

李葆嘉等
著



语义语法学导论

——基于汉语个性和语言共性的建构



中华书局
ZHONGHUA BOOK COMPANY

语义语法学导论

——基于汉语个性和语言共性的建构

李葆嘉等 著

中 华 书 局

图书在版编目(CIP)数据

语义语法学导论:基于汉语个性和语言共性的建构/
李葆嘉等著. —北京:中华书局,2007.4

ISBN 978 - 7 - 101 - 05422 - 4

I. 语… II. 李… III. 汉语 - 语法 - 研究
IV. H14

中国版本图书馆 CIP 数据核字(2006)第 147098 号

书 名	语义语法学导论——基于汉语个性和语言共性的建构
著 者	李葆嘉等
责任编辑	秦淑华
出版发行	中华书局 (北京市丰台区太平桥西里38号 100073) http://www.zhbc.com.cn E-mail: zhbc@zhbc.com.cn
印 刷	北京市白帆印务有限公司
版 次	2007年4月北京第1版 2007年4月北京第1次印刷
规 格	开本/880×1230毫米 1/32 印张20 插页2 字数593千字
印 数	1-2000册
国际书号	ISBN 978 - 7 - 101 - 05422 - 4/H · 288
定 价	48.00元

目 录

引论 论语言科学与语言技术	1
一、“语言科技”新思维的提出	1
二、计算语言学的界定要突出技术性	5
三、语言系统的计算机模式化要求	10
四、人脑语言和电脑语言的性质异同	15
五、面向语言系统模拟的语义语法学	18
第一章 语义学研究的学术背景	25
第一节 世界语言学研究的四大传统	25
第二节 欧美语言学的历史嬗变	32
第三节 欧美语言学的语义转向	49
第四节 当代语义学的主要特点	57
第五节 格语法理论与框架语义学	71
第二章 汉语句法研究的语义导入	85
第一节 汉语语感与汉语句法语义研究	85
第二节 语义成分分析与语义指向分析	90
第三节 配价语法学与格关系语法学	94
第四节 陆俭明句法语义研究的建树	104
第三章 汉语句法语义研究述评	118
第一节 马庆株的语义语法范畴研究	118
第二节 袁毓林的汉语配价语法研究	131
第三节 邵敬敏的双向语法和语义语法研究	145
第四节 陈昌来的句子语义结构研究	161
第四章 字本位观和语义句法论	175
第一节 以字为本和以义为法	175

第二节	汉语语义句法论的学术背景·····	181
第三节	汉语语义句法论的理论价值·····	187
第四节	字本位和语义句法的争议·····	191
第五章	语言信息处理对语形语法的挑战·····	200
第一节	汉语信息处理的三个方面·····	200
第二节	九十年代以来的主要论著·····	203
第三节	九十年代以来的主要项目·····	210
第四节	九十年代以来的语言工具·····	218
第五节	机译系统研制的艰难进展·····	229
第六节	九十年代以来的成就及思考·····	240
第六章	语言的本质共性是语义性·····	243
第一节	语言符号与范畴化认知能力·····	243
第二节	形态范畴的本质是语义范畴·····	246
第三节	认知语义范畴化的基本特点·····	257
第四节	汉语语法范畴的消亡和演变·····	263
第五节	语义语法学理论的普遍性·····	272
第七章	语义语法学的理论建构·····	275
第一节	语形语法学和语义语法学·····	275
第二节	语义语法学的语言观·····	290
第三节	语义语法学的语法观·····	296
第四节	语义语法学的语域观·····	309
第五节	语义语法学的方法论·····	322
第六节	语言研究的实验主义理念·····	330
第八章	词汇系统与词汇语义研究·····	336
第一节	词汇系统的语义性本质·····	337
第二节	词汇义征的本质及其分析·····	344
第三节	义征分析的难度及思考·····	348
第四节	义征分析的基本操作程序·····	364

第九章 词语搭配关系与词群义征辨析	373
第一节 词语搭配关系与语义兼容	373
第二节 “安静”和“安定”的义征辨析	375
第三节 “深厚”和“浓厚”的义征辨析	377
第四节 “熟习”和“熟悉”的义征辨析	380
第五节 “整理”和“整顿”的义征辨析	381
第六节 同核词群的义征提取方法	385
第十章 汉语类别词的性质与分类	388
第一节 “一X名”结构的剖析	388
第二节 现代汉语“量词”的甄别	391
第三节 现代汉语“量词”的分化	405
第四节 特征类别词的语义研究	409
第十一章 形貌类别词的义征提取	417
第一节 对象的筛选和义征提取步骤	417
第二节 长条状类别词的义征提取	423
第三节 平展状类别词的义征提取	437
第四节 圆弧状类别词的义征提取	445
第五节 粒块状类别词的义征提取	448
第六节 堆捆状类别词的义征提取	453
第七节 丛群状类别词的义征提取	457
第八节 组配状类别词的义征提取	460
第十二章 汉语句法语义范畴和句模研究	467
第一节 汉语句法语义范畴的研究成果	468
第二节 汉语语义句模的研究成果	492
第三节 句法语义范畴的理论探索	498
第四节 句法语义范畴研究的思考	508
第十三章 汉语语法意合网络的探索	513
第一节 谓词中枢论和句构语块观	514

第二节	语义单位和组合规则·····	516
第三节	语义句模的建构步骤·····	527
第四节	意合网络的两种图示·····	529
第十四章	以“吃”为枢纽的语义网络建构·····	534
第一节	基于义征缠绕性的语义网络理论·····	534
第二节	“吃”的语义分析和语义框架·····	539
第三节	施事角色的义场建构和义征提取·····	543
第四节	受事角色的义场建构和义征提取·····	553
第五节	外层角色的义场建构和义征提取·····	578
第六节	情态范畴的义场建构和义征提取·····	587
第七节	基于句模的语义范畴化及扩展化·····	598
第八节	语义网络模型及其基本功能·····	613
主要参考文献	·····	619
后记	·····	633

引论 论语言科学与语言技术

一、“语言科技”新思维的提出

1946年，美国科学家埃克特（J. P. Eckert）和莫希莱（J. W. Mauchly）设计并制造出世界上第一台电子计算机。虽然计算机设计的初衷是数值运算，但是在第一台计算机诞生的同年，英国工程师布斯（A. D. Booth）和美国洛克菲勒基金会副总裁韦弗尔（W. Weaver）在讨论计算机的应用范围时，就已经提到机器翻译（Machine Translation）的设想。1950年，英国数学家图灵（A. M. Turing）在《机器能思维吗？》中，也已经预见到计算机和自然语言将结下不解之缘，并且提出检验计算机智能的最好方法就是对语言信息的处理能力。然而，计算机处理语言信息的研究历程并不平坦。1964年，美国科学院成立了语言自动处理咨询委员会（Automatic Language Processing Advisory Committee），着手调查机器翻译系统的速度、质量、投资以及需求量等。在调查报告《语言与机器》（1966）中指出：机器翻译研究遇到了难以克服的语义障碍（semantic barrier），“在近期或可以预见的未来，开发出实用的机器翻译系统毫无指望”。

随着研究者普遍认识到原语和译语之间的差异不仅表现在词汇上，而且还表现在句法结构上，机器翻译研究从低迷中渐渐复苏。1976年，加拿大蒙特利尔大学与联邦政府翻译局开发出英法翻译系统 TAUM-METEO。作为唯一不经编辑处理就可以发布天气预报信息译文的机译系统，TAUM-METEO 标志着机器翻译系统在受限语言领域走向成熟。促使机器翻译系统研究掀起

第二次热潮。1977年,费根鲍姆(Feigenbaum)提出了“知识工程”(knowledge engineering)这一概念,计算机信息处理出现了从数据世界向知识世界的转移趋势。80年代以来,研制出多种类型的一大批语料库,语言知识库也应运而生。基于语料库和语言知识库的语言信息处理方法日益受到重视。进入21世纪以来,语料库语言学+搜索引擎越发成为语言信息处理的主流,在未来的几十年内这一趋势仍然可能有增无减。

1979年,日本酝酿制造第五代计算机或智能机。1982年,制订出一个“面向人工智能(AI)的第五代计算机”的十年研制规划,投入4.5亿美元的研制资金。该项目由日本通产省牵头,汇集了来自日立、东芝和三菱等公司的一批计算机技术和应用研究领域的精英,成立了新一代电子计算机研究所。1992年结束时,日本又公布了另一个为期十年的“真实世界计算机”(RWC)研制规划。第五代计算机的研制目标包括:智能接口、知识库管理、自动解决问题、能用自然语言与人类会话。尽管取得了一定进展,然而如此宏大的项目,90年代后期却不得不宣布暂停。第五代负责人在新闻发布会上宣告“这个项目已经失败。在10年时间内,要完成一项高智能系统是不可能的”。

《语言与机器》中揭示的“语义障碍”这一告诫,除了致使当时的机译热暂时消退以外,并没有引发语言信息处理领域和语言学界对语义研究领域进行与之适应的投入。第五代计算机研制搁浅的原因,除了由于当时计算机体系结构、性能等硬件难以满足实际需要、人工智能理论尚不能支撑之外,主要是自然语言的理解和生成的研究水平远远没有达到智能机的要求。人类的最高智能即语言智能,智能机的核心是自然语言处理装置。第五代计算机研制的搁浅,无疑同时宣告了现有语形语法学的成果难以满足语言能力模拟的要求,其困惑仍然是《语言与机器》中指出的语义障碍,语义研究的缺失再次成为智能机研制难以穿越的“瓶颈”。

在《新世纪将对人类产生重大影响的十大科技趋势》(新华社北京 2000 年 12 月 30 日电)的报道中,开列出一份 21 世纪的十大科技领域清单。其中包括认知神经科学和信息技术两个领域,并分别提出“探索意识、思维活动的本质”和“计算机向智能化方向发展”的任务,然而没有语言学领域的合作——关键是认知语义网络的建构,谈何思维活动、人工智能?针对科技界的这一盲区,务必确立语言科技在 21 世纪科技发展中的主要地位。

依据当代科技进步和未来社会的发展趋势,笔者在 2000 年提出“语言科技”这一概念,在《南京师范大学文学院语言科技系建设发展规划(2001~2010)》中,阐发了语言科技的内涵和外延。2001 年 6 月,南京师范大学文学院语言科学与技术系成立。2001 年 11 月,笔者向中国语言学会第 11 次年会提交了《论语言科学与语言技术》的论文(《中国语言学报》第 11 辑,商务印书馆 2003)。知识世界的载体是语言符号系统,语言信息处理的需求促使语言研究活动和语言研究成果走向技术化,当代语言学已经凸显出语言科学与语言技术的二分互补格局。所谓语言科学主要指基础性的理论语言学和描写语言学,所谓语言技术主要指面向信息处理的计算语言学或语言工程学研究。语言技术的研究任务可分为语言文本处理技术和语言系统模拟技术两大部分。简而言之,语言文本处理(processing)技术是通过编辑和编程,将印刷文本转化为可供计算机识别的电子文本的技术。语言系统模拟(simulation)技术是通过算法和编程,将自然语言的理解和生成能力移植于计算机的技术。获得语言能力的计算机才可以实现人机会话而成为名实相副的电脑。2003 年,笔者提出自然语言装置研制的新思路,将人工语言脑(artificial brain of language)研制总工程师,划分为语言基因图谱分析工程(the analysis project of language gene atlas)、认知语义网络建构工程(the construction project of cognitive semantic network)和受限

语言能力模拟工程 (the simulate project of bounded language competence) 三个前后连续的子工程, 以最终穿过智能机研制的“瓶颈”。

虽然当代语言学家不可能也不必要都转向计算语言学研究, 但是具有一定的语言信息处理意识却非常必要。无论从研究手段还是就研究目标, 面向自然语言信息处理的当代语言学研究, 其性质都应界定为语言科技。强调语言研究的技术化, 并非忽视传统语言学研究存在技术性的一面, 例如实验语音学和方言语音调查等。之所以以往的语言研究未能突出技术化, 其原因在于, 除语音研究可借助声学技术设备以外, 语言研究的大多数领域尚没有相应的技术设备。

语言科技的内涵是以理论研究为引导, 以描写研究为基础, 以应用研究为枢纽, 促使语言研究向计算机科技、数学应用、认知科学和现代教育技术领域延伸, 沟通文理工相关学科以实现语言研究过程和研究成果的技术化; 语言科技的外延包括语言工程科技、语言教育科技和语言研究科技。语言学与计算机科学相结合的语言工程科技, 研究领域是人机会话, 其目标主要是自然语言能力模拟。语言学与现代教育技术相结合的语言教育科技, 研究领域是人人会话, 其主要目标是实现语言教育的多媒体和网络化。语言研究科技是语言研究活动与计算机技术手段相结合, 研究领域是人本(文本)会话, 其主要目标是实现语言研究的技术化。语言研究科技, 既包括分析性、统计性、比较性和实验性软件的开发, 又包括语料库、词汇库、句法结构库和语言知识库的研制。

可以预言, 谁先研制出可计算性语义网络和形式化语境模型, 谁就可能首先研制出自然语言处理装置; 谁先研制出自然语言处理装置, 谁就可能首先研制出智能机; 谁先研制出智能机, 谁就处于 21 世纪信息科技的最前沿。

二、计算语言学的界定要突出技术性

1954年,美国乔治敦大学所进行的世界上首次机器翻译试验,标志着计算机科技与语言学的结合起步。立足于不同的学科视角或知识结构,在这一领域先后出现了一系列术语,如:语言工程、语言工程学、自然语言的计算机处理(工科视角)、语言信息处理(信息学视角)、数理语言学(数学视角)等。随着计算机和语言学的结合逐步深入到语言研究的多个领域,80年代出现了计算语言学(Computational Linguistics),进一步形成了计算语音学、计算词汇学、计算语法学、计算语义学等分支学科。从语言工程到计算语言学,透露出语言学立场在这门交叉学科中的强化。

尽管目前的计算语言学研究热点主要集中于文本处理技术,但是语言信息处理的最终目标是“语言能力的计算机模拟”。探索语言能力的性质和描写语言系统的结构,这些工作还得由语言学家来完成。语言信息处理或计算语言学务必以语言学为本而以计算机为用。有兴趣的语言学家首先必须具备数字化意识,了解计算机需要怎样的应用基础成果,然后才有可能将研究目标对准语言工程。

迄今为止,“计算语言学”的定义尚无一致认定,侯敏(1999)归纳为四种观点:

一、计算语言学是以计算机为工具研究语言学。任何一个学科在使用工具方面都是自由的,使用不同工具研究一个学科会带来不同特点,但不因为使用了新工具就产生了新学科。

虽然并不排除新工具的使用必然导致新学科的产生,但同样不能否认工具的变革有可能促使学科体系的革命,以致于导致新的分支或交叉学科的出现。现代自然科学之所以能够建立,无疑得益于望远镜和显微镜的技术手段。前者打开了人类认知的宏观

世界之门，后者打开了人类认知的微观世界之眼。望远镜和显微镜带来的不仅是一个工具，而是人类认知方式的巨大变革，从而引起了天文学、生物学等科学的一系列革命并产生了一系列新的学科。因此，问题在于如何使用新的工具或认知手段，如果仅仅利用计算机做语言统计工具，也许不会产生新的分支学科，但是将计算机作为语音分析和合成的工具则形成了计算语音学。

二、计算语言学是把语言学成果应用于计算机。计算机的应用领域几乎没有限度，什么学科的成果都可以在计算机上应用，因此在计算机上应用语言学的研究成果不足以建立新学科。

问题不在于什么学科的成果能在计算机上应用，而在于在计算机上所应用的学科研究成果的性质。与其他学科研究的对象迥然不同，语言学的研究对象是人类最重要的认知符号系统和知识载体，因此面向信息处理的语言成果应用于计算机足以建立新的学科。以往的语言研究面向人际交流，而计算语言学研究面向人机交流，两者具有截然不同的性质。

三、计算语言学是研究语言中的可计算问题。虽然利用可计算理论研究语言符号是建立了一个新学科，但是这种说法偏于保守，没有把计算语言学推进语言学发展的作用充分体现出来。

问题在于，面向信息处理的计算语言学研究，其特点就是揭示语言的可计算性。“推进语言学发展的作用”并非计算语言学的定义，强调研究语言中的可计算问题未必保守，反而凸显了计算语言学的特点。

四、计算语言学是基于计算机科学理论所建立的语言学理论。把计算机科学的基本思想和方法引进语言学领域，不但可以产生许多应用性课题，而且能够促使研究者从新的角度观察语言学，建立与传统语言学不同的理论和方法。

一方面，计算语言学需要理论，但本质上不是一门理论科学，同时并非所有的研究者都乐意或适合于从事理论研究；另一

方面, 计算语言学的应用性质决定了其研究的技术性特征, 而绝大多数人可能更适合——实际上也更需要语言信息处理的技术性研究。

侯敏提及, 第一和第二种是欧洲流行的广义定义, 主张计算语言学是计算机和语言学的交叉; 第三和第四种是盛行于美国的狭义定义, 主张计算语言学是计算机科学和语言学的交叉。前者仅仅把计算机当成语言研究中的一种新工具, 而后者强调计算机学科的要求和理论对语言学的影响。陈小荷(2002)认为, 计算语言学就是以计算机为手段来研究自然语言, 较严格的定义是“通过建立形式化的计算模型来处理自然语言的一门科学”。

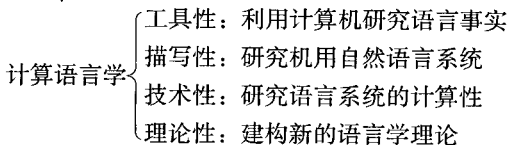
通过对目前几种理论方法以及相应语言处理技术的比较, 袁毓林(2001)讨论了计算语言学的工程主义、工具主义、认知主义、实证主义和逻辑主义等五种取向, 从而指出:

如果说科学是理论和知识体系、技术是方法和操作技巧、工程是实践和具体施行的话, 那么计算语言学就是一种工程。为了建造一个顺畅(fluent)、健壮(robust)的自然语言处理系统, 必须整合许多不同类型的知识, 诸如句法知识、语义知识、话语领域知识等, 并且要有效地用到自然语言处理系统中。正是在这一意义上, 建造处理自然语言的计算机系统跟建造其他大型的计算机系统一样, 主要是一种工程性的工作。跟其他系统建造工作一样, 计算语言学采用模块化(modularity)和建立形式模型(formal models)两种通用技术。

如果要建立形式模型来处理自然语言, 首先要完成供计算的天然语言系统的描写。这一面向人机会话的机用语言系统, 与以往面向人人会话的语言系统不同。因此机用语言系统的描写应当纳入计算语言学的研究范围, 完成了机用语言系统以后, 才能建立形式化的计算模型而使计算机获得自然语言能力。此

外，基于语言信息处理的语言学理论也是计算语言学的必要组成部分。

综上所述，计算语言学可定义为：利用计算机作为工具研究语言事实、研究机用自然语言系统、研究语言系统或语言能力的计算性，同时建构基于计算机应用、数学模型、认知科学等相关学科之上的语言学理论的交叉学科。图示如下：



虽然计算语言学的根本任务是研究人机之间的语言交际，即如何教计算机学会说话，但是从本质上来说，研究语言系统或语言能力的可计算性与利用计算机来研究语言事实是相通的，只是前者探索的是适合于人机会话的语言能力，而后者讨论的是适合于人人会话的语言规则。

依据目前的语言研究成果和信息处理技术，计算语言学可以划分为应用基础研究、应用研究、理论研究三个层面（陈小荷2002）。应用基础研究层面指语言信息处理的基础技术研究，主要包括：1. 自动分词技术：这是计算机理解自然语言的第一步。据说，在封闭性语料内，目前汉语书面语自动分词的正确率可达到95%以上。2. 词语特征标注技术：现阶段的词语特征包括词性和义项，这是句法结构理解的基础。两种标注可采用相似的计算模型，但后者要复杂得多。3. 语句分析技术：句法结构和语义结构是自然语言理解的关键技术，目前分析真实文本句子的正确率仅在40%左右。4. 语料库研制技术：语料库的研制包括建库规划、语料采集、加工深度和检索系统。目前的汉语语料库加工程度较低，所建树库很少且规模不大。5. 语言知识库建设技术：语句分析技术之所以难以取得突破，主要原因之一就是目前

尚无适合于中文信息处理的大规模语言知识库。6. 机用语言系统或面向信息处理的语言系统形式化描写, 目前尚处于探索阶段。

应用研究层面指自然语言处理软件工具的研制。现阶段的热点主要有: 1. 机器翻译工具: 半个多世纪过去了, 机译质量仍然令人失望。通行的是有限范围翻译和机器辅助翻译。2. 自动文摘工具: 微软公司的词处理器 Word 有用于英语的文摘功能, 哈工大研制的 HIT-863I 型中文自动文摘系统可按用户设定的比例压缩原文。汉语自动文摘的结果尚不令人满意。3. 自动校对工具: 存在的主要问题是误报率过高, 并且深层错误难以发现。4. 信息检索工具: 主题词检索和全文检索技术相对比较成熟, 而从语料库中获取知识的信息抽取在线工具研究方兴未艾。5. 语音识别和合成工具: 语音识别可分为词语识别(计算机口语命令)、有限词汇识别(电话订票)和无限词汇识别(将成段说话转为文字)。语音合成指将书面语转换为口语即文语转换, 主要问题是断词不当且语调刻板。

然而, 如果以为应用性或技术性明显的计算语言学仅仅是技术, 则未免失之于偏颇。虽然计算语言学是一门实验性、工程性学科, 但是仍然离不开理论和假设。计算语言学的理论研究层面, 至少包括人工智能理论和语言学理论两个方面, 例如计算机是否可能或能在何种程度上模拟人脑智能和语言能力, 计算机如何模拟人脑和语言能力, 如何寻找合适的语言计算模型等, 就是与计算语言学密切相关的人工智能理论问题。目前的人工智能研究, 主要集中在人工体能和人工技能。在语言能力模拟未能取得实质性进展以前, 尚难以企及真正的人工智能即语言智能。就语言学理论立场而言, 计算语言学研究不得不面临: 对自然语言本质属性的重新认定、面向信息处理的机用语言学理论、语言系统与数学模型的关系、语言结构和数理逻辑的关系、语言数字化处

理的可能性及其局限性、语言的异质性和受限理论、元语言研究的理论等一系列挑战。总体而言，一方面，由于牵涉的学科和知识太多，计算语言学的理论研究还相当薄弱；另一方面，与科技发展息息相关的计算语言学难以从容面对理论和方法的艰难探索。尽管计算语言学的发展动力植根于鲜明的应用性，但是同时必须认清，社会的应用要求往往导致急功近利，从而形成理论探索和应用基础研究少有问题津的误区。

三、语言系统的计算机模式化要求

从语言学的立场出发，语言系统的计算机模式化要求，首先要了解计算机需要怎样的语言描写成果。传统语言学的研究大体属于经验描写解释型，而计算语言学的研究属于实验操作技术型，自然语言系统要能进行技术化处理，首先必须实现语言的计算机模式化。冯志伟（1999）认为，自然语言处理一般经过三个过程：1. 形式化，将所研究的自然语言问题以一定的数学形式表示出来。2. 算法化，把自然语言的数学形式转换为算法形式。3. 程序化，根据自然语言的算法形式编写程序。侯敏（1999）认为，语言系统的形式化或计算机模式化必须满足三个要求：1. 高度抽象化，即从语言现象中抽象出一般规则。2. 元语言的形式化，即采取形式逻辑、数学公式、程序语言等形式语言。3. 运用过程的严密化，即运用过程必须具有数学与逻辑的严密性。

显而易见，语言的形式化是自然语言处理的先决条件。袁毓林（1993）认为形式语言至少具有三个特点：基本单元的明确性、基本运算和基本关系的明确性、运算优先级别的明确性。但是自然语言在这三方面皆不明确：1. 语法范畴的边界不明，例如语素、词与词组之间、不同词类之间的界限难以分清。2. 结构关系难以定义，如主谓、动宾等往往很难明确界定。3. 层次关系不外显，人们通常根据语感和语境等来识别结构层次。侯敏（1999）针锋