Linguistics for Knowledge Engineering

知识工程语言学

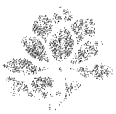
鲁 川 著

清华大学出版社

Linguistics for Knowledge Engineering

知识工程语言学

鲁 川 著



清华大学出版社
北京

内 容 简 介

本书是作者在为中国科学院研究生院和中国邮电大学研究生讲授“知识工程语言学”课程时所写的讲义的基础上完成的。这本书为人工智能、知识工程及相关专业的人员提供了所需的最必要的语言学基础理论和最有用的语言数据，包含对汉语特色的最新认识以及跟英语的对比。它凝聚了作者数十年从事知识工程领域科研工作和坚持不懈钻研语言学的心得。

本书可供从事人工智能、知识工程、自然语言理解、中文信息处理、机器翻译的研究人员参考，也可供从事对外汉语教学、汉语国际推广、英汉对比研究的教师和研究者参考，还可以作为高等院校相关专业的高年级本科生和研究生的选修课教材或参考书。

图书在版编目（CIP）数据

知识工程语言学/鲁川著．--北京：清华大学出版社，2010.6
ISBN 978-7-302-22215-6

Ⅰ．①知…　Ⅱ．①鲁…　Ⅲ．①知识工程－应用语言学　Ⅳ．①TP182

中国版本图书馆 CIP 数据核字(2010)第 038382 号

责任编辑：赵彤伟
责任校对：赵丽敏
责任印制：孟凡玉

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 刷 者：北京市世界知识印刷厂
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：185×260　　**印　张**：24　　**字　数**：566 千字
版　　次：2010 年 6 月第 1 版　　**印　次**：2010 年 6 月第 1 次印刷
印　　数：1～3000
定　　价：59.00 元

产品编号：023047-01

谨 以 此 书

献给

亲爱的祖国

庆祝中华人民共和国成立六十周年！

献给

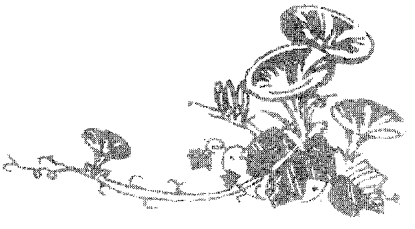
亲爱的母校

庆贺哈尔滨工业大学建校九十周年！

献给

亲爱的妈妈

祝贺梁树粹教授九十七岁寿辰！



前　言

Foreword

我要告诉读者：为什么要写这本书？这本书是怎样写成的？书中主要写了些什么？

半个世纪前的1956年初夏，中国大地上响彻了“向科学进军”的号角声，那时我正要填报高考志愿。在所能够阅读到的《知识就是力量》等科普刊物中，最令我神往的就是那刚诞生了10年的“电子计算机”。那一年招计算机专业的高等院校只有一所，就是哈尔滨工业大学。我幸运考取之后，产生了一个想法：既然“机器人”是以“计算机”为大脑的，那么这个大脑就不应该只会“计算”，还应懂得“语言”、会做“文章”。从上大学一年级开始我就在学好“高等数学”等规定课程的同时，找来中文系的各种课本默默地自学起来。

后来才知道恰在我入学的1956年暑期，在美国的达特茅斯(Dartmouth)召开的学术会议宣告了“人工智能”的诞生，也知道了新兴的“自然语言理解”等研究领域。这坚定了我对“计算机能理解人类语言”的信心，也增强了“我应该自学语言学”的恒心。我在世界“人工智能”诞生的1956年暑期考上计算机系，这个巧合是我的幸运，也决定了我一生的命运。1961年从计算机系毕业以后的20年间，我为了“人工智能”的目的，坚持自学古代汉语、现代汉语、理论语言学和外语，注意追踪国外语言学主要学派的最新成果。

1982年我慕名拜访了范继淹先生，他是中国社会科学院语言研究所著名的语言学家、中国人工智能学会的副理事长、我国“计算语言学”的开拓者之一。我向范先生表达了要报考他的“计算语言学”研究生的愿望。范先生告诉我：超过规定报考年龄的人是不准报考的。在我的恳切请求下，范先生对我进行了口试，证实了我的语言学基础知识远远超过了研究生的入学水平，答应向领导请示可否破例特招，结果没有批准。但范先生已被我的特殊经历和优异成绩打动，决定破例招收我为“不发文凭的在职研究生”，悉心辅导我学习“计算语言学硕士研究生”的全部课程。拜范先生为师之后才听别人说，早在1973年他就被确诊患了血癌，为了给国家培养急需的计算语言学人才，他分秒必争地跟死神抢夺时间而忍受着难以想象的痛苦。

1985年春天，范先生正式宣布我达到了“计算语言学硕士研究生”的结业水平。不久范先生就住进了医院，在病床上恩师对我说了他的临终嘱托：①恩师已经在给《中国语文》(1985年第5期)的稿件《无定NP主语句》中提到我的名字，在参考文献中引用了我的一篇会议论文，目的是引起语言学界的关注。让我尽早在语言学期刊上发表论文以争取语言学界的指导。②前几年恩师跟陆俭明和邢福义先生等中年语言学家发起的、有语言学大师吕叔湘和朱德熙先生出席的“现代汉语语法学术讨论会”已召开了三次。范先生知道自己不可能参加第四次讨论会了，他跟会议负责人说过，要我接替他参加讨论会，让我赶快写一篇论文在会上宣读。③让我全力筹建“计算语言学”的全国性学术团体。

含泪送别了跟血癌搏斗了12年的恩师，在各方面的大力支持下，于两年之内全面实现

了恩师的遗愿。1986年在“第四次现代汉语语法学术讨论会”上我宣读了论文《汉语句子的语义成分和语用成分》(载于论文集《语法研究和探索(四)》),得到了语言学大师吕叔湘和朱德熙先生以及诸多著名语言学家的关怀和指导,以后我陆续在语言学核心期刊上发表了多篇论文。1987年在中国中文信息学会理事长陈力为院士的支持下,我参与筹建并当选为中国中文信息学会首届“计算语言学专业委员会”主任,兼任中国人工智能学会“自然语言理解专业委员会”和“机器学习专业委员会”委员,《中文信息学报》首届编委,并在《中文信息学报》和《人工智能学报》上发表了被广为引用的几篇论文。

1990年南京大学召开徐家福教授的研究生翟成祥博士学位答辩会,答辩委员有杨芙清、陈火旺、陆汝钤、董韫美、孙钟秀、姚天顺6位教授和我共7人。进行答辩的那篇关于计算机处理中文信息的论文特别优秀,引人注目的是这篇论文有很大的篇幅是阐述翟成祥自己构建的一部信息处理用的汉语语法。在论文顺利地通过答辩之后我说:“工作量这么大的语法体系应该由专家写出来供人们选用,这样就能让博士生把他们的才华主要用在信息处理系统的程序和算法的创新上”。与会的教授们同意我的看法并建议这部“信息处理的汉语语法”由我来写,于是我就把这个重担压在了自己肩上。

1992年,应新加坡语言信息处理学界的邀请,作为“中国中文信息学会”特派的代表,董振东研究员和我在新加坡国立大学同时开设系列讲座,董先生讲授《英汉机器翻译》,我讲授《计算机对汉语的理解和生成》。此事更增加了我尽力写好这本书的动力。

初稿完成后,中国科学技术大学计算机系赵振西教授邀请我给他的研究生开设这门选修课。这本书也于1998年列入了清华大学出版社“中文信息处理丛书”的出版计划。当时我觉得再花费一年时间的琢磨就可以定稿,没想到这一琢磨竟又用了10年时间。

恰在20世纪之末,世界科技界和我国语言学界分别发生了重大突破。

20世纪末,世界科技界的重大突破是生物科学对于“人类基因图谱”的测序。这给我的启示是:①要探索并确定语言的“基因单位”;②要探索并构建语言的“排序模式”。

20世纪末,我国语言学界的重大突破是提出了汉语的“基本结构单位是字”。这给我的启示是:①语言的“基因单位”就是“基本结构单位”,汉语的“基因单位”就是“字”;②语言的“排序模式”是“基本结构单位的序模”,汉语语法研究主要是“字的排序”。

于是我彻底推翻写就的书稿而另起炉灶,推迟了向清华大学出版社交稿的时间。

我认识到,在重新改写《汉语信息语法》的同时,要积极投身到“中文信息处理和对外汉语教学”的重大科研项目中,应把亲身体验的科研实践的最新成果写进本书。从1991年到1995年的“八五”(第八个五年计划)期间,我在陈力为院士亲自领导的国家重点科研项目“中文信息处理应用平台工程(‘905’工程)”中担任《语义组合词典》课题组组长。“九五(1996—2000)”期间,在国家重点科研项目“计算机中文信息处理技术及产品开发”中从事“受限汉语处理技术及产品开发”的基础研究。“十五(2001—2006)”期间在教育部语言文字应用研究所承担的国家语言文字应用科研项目中,我担任“信息处理用规范汉字字义统计和造词模式(YB 105—48)”的项目负责人,还担任了2004年度国家社会科学基金项目“信息处理与对外汉语教学的句子语序模式(04BYY010)”的项目负责人。这些有实用价值的最新研究成果都被有选择地汲取到本书中。由于已经涉及语法之外的语音、文字、语义、语汇、语用等诸多方面,所以将书名改为《汉语信息语言学》。

在多次担任有关研究中文信息处理的博士学位答辩委员会委员或主席时,我发现博士

论文引用的英文文献远多于中文文献。这就有一个问题：如果研究生清楚地知道汉语和英语的异同，就会恰如其分地参考英文文献；如果研究生不太清楚二者的异同，就可能误入歧途。何况在知识工程的研究范围之内还有“机器翻译”和“第二语言教学”。于是我在书稿中增添了汉英对比的内容，并将汉英对比贯穿于全书。

在研究汉英对比的过程中，我所找到的最有价值的语料资源是母语为汉语者学习英语时所造的“英语病句”，以及母语为英语者学习汉语时所造的“汉语病句”。从这些珍贵的语料中我清晰地看到“母语对第二语言的负迁移”的活生生的例子，从而对汉英差异获得一些规律性的认识，并且悟出一个很重要的思路：语言信息处理跟第二语言教学应该紧密结合相互参照地进行认知语言学和对比语言学的研究。表 0-1 给出了三类语言研究的服务对象和要解决的主要问题。

表 0-1　三类语言研究的服务对象和要解决的主要问题

类别	名称	服务对象	服务对象的特征		要解决的主要问题
			人类	本族	
1	第一语言研究	本族人	是	是	使用母语的规范性和技巧性
2	第二语言研究	外族人	是	否	民族文化和认知模式的对比
3	第三语言研究	计算机	否	否	语言知识以及最普通的常识

于是我从第二语言教学跟语言信息处理应该紧密结合相互参照成果共享的角度写了若干论文，发表在《世界汉语教学》等专门研究“第二语言”的期刊上，得到了较多的引用和有关专家的关注。2004 年，北京外国语大学陈乃芳校长给我颁发了客座教授聘书并请我在北京外国语大学的国际交流学院为对外汉语专业开设了“汉英对比的认知研究”课程。

2005 年，国家重大科研项目“国家知识基础设施（NKI）研究”负责人、中国科学院计算所的博士生导师曹存根研究员邀请我给博士生开设一门课程。在计算所博士生导师、中国人工智能学会副理事长史忠植研究员的启发下，把给博士生开设的课程定名为名副其实的“知识工程语言学”。感谢曹存根研究员监听了授课的全过程，并力荐这门课程成为中国科学院研究生院的正式课程。

2006 年，在中国邮电大学博士生导师、中国人工智能学会理事长钟义信教授的支持下，我给邮电大学的研究生开设了“知识工程语言学”课程。中国人工智能学会自然语言理解专业委员会副主任王小捷教授监听了授课的全过程，并组织研究生讨论和提出意见，从而提高了讲义的水平。一些博士论文在参考文献中列入了《知识工程语言学（讲义）》。

知识工程是“研究知识表示、知识获取、知识储存、知识传播、知识运用的人脑功能并研究用计算机来模拟上述人脑功能的交叉学科”。其中“知识获取”要研究人的学习和计算机的学习，“知识传播”要研究人的语言教学和计算机的多媒体辅助教学，当然包括“第二语言教学”。这就需要一种为知识工程服务的具有交叉学科性质的语言学理论。只有把“认知语言学、对比语言学、计算语言学”紧密结合融为一体，才是时代需要的“知识工程语言学（linguistics for knowledge engineering）”。

积半个世纪刻苦学习和科研实践之经验，我从自己学习的心得笔记到给研究生授课的讲义，历经反复修改琢磨而形成了目前这本《知识工程语言学》。主要体会如下所述。

(1) 研究语言要认识到“语言是知识的编码系统”

在信息时代和知识经济社会中最重要的是“知识”，知识的表达、储存、传播、继承、创新，主要都是通过语言来实现的。语言是知识的编码系统。计算机要进行“知识处理”，必须理解人类的自然语言；人类要进行全球化的“知识交流”，必须跨越语种之间的障碍，依靠机器翻译和计算机辅助多语教学来促进“经济全球化和文化多元化”。要有为“先进生产力和先进文化前进方向”作贡献的高度责任感，就能更深入地去探索语言的奥秘。

(2) 研究语言要溯源认知模式对语言的制约

认知语言学越来越成为语言学研究的热门。不同的认知模式制约着不同的语言。

说汉语者的认知模式侧重于“整体感知、意象思维、静物为源、类比推理”。包括英国人在内的西方人的认知模式侧重于“细部感知、抽象思维、动态为源、形式推理”。

① 因侧重于“整体感知”，汉语的语音敏感单位是“音节”。

因侧重于“细部感知”，英语的语音敏感单位是“音位”。

② 因侧重于“意象思维”，汉语的文字保持着跟客观世界的“象似性”并适度抽象而形成为意合示音符号。

因侧重于“抽象思维”，英语的文字舍弃了跟客观世界的“象似性”而成为只是记录语音的高度抽象的字母符号。

③ 因侧重于“静物为源”，汉语造字时主要是用“名物性字元”构成“运动性单字”。如用名物性字元“扌(手)”构成运动性单字“打，拉，推，抓，握，抱，搂，拧，托，扶，扯，撕”。

因侧重于“动态为源”，英语造词时主要是用“运动性词素”构成“名物性单词”。如用运动性词素“duct(导)”构成名词：“con-duct-or(导游员)，in-duct-ion(归纳法)”等。

④ 因侧重于“类比推理”，汉语造字的最重要模式是“义元(意符)＋音元(声符)”；组词的最重要模式是“特征字＋类别字”。这充分显示了汉字有理据的灵活的“拆装性”(见图 0-1)。

字(马)
- 字元(马)
  - 义元(马) →单字(驯、骒、驴、驮、驰、驭、骑、驶、驾)
  - 音元(马) →单字(妈、吗、码、玛、蚂、犸、杩、祃、骂)
- 基字(马)
  - 类字(马) →字组(斑马、角马、战马、种马、木马、竹马)
  - 饰字(马) →字组(马鞍、马鞭、马蹄、马鬃、马术、马戏)

图 0-1 汉字的拆装性

因侧重于“形式推理”，英语造词时主要注重用“后缀”来表示“词性”和用“形态”来表示词的语法功能，所以英语不太注重基于“类比推理”的“特征词素＋类别词素”的复合造词模式。许多表示多元概念的词仍是难以有理分析的。

从表 0-2 可以看到，汉语所有的动物名称都有一个“通名”，而各种动物的“专名”都等于“特征＋通名”。“田鼠”是“田(里的)鼠”；“袋鼠”是“(有)袋(的)鼠”；“毒蛇”是“(有)毒(的)蛇”；“毛虫”是“(有)毛(的)虫”。英语中虽然也有较少的动物用“特征＋通名”来命名，但大多数“专名”跟“通名”毫无联系。在表 0-2 中，vole(田鼠)跟 mouse(鼠)、viper(毒蛇)跟 snake(蛇)、caterpillar(毛虫)跟 insect(虫)并无联系。

表 0-2 一些动物名称的汉英对照表

通名	专名				
鼠 mouse	田鼠 vole	豚鼠 cavy	跳鼠 jerboa	松鼠 squirrel	袋鼠 kangaroo
猴 monkey	猕猴 macaque	懒猴 loris	长尾猴 cercopithecus	卷尾猴 capuchin	眼镜猴 tarsier
蛇 snake	毒蛇 viper	环蛇 krait	赤练蛇 lateralis	四脚蛇 lizard	眼镜蛇 cobra
虫 insect	蠕虫 worm	蚜虫 aphid	毛虫 caterpillar	瓢虫 ladybug	象鼻虫 weevil
鱼 fish	鲍鱼 abalone	梭鱼 mullet	鳄鱼 crocodile	金枪鱼 tuna	凤尾鱼 anchovy
鹰 eagle	老鹰 hawk	夜鹰 nightjar	鱼鹰 cormorant	猎鹰 falcon	猫头鹰 owl

(3) 研究语言要认清各种语言的基本结构单位

汉语的基本结构单位是“字”(sinogram)。其特点是“一个字对应一个单音节·一个字对应一个方块形·一个字对应一个简单概念而多字词语对应复合概念”。汉语基本结构单位具有“基元性”,即词语的音节数跟所表示概念的复杂度具有“正相关性”,如“马”对应简单的“一元概念”,“母马”对应复合的“二元概念”,“小公马”对应复合的“三元概念”。

英语的基本结构单位是“词”(word)。其特点是“一个词不限音节数量·词与词之间有空隙·一个词既可以对应一个简单概念也可以对应复合概念”。英语的基本结构单位具有“丰富性”,即词语的音节数跟所表示概念的复杂度没有正相关性,如表示一元概念的 horse(马)、表示二元概念的 mare(母马)、表示三元概念的 colt(小公马)都是“单音节词”。

表示“一到十的基数”,汉语都是单音节;英语的 seven(七)偏偏是“双音节词”。

作为汉语基本结构单位的“字”,其特点可以概括为:“基元组装性”。

作为英语基本结构单位的“词”,其特点可以概括为:“丰富多彩性”。

(4) 研究语言要深入探索各级语言单位的排序规律

研究语言要认识到“顺序”是语言结构的最重要的形式标志。

汉语顺序依据“观察显眼性”。标准是言者按照“顺其自然”观察事物的“显眼性”。对于“主体运动的动态句”,显眼性表现在事件发展状况的先后上,“先看到的状况就先说,后看到的状况就后说”,例如“他大学一毕业就去了深圳。”先看到“毕业”就先说,后看到“去深圳”就后说。对于“主体静止的静态句”,显眼性表现在物体占有空间的大小上,“先看到显眼的大空间就先说,后看到不显眼的小空间就后说”,如“书桌上有一块橡皮。”先看到显眼的大空间“书桌”就先说,后看到不显眼的小空间“橡皮”就后说。

英语顺序依据“评估主要性”。标准是言者按照“新闻价值”评估事物的“主要性”。对于“主体运动的动态句”,评估出“事态主要性高的先说,事态主要性低的后说”,例如:He went to Shenzhen as soon as he graduated from the university. 事态“went to Shenzhen”主要性高就先说,事态“graduated”主要性低就后说。对于“主体静止的静态句”,评估出“目标物主要性高就先说,背景物主要性低就后说”,例如“There is an eraser on the desk”中的 eraser 是主要的目标物就先说,desk 是次要的背景物就后说。

(5) 研究语言要明确“五个视角”和“两个平面”

语言研究需要有全方位的五种“观察视角”,这些视角具有不同的功能,即:

① 认知的理据性；
② 语义的先决性；
③ 句法的限制性；
④ 韵律的和谐性；
⑤ 语用的实效性。

语言研究需要有能对句子构件进行“剖析”(parsing)的“两个平面”：
① 跟显性的、有序的、省略的、一维的“表层结构”一致的“句法平面”；
② 跟隐性的、无序的、完整的、多维的“里层结构”一致的“语义平面”。

(6) 语言研究中如有较大的新发现，在必要时该用“新术语”就得用，不要对“旧术语”修修补补，那样可能永远也说不明白

举例而言，在汉语的“主语、宾语”问题上，虽然有的语言学家已察觉汉语跟英语有重大的差异，但还是在英语术语“Subject(主语)、Object(宾语)”的基础上修修补补，迄今仍然说明不了“台上坐着主席团”和“王冕七岁上死了父亲”中，何为主语？何为宾语？其实吕叔湘先生在1946年的《从主语、宾语的分别谈国语句子的分析》中就说过：不妨不用“主语”和“宾语”这两个名称，改用“起语”和“止语”。遗憾的是，我国语言学界迄今并没有使用“起语”和“止语”。

英语是“Subject(主语)、Object(宾语)”型语言。主语跟“谓语动词”保持“人称”和“数”的“一致关系”。英语主动句的主语都是动词的“主体”，人称代词作“主语”要用“主格”。所有的“宾语”都是动词的“客体”，人称代词作“宾语”要用“宾格”。例如，She loves him (她爱他)中的 She(她[主格])既是语义上的“主体”，也是句法上的“主语”；而 him(他[宾格])既是语义上的“客体”，也是句法上的“宾语”。英语的“主语、宾语”是句法跟语义保持统一的句子成分。

汉语是“起语(initial)、止语(final)”型语言。“起语”跟谓语动词无任何“一致关系”，只是“说话的起始点”，即这句话是说“起语”的。“王冕七岁上死了父亲”是说“王冕”的，不是说“父亲”的。“起语”不限于“主体”，总在句子的开头，不准在“谓语动词”之后。“止语”是“说话的终止点”，即话说到“止语”就完了。“王冕七岁上死了父亲”说到“父亲”就说完了。句法的“止语”如果从语用视角来看就是“句尾焦点”。句子的“谓语动词”后最多只有两个“语块”，即“补语”和“止语”，可容纳“邻体”和“客体”，如果有空缺则可容纳“止语化”(finalization)的成分，绝不容纳任何“状语”。总而言之，汉语的“起语、止语”是句法跟语用保持统一的句子成分。

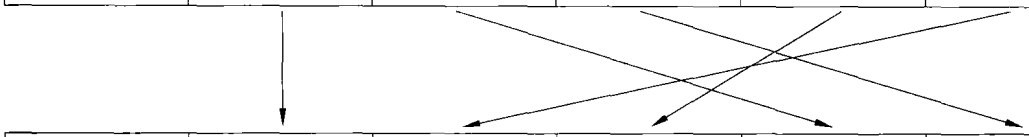
(7) 研究语言要重视“语义网络”和基于“语义角色”的排序模式

进入新世纪，国内外语言学界更加重视“语义研究”。在语言信息处理和第二语言教学中都显示了“语义网络”和“语义角色”的重要作用。2001年本书作者在商务印书馆出版了《汉语语法的意合网络》，被许多高校和科研院所指定为研究生必读书目，参见文献(鲁川，2001)。近年来，作者在“语义网络和语义角色”方面又有新的研究成果。跟那本《汉语语法的意合网络》相比，本书的研究已从“按语义角色来分析语义关系”发展到“按语义角色来排列语言顺序”。

机器翻译先形成源语的“语义网络”，在各种“语义角色”的节点中填入译语的对应词，再

“按译语的语义角色排序规则”列表，就生成了合格的译语句子，如图 0-2 所示。

Example 1	Our graduates	consulted	the data base	in the library	last Friday
Example 2	Professor Li	reported	his discovery	at the meeting	this morning
Example 3	The policeman	caught	a thief	on the train	yesterday
Semantic role	*Subjective*	*Verbal*	*Objective*	*Location*	*Time*



语义角色	主体	时间	处所	述谓	客体
译语角色序号	①	②	③	④	⑤
例句 1	我们的研究生	上星期五	在图书馆	查阅了	数据库
例句 2	李教授	今天上午	在会上	报告了	他的发现
例句 3	那个警察	昨天	在列车上	抓了	一个小偷

图 0-2　英汉“填表式翻译”示意图

从图 0-2 可知这种“句模”对比：英语“S＋V＋O＋L＋T”；汉语“S＋T＋L＋V＋O”。

语言类型学指出：日语是 SOV 语言；英语是 SVO 语言；汉语也是 SVO 语言。如果引入“状语”A(adverbial)，就可以区别出：英语是 SVOA 语言；汉语是 SAVO 语言。

有关汉语句子的全面的“按语义角色来排列语言顺序”的模式见表 0-3。

表 0-3　按语义角色来排列汉语句子顺序总表

预想部分						中枢	待晓部分	
起因		环境		状况			后果	
主体	情由	时间	空间	同事	涉事		量度/邻体	邻体/客体
			范围	用事			客体	系体
①	②	③	④	⑤	⑥	⑦	⑧	⑨
王汶		今年	在故乡	跟全家		**欢度**	春节	
钱科长	为了搞配件	这个月		乘卡车	从广州	**跑了**	三趟	深圳
李华	受委托	昨天	在医院	恭敬地	替病人	**送**	周大夫	一面锦旗
歹徒们		上周末	在郊区	持枪		**抢了**	储蓄所	几万现金
爷爷	照奶奶的意见	星期天	在瓜田	以低价	给全家	**买了**	瓜农	一车西瓜
郭厂长	为了提高效率	今天	在厂部	亲自	从机房	**搬**	到办公室	两台电脑
哥哥	按妈妈的嘱咐	每天清晨	在厨房	用铝锅	给妹妹	**热**	一次	牛奶
董事长	为开拓市场	星期一	在会上	隆重地		**聘任**	李华博士	为总经理

针对表 0-3 中的这些问题：怎样确定各种“语义角色”？为什么这种语义角色在“前”而

那种语义角色在“后”？为什么同一种语义角色在汉语中应该排在谓语动词之“前”而在英语中应排在谓语动词之“后”？本书提供了有理有据的详细论述。

（8）研究语言要有足够的精选的“语言资料”和基于数据库的统计数字

本书是依靠作者长期积累的“语料库”和“语言知识库”而写成的。在此愿意把自己认为重要的资料和数据推荐给广大的读者。用数据库来研究语言，有时候计算机就可帮助人们发现一些有规律的语言现象。例如，从数据库中把“月”作偏旁的字都调出来而列于屏幕上，左偏旁的有：“肌肋肛肝肚肘肤肠肺肢股胆脚腿”；右偏旁的有：“明朗期朝朔望”。于是认识到：“月”作左偏旁，义为“肉”，造字较多；“月”作右偏旁，义为“月亮或时期”，造字较少。本书提供的丰富资料有：“汉字最常用一千字表”、“汉字的一百个常用义元表”、“汉字的三百个常用音元表”、“现代汉语基本句模表”和“汉语句子按语义角色排序总表”等。这些资料使本书兼具理论书和工具书的特色。

（9）研究一种语言的特点要从跟其他语言的对比中求索，从“A 语言有而 B 语言无”，或“B 语言有而 A 语言无”的语言现象中往往能找到突破口

当中国教师给外国学生讲“学汉语一定要学会‘把’字句”时，英国学生在心里说：“我们英语就没有‘把’字句”。当英语教师给中国学生讲：“学英语要学会定冠词 the”的时候，中国学生在心里说：“我们汉语就不需要定冠词 the”。其实，正是“把字句”跟“定冠词 the”有着极其深刻的联系。相关信息可参见表 0-4，更详细的解释可参见第 10 章。

表 0-4 汉语“把字句”跟英语“定冠词 the”的联系表

例	英 语	汉 语	简 单 解 释
1	Give me ***a*** novel.	给我**一本**小说。	在火车上对带有多本小说的同伴说的，什么小说都行。
2	Give me ***that*** novel.	给我**那本**小说。	在书店买书时对售货员说的，小说是顾客手指的那本。
3	Give me ***the*** novel.	**把**小说给我。	对借走小说该还的人说的，听话的人明白指的是哪本。

把表 0-4 中第 3 例英语句子译成“给我**这本**小说”或把第 3 例汉语句子译成“Give me ***this*** novel”都不对。因为 **this** 跟 **that** 都是“指示代词”，区别在于 **this** 是“近指”而 **that** 是“远指”。所以，“Give me ***this*** novel”仍然是在书店买书时对售货员说的话，指的是离顾客很近的小说，并不能表示“**把**小说给我”的意思。“**把**小说给我”中的“小说”并不在谈话双方的眼前，既不能“近指”也不能“远指”，而是索书人和借书人心中都明白的都确指的那本应该归还但是并没归还的“小说（***the*** novel）”。

本书作者是我国 20 世纪 50 年代培养的第一批计算机专业大学生，半个世纪以来一直奋战在人工智能、知识工程、自然语言理解、机器翻译、计算语言学和对外汉语教学的科研第一线。为了事业的需要，本书作者长期坚持不懈地自学和拜师求教，从而掌握了较深厚的现代汉语、古代汉语、英语、语言文字学的知识，发表了相当数量的探索汉语、汉字特点和汉英对比的论著，被学界公认为语言学家和我国计算语言学的学术带头人之一。本书是作者为自己、知识工程的科研工作者和对外汉语教学与研究者精选的语言知识库，是作者为研究生授课的讲义。本书也可以比喻为作者的“学术自传”。

本书只讨论从事“知识工程”的人应具备的那些“语言学”知识，不涉及信息处理的技术、

程序和算法，全书通俗易懂。对于关心语言信息处理的文科师生来说，不仅可了解“知识工程”需要什么样的“语言学”知识，也可找到在为语言产业提供“语言学”研究成果方面您自己能作出哪些贡献。至于关心世界汉语教学和我国外语教学的读者，因本书明确表示“知识工程包括跨语种知识交流、知识学习和知识传授”，您当然能在书中找到感兴趣的章节和具有重要参考价值的内容。

感谢国内外语言学界和计算语言学界的各位老师，没有他们的指导和启发就没有本书。

感谢“信息处理用规范汉字字义统计和造词模式”和“信息处理与对外汉语教学的句子语序模式”两个项目组的全体成员，特别感谢项目组主要成员王玉菊研究员。

感谢司富珍教授审阅了本书的全部初稿并提出了宝贵的修改意见。

感谢清华大学出版社对本书的热情和耐心以及为本书的出版付出的辛劳。

由于作者水平所限，书中有一些不成熟的论点，特别是有一些跟语言学界流行理论不同的说法，可能是作者的错误认识，诚挚地欢迎批评指正。

宝贵的批评意见请发电子邮件至：Luch111@163.com。

谢谢！

鲁　川

2007 年 6 月 28 日

于教育部语言文字应用研究所计算语言学研究室

目　录

Contents

上篇　导　论

中篇 语言的库存单位