现代语言学丛书

主编　王宗炎　戴炜栋

自然语言处理简明教程

A Concise Course of Natural Language Processing

冯志伟　著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

图书在版编目(CIP)数据

自然语言处理简明教程 / 冯志伟著. —上海：上海外语教育出版社,2012
(现代语言学丛书)
ISBN 978-7-5446-2785-6

Ⅰ. ①自… Ⅱ. ① 冯… Ⅲ. ①自然语言处理-教材 Ⅳ. ① TP391

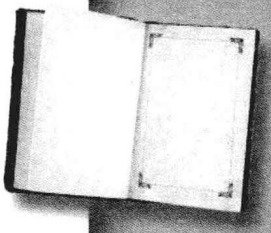
中国版本图书馆 CIP 数据核字(2012)第 122107 号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com.cn http://www.sflep.com
责任编辑：蒋浚浚

印　　刷：上海信老印刷厂
开　　本：890×1240 1/32 印张 30.625 字数 875 千字
版　　次：2012 年 9 月第 1 版 2012 年 9 月第 1 次印刷
印　　数：3 100 册

书　　号：ISBN 978-7-5446-2785-6 / H • 1346
定　　价：68.00 元

“现代语言学丛书”(修订版)
编委名单



『现代语言学丛书』修订说明

外教社“现代语言学丛书”自20世纪80年代面世以来，在语言学界产生了深远的影响，深受国内外广大读者的赞誉。这套丛书的作者均为我国语言学界知名专家和学者，在语言学教学和研究领域成就斐然。丛书深入、系统地介绍了现代语言学各领域的基本理论、研究方法和学术成果，为推动我国的语言学研究和外语教学作出了积极的贡献。

随着语言科学的不断发展，语言学应用的范围也越加宽泛。作为一门迅速发展的学科，近年来，现代语言学在研究语言结构、语言运用、语言的社会功能和历史发展等领域，新理论、新方法、新成果和新动向层出不穷，研究的内涵逐步深入，外延也不断拓宽，成为近半个世纪以来发展最快、变化最大的人文学科之一。

为使国内外广大读者及时了解现代语言学各个领域的最新发展态势，外教社对“现代语言学丛书”陆续进行修订和扩充。新版丛书在对原有的学术精华进行补充和完善的基础上，广泛吸纳近20年来国内外语言学领域的最新研究成果，融“经典”与“创新”为一体，从而更具有学术性、科学性和实用性。

作为开放系列丛书，这套丛书将与时俱进，不断丰富学科内容，拓宽研究领域，为广大读者展现现代语言学的各项前沿成果，从而更有力地推动这一学科的建设与发展。

上海外语教育出版社

2010年8月

总序

现代语言学丛书

(修订版)

“现代语言学丛书”自20世纪80年代陆续推出之后，在业内产生了深远的影响。该套丛书的编委会委员和编写者均为学界知名专家学者，在语言学的不同领域取得了很大成就。正是他们的辛勤努力使得丛书具备普及与提高相结合、引进与本土化相融合的特色，而丛书前沿性的学术内容、深入浅出的理论阐释、科学规范的研究方法等使高等院校的师生、外事外贸单位的翻译、新闻出版界的编辑等语言工作者和学习者受益匪浅，得到他们广泛的认同和喜爱，为推动我国语言学的研究和发展作出了积极的贡献。

近20年来，现代语言学作为发展最快的学科之一，有许多新发现和新成果，需要进行多角度、多层次、全方位的研究。目前人文科学、社会科学和自然科学等的渗透使得语言学的分支更加丰富，出现了越来越多的交叉学科。语言学家的研究视野也得以逐步拓宽，探索更加深入，研究观念不断更新，研究范式更加多样化。为了更加充分地反映这一发展趋势，及时向广大读者反馈语言学及相关学科的最新研究成果，我们在征求编委会委员、广大教师和学生意见的基础上，对“现代语言学丛书”进行修订，力求全方位呈现该学科领域的新理论、新观点、新方法、新结论。

该丛书修订版一方面保留了原版编者权威、内容全面、编辑规范的特点，另一方面突出“经典”和“新颖”两个特色，注重学术历史积淀与社会发展的契合，使丛书更加具有学术性、科学性和实用性。这套丛书仍然是开放的，将陆续出版语言学及相关学科的权威研究成果，以促进我国的语言学研究的学科建设。首批推出的系列著作涉及语言学科的不同层面，涵盖学科研究的前沿内容和最新成果，如《语言学新视角》、《“人本语义学”十论》、《语言系统及其运作》（修

订本)、《现代语言学的特点和发展趋势》(修订本)、《比较词源研究》等。

作为人类交流的工具和文化的载体,语言的重要性决定了语言学的重要性。语言学的发展不仅受到各个学科的影响,也同时影响到其他各学科的发展。只有充分了解该学科的最新研究态势,切实关注语言学科的发展,才能更好地了解语言,运用语言。相信在业内专家学者和广大读者的支持下,“现代语言学丛书”修订版将充分发挥良好的学术影响,为语言学及相关学科的进一步发展作出更大贡献。

高等学校外语专业教学指导委员会主任委员

戴炜栋

2010 年 9 月

《现代语言学丛书》
编委名单(原)

主　编　王宗炎

副主编　戴炜栋

编　委(按姓氏笔划为序)

王宗炎　(中山大学)
王彤福　(上海外语教育出版社)
王德春　(上海外国语大学)
伍铁平　(北京师范大学)
张日昇　(香港理工大学)
赵世开　(中国社科院语言研究所)
胡壮麟　(北京大学)
桂灿昆　(广东外语外贸大学)
桂诗春　(广东外语外贸大学)
戚雨村　(上海外国语大学)
缪锦安　(香港大学)
戴炜栋　(上海外国语大学)

总序（原）

为什么出版《现代语言学丛书》？

因为我们感到，中国现代化包括许多方面的工作，其中之一是语言学研究的现代化。我们希望这一套丛书的出版，会有助于这一工作的开展。

近几十年来，国外语言学的研究进展很快。一方面，关于语言的内部结构，出现了各种理论和模式；另一方面，从各种不同的学科去研究语言，产生了诸如人类语言学、社会语言学、心理语言学、神经语言学、计算语言学等多科性研究。了解和介绍这两方面的理论、模式、实验和数据，供我国语言研究者参考，从而为语言学研究的现代化出一点力，这是我们的希望。

要做到语言学研究的现代化是不容易的。首先要对国外新的语言学理论加以分析和比较，作出我们自己的判断；更重要的是要结合汉语的研究

加以验证，写出结合中国实际的论著。我们这里先做第一步工作。

中国语言学史上，不乏利用外国的语言理论，为汉语研究开辟新路的例子。郑樵说："切韵之学，起自西域。"马建忠以拉丁文法为范式，写出了《马氏文通》。赵元任、罗常培等前辈先生运用描写语言学的方法，为我国方言调查做出了典范。近时汉语语法学家利用国外语言学的研究方法，使语法现象的分类和范畴的描写更有理据，更为精确。先行者研究外国语言理论的态度，永远是值得我们学习的。

作为第一步，我们打算出版15至20种书。以普及为主，逐步提高，以引进为主，同时注意结合我国的实际。我们希望和国内语言学界同志共同努力，填补我国语言学科中的一些空白点。

我们心目中的读者，是高等学校中文、外文和其他文史专业的师生，翻译界、新闻出版界人士，中学语文教师，以及一般语文工作者和爱好者。我们将力求用明白易懂的语言介绍新的学说和理论。

我们将注意国外新出的语言学文献，为中国的语言学的现代化尽快提供信息。我们的力量还很薄弱，我们要努力去做，并热诚希望国内语言学者和语文工作者给予指导、批评和支持。

《现代语言学丛书》编委会

1982年11月初稿

1984年5月修改稿

参考文献：1. 冯志伟著:《自然语言的计算机处理》,上海外语教育出版社,1996 年。
2. 冯志伟著:《计算语言学基础》,商务印书馆,2001 年。
3. 冯志伟著:《自然语言处理的形式模型》,中国科学技术大学出版社,2009 年。
4. R. Mitkov 主编,Oxford Handbook of Computational Linguistics, 外语教学与研究出版社、牛津大学出版社,2009 年。
5. B. Partee 等,Mathematical Methods in Linguistics, 世界图书出版公司,2009 年。

前言

自然语言处理(Natural Language Processing，简称 NLP)，就是以电子计算机为工具，对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。这项技术现在已经形成一门专门的边缘性交叉性学科，它涉及语言学、数学和计算机科学，横跨文科、理科和工科三大知识领域。自然语言处理的目的在于建立各种自然语言处理系统，如机器翻译系统、自然语言理解系统、信息自动检索系统、信息自动抽取系统、文本信息挖掘系统、术语数据库系统、计算机辅助教学系统、语音自动识别系统、语音自动合成系统、文字自动识别系统等。

自然语言处理是语言文字应用的一个新课题，从语言学的观点来看，我们可以把它作为应用语言学的一个分支。

自然语言处理又是**人工智能**(Artificial Intelligent，简称 AI)的一个主要内容，它是电子计算机模拟人类智能的一个重要方面。因此，自然语言处理还是研制智能化的电子计算机的一项基础性工作。目前，科学技术的发展突飞猛进，信息的数量与日俱增，电子计算机技术得到越来越广泛的运用。世界性的互联网(World Wide Web，简称 WWW)已经联成，并向语义互联网(semantic web)这个更高的、更加智能化的方向发展。智能化的电子计算机和智能化的互联网已经不是虚无缥缈的幻想，而是指日可待的现实。当前，美国、英国、日本等发达国家，都投入大量的人力、物力和财力，把智能化电子计算机和智能化互联网的研制放在十分突出的地位，这对于人类社会将产生

不可估量的影响。它同人类历史上语言的出现、文字的创造、造纸技术的发明以及印刷技术的发明一样，将成为人类文明史上的又一件大事。

自然语言是人类区别于其他动物的重要标志之一。人借助于自然语言交流思想，互相了解，组成社会；人还借助自然语言进行思维活动，认识事物的本质和规律，创造了人类的物质文明和精神文明。

自然语言是人脑的高级功能之一。心理学研究表明，人脑的语言功能具有一侧化的性质，它主要定位在大脑左半球，由大脑左半球所控制。因此，自然语言是人类特有的一种最重要的智能，智能化电子计算机和智能化互联网的研究离不开自然语言处理，自然语言处理的研究水平，在智能化计算机和智能化互联网的研制中，起着举足轻重的作用。我们中国的自然语言处理工作者，应该站在电子计算机和互联网的智能化这样的高度，以战略的眼光来看待自然语言处理技术的研究，把我国的自然语言处理提高到一个新的水平。

在计算机软件中，早已设计了许多人工语言，如 BASIC、PASCAL、COBOL、PROLOG、LISP 等程序设计语言，这些人工语言与自然语言一样，都遵循着形式语言的规律和法则。美国语言学家乔姆斯基(N. Chomsky)的形式语言理论，既适用于人工语言，也适用于自然语言，这有力地说明，自然语言与人工语言之间，在形式描述方面，确实存在着某些共同的性质。正如美国著名的逻辑学家蒙塔古(R. H. Montague)在《英语作为一种形式语言》一文中所说的："我并不认为形式语言和自然语言在理论上存在着重要的区别。"

但是，自然语言毕竟是人类历史长期发展而约定俗成的产物，它带着几千年人类历史的痕迹，比人工语言要复杂得多，因而用计算机处理起来也就困难得多。

自然语言起码在下面四个方面与人工语言大相径庭：