电脑语言学基础

罗安源／主编

中央民族大学出版社

电脑语言学基础

中央民族大学康赛电脑语言学研究中心编

中央民族大学出版社

责任编辑：葛小冲
封面设计：赵秀琴
责任印制：陈立彬

图书在版编目（CIP）数据

电脑语言学基础/罗安源主编．-北京：中央民族大学出版社，1998.3
ISBN 7-81056-114-6

Ⅰ．电… Ⅱ．罗… Ⅲ．语言学-研究-计算机应用Ⅳ．H 0-39

中国版本图书馆 CIP 数据核字（98）第 01302 号

电脑语言学基础

中央民族大学康赛电脑语言学研究中心编

*

中央民族大学出版社出版发行
（北京白石桥路 27 号）
（邮编：100081　电话：68472815）
新华书店北京发行所经销
北京东方印刷厂印刷

850×1168 毫米　32 开本　6.625 印张　163 千字
1998 年 3 月第 1 版　1998 年 3 月第 1 次印刷
印数：01—2000 册

ISBN 7-81056-114-6/H·14
定价：10.00 元

内 容 提 要

本书分为“绪言”，“电脑构成及使用”，“电脑与文字处理”，“电脑与语音处理”，“电脑与词汇处理”，“电脑与句法处理”，“语言分析软件示例”等7章，最后附有国际音标练习。全书粗略介绍了电脑语言学的兴起、发展、成就及电脑语言学的研究对象和任务，着重阐述了电脑语言学研究、处理人类语言和文字的一些基础知识与方法。从20世纪60年代开始逐渐发展起来、由多种学科参与其成的电脑语言学，是集电脑科学、语言学、心理学、声学、生理学、数学、社会学等有关理论、技术、方法于一体的交叉学科。它的基本目的，是要应用电脑科学、语言学及其他有关学科，来研究、处理作为人类最重要交际工具和思维工具的自然语言。它的具体任务，是开展对人工智能、自动翻译、情报和文献检索、言语识别和言语合成、词语和风格统计、文字信息处理、语言结构分析、程序教学等等问题的研究。我国是一个多民族、多语种、多文种、多文化的社会主义大家庭。现在56个民族所使用的语言在80种以上，正在使用的文字有21种，历史上流行过、现能见之于文献的文字有13种。因此，电脑语言学在我国大有用武之地。积极开展电脑语言学的研究，是高等学校、科研单位提高教学、科研水平的必要措施。中央民族大学与美国世界少数民族语文研究院协作，率先在民族院校中倡导开展电脑语言学的教学与研究，已经取得了明显的效果。为了尽快在师生中普及电脑语言学的基础知识和基本技能，中央民族大学中国少数民族语言文学学院康赛电脑语言学研究中心的几位成员，包括语言学教授、副教授、讲师、博士、硕士，根据在边学边教之中所得到的一些体会，编写了这本《电脑语言学基础》。本书内容简明扼要，叙述深浅得宜，适合各高等学校和科研单位，特别适合文科高等学校的有关专业使用。

电脑语言学基础

主编　罗安源
编撰

此为试读，需要完整PDF请访问：www.ertongbook.com

前　记

受中央民族大学中国少数民族语言文学学院院长戴庆厦教授和语言学系负责同志的委托，我们“康赛 CUNSII 电脑语言学研究中心”的全体同仁，集体编写了这本《电脑语言学基础》。

电脑语言学是一门新兴的交叉学科，近几年来，我们一直在美国“世界少数民族语文研究院”同行的帮助下进行学习。为了尽快在我校有关专业师生中普及电脑语言学的基础知识和基本技能，我们只好采取边学边教的办法，来促进我们的工作。按照原定的计划，本书的编写事宜由主编人筹措，由康赛电脑语言学研究中心全体成员共同撰稿。具体进程是，1996 年 10 月以前拿出总体设计方案和编写纲目，年底以前进行集体讨论并分领写作任务。1997 年 6 月底交齐全部初稿，7 月底再由主编人统一审定付印。现在虽然已经如期完稿，但是总觉未能遂心如意。

在编写过程中，我们拜读了不少专家学者的论著和译作，不但从中得益很多，而且直接引用了一些专业性很强的内容。现特将这些论著和译作按出版时间排列于全书之后，并向各位著译者表示衷心的感谢。掠美之处，尚希谅察。

需要特别声明的是，目前我们呈献的这本《电脑语言学基础》，只算是应急的教材，而不是深入的专论。限于时间和水平，缺点和错误定不可免，我们衷心祈盼得到各方面专家和朋友的指教。来函请寄“100081 北京白石桥路 27 号中央民族大学康赛电脑语言学研究中心”，来电请直拨“（010）68932367”。

本书在出版过程中，葛小冲同志精心加以编校，纠正了原稿中一些不妥之处，谨此铭记致谢。

《电脑语言学基础》编者谨识

1997 年 7 月 31 日

序

由中央民族大学康赛电脑语言学研究中心组织编写、由罗安源教授担任主编的《电脑语言学基础》一书已定稿发排，我感到非常高兴。因为这是一部具有开创性的、并具有理论意义和实用价值的好教材，其出版及时解决了我校文科基地班电脑语言学教材之急需，还为我校民族语言文学专业的教材建设增添了一项新内容。

电脑在本世纪中叶出现后，很快就应用到语言研究中去，并形成了一个交叉的新学科——电脑语言学。电脑在语言研究中应用，大大地改进了语言研究的手段，使语言研究进入了一个新的领域。人们可以利用电脑来储存语言材料，进行语音、语法、词汇的分析，还可以进行各种统计和检索，还可以进行语言翻译等。它使人们对语言现象的认识更加细致、准确，并能检验传统的“口耳”之学的准确性。因此，使用电脑来分析研究语言将是今后语言研究的一个重要手段，必将为语言学研究开辟广阔的前景。

我国是一个多民族、多语种、多文种的国家，各民族的语言文字丰富多彩，少数民族文献浩如烟海，语言文字的研究任务非常繁重。近年来的语言研究成果表明，电脑应用在少数民族语言教学、研究中大有可为，势在必行，是深化和加速少数民族语言研究的必由之路。

为适应语言学学科建设以及少数民族语言教学科研发展的需要，中央民族大学于1992年建立了“康赛电脑语言学研究中心”，并组织了部分教师从事电脑语言学研究。五年来，该中心在美国电脑语言学家的帮助下开展了几个课题的科学研究。到目前为止，已完成了国家教委博士点课题“汉藏语言电脑统计分析与语音实验研究”，彝文信息化处理的研究工作正在进行中。先后培训了5批学员，经过培训的学员均已掌握了利用电脑和语音分析仪器进行语料统计、语音实验以及辞书编排等技能。我校语言文学学院的教师和博士生，大多受过电脑语言学的培训，他们学成后也极大地改善了研究手段，提高了教学和科研的效率和质量。与此同时，该中心还为研究生、本科生开设了“电脑语言学”讲座，并为文科基地班开设了“电脑语言学基础”课程。今后，该中心除了继续培训学员，为本科生、研究生开设“电脑语言学基础”课程外，还将继续开展电脑语言学的科研工作。拟建立中国少数民族语文数据库、音档，加强少数民族语文信息化处理研究，建立一个资料全、使用方便的“国内外中国民族语文论著目录检索系统”，开展各种民族语言的专题研究。

参加这部教材编写的教师，除汉语文外至少还熟悉一种少数民族语言。他们都有多年从事少数民族语言教学研究的经验，并较好地掌握了电脑语言学基本的理论、知识、技能。应该说，这部分教师从事电脑语言学研究是有很好的条件的，是能够在少数民族语言这块沃

土上大展电脑语言学宏图的。在这部教材里，他们已把自己这几年的研究心得写了进去。当然，作为一门新兴学科，现有的认识还是很不够的，有待今后一点一滴地再充实、提高、完善。不过，基础有了，以后的事情就好办一些。万事开头难，我不能不佩服这些教师的创业精神，因为我知道，要形成这样一部系统的教材，他们不知遇到了多少难题，不知啃了多少“硬骨头”，没有事业心的支持是不可能做好的。

戴庆厦

1997.7.25

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第一章 绪 言

第一节 电脑科学与语言科学的不解之缘

一、电脑的出现与自动翻译的起步

电脑科学是年轻的科学，语言科学既是古老的科学又是年轻的科学。有趣的是，电脑科学从产生的时候起，就跟语言科学结下了不解之缘。

世界上早就有人发明了数字式计算工具。1642 年，法国人帕斯卡制成了一台加法计算机。1672 年德国人莱布尼兹制成了一台加减乘除法计算机。这两种计算机都是数字式计算工具，基本方式是用齿轮表示数，用手摇动完成计算，它们的功能跟古老的算盘差不多。直到距今 50 年以前的 1946 年，美国宾夕法尼亚大学才制造出了世界上的第一台电子计算机（命名为“ENIAC”），后来就有人把电子计算机称为电脑。

电脑的出现，扩展了语言学者的视野。语言研究精确化的设想，如语言文字的计量分析，可以在电脑的辅助下实现。有了电脑之后，一些学者就提出了用电脑进行语言自动翻译的想法。起初，自动翻译的基本思路，是在两种语言之间实现一对一的自动翻译，把两种语言的词语编制成一部尽可能完备的双语词典，用类似译码的办法，把一种语言翻译成另一种语言。1952 年，在美国麻省理工学院召开了第一次自动翻译会议。1954 年初，美国乔治敦大学在国际商用机器公司的协助下，用 IBM-701 机进行了世界上第一次自动翻译试验，揭开了自动翻译研究的序幕。

自动翻译的想法和初步成就引人入胜。但是这种自动翻译只着眼于词对词的对译，而两种语言之间的翻译，并非仅仅把彼此的词语对译一番就算完事儿了。因为两种语言之间的差异，除了

表现在词语上以外，还表现在句法结构、修辞色彩以及文化背景上。只将词与词对译的自动翻译，其译文效果几乎是失败的。失败为成功之母，人们从中得到启发，要搞好自动翻译，必须下足两方面的功夫。既要搞好电脑科学的研究，又要搞好语言理论的研究。尤其是语言处理理论的研究，乃是自动翻译的基础。

自动翻译的艰难曲折，促使电脑语言学从语言计量方面的研究，拓展到语言结构方面的研究。

电脑是人创造的。要让电脑运行，还得靠人来编制程序。电脑之所以能够运行，是因为有用本机固有机器语言编写的程序。但是汇编语言或高级语言编写的程序，必须通过编译（或解释），把汇编或高级语言编写的源程序，变成用机器语言编制的目标语言，机器才能执行。为了使这种编译达到等价的要求，电脑的编译系统必须完成类似于人脑编译的过程，那就是从“词类分析”，到“语法分析”，到“语义分析”，到“生成代码”，到“修辞优化”的过程。这就注定了电脑科学的命运，必须与语言科学结下不解之缘。电脑科学软件的设计，无可奈何地要建立在现代形式语言学观点的理论基础上。

二、电脑与人之间要实现语言交际

语言学理论，特别是形式语言学的理论，对电脑语言的形成发挥了重要作用。电脑科学专家从语言科学专家对自然语言所作的形式数学模型分析中，找到了用字母表、词法表、句法结构和语义解释的电脑“高级语言”的基本形式。从60年代中期起，电脑语言学界开始了以“自然语言理解”（natural language understanding）为中心课题的研究。自然语言理解课题所要探索的，是如何使电脑懂得作为人类社会最重要交际工具的自然语言。其根本目的是要模拟人类语言交际的过程，建立人与电脑之间用自然语言交际的模型。

第一代自然语言理解系统，只要求输入的句子适合系统中的

此为试读，需要完整PDF请访问：www.ertongbook.com