

计算语言学

俞士汶 主编

概论

JISUAN YUYANXUE GAILUN

应用语言学系列教材

总主编 陈章太 于根元

商务印书馆

应用语言学系列教材

总主编 陈章太 于根元

计算语言学概论

主编 俞士汶

编者 ~~常宝宝~~ ~~詹卫东~~

商 务 印 书 馆

2003 年·北京

图书在版编目(CIP)数据

计算语言学概论/俞士汶主编. —北京:商务印书馆, 2003

(应用语言学系列教材)

ISBN 7-100-03796-4

I. 计… II. 俞… III. 数理语言学—教材 IV. H087

中国版本图书馆 CIP 数据核字(2003)第 039930 号

所有权利保留。

未经许可,不得以任何方式使用。

应用语言学系列教材

JISUÀNYŪYÁN XUÉ GÀILÙN

计算语言学概论

俞士汶 主编

商务印书馆出版

(北京王府井大街36号 邮政编码 100710)

商务印书馆发行

北京冠中印刷厂印刷

ISBN 7-100-03796-4/H·962

2003年9月第1版

开本 850×1168 1/32

2003年9月北京第1次印刷

印张 11³/₄

定价: 20.00 元

应用语言学系列教材

策划 北京广播学院播音主持艺术学院
商务印书馆汉语编辑室

编辑出版委员会

主任 陈章太 于根元

委员 曹先擢 戴庆厦 李如龙 李晓华

王维新 俞士汶 袁 晖 张 颂

张万起 赵金铭 周洪波 庄文中

总 序

应用语言学事实的存在源远流长，而作为一门独立的实验性学科其历史则较短。1870年波兰语言学家J. N. 博杜恩·德·库尔德内提出“应用语言学”这一术语，经过近百年的酝酿、探索，到20世纪40~50年代，第二次世界大战后，社会、科技、文化、教育等的发展，对语言在各方面的应用提出了诸多课题，于是一门多学科结合的应用语言学才作为独立学科建立起来，60年代开始进入发展时期，其标志是1964年第一届国际应用语言学会在法国召开和国际应用语言学协会的成立。弗赖斯(C. Fries)、拉多(R. Lado)、弗格森(C. Ferguson)、里弗斯(W. Rivers)等对应用语言学的建立、发展做出了重要贡献。20世纪60年代以来，国外在语言信息处理和计算语言学、社会语言学、语言规划及语言教学理论、方法和手段等方面都比较先进。语言信息处理和计算语言学是随着语言学与计算机的紧密结合而迅速发展起来的，沃古瓦(B. Vauquois)、伍兹(W. Woods)、维诺格拉德(T. Winograd)、海斯(D. G. Hays)、巴希勒(Bar-Hillel)、马丁·凯依(Martn Kay)等贡献很大。社会语言学在欧、美一些国家先发展起来，拉波夫(W. Labov)、甘柏兹(J. J. Gumperz)、海姆斯(D. H. Hymes)、罗曼(S. Romaine)、特鲁吉尔(P. Trudgill)等人很有贡献，他们以其研究成果及方法论为支柱形成了各自

的特色,在社会上影响较大。语言规划方面,美国、法国和俄罗斯也比较领先,威因里希(Vreil weinrich)、费什曼(S. Fishman)、克洛斯(Kloss)等发挥了重要作用。从总体上看,国外应用语言学发展很快,但在理论上还比较薄弱,这有多方面原因:一个是学科形成到现在的时间还不太长,学科基础不够深厚,理论研究自然比较薄弱;另一个是应用语言学涵盖的内容繁多,不容易进行总的理论概括和总结。加强应用语言学各分支学科方法论的综合梳理与总结,加强应用语言学基本理论研究,将是应用语言学今后在理论建设方面的重要任务。中国应用语言学形成于20世纪80~90年代,其标志是1984年语言文字应用研究所的成立和1992年《语言文字应用》杂志创刊。1995年召开“首届全国语言文字应用学术研讨会”,会上的学术成果丰富多彩,充分显示我国的应用语言学已经有了一定基础。我国应用语言学能在20世纪80~90年代成为独立学科,主要有两方面原因:一是社会发展的需要,二是受国外的影响。这时,我国实行改革开放,加快现代化建设步伐,强调新兴学科和交叉学科的建设,更加重视应用,强调科学研究与现代化建设紧密结合,而国外的应用语言学已经比较成熟,并成为热门学科,我们引进了这门新兴、实用的学科及国外的先进成果。我国应用语言学自建立以来,已经有了长足的发展,相继成立了一批专门的研究机构,创办了语言应用、语言教学和语言信息处理等专业刊物,筹备成立学术团体,建立了一批硕士点和博士点,并逐步建立了学科基地,在高等院校开设了专业课程,培养了一批专业人才,建立了专业队伍,取得了不少有价值的成果。进入新世纪,我国应用语言学在研究、解决语言应用的实际问题,正确引导我国语言生活,满

足社会、文教、科技的需求等方面,应当做出更大的成绩,同时为世界应用语言学的发展做出贡献。

应用语言学是一门独立的交叉学科,分广义、狭义两种。狭义的应用语言学研究语言教学。中国一般取广义的,也有取狭义的。我们取广义的,包括语言应用的各个方面,范围是开放的,具体包括四大部分:一是语言教学,主要研究第二语言教学或外语教学;二是语言规划,主要研究语言地位问题和语言文字规范化、标准化;三是广义的社会语言学,研究语言同社会的关系和语言的社会应用;四是语言本体和本体语言学同现代科技的关系,例如语言信息处理和计算语言学。应用语言学是语言本体和本体语言学同相关学科发生关系的学科。作为交叉学科,它和本体语言学具有交融性。如,语法学谈语法结构属于本体研究,而谈语法运用就跨界了,谈语法规则基本上属于应用语言学了。应用语言学既注重语言应用规律和基础理论的研究,更着眼于语言的实际应用研究;既注重思辨性的探讨,更立足于自然观察、社会调查和科学实验。

正是由于应用语言学的理论比较薄弱,所以过去国内外有些人认为,应用语言学没有什么理论,它只是对语言学理论的运用。这种认识显然是不全面的,没有理论怎么能成为一门学科?事实上应用语言学有自己的理论,如交际理论、动态理论、中介理论和人文性理论,尽管这些理论还不够成熟,也还有待进一步完善。交际理论是居总纲位置的理论,是本体语言学和应用语言学结合的纽带。交际是一种活动,语言存在于交际活动之中。因此,语言是动态的,是不断发展变化的,动态性是语言的本质。其实人们早就看到了语言的动态性,问题是如何看待语言的动

态和静态的关系和地位。由于结构主义语言学的影响,很长一个时期里,人们习惯于把静态看作是语言的本质特征,认为动态是对静态使用的表现。语言的动态理论则认为,物体运动速度不同,速度相对比较慢的叫稳态,速度相对比较快的叫动态,动态是语言的主导方面。而所谓的静态只是为了研究、说明、解释的需要而提出的一种状态。中介理论一开始是从语言教学中来的,现在不仅语言教学讲中介,社会语言学也重视中介。还有人文性理论。这三者都是语言交际理论下位层次的理论,此外还有层次理论、潜显理论等。与理论相关联的是方法,应用语言学的方法有综合的,也有本学科常用的,总体上是兼容的;其实有些方法很难截然分开是属于本体语言学或其他学科还是应用语言学的,只是在研究中有时会有所侧重。另外,方法和方法论也有层次,在一个层次里是方法,在另一个层次里它可能又是方法论。应用语言学的交叉学科性质,决定它的方法主要是兼容综合的。有传统语言学和其他学科的方法,如比较法,它从本体语言学中来,在应用语言学中也非常重要。应用语言学也有自己比较常用的方法,比如实验法和计量法,这些方法在传统语言学 and 有些学科中也使用。现在就我国来看,实验法在应用语言学中用得不够好。计量法在我国应用语言学中有的用得比较好,但总体上用得还比较少,而国外这方面用得很好。还有调查法,在传统语言学 and 有些学科中也用,但应用语言学中用得更多一些。可以说,实验法、调查法和计量法是应用语言学的基本方法。应用语言学的方法论主要是实证论,特别强调使用的方法要受到实践的检验。

我国应用语言学研究有自己的特色。我国应用语言学建立的背景是：那时语言教学需要加强，尤其是对外汉语教学和双语教学；语言文字工作决策要进一步科学化；语言研究要更好地为现代化建设服务；新兴的交叉学科要进一步发展。语用所就是随着这几个条件成立和发展的，我国的应用语言学实际上也是随着这几个条件产生和发展，并为这些条件的发展做出了贡献。我国的应用语言学十多年来与上面几个方面的需要和发展同命运，与我国的改革开放同命运，得到了迅速的发展。跟国外相比较，我们的应用语言学比较注意研究各分支学科的实际问题和总规律，而且认为应用语言学不只是理论的应用，本身也有理论。在理论方面，有从国外借鉴而有所发展的，有主要是我国提出的。在分支学科方面，尤其是语言规划、语言与社会文化、语言信息处理等，我们也有许多特点。我国应用语言学学科一建立，基本上就是取广义的，在研究中跟语言实际紧密结合，力求解决我国语言应用、语言教学和语言生活中的实际问题。我国的语言应用研究还跟本体语言研究密切结合。一个时期我国的本体语言学很受重视，现在研究应用语言学的有一批是从本体语言研究中来的，他们兼做本体和应用两方面的研究，这对应用语言学的发展十分有利。有比较好的本体语言研究做基础，这是我国自古以来就有的好传统，也是我国语言应用研究的一个特点和长处。

近些年来，我国的应用语言学得到了比较快的发展，这是令人鼓舞的。一门学科的发展，要具备两个相互推动的基本条件：一是社会发展的需要，这是外部条件；二是及时满足社会发展的需要，这是内部条件。从外部条件看，我国社会条件的改善，社

会生活的发展,科学技术的现代化,改革开放的深化等等,对我国的应用语言学提出了许多要求,同时也为其发展创造了良好的条件。从内部条件看,重视、关心、参加语言应用研究,是我国自古以来语言研究的传统。同时,我们还有了一支良好的研究队伍。我国的应用语言学正遇到好时机,在今后时期内将会有较大的发展,理论和方法方面也会有所加强和完善。但就目前来看,我国的应用语言学在理论建设、方法研究、队伍建设及学科的整体规划方面,还存在不足,学科的发展有一定的局限性。理论建设的不足,跟我们的语言应用研究工作急于求成,追求立竿见影的效果有很大关系。语言应用是个庞大而复杂的系统工程,启动起来需要一定时间,而要停下来,由于惯性作用也需要时间,所以语言工作有时要适度超前,做好促进工作。对队伍建设的重视不够,也跟过去对语言应用的认识不足有关。在方法、方法论建设方面,现在做得不够好、研究不够深入,计量、实验等方法用得较少,还有为方法而方法的情况,采用的方法不能很好地说明问题。对应用语言学学科建设的总体规划不够完整、不够清晰、不够具体,这也是比较重要的问题。存在上述问题,既有历史的也有现实的原因。历史上,我国的语言学理论建设有较大局限,科学方法没能很好发展起来。现实中,与现代科技的结合比较薄弱。今后我们要在这些方面加倍努力。应用语言学是新兴的交叉学科,因此队伍建设要重视跨学科人才的培养,加强两栖人才、复合型人才的培养。作为现代应用语言学研究,应该懂语言学又懂现代科技及其他相关学科。学科建设,还要有学科带头人、学术领袖和高层次的学术群体。纵观我国学科发展情况,学科带头人或学术领袖起了很大作用。学术群体不

一定是地域性的,可以跨地域跨学科,既有共同点,又有个性,群体与群体之间是合作的、交融的。我们还要营造良好的学术空气,要提倡学术民主,在这方面我们和国外差距较大。应用语言学搞得好不好,要看我们对问题的解决程度,对学科发展的贡献和对社会发展的贡献,这是学科发展的立足点。

学科建设的一个重要方面是教材建设。在1998年第二届全国语言文字应用学术研讨会上,许嘉璐先生对应用语言学学科今后的发展谈了三点意见:一是要加强理论建设,二是要使应用语言学进入大学课堂,三是研究人员要实行知识更新。会后,我们讨论了应用语言学进入大学课堂问题,想组织人员编写应用语言学系列教材。不久,北京广播学院播音主持艺术学院和商务印书馆汉语编辑室共同策划此事。1999年10月,在北京广播学院召开了第四次应用语言学学术研讨会,会议的主题之一就是应用语言学教材建设。陈章太在会上作了“关于编写应用语言学系列教材的几个问题”的报告。陈章太说:“应用语言学在国际、国内建立以后,其基础知识、基本理论与主要方法,以及新的研究成果,应当编入有关的教材当中;而在现代化建设进程中,应用语言学在解决社会、政治、经济、教育、文化、科技等对它提出的问题时,取得了明显的成绩,这也要通过教材把它总结、肯定下来,以便传授给众多的专业受业者,并促进本学科的不断发展。然而我国至今还没有成体系的应用语言学教材,更谈不上教材的多样与选择了,这不能不说是一种不小的损失和遗憾!摆在我们应用语言学者面前的一项紧迫任务,就是充分发挥集体的力量和有利的条件,尽快编写、出版一套具有中国特色的有较高水平的应用语言学教材。”(《语言文字应用》2000年第1

期)这次研讨会之前,已经策划快一年了。会议结束的第二天,在商务印书馆召开了总主编、分册主编、策划人员等参加的工作会议。陈章太就编写应用语言学系列教材的缘起、特色、编写人员、具体要求和第一批书目等问题做了具体说明。计划第一批编写10本,都叫概论。总主编是我们两人。于根元主编《应用语言学概论》,俞士汶主编《计算语言学概论》,庄文中主编《中小学语言教学概论》,赵金铭主编《对外汉语教学概论》,陈章太主编《语言规划概论》,戴庆厦主编《社会语言学概论》,张颂主编《传播语言概论》,李如龙主编《文化语言学概论》,曹先擢主编《中国辞书学概论》,袁晖主编《汉语语体概论》。会议决定成立这套教材编委会,由总主编、分主编和有关人员组成。各分册编者请有实力的本学科学术带头人担任,组织编写人员时应注意老中青结合、教学人员与科研人员结合。每一分册的编写组拟出大纲,其框架和内容经过编委会详细讨论,最后由总主编和分主编审订。会议商定2002年陆续完成初稿,由分主编统稿,总主编审稿,最后由分主编定稿,2003年分期出版。期间又召开了几次编委会,主要讨论全书的内容、写法、体例,及有关的编务,交流各书大纲和编写过程中的问题。编写、出版工作大体按计划进行。这套教材的体例总的来说是统一的,第一章是“绪论”,最后一章是“回顾与展望”。一般每节后面配有思考和练习题,章后附有参考文献,书末附有术语表。各书有自己的内容,但有的内容是共有的,共有的内容在观点和写法上会有一些不一致的地方,可以说是大同小异吧,这是很自然的事儿,可能有利于学术的发展。由于我们的水平和经验所限,这套教材的编写还有不尽如人意的地方,敬请读者朋友不吝赐教。两年后,我

们将根据读者意见和学科发展及教学实际的需 要,再进行认真的修订。

最后我们要衷心感谢参加这套教材编写的全体同仁,正是由于他们的辛勤劳动,编写工作才得以顺利完成。我们还应该真诚感谢商务印书馆的领导和有关同志,正是由于他们的最大支持和认真督促,这套教材才能够如期与读者见面。

陈章太 于根元

2003 年 4 月

目 录

第一章 绪论:什么是计算语言学	1
第一节 计算语言学的研究对象	2
第二节 计算语言学的研究方法	9
第三节 计算语言学的实际应用	12
第四节 小结	14
第二章 语言知识的形式化表达	16
第一节 语言与语言知识	16
第二节 形式化表达手段	24
第三节 语法知识的形式化表述理论体系	39
第四节 语义知识的形式化表述理论体系	53
第五节 语篇知识的形式化表述理论体系	66
第六节 小结	72
第三章 语料库:语言知识的另一种表示形式	76
第一节 语料库研究概况	76
第二节 语料的收集与加工	83
第三节 语料库的应用	99
第四节 小结	108
第四章 词法分析	110
第一节 概述	110
第二节 “词”的识别	113

第三节	词性标注	135
第四节	词义标注	154
第五节	小结	170
第五章	句法分析	173
第一节	句法分析导引	173
第二节	广义 LR 分析算法	182
第三节	基于栈图的分析技术	205
第四节	其他句法分析技术	214
第五节	小结	234
第六章	机器翻译	237
第一节	机器翻译概述	237
第二节	基于规则的机器翻译	250
第三节	基于语料库的机器翻译以及 混合式机器翻译	261
第四节	机器翻译的困难、对策和评价	275
第五节	小结	285
第七章	面向文本的智能信息处理	288
第一节	信息检索	288
第二节	信息提取	307
第三节	文本自动分类	323
第四节	小结	338
术语表		341
后记		355

第一章 绪论：什么是 计算语言学

先来看下面这一小段以汉语进行的对话。

甲：听说吴妈打赢了阿 Q。

乙：不错，阿 Q 确实被吴妈打败了。

甲：这个结果有些出人意料。

乙：阿 Q 是大意失荆州，怪不得别人。

这段对话也许从未在真实世界中发生过，但懂汉语的人很容易明白它的意思。如果这段对话是由两个说汉语的中国人讲出来的（比如两个相声演员），这一点也不稀奇，人们会觉得这“太过平常”了。但如果上面这段对话发生在一个人跟一台机器之间，甚至发生在一台机器跟另一台机器之间，可能人们就会“啧啧称奇”了。如果像这样的对话不仅仅只是说上四句就结束，而是滔滔不绝地说下去，那恐怕就要用“匪夷所思”来形容人们的感受了。

从某种意义上说，人类文明的发展历史或许可以表述为：将过去在人们看来“匪夷所思”的事情，变成现在人们容易理解，甚至“习以为常”的事情。那么，上面谈到的今日“匪夷所思”之事，是否能成为明日的现实？而要让这“匪夷所思”之事变成现实，人们又应该做怎样切实的努力呢？

上面提出的问题,实际上又可以转化成一系列相关的问题来加以考虑:

(1)人用来交际的“语言”具有什么样的性质?

(2)人用来交际的“语言”跟机器可以理解的语言有什么样的关系?

(3)人是如何运用“语言”进行交际的?

(4)人运用“语言”进行交际的“过程”是否可以描述为一个“机械的过程”?

(5)什么叫做“理解”一种语言?

(6)如何从“内在的交际意图”到“外显的语言表达”?

.....

大致说来,正是对上述问题以及相关的延伸问题的探索,形成了计算语言学这一交叉边缘学科。如果要用相对严谨和概括的说法来表述的话,可以说:计算语言学(Computational Linguistics)指的是这样一门学科,它通过建立形式化的数学模型来分析、处理自然语言,并在计算机上用程序来实现分析和处理的过程,从而达到以机器来模拟人的全部或者部分语言能力的目的。

上面对计算语言学的界说可以分解为以下几个方面来做进一步的阐述。

第一节 计算语言学的研究对象

从“计算语言学”这个名称上可以看出,这门学科的研究对象涉及“计算”与“语言”两个方面。计算语言学的研究工作一方面可