

后 记

计算语言学是植根于计算机科学、语言学与数学等多学科沃土而成长起来的一门新兴学科，它的研究内容主要是自然语言信息处理，也就是人类语言活动中，信息成分的发现、提取、存储、加工与传输。

计算语言学的研究成果可在机器翻译、文献检索、信息提取、文本管理、自然语言人机接口等重要领域得到应用。近年来，因特网（Internet）日益普及，社会对语言信息处理技术的需求日趋迫切。计算语言学作为语言信息处理技术的主要基础学科，其重要意义已经得到学术界和产业界越来越多的认可，在欧美已经有不少大学设立了计算语言学系或专业，从本科起就组织计算语言学的教学工作。

本教材从“基础”、“算法”和“应用”三个方面，较为系统地介绍计算语言学的基础理论、相关的自然语言处理技术和应用。

基础部分（包括一、二、三章）主要讲授计算语言学的数学和语言学基本概念，计算语言学和自然语言处理技术的关系，自然语言处理的基本流程等内容。此外，作为计算语言学的重要分支之一，语料库语言学得到了较大发展，有关语料库语言学的基础概念、基本理论、近年来的一些发展概况在第三章中加以介绍。这部分内容的教学目标是使学生在不涉及技术细节的情况下，对计算语言学有一个初步的，同时也是较为全面的理解和掌握。

算法部分（包括第四、五章）主要讲授计算语言学的常用技术和算法。从处理对象来讲，主要包括词法层面的分析技术、句法层面的分析技术和语义层面的处理技术；从处理方法来讲，既有传统的规则方法，也有基于语料库的统计方法的介绍。这部分内容的主旨是希望通过本章学习后，使学生能对计算语言学的领域中的一些主流技术（比如隐马尔可夫模型在词性标注中的应用，GLR 算法，部分分析技术等等）有一个概要的认识，并能运用这些技术进行计算语言学相关的工作实践。

应用部分（包括第六、七章）主要讲授自然语言处理应用系统。需要说明的是，基于语言信息处理技术的应用系统很多，本教材只是重点介绍一些常见的系统，介绍这些系统的工作机理、发展、取得的成绩，也客观介绍这些系统存在的问题和困难。主要包括：机器翻译系统，信息检索系统、信息提取系统，文本分类系统等。

编写工作的具体分工如下：由俞士汶组稿，提出大纲，詹卫东撰写第一、二、三、四章，常宝宝撰写第五、六、七章。最后由俞士汶审订统稿。

国家语言文字工作委员会冯志伟研究员、北京大学计算机科学技术系博士生咎红英同学、孟涛同学、北大数学学院博士生陈阳同学曾先后阅读了全部或部分初稿并提出不少改进意见，在此向他们表示诚挚的谢意。

本书作为教材已经在北大计算机系（面向理科学生）和北大中文系（面向文科学生）的研究生课程《计算语言学概论》课上多次试讲，从教学实践中得到许多宝贵的意见和建议，在此向听课的同学表示感谢。

由于水平所限，本教材难免存在缺陷与谬误，恳请广大读者朋友和专家学者予以批评指正。

编 者

2003 年 3 月