汉语语法信息词典》在语言信息处理各个领域的应用以及作者关于建设综合型语言知识库的设想和探索自然语言理解理论模型的考虑。作者期望通过本章的介绍能使更多的自然语言处理的研究者产生应用这部词典的欲望，并吸引更多的青年学者加入到自然语言处理研究的队伍中来，从而使语言信息处理研究能够攀登上一个新的高峰。

本书脱稿之际，回顾《现代汉语语法信息词典》开发的历程，作者对这项工作的所有支持者、指教者、帮助者的感激之情油然而生。首先感谢电子工业部、国家自然科学基金委员会、国家技术监督局等主管部门提供的资助和北京大学提供的基本条件。作者及所有参与实际开发工作的研究人员都是北京大学的学子，为能在北京大学这样一个文理学科交汇、学术思想活跃的科学殿堂里从事科学研究并取得些微成果而感到幸福。北京大学计算语言学研究所所长、中国科学院院士杨芙清教授一向支持这项研究并十分爱护已取得的成果。作者当然不会忘记参加“七五”科技攻关成果“现代汉语词语语法信息库”和“八五”科技攻关成果“现代汉语语法词典”鉴定的专家们，他们是陈力为、曹右琦、常宝儒、董振东、黄昌宁、冯志伟、史有为、吴蔚天、许卓群、叶蜚声、袁琦、张普、赵淑华等先生。国内信息处理学