新一代人工智能系列教材

自然语言处理

刘挺 秦兵 赵军 黄萱菁 车万翔 编著

高等教育出版社

新一代人工智能系列教材

自然语言处理

刘挺 秦兵 赵军 黄萱菁 车万翔 编著

高等教育出版社·北京

内容提要

本书在新一代人工智能背景下，重点介绍自然语言处理的基础知识、主要的经典技术、前沿技术及应用。本书分为四部分内容进行阐述：基础知识、自然语言处理的基础技术、自然语言处理的应用技术、自然语言处理的应用系统。首先，在基础知识部分，介绍了语言学理论和机器学习理论两个方面的基础。其次，介绍自然语言处理中的基础技术，包括语言模型、词法分析、句法分析和语义分析等内容。然后，通过文本的信息抽取、篇章分析、情感分析及文本生成等任务，介绍自然语言处理的应用技术。最后，在自然语言处理的应用系统方面，详细介绍了问答系统、对话系统和机器翻译中的相关技术和系统结构等内容。

本书可作为人工智能专业、智能科学与技术专业以及计算机类相关专业的本科生及研究生学习自然语言处理的教材，也可以作为自然语言方向相关研究人员的参考用书。

自然语言处理

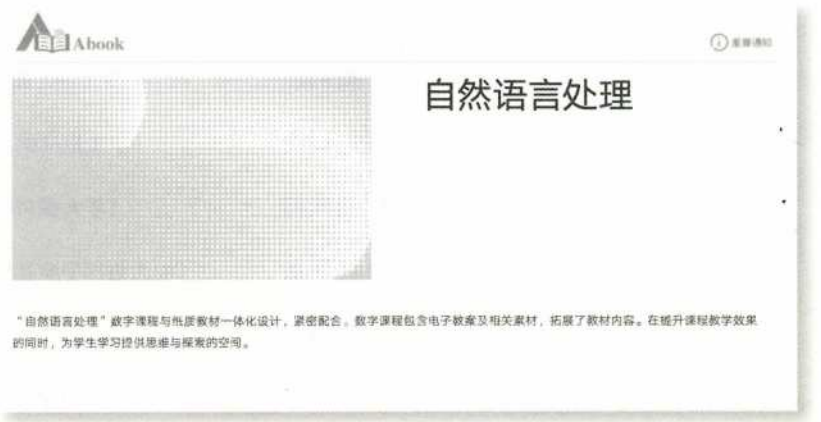
刘　挺　秦　兵
赵　军　黄萱菁
车万翔

1　计算机访问http://abook.hep.com.cn/187935，或手机扫描二维码、下载并安装Abook应用。

2　注册并登录，进入“我的课程”。

3　输入封底数字课程账号（20位密码，刮开涂层可见），或通过Abook应用扫描封底数字课程账号二维码，完成课程绑定。

4　单击“进入课程”按钮，开始本数字课程的学习。



课程绑定后一年为数字课程使用有效期。受硬件限制，部分内容无法在手机端显示，请按提示通过计算机访问学习。

如有使用问题，请发邮件至abook@hep.com.cn。

http://abook.hep.com.cn/187935

前言

1956年的达特茅斯会议上，正式提出了人工智能的概念。2016年美国发布《国家人工智能研究和发展战略计划》，2017年我国国务院印发《新一代人工智能发展规划》。我国人工智能科技目前处于国际前沿的水平，且有望引领该方向的未来发展。

世界范围内人工智能方面的竞争归根结底是人才的竞争。高等院校是科技成果的一个重要孵化器，而人才培养则是科技发展背后的不竭动力。如何在当前人工智能快速发展的条件下，建立起完善的人工智能人才培养体系，决定了未来我们国家在人工智能方向上发展的广度和高度，以及能否处于并保持国际领先的科技水平。

目前学者们普遍对人工智能发展阶段的认识是从运算智能到感知智能再到认知智能的过程，而目前正处于认知智能的探索和突破阶段。自然语言处理属于认知智能的研究范畴，被誉为“人工智能皇冠上的明珠”，因而受到了学术界和工业界的广泛关注，同时也被称为深度学习的下一个突破点。基于以上情况及条件，以及恰逢全国各地高等院校纷纷设立人工智能专业之际，笔者团队意识到组织编写一本适合本科生学习的自然语言处理方面的教材，将成为人工智能方向人才培养体系的一个重要基础。

在选择本书应包含哪些内容的时候，笔者组织了国内顶尖自然语言处理研究团队的人员进行研讨并设计本书的结构和大纲。在内容选择方面，本书包含了语言学理论和机器学习理论两个方面的基础，同时对于近几年随着深度学习的发展而逐渐发展的文本表示学习方面的内容进行了详细的介绍并作为一项自然语言处理的底层技术。在自然语言处理的基础技术方面，语言模型、词法分析、句法分析和语义分析等内容必然包含在本书之中，在介绍具体技术的同时，对该项技术涉及的其他任务的相关技术发展以及当前较为新颖的方法及模型进行了概括性介绍。进一步，文本的信息抽取、篇章分析、情感分析及文本生成任务作为自然语言处理应用技术的典型代表在本书中进行了详细的介绍。最后，在自然语言处理的应用系统方面，本书详细介绍了问答系统、对话系统和机器翻译中的相关技术和系统结构等内容。

在撰写的过程之中，笔者团队着重考虑本书的教材属性以及面向学生的知识储备水平，尽量做到深入浅出和内容的完备性，同时对于需要延伸阅读的内容以及具有前期基础的同学，同样给出了相关资料的出处。本书的课后习题设计方面也经过多次研讨，对于不同章节要求也不同，如对于基础理论方面的习题以理解概念为主，对于基础技术方面的习题则注重任务的理解和相互间的关联，同时强调对于代表性方法和模型的理解，对于应用方法和应用系统方面的习题则涉及具体的技术和系统结构的实现等内容。

本书为主要面向本科生的自然语言处理课程教材。教学学时安排上，可以覆盖大学分（72学时）类专业必修课程或专业选修课（32学时）。作为专业必修课的教材

时，除了全部章节的课程授课外，可以安排针对自然语言基础研究及应用技术上的实验课时（建议24学时）；作为专业选修课的教材时，可以在授课时以自然语言处理的基础部分为主要授课内容，对于自然语言处理的应用系统方面的内容可适当介绍相关应用技术，具体学时分布及内容安排以实际授课教师根据教学需要为准。

本书第1章由哈尔滨工业大学张伟男副教授编写；第2章由北京大学詹卫东教授编写；第3章由南京大学戴新宇教授编写；第4章由复旦大学邱锡鹏教授编写；第5章、第6章由美国麻省理工学院（MIT）郭江博士后编写；第7章由复旦大学张奇教授编写；第8章由天津大学张梅山副教授编写；第9章由剑桥大学孙薇薇副教授编写；第10章由哈尔滨工业大学刘铭教授、丁效副教授、清华大学刘知远副教授编写；第11章由首都师范大学宋巍副教授编写；第12章由哈尔滨工业大学（深圳）徐睿峰教授、中科院自动化所刘康研究员、哈尔滨工业大学赵妍妍副教授编写；第13章由哈尔滨工业大学冯骁骋副教授、北京大学万小军教授编写；第14章由中科院自动化所何世柱副研究员、哈尔滨工业大学张宇教授编写；第15章由哈尔滨工业大学张伟男副教授、清华大学黄民烈副教授编写，第16章由中科院自动化所张家俊研究员编写。

全书由哈尔滨工业大学刘挺教授统稿并负责整理第1章、第2章和第16章，哈尔滨工业大学秦兵教授负责整理第10—13章，中科院自动化所赵军研究员负责整理第14—15章，复旦大学黄萱菁教授负责整理第3—4章，哈尔滨工业大学车万翔教授负责整理第5—9章。感谢高等教育出版社在本书写作过程中提供的帮助，感谢小米公司自然语言处理首席科学家王斌博士审阅了全书并给出了宝贵意见，本书的编写还参阅了大量的著作和相关文献，在此一并表示衷心的感谢！

最后，尽管本书在撰写的过程中始终保持认真细致的态度，但错误在所难免，敬请各位读者指正及反馈，我们将不断完善。同时，衷心地希望本书的出版能够为人工智能方向人才培养提供帮助和支持，助力我国人工智能及相关方向的整体发展和进步。

编者

2020年12月

目录

第2章　现代语言学基础

2

诺姆·乔姆斯基指出：研究人类的语言，就是探讨所谓“人类的本质”，也就是探讨迄今所知为人类独有的心理属性。

语言，作为人类社会特有的一种现象，在日常生活中如空气和水一样常见。但有关语言的诸多问题，仍然是科学界的未解之谜。例如：语言从何而来？语言跟思维是什么关系？语言能力是独立于人类的认知能力还是人类认知能力的一部分？儿童为何能在很短时间内掌握母语，而大多数成人学习外语却非常困难？等等。

本章分为4节：第1节简述现代语言学的学科来源与分支发展概况，介绍语言学者对人类语言基本性质的思考；第2节详细说明自然语言系统的层级组织方式、语言系统各级单位的基本范畴划分，以及现代语言学用于描述和解释各个层面语言现象（语音、词汇、句法、语义、语用等）的知识模型；第3节简要探讨语言系统的歧义性和创造性，自然语言的歧义性和创造性，对计算机的自然语言处理构成了巨大挑战；第4节介绍为自然语言处理提供支持的语言知识资源，包括语言知识库和标注语料库两大类型。

2.1　语言学与人类的语言

2.1.1　现代语言学的起源及学科分支

人类把语言作为学问研究的对象，在各个古代文明中均有体现。关于语言的早期研究具有鲜明的实用工具目的，人们关心的是古典文本的释义以及修辞问题。这使得有关语言的学问长期以来仅以其人文性的一面融入各个文明的学术传统。直到大航海时代开启了殖民探险和大范围的人口迁移，使得语言之间的直接接触大量增加。伴随着现代科学思想渗透到人类社会的方方面面，学术界逐渐开始从科学（而非仅仅是人文）的角度来审视语言的群体（而非仅仅是个别的语言），通过对语言形态和基本词汇的对应关系考察，根据相似确立语言间的同源关系，根据差异划分语言谱系树。如同生物学的谱系分类研究那样，人类对语言的认识，才开始步入科学的轨道，不再停留在个体语言的具体表达内容的注释、翻译和修辞润色上，而是迈上了宏观俯瞰的层面，把语言的接触、融合、变异，放在进化史观的时空大格局下加以审视，举起历史比较语言学的旗帜，拉开了科学意义上现代语言学的序幕。

自此之后再到20世纪初叶，以1916年索绪尔的《普通语言学教程》出版为标志，语言学正式开启了结构主义方法论的时代，彻底丢弃了所谓“经学之附庸”的标签，开始走上独立学科的发展道路。跟化学研究物质的结构与化合现象的思想一样，学者们开始全方位地将语言整体视为一个系统，而不仅仅是个别孤立的现象，并对语言成分以及各成分间的关系展开分析，构建科学意义上的语言学理论体系。可以说，

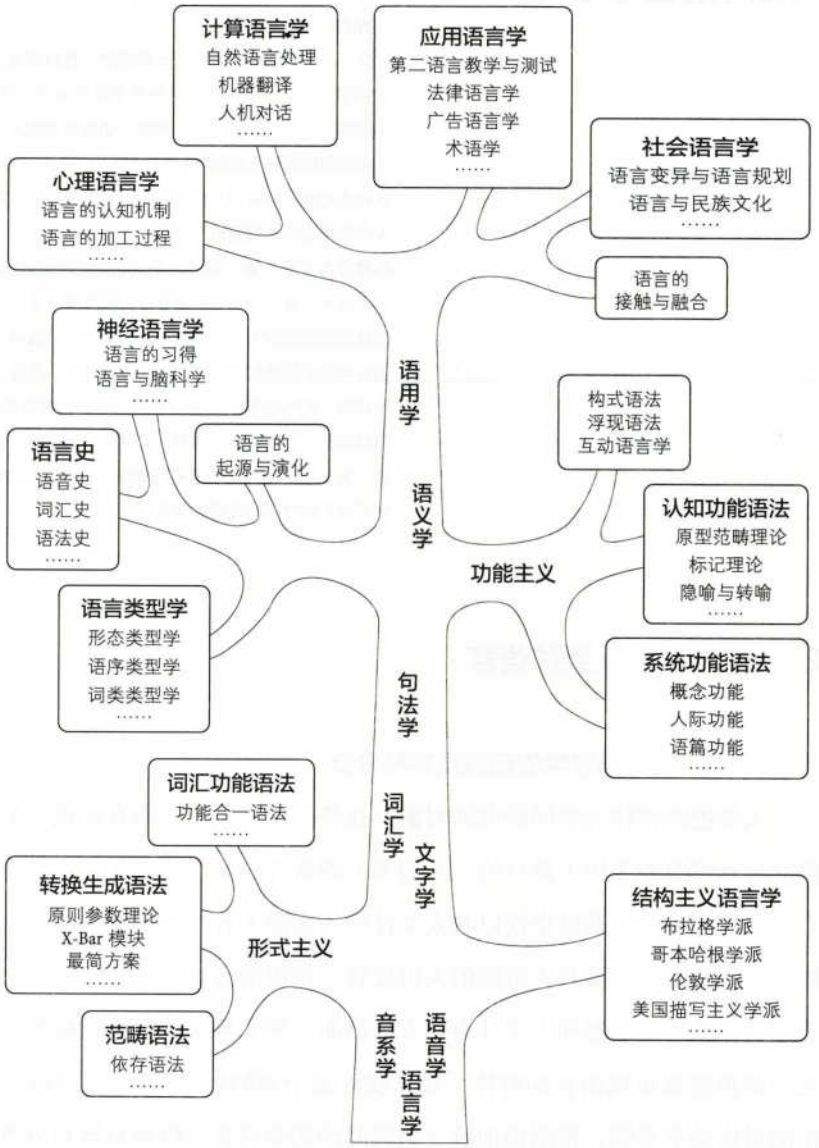


图2.1 现代语言学分支流派简化示意图

整个20世纪，是语言学大发展的100年。这100年间，人类对语言的认识，无论是视角的广度，还是触及的深度，都远远超过历史上的任何时代。语言的里里外外、方方面面，都成为科学研究的对象。语言学者提出了观察充分、描写充分、解释充分的研究目标，从不同角度，用不同的方法，在不同层面上尝试揭开沉睡多年的人类语言的奥秘，渐次发展出如图2.1所示的现代语言学分支和流派。

图2.1所示的语言学生长树图，只是极为粗略地勾勒了现代语言学轮廓。语言学的实际发展状况除了图2.1展示的枝繁叶茂之外，更多的还是细节上的藤蔓错杂、纠

缠不清。不过，其中有一条生长脉络，却因其突出的符号形式科学背景和鲜明的信息技术色彩而格外引人注目，这就是图2.1顶端所展示的计算语言学。站在21世纪的今天回望计算语言学这一脉枝叶的发展历程，来时之路大体清晰：第一阶段，以语言成分的结构关系为切入点，像化学家的工作那样，分析语言系统各级单位的性质（结构主义语言学）；第二阶段，以语言的无限能产性为旨归，像数学家的工作那样，构建形式规则，约束语言单位的组合与变换（以转换生成语法为代表的形式主义语言学）；第三阶段，以帮助机器模拟人的自然语言能力为应用目标，像计算机科学家的工作那样，为计算机提供自然语言处理的理论、模型、算法和数据。

2.1.2 人类语言的符号性与层级性

一般人很容易把语言等同于听到的话或看到的语句和文章。但实际上，后者只是语言的外在表现，而非语言本身。人类认识到这一点的历史并不长。20世纪初现代语言学问世之时，语言学家才开始明确区分“语言”（language）和“言语”（parole）这一对概念，来说明人脑内在的“语言功能”或者说“语言能力”，与社会交际中外在的“言语表现”之间的不同。前者是抽象的无法直接感知的心理或认知模型，后者是具体的可以诉诸听觉或视觉感知的物理实体。这种关系可以用抽象的围棋规则、棋理，与具体的围棋棋局来类比。棋局是可以观察到的游戏表象，而其背后隐藏的围棋规则和胜负之道（棋理），是无法直接观察到的，但有可能根据棋局以及棋手下棋落子的过程去推知。

尽管语言学界区分了“语言”和“言语”，但一方面二者的联系实在太过紧密，在大多数宽泛使用的场合，人们仍然无意识地把二者混为一谈。在思考“语言的性质是什么”这一问题时，人们给出的答案往往是通过观察言语而体会到的性质；另一方面，由于对人脑的了解太少，目前还缺乏比较系统的可靠的对语言生物学基础的深入认识，这也使得现在对语言的认识，集中于外在的言语表现，即交际过程和结果（有声言语及其文字记录）方面，而对大脑如何产生并理解语言，缺乏足够的了解。

广为接受的有关语言的定性概括是从功用的角度描述语言，即语言是人类社会的交际工具。布拉格学派的代表人物，语言学家雅各布森综合前人对语言交际功能的认识，提出了一个语言交际功能六要素的说法，影响广泛[4]。图2.2是六要素的概括示意图。这既是对典型通信系统参与要素的概括，也是对人类语言的外在表现，即语言的不同社会功能比较恰当的划分。按照雅各布森的看法，理解语言作为交际工具的人类交际系统，涉及下面这六个要素，对应六个功能（这里的序号对应图2.2中的标号）。

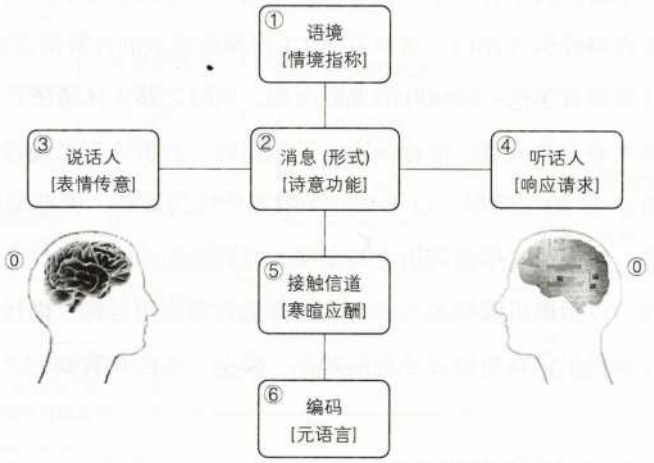


图2.2 根据雅各布森语言交际功能六要素理论绘制的示意图

（1）语境（context）要素：通过指称（referential），将语言符号跟外部世界联系起来，通过描述一个情境（语境或上下文），为交际提供一个背景。

（2）消息（message）要素：通过符号的形式本身，引起诗意的感受。

（3）说话人（speaker/addresser）要素：把说话人掌握的情况、当前的感受和需求，传达出去。

（4）听话人（receiver/addresser）要素：听话人对说话人的信息做出反应，如响应命令，回答疑问等。

（5）接触信道（contact/channel）要素：人与人之间通过寒暄应酬，建立并维护社会关系。

（6）编码（code）要素：人们可以用语言谈论语言，语言可以编码自己，可以看作是自身的元语言。

上述六个要素及其涉及的功能，高度概括了语言在人类社会中所起的作用。图2.2中的说话人和听话人处标记了序号0，暗示人类对语言系统的关注仅触摸到了语言的外在表现，对于大脑深处的语言奥秘，还是所知甚少。

图2.2的语言通信系统模型不仅可以用来描画人类交际的基本面貌，也可以据此观察动物的信息传递情况。关于动物也有语言的说法并不少见，常举的例子包括鸟类（鹦鹉）、蜜蜂、海豚、黑猩猩等。在语言学者早期关于语言起源的讨论中，甚至有人想象“语言诞生于人类的求爱期……最初说出的言语有点儿像屋顶上的猫咪在夜色下吟诵的爱情诗，又有点儿像夜莺唱出旋律优美的爱情之歌”（奥托·叶斯派森《语言的本质、发展和起源》）。这是赋予了动物声音和行为以额外的意义，超出了那些声音和行为本身。正如哲学家罗素所言，“一只狗，无论它叫得多么起劲，却无法告诉你，

它父母虽穷但诚实”。能发出声音，并且能传递信息的动物很多，但是，迄今还没有证据显示，动物以声音或其他行为传递信息的方式，能够跟人类语言的复杂程度相提并论。在图2.2所示的六个要素中，动物的信息传递完全不涉及（1）（2）和（5）（6）。而在（3）（4）两个要素上，动物信息传递表现为“直来直去”的固定信号信息传递模式，例如，动物学家观察到雄蜘蛛会以一套固定的动作向雌蜘蛛传递求爱信息；有一种意大利蜜蜂，可以用三种舞蹈动作：圆形舞、镰形舞、摆尾舞，来表示蜜源距离蜂巢的距离远近差别。这些信息传递方式，要么仅限于一个交际目的，要么是在一个目的下仅有十分有限的取值可选。跟人类语言无限丰富的表达可能性无法同日而语。

除自然有声语言外，人类也发明了一些非自然语言的信息传递系统，如烽火、旗语、交通灯系统、人类的体态、手势语等，人类还将一些跟沟通、表达思想有关系的系统赋予语言的名称，如音乐语言、绘画语言、建筑语言、数学语言、计算机语言等，此外还有诸如“爱是人类共同的语言”这样的修辞性表达，都可以从图2.2所示的交际系统六要素去分析和理解。

上述这样的自然语言交际系统，从何而来？线索之一是从沟通的动机角度溯源。有学者提出了“以手指物及比划示意”对沟通的意义，并仔细分析了人类沟通的三种基本动机：（1）求助（即要别人去做自己想叫他们做的事）；（2）助人（即我要你知道某事，因为该事对你有帮助）；（3）分享（即我要你有某种感觉，这样我们可以一起分享情感/意见/态度）。在进化过程中，从“指物”到“言说”，沟通的这三种基本动机和复杂的现实生存环境交织在一起，通过人类特有的文化进化（而不仅仅是生物进化）途径，推动着人类社会逐渐发展出复杂的语言系统，以有穷符号编码的形式，来实现无穷的沟通意图[8]。

上述对语言系统社会功能的观察当然主要来自那些人们熟悉的自然语言，不过，语言调查表明，无论一种语言是否发展出书面语的形式，已知的人类语言均有图2.2所示的交际要素及其对应功能表现。在探索语言本质属性的努力中，也包括对世界上所有人类语言的不断收集和系统整理。除了带有普查性质的语言记录和保护工作外，语言学家（特别是语言类型学家）以及人类学家出于研究目的，对很多语言做了更为深入的调查和基本事实记录[2]。尽管人类语言表面差异巨大，但多数的现代语言学者相信语言之间的内在共性大于表面的差异。作为典型的复杂自适应系统（complex adaptive system），人类每个族群的自然语言，都是在历史文化的进程中，由大量个体或积极或被动地相互竞争与合作，在没有一个权威做统一规划的情况下，通过无数的彼此相互作用和相互适应形成的整体上相对有序的系统。现代语言学理论的主要目

标是寻找普遍语法（universal grammar），即用统一的原则来描写所有人类语言的性质，包括不同语言的共性和个性两个方面。对此的研究和探讨仍然有很长的路要走，因为语言是人脑的产物，对语言本性的了解，最终必然涉及人脑奥秘的揭示。这或许意味着，要说清楚人类全部语言的共性细节，仅靠观察言语表现，很难完全实现这一目标。不过，在过去100年现代语言学的发展中，已逐步形成了一套相对统一的，可以大体适用于不同的具体语言，用来描写和分析语言事实，解释语言现象的理论框架。而之所以能达成这一点，根本原因在于，人类语言都是声音（形式）与意义（内容）相结合的符号系统。在符号性、层级性、结构性、组合性与聚合性等基础层面，人类语言具有广泛的系统共性。

人类语言系统的符号性表现为3个基本特点。

（1）语言系统的**符号是形式和意义的对应结合体**，在语言学术语中，通常把符号的形式称为“**能指**”（signifier），符号的意义称为“**所指**”（signified）。例如，英语中“eye[aɪ]”代表了符号形式（书写形式和语音形式），其所指则是“眼睛”  这一事物。

（2）**符号能指与所指之间的对应关系是约定的，任意的**，并非自然联系。在英语中[aɪ]代表的声音对应着eye（“眼睛”）这一事物，而在汉语系统中，[aɪ]这个声音代表的则是“爱”（love）这个行为概念。用什么样的声音代表什么事物或者行为，在不同语言中，完全是任意约定的。在一种语言系统中，一个声音（能指）与其所对应的意义（所指）的关系一旦约定，被社会认可，则这种约定关系在一定时间内将相对稳定，不能任意取消或改变。而随着使用的增多，在使用过程中，个别符号的形式——意义对应的约定关系，也可以改变。因此，语言符号系统在保持整体相对稳定的同时，又处在个体不断变化重新组织的动态中，可以说是一种宏观形态的稳定与微观形态的变化动态平衡的状态。

（3）能指（形式）和所指（意义）配对的**基本符号单位可以进一步分解为无意义的基本语音单位**。人类语言符号系统实际上是一个两层架构。形式和意义配对的基本符号层之上，人类可以将基本符号单位组合起来表达无穷的意思，在形式和意义配对的基本符号层之下，还有一个纯物理信号的声音记号层。语言系统从无意义的基本语音单位的排列组合开始，形成了有意义的语言符号，再进一步组合繁衍出无穷无尽的语义。

现代语言学在仔细观察和深入分析自然语言的这种层级组织方式的基础上，逐

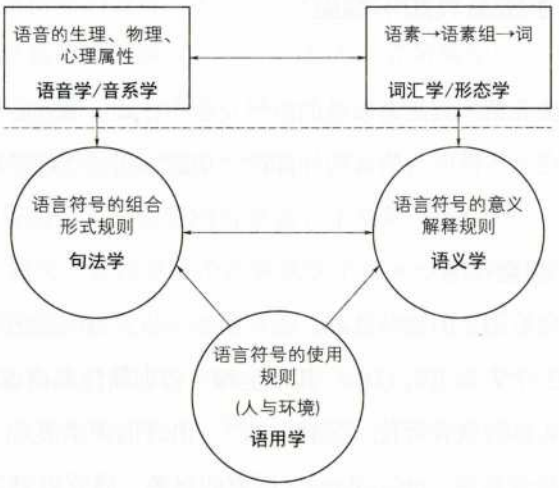


图2.3 语言学本体研究的五大核心模块

渐形成了语言学的5个核心学科，即图2.1语言学分支树的主干部分[①]：（1）语音学/音系学，研究语音的生理、物理（声学）和社会心理属性；（2）词汇学/形态学，研究语言系统的基本意义单位语素和词的构造模式、产生机制、形态和意义变化；（3）句法学，研究语言符号的形式组合模式和约束条件；（4）语义学，研究语言符号的语义解释模式和约束条件；（5）语用学，在真实的语言使用场景下，研究人的表达意图和语言形式以及信息传递方式之间的关系。这5个方面既相对独立，又相互紧密联系，共同构成了人类语言知识系统的总体框架，如图2.3所示。

从语言系统的分层机制来说，语音学/音系学不涉及符号的意义，只涉及语音符号的形式区别，是相对独立的一层；其他的词汇、句法、语义、语用，研究的核心都是各级语言符号单位的形式和意义的对应关系问题。而从所研究单位的数量角度来说，语音学和音系学研究的对象是数量有限（绝对数量也不多）的语音符号。词汇学和形态学的研究对象是数量相对有限（但绝对数量较为庞大）的语素和词汇符号。图2.3将语音学/音系学和词汇学/形态学跟其他三个模块句法学、语义学和语用学用虚线隔开，意指虚线上面两个模块的分析对象是数量上相对封闭的语言单位，其组合模式也相对有限；虚线下面三个模块则涉及语言的全部，是开放的无限的语言单位。

现代语言学的主要目标就是揭示语言系统各个层面的有限规则，充分描写并充分解释从语音到词汇到句法到语义到语用的各个层级上的语言现象，来实现“语言是有限规则的无限应用”这一语言哲学信条。

① 图2.1的树中还列出了文字学，不过，文字学主要是针对象形文字系统的研究。对拼音文字系统而言，文字学缺乏普遍性。

2.2 语言系统及其知识模型

本节主要介绍人类语言系统的组织及运作方式、语言各单位的层级结构，并试图在语言的内在本体性质与语言的外部社会功能之间建立起联系。

2.2.1 语音系统

语音是自然语言的物理基础。现代语音学将人类语言的声音系统分为生理、物理和心理属性三个方面进行分析，其中生理、物理属性是语音学（phonetics[①]）的研究对象，研究人体的发音部位、发音方法[②]、语音的声学表现（物理特征）等；语音的心理属性，是音系学（phonology）研究的对象，研究以对立互补关系为原则，通过建立语音的区别特征体系，来确定语言的基本音位单位，归纳音位组合模式和分布条件，即音节规则。在以上针对语音基本静态单位的系统分析基础上，还要进一步讨论音节在实际语音流中的动态语流音变现象。

语音跟其他声音一样，在声学中都以声波来表示。声波可以用周期、频率、振幅等物理量加以描述。具体到语音，通常对应到4个语音声学要素：**音高**、**音强**、**音长**、**音质**。其中音高的差异是声波频率不同造成的，语音中声调的不同主要就由音高决定。音强取决于声波的振幅，表现为声音的强弱，可以帮助形成语音中轻重音的差异。音长是语音持续时间的长短，可以形成语音中长音和短音的对立。音质也叫音色，是语音最重要的物理要素，大部分语音的不同，主要是音质的不同。音质（音色）由振动体、振动方式、共鸣腔形状等多个因素综合决定。通常人们能区分不同人的声音，就是因为人的发音器官（比如声带、口腔）的生理特征有差异，造成音色不同。同一个人发不同的音（比如“雨”和“衣”），因为发音器官和发音方式不同，改变了自身共鸣腔的形状，造成声音不同。

人类语言有两个最基本的语音类：**元音**（vowel）和**辅音**（consonant）。发元音时气流通过声腔没有遇到阻碍，因此可以独立且持续地发出元音（如a），使得元音听起来清晰响亮；发辅音时气流会在声腔的某个位置遇到不同程度的阻碍，因此辅音很难持续，听感上也远不如元音清晰响亮。元音和辅音统称为**音素**（phone），是听感上人能感知的最小语音片段（因此也用segment这个术语）。音素是从物理层面对语音单位的描述。从声学角度看，语音之间的差异可能很大，但从心理角度，人

① 在英语术语中也经常宽泛地使用speech来表示语音学。

② 完整的语音生理系统除了发音功能当然还要有听音感知功能，涉及听音器官人耳的构造、听辨工作机制等。限于篇幅，这里从略。

们却不会关注所有的差异，人的语音处理系统有一套机制，可以删繁就简，只保留那些值得注意的差异。换言之，人对客观声音的感知，可以是“很主观”的。例如北京话中，“榴梿”（liúlián）和“牛年”（niúnián）的发音差异“明显”，但在武汉话中，这两个词当地人要么发成“榴梿”，要么是“牛年”，当地人音感上没有区别。可见，在心理层面，人对物理的音素进行了重新分类处理，这就形成了心理上的**音位**（phoneme）概念。音位是一个语音系统中有区分意义作用的最小语音单位。两个物理上属于不同音质的音素，如果在语音系统中不承担区别的作用，就被看作是同一个音位，例如在武汉话中，l和n就合并为同一音位（可以任选其中一个符号作为该音位的记号）。而在北京话中，l和n就是两个不同的音位，可以区分像“牛累了”和“流泪了”、“梨水”和“泥水”这样的语词，在北京话的记音系统中就必须用不同的符号来区别。

不难看出，谈论音素时，可以跨语言（方言）或者说不依赖一种具体语言，但在谈论音位时，则必须依赖一种具体的语言（方言）才行。音位是从语音心理系统的角度对音素的划分和认定。

在分析具体语音系统的音位过程中，很自然地，学界对于语音之间的区别有了更深入的认识，逐渐发展出一套用**区别特征**来定义音位的办法。比如“瀑布”（pùbù）中，“瀑”和“布”两个音的区别是辅音声母p和b的区别，而这两个音的特征分别是[+双唇，+闭塞，+清音，+送气]和[+双唇，+闭塞，+清音，-送气]。通过这种描写方式，两个音的共同点和差异就很清楚了。p和b在发音部位（双唇位置形成阻塞）上是一样的，发音方法上，都是闭塞音（爆破音）和清音（声带不振动），只有送气和不送气这一点构成对立。通过将音位还原成更小的区别特征来定义的方法，不仅可以更精细地描写和比较一个语音系统内部不同音位间的关系，还可以跨语言地比较不同语音系统的特点。这种方法不仅在音系学研究中受到重视，还推广到了词汇、句法、语义学的研究中。

音位分析的主要目的是得到一个语音系统的基本单位，在此基础上，归纳音位组合为**音节**的模式。音节是人能够自然感知到的最小语音片段。以汉语来说，人听到一句话经常会说这句话中包含几个字，从语音学来讲，一个字就是汉语的一个音节（字内部的语音片段很难自然感知到，需经仔细对比分析才能察觉）。对音节的感知是跟语音系统紧密相关的。跨语言（方言）的比较很容易体会到这一点。英国球员David Backham的名字在普通话中译为大卫·贝克汉姆，在广东话中则是大卫·碧咸。在普通话语音系统中，因为没有闭塞音结尾的音节，因此Back对应为“贝克”两个音节，ham对应为“汉姆”两个音节，而在粤语中，因为有塞音结尾的入声，可以将

Back对应为“碧”一个音节，ham对应为“咸”一个音节。可见，对音节的感知，是强烈依赖具体的语音系统的。

在一个语音系统内部，对音节的界限则可以从发音和听觉两方面来感知。例如，普通话中“d-a-i”这三个音素连读，可能是“大-衣”两个音节，也可能是“带”一个音节。从发音来说，前者会有两次肌肉紧张，后者只有一次；从听感来说，前者有一个响度起伏，分隔了两个音节，后者没有响度起伏，只有一个音节。

在实际的语音流中，音节受到前后环境的影响，实际读音跟作为独立单音节时的发音相比，可能发生变异，这就是语流中的共时**语流音变**。语流音变的类型很多，包括同化、异化、弱化、合音、增音、减音、脱漏、转换等。英语中常见的going to说成gonna，want to说成wanna，got to说成gotta，都是很典型的合音的例子。汉语中也存在很多语流音变的情况，北京话中“这一”说成zhei，“那一”说成“nei”也都是合音的例子。此外，北京话中如果两个上声调音节连读，前一个音节听起来会变成像是阳平调，例如“想买”的“想”，听起来像是“祥”。如果一个上声调音节和另一个非上声调音节连读，则前面这个上声调音节听起来只念了一半，用音调五度标调法[①]记录的话，就是[214]调，如“买书”的“买”。

从语言学科内部来讲，语音系统相对独立于语言系统的其他部分。但在现实中，语音跟语言中的其他模块都会发生联系：词层面有轻重音；语法层面，也有通过超音质特征来区分语法意义的诸多现象。例如，（1）**停顿**可以区分不同的语法结构：“牛奶饼干”没有停顿时是修饰性结构（牛奶味的饼干），有停顿时是并列结构（牛奶和饼干）；（2）**重音**也可以区分不同的结构：“讲得清楚”，重音若在“讲”上，是表达可能情态，相当于“能讲清楚”；重音若在“清楚”上，是表达对结果状态的评价，相当于“讲的条理很清楚”；（3）**语调**可以区分疑问、感叹和陈述等不同句类，在传递信息时实现不同的交际功能。

2.2.2 词汇系统

从语音系统到词汇系统，是一个巨大的飞跃——符号的意义登场了。音节（形式）绑定了意义（内容）之后，就从语音系统跃升而成为词汇系统中的单位。比照语音系统中的最小单位音素，词汇系统的最小单位称为**语素或词素**（morpheme）。语素一般定义为：最小的音义结合体。**词**由语素构成。词一般定义为：最小的能独立使用的音义结合体。不同于语素，词的内涵中增加了一项特征约束“能独立使用”。词可

① 普通话四个声调的五度标调法分别是：阴平[55]，阳平[35]，上升[214]，去声[51]。如：妈[ma55]，麻[ma35]，马[ma214]，骂[ma51]。

以由一个语素构成，如“人”，既是一个语素，也是一个词；也可以由两个或多个语素构成，如“人群”，包含“人”和“群”两个语素；“人民币”包含“人、民、币”三个语素[①]。词汇学的研究任务就是分析词与语素的形式和意义关系、词与词之间的意义联系（如同义、反义、上下位关系）等。本节仅讨论词的内部构造模式，有关词义以及词义之间的联系，在2.2.4节“语义系统”中再谈。

为了便于说明词的构造，即**构词法**[②]，需要先简要分析一下语素的类型。根据构词时的功能和地位差异，语素可以分为**词根**、**词缀**、**词尾**。词根语素有独立的实在意义，往往是决定一个词词义的主要因素，通常参与构词时的位置不固定，可前可后如“人、民”等就是词根语素。词缀是黏着语素，没有独立的实在意义，但可以辅助表义，例如，英语中的“un-、in-”等前缀（prefix），可表示否定义；英语中的“-er”，汉语中的“-者”等后缀（suffix），都可以标记指人范畴[③]。词尾（inflection）跟后缀的性质类似，也是黏着语素，不过词尾比后缀更抽象，一般表示语法意义。例如，英语中的-s附着在名词后表复数，-ed附着在动词后表示过去时等。汉语中“桌子、椅子、刀子、胖子”中的“子”尾，标记了这些词均为名词，性质接近词尾。不过，汉语没有成系统的词尾语素，词尾跟后缀没有明显的区别。

在语素类型划分的基础上，可以把词分为**单纯词**和**合成词**两大类。单纯词就是仅由一个语素构成的词，合成词则是由两个以上语素构成的词。以汉语为例，单纯词和合成词下面还可以划分出不同的小类，如表2.1所示。合成词中的复合词是由两个或两个以上的词根组合而成，词根之间的关系比较多样，因而复合词的小类较多。各构词类型的含义，可以通过表2.1中的示例体会，如“并列式”中“声音”的两个词根语素“声”和“音”是近义语素，二者并列组合为词；“复量式”是一种特殊的并列复合词，其中两个词根语素都是可以单用作量词的，如“场”和“次”单独都可用作量词，组合在一起构成复合量词，因而称为“复量式合成词”；“名量式”中“纸张”的两个词根语素“纸”为名词性语素，“张”为量词性语素，组合在一起构成“名+量”型复合词。限于篇幅，这里不再详述。一个语言中的日常用词相对稳定有限，表

① 注意，从语言学理论上讲，“人群”和“人民币”中的“人”身份是语素，而不是词。从逻辑自洽的角度说，词包含语素，但词不能包含词。词跟语素的定义很接近，区别在于能否独立使用。但是，能否独立使用是相对的，没有清晰的界定标准。这样的定义在操作层面不易把握。好在理论表述上，不难做到自洽。因为一个语言单位可以被同时赋予词和语素双重身份，当它自由时，它是词；当它不自由时，它是语素。反之亦然。

② 构词法在英语术语中对应的有两个不同的名称，一个是morphology，通常译为形态学，因为英语的构词跟形态语素关系密切，研究词的形态变化，除研究形态语素的语法意义外，同时也就是在研究构词。另一个是word formation，这个对应回汉语，有两种说法，一个是构词法，另一个是造词法。前者侧重分析词的内部构造模式，后者侧重分析新词如何产生。

③ 有的语言中有居于单词内部的词缀，即中缀（infix）。常见于南岛语系和南亚语系。例如，他加禄语（Tagalog）中的中缀-um-插入到动词takbo（跑，run）中，形成的tumakbo，就表示“跑了”（相当于英语run的ran形式，即简单过去式）。不过关于中缀是否需要独立成为一个语素范畴，语言学界存在争议。

2.2对汉语中6万多普通词汇，按照表2.1所示的构词类型进行了统计，可以据此大致了解汉语词汇系统中不同构词类型所占比例情况。

表2.1 汉语构词类型表

大类	中类	小类	示例
单纯词		1. 单音节单纯词	山 水 花 鸟 虫 鱼 的
	联绵词	2. 双声联绵词	琉璃 淋漓 吩咐 秋千 恍惚
		3. 叠韵联绵词	玫瑰 从容 腼腆 唠叨 徘徊
		4. 其他联绵词	蝙蝠 妯娌 狼狈 犹豫 玻璃
		5. 译音词	沙发 拷贝 摩托 巧克力 乌托邦
		6. 叠音词	猩猩 太太 姥姥 皑皑 悄悄
		7. 拟音词	扑通 哗啦 噼啪 乒乓 哎呀
合成词	复合	8. 并列式	声音 孤独 头绪 根本 制造
		9. 复量式	场次 架次 批次 篇部 部集
		10. 名量式	纸张 车辆 花朵 船只 枪支
		11. 数量式	一些 一丝 一线 一番 一点儿
		12. 方所式	野外 眼下 眼前 身上 天底下
		13. 定中式	草帽 货车 红旗 摇篮 试卷
		14. 状中式	飞快 重视 小看 雪白 肤浅
		15. 支配式	关心 留意 惊人 抱怨 怀疑
		16. 介宾式	从小 从前 从此 以后 沿途
		17. 连动式	查封 抽调 借用 逼供 劝降
		18. 补充式	改善 纠正 冻僵 证明 推翻
		19. 陈述式	性急 手软 肉麻 心疼 胆怯
	重叠	20. AA式重叠词	偏偏 常常 万万 舅舅 久久
		21. AABB式重叠词	骂骂咧咧 婆婆妈妈 形形色色
	附加	22. 后缀附加	桌子 椅子 胖乎乎 黏糊糊
		23. 前缀附加	老虎 老鼠 老师 阿姨 阿婆

表2.2 汉语构词类型比例统计表

序号	构词结构	总计	占比	序号	构词结构	总计	占比
1	定中式	26 377	40.41%	6	单纯词	2 344	3.59%
2	并列式	13 552	20.76%	7	连动式	2 112	3.24%
3	支配式	9 282	14.22%	8	陈述式	1 441	2.21%
4	状中式	4 816	7.38%	9	补充式	1 028	1.57%
5	后缀附加	2 732	4.19%	10	前缀附加	719	1.10%

续表

序号	构词结构	总计	占比	序号	构词结构	总计	占比
11	重叠词	312	0.48%	15	数量式	71	0.11%
12	方所式	218	0.33%	16	复量式	23	0.04%
13	介宾式	164	0.25%		合计	65 274	100.00%
14	名量式	83	0.13%				

（统计数据来源：北京大学计算语言学研究所《中文概念词典》构词模式标注，2016年）

表2.2的统计显示，汉语中大多数词都是合成词，合成词中又以**复合词**为主。汉语复合词的结构关系，跟词组组合时的结构关系基本一致。可以说，汉语基本上采用了同一套组合模式来处理复合词的构造和词组的构造（参见2.2.3节图2.3展示的汉语部分词组结构关系类型）。也正是这个原因，从单位划界角度来说，汉语中词跟词组的界限有一定的模糊性。例如，“开心”跟“开车”是不同性质的语言单位。尽管二者从语素序列上看很相似，但“开心”是词，“开车”是词组。因为前者两个语素结合更紧密，中间难以插入其他成分，更应视作一个整体；后者两个语素相对独立，“开”和“车”之间可以插入别的成分（如“开我的车、开了半天车”等），因此“开”和“车”应视为词，“开车”即为词组。再例如，“忘记”既可以是词，也可以是词组。当“忘记”作为词使用时，其中的语素“记”对整个词义不起任何作用，忘记=忘；当“忘记”作为词组使用时，“忘”和“记”都有表义功能，相当于说“忘记记录（了）”。

现代汉语以双音节词汇为主（在常用词汇中占比四分之三），但也有越来越多的三音节甚至四音节以上的词汇。这就涉及词的内部构造层次问题。多音节复合词的内部构造层次跟词组的构造方式基本一致（有关词组层次构造的分析，参见2.2.3节）。图2.4所示是几个示例。

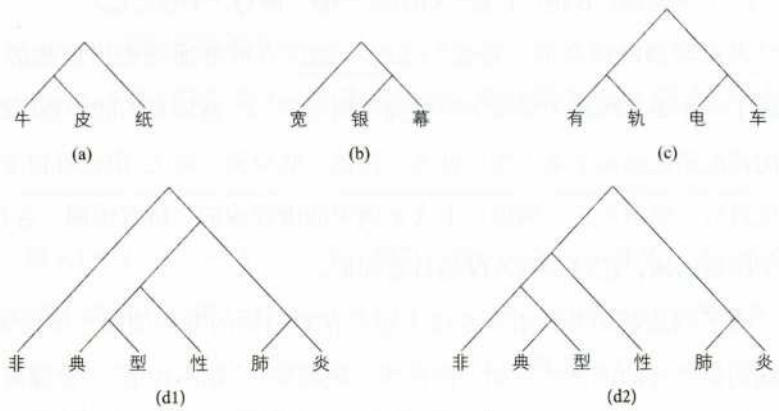


图2.4 汉语复合词内部层次构造示例

“非典型性肺炎”这个复合词按照图2.4中d1的层次切分是错的，按照d2是对的。d1中“非”跟“肺炎”发生直接的意义联系，跟复合词整体词义不符。d2中“非”跟“典型”发生直接的意义联系，跟复合词整体词义相符。

从根本上讲，研究一个语言单位，是希望从形式推知意义。就词而言，就是要从语素的意义，能推知词的整体意义。但语言符号形义结合的任意性（约定俗成）特点，从语素开始，到词这一级，仍然延续。从语素的意义到词的意义，有的有明显的联系，有的则看不出联系。例如，“观看”跟“观”和“看”的意思都比较接近，而“漂亮”跟“漂”和“亮”分别有什么关系？“东西”跟“东”和“西”分别又有什么关系？就很不清楚。研究汉语的构词法，主要作用在尽可能全面地描写词的构造模式，而对于词义的解释和预测性相对较弱。

下面再以汉语为例简要介绍创造新词的一些手段，即所谓的**造词法**。

（1）仿词：一般是仿照已有词语的结构，保留词语中的一些成分，替换另一些成分造出的临时词。例如：“一个阔人说要读经，嗡的一阵一群**狭人**也说要读经。岂但‘读’而已矣哉，据说还可以‘救国’哩。”（鲁迅《这个与那个》）

（2）缩略：一些较长的表达形式，因其比较常用，可以通过不同方式缩略为较短的形式，从而造出新词。例如：非典型性肺炎，缩略为“非典”；北京大学，缩略为“北大”；北京大学第三医院，缩略为“北医三院”；高端、大气、上档次，缩略为“高大上”。汉语中还有一类自古以来能产性就很强的，由数词开头缩合成词的缩略词模式，如古代就有的“五官、五谷”，现代的“四化、八荣八耻”等。

（3）谐音：童鞋（同学）、妹纸（妹子）、杯具（悲剧）、灰常（非常）。

（4）合音：酱紫（这样子）、表（不要）。

（5）拼音：哥屋恩（滚）、吃屋恩（蠢）。

（6）译音：幽默、逻辑，粉丝、锅庄、唐卡。

（7）字母词：B超，U盘、GB码、3D、阿Q、卡拉OK。

来自英语的译音词“粉丝”（fans）跟原有的普通词汇指食物的“粉丝”同形，造成了一个多义形式（语言学中称为“同形词”）。新词有的能得到广泛使用而进入语言的词汇系统稳定下来（如“博客、高铁、脱口秀”等），但也有很多只是昙花一现，热度过后，就消失了。例如，上文示例中的谐音构词、拼音构词、合音构词，基本都限于网络语体，还没有进入汉语日常词汇。

除了创造新词外，语言系统中还存在对旧词的变形使用，使得实际语言表达中出现词典之外的词形式，如“散散步、跳跳舞”“相不相信”“连澡都不洗”“一个丫头的命，却成天操着主子的心。”“这个默可不是谁都能幽的，官老爷能幽的默，老百

姓不见得能幽。”“这**傲**相当**骄**。我看到这个说法的时候**震**了一**惊**。”等。其中“散步、跳舞、相信、洗澡、操心、幽默、骄傲、震惊”等词语，或者词中语素同形重复，或者两个语素被其他成分隔开，或者改变语素原来顺序后再被其他成分隔开，都是临时把语素升级为词，以词的身份参与组合。这跟上文提到的汉语中词跟词组界限比较模糊的情况类似，词跟语素的界限也有一定的模糊性。

词汇学研究中还非常关注从语言的历史发展角度观察“新词”的产生。语法化的研究者相信“今日之词法，皆为昨日之句法”。即今天的词，可能在过去是一个词组（语法结构）。这里仅举一例说明。汉语中“一律”这个词，现在是副词，表示“相同、无例外”的意思。但在汉语史上，“一律”早先是一个“数词+名词”的词组，其中“律”的原义是“法则、律条”（这个意思现在仍用于“法律”这个词中），“一律”就是“同一法则”的意思。后来在使用中发生句法分布位置的变化，逐渐由谓语位置，前移到状语位置，至南宋时发展出副词的用法和词义，到明清已经固定下来。

例2.1 （1）其以为音也，**一律**而生五音，十二律而为六十音。（西汉《淮南子》）

（2）今有司以为予告得归，赐告不得，是**一律**两科，失省刑之意。（《汉书·冯奉世附传》）

（3）侗者，同也，於物同然**一律**，無所識別之謂。（南宋《朱子语类》）

（4）或流于申韩，或归于黄老，或有体而无用，或有用而无体，不可**一律**观。（南宋《朱子语类》）

（5）今之人传得法时，便授与人，更不问他人肥与瘠，怯与壮。但是**一律**教他，未有不败、不成病痛者。（南宋《朱子语类》）

（6）商功父赋性慷慨，将着贾家之物作为己财，**一律**挥霍。（明《二刻拍案惊奇》）

（7）吩咐家丁，凡来道喜的，都**一律**挡驾。（清《二十年目睹之怪现状》）

例2.1中，（1）—（3）“一律”都是“数词+名词”词组，（1）中指音律，（2）中指法律，都是“一律”的本义；（3）中“一律”的语义引申为抽象的“相同”义。（4）—（7）中，“相同、无例外”这个引申义逐渐固定下来，并且语法位置常居动词前，成为状语性成分，凝固为双音节副词。

2.2.3 句法系统

理想的句法系统有两个作用：① 检查什么样的词序列是合法的句子，什么样的词序列是不合法的；② 对于合法的词序列，分析句子中各个词之间的关系，建构句子的内部结构，为理解句子的语义做准备。这是从理解句子的角度对句法系统的描述。如果从生成句子的角度看，说话人的大脑基于跟理解时所用相同的句法系统，对表义所需的词语进行正确排序，输出合法且能恰当表达意思的句子。

例2.2 （1）她 从 东京 来 。
（2）她 从 来 东京 。
（3）她 从 来 东京 到 现在 就 没 笑 过 。

例2.2中，（1）和（2）的词语是相同的，不过词的顺序不同。（1）是汉语中能说的句子，（2）不成立。句法系统要对（1）和（2）做出区分。（3）说明，句法系统不能仅仅从表层线性序列的角度对（2）的不合语法性做出判断，因为（3）的前4个词正是跟（2）完全相同的词序列。显然，后续词语的出现，可以使得（3）成为汉语中合法的句子。

词与词组合成**词组**，也称为**短语**（phrase），短语可以再跟短语组合成更大的短语，即可以嵌套。如果同类型的短语自我嵌套，就形成递归（recursive）结构。例如，“阿伦的同党的邻居的亲戚的孩子”，就是一个由多项名词组合的递归结构，名词组自我嵌套，从小的名词组形成大的名词组。尽管一般情况下句子长度有限，但人们普遍有一种语感，即无论多长的句子，还可以再在其上添加新的词语，使其增长。一方面，全体句子的数量似乎是无限多的，另一方面，一个句子的理论长度似乎也可以是无限长的。

为了刻画句子的这种嵌套、递归的组合特点，现代语言学普遍采用短语树的形式来表示句子的句法结构。图2.5所示是例2.2（1）的句法结构树的表示。在例2.2（1）的四个词中，第一个词“她”并不直接跟其后的“从”发生组合关系，而是后面三个词组成“从东京来”之后，才跟“她”组合成最终的整句“她从东京来”。树结构可以准确表达词语组合的先后顺序。

树是一种递归结构，树节点的子节点在语言学上一般称为“**直接成分**”（immediate constituent），从根节点S开始，每个直接成分都可以再分解为更多的直接成分，直到不能分解（即词）为止。

图2.5所示的树结构中每个节点要么是一个直接成分，要么是两个直接成分，这

体现了语言学分析句法结构时倡导的二分原则，即对于一个可分的语言单位，在分解其直接成分时，应尽可能一分为二。从句法系统的整体来看，以二分法得到直接成分，有两个优点。① 可以更好地反映词语组合的先后顺序差异。如果图2.5的树结构表示为S下面直接包括4个子节点，即对应句中的4个词，那就跟表层线性词序没有任何区别了，无法说明“从”是先跟“东京”组合而不是先跟“她”组合。② 可以使得直接成分在整个句法系统中的重用可能性最大化。在叶子节点相同的情况下，二分支结构的层数多于多分支结构，以二分方式划分得到的直接成分独立性更强（或者说依赖性更弱），更易于重用。

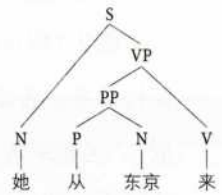


图2.5 句子结构树

图2.6所示是一棵不完整树，缺少根节点（以*表示），对应着例2.2（2）。通过图2.6和图2.5的差异，可以反映例2.2（2）和例2.2（1）的差异，即凡能表示为树结构的词序列是合语法的，凡不能表示为树结构的词序列是不合语法的。

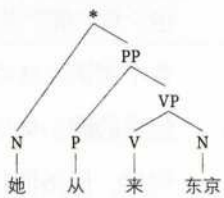


图2.6 不完整结构

要判断词组之间能否组合形成树结构，就涉及树节点的范畴划分问题。图2.6中PP代表介词性词组，在汉语的句法结构中，PP是向右组合（即前置型词组[①]），很少能向左组合，因此无法跟左边的“她”形成更大的单位。要构建一个语言的句法系统，理论上就需要对全部词语的组合可能性进行分析，归纳像PP这样的短语范畴，理清楚哪些范畴可以组合，哪些范畴不能组合。如果可以组合，以什么样的结构关系组合。

图2.5所示的层级树结构是语言系统普遍存在的结构形式。词的内部结构、语音系统中音节的内部结构，也都能采用树结构来表示。跟词汇系统和语音系统单位的树结构相比，句法系统中句法树的嵌套层级可以多得多，理论上达到无限，而词和音节的树结构只有很有限的嵌套。另外，句法组合模式的数量庞大，更需要在树节点上标记范畴以作区分。词语和音节的内部结构，往往只画出层级树图，并不需要在节点上标记范畴。

短语类和词类，是句法系统的基础类别。理论上，可以先定词类，再根据词类定

① 介词（preposition）的字面义就是“前置词”。英语和汉语中介词都是向右组合，比如：在实验室，in the lab，但组成介词词组PP（preposition phrase）后，汉语的介词词组仍然主要前置，向右跟其他词组组合，而英语的介词词组一般后置，向左跟其他词组组合，比如：他在实验室写程序，He was coding in the lab。

短语的类，例如，所有的短语都有一个中心，短语类就继承其中心词的词类。也可以先定短语的类，再根据词能出现在哪些类的短语中，不能出现在哪些类的短语中（即词的分布能力差异），区分出不同的词类。语言学者一般根据意义和分布两方面的标准，再系统地分析语料中词和短语的分布异同后，最终确定一个语言的基本语法范畴体系。

下面通过一个句子的结构树图来展示汉语的部分主要短语类和词类。

图2.7所示的句法结构树遵循二分原则分析整句的层次构造，其中包含11种短语结构关系，各结构关系都由两个直接成分构成，如“主谓结构”由主语项加谓语项构成，“述宾结构”由述语项加宾语项构成，“述补结构”由述语项加补语项构成等。这些结构之间的区分，母语者一般都有语感直觉。例如，“阿Q-吃饭、阿Q-喝酒”是主谓结构，“吃-茴香豆、喝-黄酒”是述宾结构，“回家-吃饭”是复谓结构，“吃-饱、喝-醉”是述补结构。在确定了基本短语结构关系的基础上，就可以进一步审核各个词语在这些结构中的分布，确定词类范畴，例如，名词就是经常处于主语、宾语位置的词，动词就是经常处于谓语、述语位置的词，形容词就是经常处于补语、谓语位置，但不能带宾语的词。图2.7例句中包含a（形容词）、n（名词）、v（动词）等主要实词，以及c（连词）、d（副词）、p（介词）、de（的）、le（了）等虚词。“的、了”在汉语语法学中常归入“助词”类，但从分布角度来说，这些助词缺乏共性，个性特点很强，分别独立出来，更能体现其分布特点。在词类范畴确定的基础上，可以由短语词的中心词进一步确定短语的范畴，如np（名词性短语）以n为中心词、vp（动

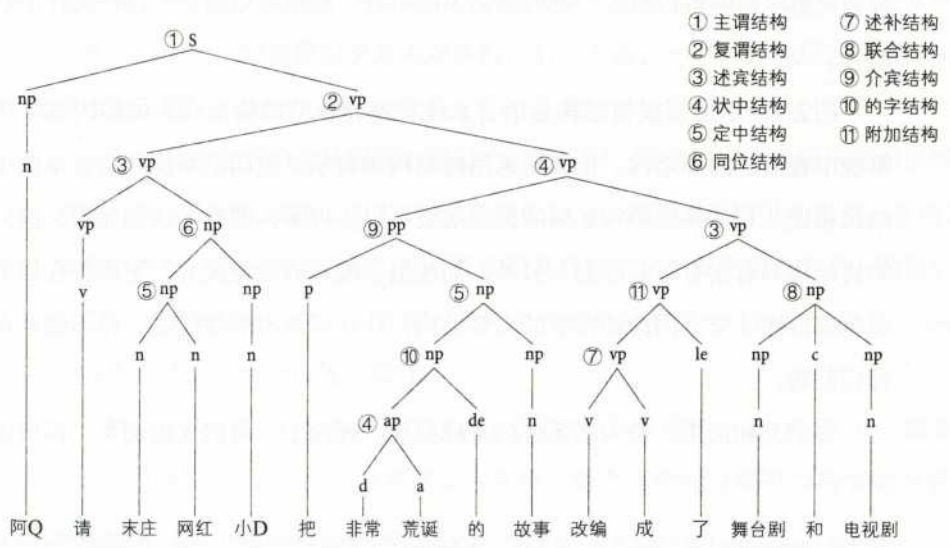


图2.7 汉语词类、短语类、结构关系示例

词性短语）以v为中心词、ap（形容词性短语）以a为中心词、pp（介词性短语）以p为中心词等。

现代语言学基本按照上述思路确定词和短语的类别，并给出短语的结构模式，表2.3是对图2.4中一些结构模式的描述。其中→表示其左边的符号可以分解为右边的符号，反映了树节点及其直接成分。一个语言的句法系统，就是像表2.3这样的短语结构模式的完整列表。

表2.3　汉语短语结构模式示例

短语结构模式	说明	实例
ap → d a	形容词短语由副词（d）加形容词（a）组成	非常 + 荒诞
np → np np	名词性短语由其自身递归组成	末庄网红 + 小D
vp → pp vp	动词性短语由介词短语加动词性短语组成	把非常……故事 + 改编成了……
vp → vp le	动词性短语由动词性短语加“了”组成	改编成 + 了

关于一个句子的句法结构树具体该如何画，一种语言中需要定义多少词类，多少短语类，存在许多不同的处理方式。不过，从共性的角度来说，无论句法结构树的具体表现如何不同，句法结构树都需要包含4个要素，才能较为全面地反映一个句子的结构面貌：**层次**、**关系**、**范畴**、**中心**。① 句子不是表面的词语线性序列，句子是有层次的，层次是句子的基本结构特征；② 各个层次上直接成分之间的组合存在不同的关系，句子中哪些词之间没有关系，哪些词之间有关系，具体是什么关系，也是句子的基本结构特征；③ 由于句子的无限性，句子的层次构造需要在对词和短语进行分类（范畴化）的基础上进行描述，才能做到以简驭繁，以有限表达无限；④ 两个直接成分组合时，往往有一个是中心，中心成分的范畴属性更能代表整个结构的范畴性质。这4个要素中，层次和关系，是最基础的两个要素，独立于语法理论体系，或者说，任何一个语法理论体系，在描述自然语言句子的构造时，必然要考虑这两个要素。范畴和中心，则是为了知识表示的概括性和便利性所做的进一步理论假设，是补充性的要素。不同的句法理论体系设计，对于句法结构“范畴”和句法结构“中心”的认识，可能存在比较大的差异，使得不同人画出来的句法结构树看上去不一样。即使在同一语法理论体系内部，不同学者也可能存在不同的认识。例如，汉语语法学界对于“这本书的出版”中的“出版”属于动词范畴还是名词范畴，就有争议。再例如，“弄-坏”中哪个词是结构中心？“同意去-的”中，是“同意去”是结构中心，还是“的”是结构中心？都存在不同看法。

除上面的短语结构树外，自然语言处理领域也常采用依存句法树模型来表示句

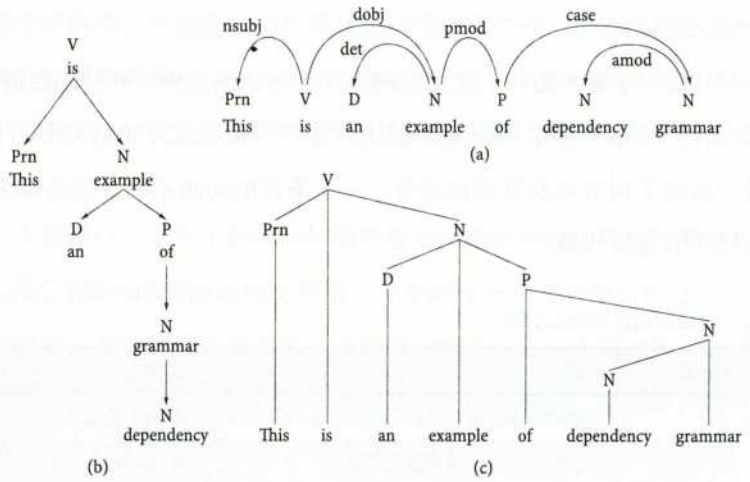


图2.8 句子结构的依存树表示示例

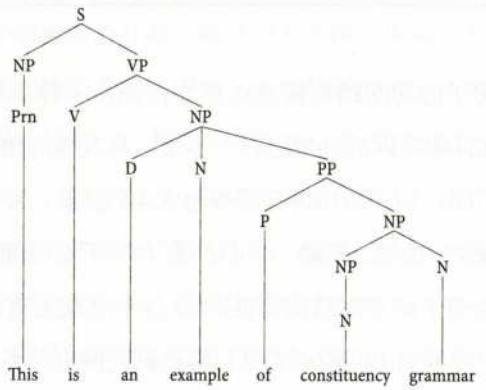


图2.9 句子结构的短语层级树表示示例

子的结构。图2.8和图2.9所示是依存句法结构树跟短语结构句法树的简单比较[1]。依存句法树模型最初的树图没有显式地描述词组（短语）范畴。如图2.8（a）、图2.8（b）所示①，但在后期的发展中，则逐渐采用图2.8（c）所示的树图表示，虽在树节点上未采用短语范畴标记，但实质上，图2.8（c）上树节点的词类标记等同于短语标记，跟图2.9所示的短语结构树达到完全相同的句法结构表达效力。从句法结构表示的四要素来看，早期的依存语法模型只关注“关系”和“中心”，后期则增加了“层次”和“范畴”，4个要素在句法树上都能加以呈现时，依存语法树跟短语结构语法树就等价了。

① 图2.8（a）依存关系弧上的标签是编者所加。nsubj代表名词性成分主语，dobj代表直接宾语，det代表限定关系，pmod代表介词性修饰成分，case代表介词引导关系，amod代表修饰关系。

2.2.4　语义系统

对于语义系统的作用，一种理解方式是，在句子结构作为输入时，负责输出句子的语义解释，这是句法系统和语义系统串行的模型。另一种理解方式是，语义系统和句法系统同时在解读句子的过程中起作用，二者并无先后顺序关系。在判断词的线性序列是否构成合法的句子，以及以何种结构方式构成句子时，并不仅仅是调用句法系统，而是已经并行地利用语义系统在进行分析审查工作了。

无论是上述哪一种看法，要清晰地揭示出语义系统的样貌，最直接也是最大的困难是，在基础概念层面，搞清楚“语义的本质是什么”；在知识表示层面，搞清楚“语义应该长什么样”。学术界提出过多种关于意义的理论，如指称论、意念论、行为–环境论、验证论、真值条件论、用法论、境况论[9]。这些有关意义的不同表述，无不是在尝试回答上面两个问题。虽触及意义的诸多方面，但仍难形成定论。

在关于意义本质的讨论中，图2.10所示的“语义三角”图影响广泛，其中A代表了语言符号形式，A的意义是通过B（心智、概念、思想）而指向C（世界上的所指物）。

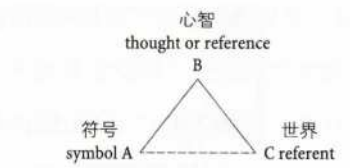


图2.10 语义三角图

把意义归结为指称，是很自然的想法，但也面临明显的困难，如“土豆”跟“马铃薯”、“花果山的美猴王”跟“大闹天宫的弼马温”、“周海婴的父亲”和“许广平的丈夫”指称的事物相同，但这些词和短语表达的意义是否相同呢？折中的答案可能既有同也有异。

语义分析的困难让20世纪上半叶的语言学家望而却步，绕道而行，甚至到20世纪50年代乔姆斯基开创转换生成语法之初，也将目光聚焦在语言形式层面的句法结构上，而避免过多探讨语义。但显然语义又是语言研究的目的所在，不朝着目标前进的研究即便走得再远，也很难彰显其价值。随着研究的深入和推进，现代语言学的各个流派更多地自觉将语义置于研究的中心位置加以关照，在对语义的结构化表示、语言系统不同层面语义表现的挖掘上，取得了一定进展，特别是将人和实际交际场景中的因素也纳入语义分析的范围，并最终独立出来，成为专门的语用学研究（参见2.2.5）。从某种意义上说，语义学和语用学的区别大致可以概括为，对句子语义的分析，前者着眼于内部成分的组合制约，后者着眼于外部环境因素的影响。一个由内求义，一个向外求义。在追寻意义之旅中，“**意义组合原则**”（principle of compositionality）体现了“由内求义”的思路，“**意义情境原则**”（principle of contextuality）体现了“向外求义”的思路。随着对意义研究的深入，后者的重要性似乎大有超过前者之势。“意义情境原则”不仅体现在“观词伴，知词义”的词汇层

面[5]，同样也可以拓展到句义。理解句子的语义，不仅需要看其组成部分的语义如何组合，还需要看句子的上下文，以及句子的使用场合、使用者等句子外的因素。这些都是语用学讨论的主题，本节主要从“组合性原则”的角度“向内求义”。

根据意义的组合性原则，句子的语义是词义和结构意义的综合。例如，“狗咬猫”跟“狗咬人”的意思不同，是因为句中有两个词“猫”跟“人”的意思不同。而“人咬狗”和“狗咬人”的意思不同，则是因为两个句子的结构不同——两句所含词语完全相同，但词序不同。

从人的一般认识来说，对于句子语义的通俗表示，主要有两种方式：一种是给出一个句子的其他同义形式。另一种是用回答问题的方式来表达对句子语义的理解。

下面两组句子展示了第一种释义方式。

例2.3	（1）阿伦**告诉**吴姐他**下岗**了	（2）阿伦**跟**吴姐**说**他**失业**了
例2.4	（1）阿伦是有时间谈恋爱的	（2）谈恋爱阿伦是有时间的

例2.3（1）的意思可以用例2.3（2）来说明。例2.3（2）跟例2.3（1）中所含词语不完全相同，但句义基本一样。可以说，例2.3(2）通过同义词语替换表示了例2.3（1）的句义，“失业”替换了“下岗”，“跟……说”替换了“告诉”，替换后不改变句义。例2.4（1）跟例2.4（2）两句的用词完全相同，但词序不同。例2.4（1）和例2.4（2）的句法结构有同义变换关系，保证了两句同义。在保持句型不变的情况下替换例2.4（1）句法结构中的具体词语，可以观察到替换前后两个句法结构之间的语义同一性。下面是跟例2.4同型的更多例句。

例2.5	a（1）阿伦是有资格拿奖学金的	b（1）拿奖学金阿伦是有资格的
	a（2）阿伦是有把握考第一的	b（2）考第一阿伦是有把握的
	a（3）阿伦是有办法发论文的	b（3）发论文阿伦是有办法的

上面这种方式也可以看作是**在符号系统内部寻找一个句子与其他句子形式之间的对应关系**，语义学研究中常以这种方式来分析同义句，同时也分化多义句。例如，例2.5中a、b两组句子构成同义关系，但实际上a组句子形式本身是多义的，并不是所有符合a组句子结构形式的实例，都可以同义变换为b组的句子结构形式。

例2.6	a（1）阿伦是有人陷害入狱的	b（1）陷害入狱阿伦是有人的
	a（2）阿伦是有大人物撑腰的	b（2）撑腰阿伦是有大人物的

例2.6跟例2.5的句子结构形式在词类序列层面看是一样的，a组例句的模式为：N_1+是+有+N_2+V+的，b组例句的模式为：V+N_1+是+有+N_2+的。例2.5中，a和b组句子的实例可以同义变换，例2.6中，a和b组句子的实例不能同义替换，例2.6中的b组句子形式都不是合法的汉语句子。稍加观察可以发现，例2.5和例2.6的差异，主要在于结构中N_1和V的语义关系不同：例2.5中，N_1和V都是主动关系，如“阿伦-谈恋爱、阿伦-考第一”；例2.6中，N_1和V则是被动关系，如“阿伦-被陷害入狱、阿伦-被撑腰”。这种隐性的词语间语义关系，跟词义一样，也是句义的一部分。通过句子的形式变换操作，可以揭示出这类语义。

给句子释义的另一种方式是回答问题。这种方式是**跳到了符号系统之外，试图在句子的符号系统和外部世界（包括物理世界和人的抽象概念世界）之间建立对应关系**。对于一般常见句子而言，句子传递的信息通常可以归纳为“五个W和一个H”，即Who did what to whom when where and how？（谁在何时何地以何方式对谁做了什么？）此外，还可以针对句子整体发问：这句话说的是真的吗？即判断句子所代表命题的真假。要回答后一个问题，需要从回答前面的“五个W和一个H”问题入手。需要把句子的符号跟外部世界的实体、动作行为、关系、性状等物理和心理对象物建立起映射关系。人的语义系统中预先就要将可能存在的关系存储起来[①]。对此当代语义学提出的主要方案是，通过描写动词的论元角色及其句法配置[②]，来建立句子中主要词语（实词）跟外部世界（概念世界）中各种对象物之间的关系。表2.4是这种**“论元结构”**语义知识表示法的示例。例2.7中的句子展示了表2.4中动词“咬”的各种语义角色在句中的不同句法配位。

表2.4 动词（“咬”）论元角色及其句法配置示意

词语	语义角色	句法配置模式
咬	施事：动物 受事：动物或具体事物 受事部件：身体部位 动量：数词+下|数词+口 工具：牙|嘴 结果：伤|破|碎|坏…… 时间：时间词 空间：方位处所词	施事＋咬＋受事 施事＋咬＋受事＋的＋受事部件 施事＋咬＋受事＋的＋受事部件＋动量 施事＋把＋受事＋咬＋结果 施事＋用＋工具＋咬＋受事 受事＋被＋施事＋咬 受事＋被＋施事＋用＋工具＋咬 ……

① 从某种意义上说，这相当于在构建语义系统的知识图谱。

② 这种模式当然也可以推广到形容词、名词等其他实词词类。描述对象不同，但内在的思路是一样的。有的学者就主张以名词为核心，将动词看作是围绕名词的语义角色，来描述词间关系。

例2.7 （1）吴姐的狗咬了阿伦的大腿两口。

（2）阿伦的大腿被吴姐的狗咬了两口。

（3）阿伦被吴姐的狗咬了大腿两口。

（4）？阿伦被吴姐的狗咬了两口大腿。

（5）*吴姐的狗把阿伦咬了两口大腿。

例2.7中“吴姐的狗”是“咬”的施事，“阿伦”是受事，“大腿”是受事的部件（身体部位）。“两口”是动作的量。例2.7（1）、2.7（2）、2.7（3）都是汉语中表达同样的语义角色关系可以选用的句法配位形式，例2.7（4）的可接受程度有一定疑问，打了？号。例2.7（5）在汉语中是不大能接受的句法配位形式，打了*号。

这种语义表征模式的假设是，可以在词汇层面，把词汇间各种可能的组合关系都预先表示出来，可以想象人脑中存储着一张巨大的词语语义关系表，在碰到实际句子的词序列时，在句法结构分析的同时，查词义关系表，将句子中的词语（主要是动词与名词），跟表中的角色及其约束条件相匹配，匹配成功的那些关系，就作为句子语义的表示输出了。

理论上，这样操作并无问题，但实际上存在的困难是词语间的组合关系本身几乎是无限的。构造一张有限的词义关系表，离真实句子中任意两个词之间的语义关系的分析，相距甚远。例如，动词论元结构语义表示中，并没有同时考虑动词与动词之间的语义关系。而下面的例子可以说明动词之间的语义联系对句义理解的重要性。

例2.8 北京梁思成故居被拆除，有关部门说是“维修性拆除”，针对这种明显矛盾的说法，有相声演员嘲讽道：应该将说这种话的人进行“保护性枪毙、治疗性活埋”。并调侃：以后街上有人持刀抢钱，得算是“理财性抢劫”。（来源：徐德亮相声作品）

例2.8是相声演员对不合理的拆除古建筑的现象进行讽刺的一段台词。其中凸显了4对动词之间的关系：“维修—拆除”“保护—枪毙”“治疗—活埋”“理财—抢劫”。这4对动词的共性是，前后两个词的词义中有矛盾的成分：前三对，是动词所代表动作行为的目的相反，最后一个，是动作行为的目的相同，但方式不同，一个合法，一个非法。这几对动词词义对比冲突尖锐，因而在相声台词中达到了很好的艺术效果。

除像例2.8展示的动词词义的对立关系外，动词之间的关系还跟事件的时间状态有关。

例2.9 （1）阿伦**后悔**学人工智能。——阿伦已经学了人工智能。

（2）阿伦**打算**学人工智能。——阿伦还没有学人工智能。

（3）阿伦**喜欢**学人工智能。——不清楚阿伦有没有学人工智能。

例2.9中“后悔、打算、喜欢”三个动词，都可以带小句宾语，不过，它们对后续小句宾语所表达事件的时间状态有不同影响。“后悔”蕴含了后续小句事件是已经发生的事情，“打算”则相反，蕴含后续小句事件是未然事件。“喜欢”两种状态均可。

以上讨论的语义现象基本是将句义归结到句中实词本身的语义及实词间组合语义，主要反映的是句子对应的命题的客观意义。还有的语义现象超出了这个范围。下面再看一些例子。

例2.10 （1）吴姐知道阿伦和小丁都是绍兴人。

（2）吴姐过生日，阿伦和小丁都送了礼物。

例2.11 （1）吴姐说：阿伦离开老家三天了。

（2）吴姐说：阿伦离开老家都三天了。

例2.10（1）中“都”的语义很虚，可以去掉，不影响整句的意思。例2.10（2）中“都”相当于“分别”，如果去掉，会造成句义差异：有“都”的情况下，例2.10（2）一般会理解为阿伦和小丁各自送了一份礼物给吴姐，因此是两份礼物；去掉“都”后，则更倾向于理解为阿伦和小丁合送了一份礼物。

例2.11（1）和例2.11（2）的基本意义（客观义）是一样的，即阿伦不在老家，且阿伦不在老家这一状态已持续三天。但例2.11（2）还多了一层意思，“都”用在这里，传递了一个信息，即说话人吴姐认为：三天是一个比较长的时间。这个意思是对客观事件“阿伦离开老家已经三天了”的一个“主观”评价，是**主观语义**，表达了说话人对一个事件的态度和评价。

例2.10和例2.11显示的“都”对句义的影响，跟前面实词对句义的影响不同，虚词对句子语义的影响，并不一定是通过虚词跟句中某个或某些词语发生直接联系来表示的，虚词对句义的影响，往往是作用于整句。

例2.12 （1）在昨天的“末庄之夜”晚会上，吴姐甚至**吻**了阿伦。

（2）在昨天的“末庄之夜”晚会上，吴姐甚至吻了**阿伦**。

例2.12（1）和例2.12（2）两句，书面上形式相同，口语中重音不同，例2.12（1）句重音在“吻”上，例2.12（2）句重音在“阿伦”上，两句适用的背景语境不同，整句的意思因而有别。两句的句义共性是：“吴姐吻阿伦”是极小概率事件，而这个极小概率事情竟然发生了；两句的句义差别是：例2.12（1）句强调吴姐对阿伦实施“吻”的行为，是让人意外的（暗示吴姐虽然跟阿伦有一定的关系，但两人的关系远未达到可以接吻的亲密程度）；例2.12（2）句强调吴姐吻的对象是阿伦，这是让人意外的（暗示吴姐最不可能吻的是阿伦，言下之意还包括吴姐可能吻了除阿伦之外的其他人）。

例2.12展示了**焦点**（focus）成分对句义的影响，尽管词语及语序完全相同，但通过重音的语音手段赋予句中不同词语以焦点身份，使得整句句义在基本命题义（字面义）相同之外，言外之意又有所不同。这一意义无法单纯通过词义的组合加以表达。

例2.13　（1）吴姐：小丁的前女友为什么跟小丁分手？
　　（2）阿伦：小丁啥时候有过前女友？

例2.13两句表面形式都是问句，但却构成了一对问答关系。吴姐的问题集中在句子谓语部分“为什么分手？”，阿伦的应答句却与这个问题没有直接的语义关联，而是对上句的主语部分“小丁的前女友”进行了反问：“小丁啥时候有过前女友”。显然，这个反问意味着上一句中包含一个语义“小丁有前女友”。这个语义是例2.13（1）句的预设义，不在句子的表层形式中，而是由句子主语“小丁的前女友”这个结构带来的。“小丁的前女友”是汉语中的定中结构（参见2.2.3），这个结构中的定语和中心语两项成分之间，有表示“存在（有）”的语义关系，如“我的儿子——我有儿子”“公司的制度——公司有制度”“香蕉皮——香蕉有皮”等。这是定语和中心语之间有领属关系义的定中结构都包含的语义，是结构意义的一部分，不是由词义本身组合来的意义。

例2.14　（1）你罚你的款，他违他的章。
　　（2）你走你的阳关道，他走他的独木桥。
　　（3）你说你的，他干他的。

例2.14的3个句子是同一个模式，每句都包含两个小句，句型相同，都是“代词+动词+代词+的+（名词）”模式，而且小句中两个代词必须同形。这个句式有一

个明显的意义：互不干涉、互不影响、各干各的。这个意义并不是句中任何一个词语带来的，而是附着在整个结构上的。任何像“罚款、违章、走、说、干、阳关道、独木桥”这样的动词和名词填入到这一模式中给动词和名词预留的空位上，整个句子就都会有“互不干涉、互不影响、各干各的”这样的意思，而不受具体所填入动词和名词的影响。

例2.14和例2.13展示的语义现象一样，都是结构带有语义，跟结构中的具体词语的词义无关。相对来说，例2.14的结构义更具体，而例2.13的结构义更为抽象。相应地，例2.14对其组成成分的形式要求也就更高（如结构同形、代词同形等约束），例2.13对其组成成分的要求相对宽泛。

2.2.5 语用系统

人类语言的语用系统亦是为完成传递信息这个总目标服务的。回顾一下2.1.2中图2.2的交际6要素，交际行为并不仅仅是符号形式本身，除了符号信息流的语义外，交际双方的认知状态（包括各自的已知信息、未知信息、共享信息等）、社会身份（性别相同或有别，地位平等或有差异等）、所处的时空环境等，都可能影响到语言表达形式的选择和语义解读。

例2.15 阿伦对吴姐说：我喜欢你做饭时的背影。

可能的意思1：阿伦认为：吴姐做饭的时候，很有魅力。

可能的意思2：阿伦对吴姐说恭维话，想追求吴姐。

可能的意思3：阿伦和吴姐已结婚多年，阿伦夸吴姐的目的是让吴姐去做饭。

例2.15的交际双方是阿伦和吴姐。意思1是阿伦所说句子的字面意义。意思2是当两人没有婚姻关系时阿伦说这个句子可能的意图。意思3是当两人已有婚姻关系时阿伦说这个句子可能的意图。一句话的字面意思可以由语义系统中有关词义、结构语义的组合得到，但在交际中，除字面意义之外，还会激活听话人对说话人表达意图的猜测，甚至第三方（听众或读者）对听说双方意图的猜测。显然，表达意图的判定，需要结合交际双方的社会关系和时空环境，才能确定。

从语言使用时的外部语用环境来看句子在交际中实际传递的信息、发挥的功能，就是调用语用系统对句子的语义做出更全面的解释。语用系统的本质是基于经验的逻辑推理。

语用分析并非自然语言处理的传统任务，然而，随着人机对话等研究的兴起，自

然语言处理领域也开始逐步关注语用分析，但是目前的研究还比较浅层。限于篇幅，这里不再对语用系统深入展开。

2.3 语言的歧义性与创造性

2.2节简要勾勒了自然语言系统的基本面貌。这一节介绍自然语言的歧义性和创造性。这两方面，也正是需要对语言进行深度理解的自然语言处理系统面临的主要挑战。

2.3.1 歧义性

一种语言语法系统里的错综复杂和精细奥妙之处往往在歧义现象里得到反映[11]。自然语言语素之上的各级单位，既存在一个形式对应一个意义的简单情况，也存在一个形式对应多个意义的复杂情况，即歧义现象。前者如“自行车”，只有一种意思，指一种靠人蹬踏骑行的两轮交通工具。后者如“便衣”，就有两个意思：一是指（不穿职业装而身着便衣工作的）警察或军人；二是指跟职业装（通常是警察或武装部队制服）相对的普通服装。“便衣”的这两个意思分属服装和人的职业身份两个截然不同的范畴，会对句义的理解造成明显的影响。

例2.16 （1）巷子里人声嘈杂，既有军警，也有**便衣**，簇拥着一个矮胖子。
（2）下班后，他换上**便衣**出了门，直奔火车站。
（3）少帅的**便衣**并没有能帮助他脱险。

例2.16（1）中的“便衣”指人，例2.16（2）中指衣服，例2.16（3）中则不清楚，可能指人，也可能指衣服。只是从一般常理来说，例2.16（3）中“便衣”指人的可能性大，但在特定语境中，也有可能指衣服。例2.16（3）因无法消除其中多义词“便衣”的歧义，整个句子也就有了歧义。

除了多义词，语言系统的歧义性更多地表现在词的组合使用中，即使单义词，在组合时也可能会在语言系统的各个层次上发生歧义现象。

上文在讨论句法系统时已经指出，语言单位在组合时，表面是线性接续关系（仅单向组合），但实质上是层次结构关系（可双向组合）。因此，从线性层面看，一个语言单位，就有向前还是向后组合的问题。这可以称为语言单位组合中的**边界歧义**。

例2.17　30日，港务区管委会证实了此事："我们这块地是有项目的地，是要拆迁的。"

例2.18　春节红包大战又来，天上掉20亿网民平均能捡3块多。

例2.19　阿伦住在一个有很多富人的小区。

例2.20　吴姐工作的地方有很多富人的房子。

例2.17中"有项目的地"的内部成分是："有+项目+的+地"这四个词，而不是"有+项+目的地"这三个词。"目的地"涉及词语划分的边界问题。

例2.18中的"20亿网民"不是一个结构体，结构边界是"天上掉20亿"+"网民平均能捡3块多"。

例2.19和例2.20中有一个词类序列相同的词串"有+很多+n+的+n"，但例2.19的结构边界划分是："有很多富人的+小区"，例2.20的结构边界划分是："有+很多富人的房子"。

在结构边界已经确定的情况下，语言成分的组合还可能发生**句法结构关系歧义**、**语义关系歧义**、**语义指向歧义**等不同类型的**关系歧义**。例如：

例2.21　阿伦叫吴姐去了。

例2.22　没想到阿伦离开末庄后，最担心的是吴姐。

例2.23　（1）他老爹从小教育他就是用的他爷爷的事迹。
（2）他老爹从小就跟着他爷爷的拜把兄弟学木匠。

例2.21是句法结构关系歧义。"叫吴姐去"是两个动词性成分连用的复谓结构，但具体又可分化为两种结构关系，一种是递系复谓，"叫吴姐+去"="叫吴姐 并且 吴姐去"，另一种是倒置的连动式复谓结构，"叫吴姐+去"="去+叫吴姐"。前一种结构关系下，"叫"的语义相当于"让"，后一种结构关系下，"叫"的语义相当于"找"。

例2.22是语义关系歧义。"最担心的"是一个"vp+的"结构，结构里的vp主要

动词是“担心”，是一个二元动词，需要两个名词分别充任施事和受事角色才能使句子的基本意义明确、完整（X担心Y），而例2.22中的“vp+的”结构里，“担心”前后的两个角色位置都是空位状态，这就造成有两种填位的可能性：阿伦最担心的是吴姐或最担心阿伦的是吴姐，两种填位，在句法结构层面都可以接受。这样，在语义关系上，就分别对应两种语义关系配置：阿伦最担心吴姐或吴姐最担心阿伦。

例2.23是语义指向歧义，即词语的远距离语义关系歧义。例2.23（1）和例2.23（2）中的“从小”在线性位置上相同，但例2.23（1）中“从小”指“他”，例2.23（2）中“从小”指“他老爹”。

有的句子表现的歧义不是句子本身字面意义层面的多义，而是言外之意的多义，需要根据句子以外情境来推断意思。

例2.24 （1）关于酒店的装修，阿伦有意见。

（2）安德森对小丁说，阿伦正在说你呢。

（3）安德森指着阿伦跟小丁，说：这两人真是没话说。

例2.24（1）的“有意见”可以是正面的意见，也可以是负面的意见。例2.24（2）的“说你”可以是中性的谈论，也可以是“说坏话”。例2.24（3）的“真是没话说”的意思更复杂，可以指阿伦跟小丁关系不好，两人无话可说；也可以指两人都是好哥们、讲义气；也可以指两人之间关系很紧密，是一个小团体。

下面例2.25（1）的歧义在书面上要依赖上下文才能判别。在口语中，依靠句中不同词语上的重音标记，可以自然地分化歧义。

例2.25 （1）这道题你都不会做。

（2）这道题很容易，你不会做这道题，别的题就更别指望会做了。

（3）这道题很难，你不会做这道题，别人就更不可能会做了。

例2.25（1）表达例2.25（2）的意思时，口语中重读“这道题”；例2.25（1）表达例2.25（3）的意思时，重读“你”。书面上重音信息丢失，如果没有上下文，就看不出例2.25（1）到底表达例2.25（2）的意思还是表达例2.25（3）的意思。因此，在书面上，例2.25（1）是有歧义的。在口语中，例2.25（1）往往通过重音不同而消解了歧义。

2.3.2　创造性

人类语言符号来自约定俗成，它一方面形成惯例，一方面又在不断创新。为新生事物命名（约定）是最常见的创造。对一般人来说，也许像诗歌那样的语言创造让人印象深刻，更容易引起注意，例如“城市是几百万人一起孤独生活的一个地方”（梭罗），“夸张是发了脾气的真理”（纪伯伦）。但事实上，每个人的话语中都蕴藏着无限的创造可能性。创新是自然语言系统的本性之一，它源于人们交际中三个普遍的动机，甚至在无意识的自然状态下发生。这三个动机一是节省编码；二是吸引注意（诗圣杜甫所说的“语不惊人死不休”）；三是游戏娱乐。这些动机有时是浑然一体的。

在已经存在一种“形式—意义”对应关系的表达手段情况下，说话者出于**节省编码**的目的（同时也可以伴随着其他社会目的，如避免直白，隐藏意图等），可以对原有表达形式中的部分符号进行删减，从而形成新的表达形式。

例2.26　记者称阿伦是自愿表态生活幸福的，全体村民都不存在“被幸福”的情况。

例2.26中的“被幸福”是近十年来汉语中新兴的一种表达格式。汉语语法系统中原本就有的“被+v”结构要求“被”后的动词是及物动词，如“被批评教育、被开除、被发现”等。但在网络语境中，有人创造性地使用“被自杀”表示一个人不是真的自杀（怀疑是被杀害），而被人说成是自杀。之后，出现了许多新型的“被+v”结构用法，如“被及格、被就业”等，不及物动词“及格、就业”都进入到这个结构中，打破了原先“被”字结构排斥不及物动词的约束。由于这个结构有强烈的表达效果，越来越多的使用者逐渐把“被”后的词语范畴从动词扩展到形容词，再到名词和数词，形成“被+X”的新兴表达形式（X代表范畴泛化）。语义通常是表示X并非真实情况，而是被人说成了X这种情况。显然，这样的意思，用原有的结构去表示，编码很长，而用“被+X”格式表达，编码简短，表达效果突出，带有调侃、讽刺等意味，容易引起人们关注。

语言系统中实际上还有很多类似的因省略而造成的新组合形式，在长期使用中，已经由“新”转“旧”，不被人注意了。如“参与意见、催稿子、敲了几个**字**”等，这些“动+名”的组合，已经是日常语言中常用的说法，但仔细分析不难发现，组合中动词跟名词的语义并不是直接发生关系的，这些“动+名”组合是编码更长的“动$_1$+名$_1$+动$_2$+名$_2$”组合的缩略形式：

例2.27 （1）参与意见=参与+讨论+提出+意见

（2）催稿子=催+人+交+稿子

（3）敲字=敲+键盘+输入+字

例2.27这些例子由两个动词词组连用的复谓结构，“压缩”成一个动词词组的述宾结构，“参与意见、催稿子、敲字”等，成了跟普通的“参与基层工作、写稿子、敲桌子”形式上完全相同的述宾结构。

值得注意的是，语言系统的创造性在带来新的表达形式的同时，增加了新形式与旧形式发生歧义的可能性。例如“被升职了”“被涨工资了”，在新兴“被动”表虚假描述义，跟原有的“被+及物动词”表达真的事件被动关系之间，就可能存在歧义。从这个角度说，语言系统的创新性和歧义性，是高度相关的。这一点，在旧形新义这种创新类型上，表现得更为直接。下面是一个旧形新义的例子。

新媒体标题中常用的一种表达格式“有一种X叫Y”就是利用原有形式，通过隐喻和转喻的认知机制，增加新义，使“有一种X叫Y”格式产生了一种创新用法。

例2.28 （1）有一种毒药叫砒霜。

（2）有一种毒药叫成功。

（3）有一种爱叫放手。

（4）有一种误差叫数据造假。

隐喻是基于相似性在概念范畴之间建立联系，**转喻**是基于相关性在概念范畴之间建立联系。例2.28（1）是“有一种X叫Y”的原有普通用法，表达一种认识，常用于介绍、说明性语境，传达客观信息：Y是X的下位概念，二者构成逻辑上的所谓“种属”关系。例2.28（2）和（3）中的X和Y显然不再是这种关系。例2.28（2）中“成功”让人渴望甚至不择手段地去追求，获得了成功之后，又可能让人得意忘形，从高峰跌入深谷。由此建立了“成功”跟“毒药”之间的相似性：成功和毒药，都让人成瘾，让人走向毁灭。于是，借用“有一种X叫Y”这种格式，把“成功”归入“毒药”的“下位”范畴，就表达了说话者的一个新观点。从这个角度说，可以把这个格式命名为“报道新知”格式。例2.28（3）中“放手”意味着失去，失去对方，但目的却是让对方获得幸福，通过放手这种方式，表达更深沉的爱，由此建立的两个范畴之间的新的相关性，使例2.28（3）产生了“报道新知”的表达效果。例2.28（4）的“数据造假”和“误差”也有明显的相似性，二者都是表面数据与真实数据不一

致，因而也可以像例2.28（2）那样，以隐喻机制进入这一格式，表达新义。

从认知图式（schema）的角度看，“有一种X叫Y”从普通的表示上下位关系（Y是一种X）的表义格式，发展出表示等同关系（Y=X）的新义，达到“报道（主观）新知”的效果，是基于同一图式，从不同的角度解读实现的。对于X和Y的关系，如图2.11所示，可以有两种解读方式：① 静态集合解读，Y是X中的一个元素；② 动态聚焦解读，远看是X，近看实则是Y，即X和Y是一个事物的“表（表象）—里（实质）”关系。从说话人的角度讲，用这个格式表达，还传递出一个信息，即“一般人说的Y实质上是X”，这个格式中X和Y两个范畴的关系，形成了“名义上有别，而实质上相同”的关系。“叫”是命名动词，因此，“有一种X叫Y”也可以说是一种命名构式。它表达了一种“说法”。这种说法的表达意图在于：说话人针对已有的成见，要提出一个新观点，确认关于X和Y之间的新关系。在这个格式产生出新义的开始阶段，使用者还兼顾旧义“Y是一种X”和新义“Y=X”，例如例2.28（2）和（3），两种解读都可以成立，“成功是一种毒药”“成功=毒药”“放手是一种爱”“放手=爱”，随着使用的增多，这个格式的语义功能可以发生进一步的固化，即只强调新义，而不兼顾旧义。而例2.28（4），语义是强调：误差（表象）其实就是数据造假（实质），而不再兼顾旧义“数据造假是一种误差”。

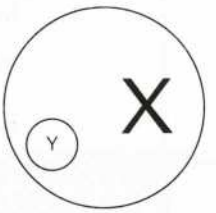


图2.11 X和Y的关系示意

语言在传递信息的同时，还可以对编码形式本身进行改造，伴随游戏和娱乐性质。例如下面这些虚拟的微信聊天场景，都是语言（文字）游戏的例子。

例2.29　阿伦：我做的菜好吃吗？

吴姐：嗯，厨（tai）艺（nan）不（chi）错（le）。

例2.30　小丁：你还好吗？

阿伦：~~sad broken defeated crushed lonely~~，我很好。

例2.31　安德森：群主，公司新版主页过两天发布，帮写个预告呗。

小丁：本网页正在~~积极破坏~~建设中，欢迎访客~~狠狠拍砖~~提出宝贵意见和建议。

例2.32 7H15 M3554G3 53RV35 7O PR0V3 H0W 0UR M1ND5 C4N D0 4M4Z1NG 7H1NG5! 1MPR3551V3 7H1NG5! 1N 7H3 B3G1NN1NG 17 WA5 H4RD BU7 N0W, 0N 7H15 LIN3 Y0UR M1ND 1S R34D1NG 17 4U70M471C4LLY W17H 0U7 3V3N 7H1NK1NG 4B0U7 17, B3 PROUD! 0NLY C3R741N P30PL3 C4N R3AD 7H15. PL3453 F0RW4RD 1F U C4N R34D 7H15.

例2.33 研表究明，汉字序顺并不定一影阅响读。比如当你看完这句话后，才发这现里的字全是都乱的。

例2.29 ~ 例2.31打破了口语中语音流的线性约束，在书面上通过加注或删除的方式，形成了符号穿插接续的形式，使得表面的一句话中实际上包含两句话，并且两句话语义相对立，形成冲突，故意制造出跟掩耳盗铃相似的表达效果。

例2.32用数字代替形近的字母组成英语单词，如“7HI5”实为“THIS”，“M3554G3”实为“MESSAGE”，等等。例2.33打乱了正常的汉字顺序，如“研究表明”写成“研表究明”。这些字符及其序列，都是新的“形-义”配对，是对语言系统已有的编码模式的创新。

2.4 语言知识资源

计算语言学和自然语言处理技术的发展，对推动语言学研究从理论分析扩展到语言工程资源建设，起到了很大的作用。区别于面向人的语言学研究，面向计算机的语言研究要求将语言学的理论研究成果转化为形式化和大规模数据化的语言知识资源。

语言知识资源目前主要包括**语言学知识库**和**标注语料库**两种形式。前者是记录语言学专家知识的数据库，主要是刻画词汇层的句法和语义知识，对具体语言的词汇表中的词语，按照句法或语义知识表示规范，逐条进行信息标识，从理论上讲，要求对词语的**全部可能用法**进行描述，一般可以把这种描述称为**type（类型）层的知识表示**。后者是对真实语料文本中的各级语言单位（词、词组、句子等），根据某种语法语义理论体系，标注语言成分在实际使用中表现出来的句法和语义属性特征，这是对词语的**每一次具体用例**的性质进行描写，一般可以把这种描述称为**token（实例）层的知识表示**。标注语料库可以供机器学习作为训练数据或测试数据使用，也可以为语言学的

定量研究提供支持。这两类资源通常都是独立于具体应用程序的，理论上可以供不同应用场景下的自然语言处理系统使用①。

知识库对词语句法信息的描写，就是描述一个词语跟其他词语发生句法结构组合关系的可能性[10]。表2.5是对汉语三个动词“交往、郊游、浇”的部分语法信息的描述。

表2.5 汉语词语语法信息表示示例

词语	有___	___名	名___	___宾	时态	重叠	V-V	……
交往	+	+	+	-	着了过	-	-	……
郊游	-	+	-	-	了过	-	-	……
浇	-	-	-	+	着了过	+	+	……

表2.4中“有__”描述动词是否可以跟“有”组合为合法结构，“有交往”可以，但“有郊游、有浇”在汉语普通话中均不成立。“___名”描述动词是否可以直接修饰名词，“交往、郊游”可以（如“交往时间、郊游地点”），因此这项特征的取值标记为+。“浇”不能直接修饰名词，标记为-。“名__”描述动词是否可以直接受名词修饰，“交往”可以（如“理性交往”），取值为+，“郊游、浇”不行，取值为-。“___宾”描述动词是否能带宾语，“交往、郊游”均为不及物动词，不能带宾语，取值为-。“浇”是及物动词，能带宾语，取值为+。“时态”描述动词是否能后附汉语的时态助词“着、了、过”，“交往、浇”全部可以，取值为“着了过”，“郊游”不能后附“着”，取值为“了过”。“重叠”描述一个动词是否有重叠形式，“v-v”描述动词是否有“v-v”形式，以上两项只有“浇”可以（如“浇浇水”“浇一浇水”），取值为+。“交往、郊游”这两项的取值均为-。

理论上讲，语法知识库的目标是描写一个词语的全部分布位置。但目前的知识库很难做到这一点，一般只能像表2.4所展示的那样，按照两两组合的模式，来评估一个词语的分布可能性。这有点像二元模型（bigram），前后接续的两个元素之间是有约束的，但如果超过两个元素，第一个元素就无法对第三个元素有制约作用，只能通过先影响第二个元素再间接影响第三个元素。

① 还有一些资源是跟句法语义分析程序（过程性系统）紧密绑定的，因而不大容易独立用在其他的自然语言处理系统中。例如法国巴黎第七大学开发的INTEX句法语义分析系统就内置了基于语言学家Maruice Gross的词汇语法理论的词典和语法。德国人工智能研究中心（DFKI）、挪威奥斯陆大学、美国斯坦福大学CSLI语言工程实验室等多家单位基于HPSG语法理论开发的DELPHI-IN深度语言处理系统中，也包含了HPSG语法资源ERG（英语资源语法）。

例2.34 （1）也许 抽烟 的 不 怕 烟味
（2）一直 抽烟 的 不 怕 烟味

例2.34中（1）、（2）的前3个词是同词类序列：副词+动词+的，但是，从结构和语音停顿上讲，两句有区别，例2.34（1）中“抽烟的”先跟“不怕烟味”组成一个结构体，然后才跟“也许”组合；例2.34（2）中“一直”跟“抽烟”先组成一个结构体，然后再跟“的”组合，整体充当主语，再跟“不怕烟味”（充当谓语）组合成句。例2.34（1）中“也许”后明显停顿（长于该句其他词间停顿），例2.34（2）中“抽烟的”后明显停顿（长于该句其他词中间停顿）。从词的分布能力描述角度讲，汉语的知识中包括一条：“也许+抽烟+的”三个词项组合，一般不构成一个结构体，而“一直+抽烟+的”三个词项组合，一般构成一个结构体，这个知识涉及同时考虑三个词项共现时的分布能力描述，而当前的语法知识信息描述，一般都建立在语法单位两项组合的框架基础上，离理想中的“全面描述一个语言单位的分布能力”，包括能分布在哪些位置，以及不能分布在哪些位置，还有相当大的差距。前文已经分析过，自然语言句子的结构是层级性的树状结构，因此，描写一个词的分布特征，从根本上讲，应该在树状结构的框架下描述，而目前的语言知识资源中对词语分布特征的描述，都是基于二元组合来描述的，或者说，试图通过在最小二叉树的结构框架下描述一个词语的全部组合能力，来间接地预测一个词在可能的句子树结构上的分布状况，这种方法只能反映词语的部分分布特征信息，无法刻画词语的全部分布信息，反映词语用法的全貌。

在知识库中对词语的语义信息进行描写，概括而言是描述一个词语可能跟哪些词语发生语义上的联系，以及以何种关系发生联系，如上下位、部分-整体、同义反义、施事、受事、工具、处所等。表2.6是一些词语的语义信息描述示例。

表2.6 语义信息示例

词语	概念编号	语义描述
打	015492	weave|辫编
打	017144	exercise|锻炼,sport|体育
打对折	017317	subtract|削减,patient=price|价格,commercial|商,（range|幅度=50%）
医生	160648	human|人：HostOf={Occupation|职位},domain={medical|医},{doctor|医治：agent={~}}
医院	160682	InstitutePlace|场所：domain={medical|医},{doctor|医治：content={disease|疾病},location={~}}
患者	063820	human|人：domain={medical|医},{SufferFrom|罹患：experiencer={~}},{doctor|医治：patient={~}}

续表

词语	概念编号	语义描述
药费	158761	expenditure|费用：{buy|买：cost={~},possession={medicine|药物}}
车前子	021705	FlowerGrass|花草：MaterialOf={medicine|药物}

一个多义词的不同词义用不同的概念编号表示。打_015492代表了“打毛衣”中的“打”，它属于“辫编”这个概念语义范畴。打_017144代表了“打篮球、打太极拳”中的“打”，它属于“锻炼”和“体育”两个概念语义范畴。“打对折”是“削减”的同一范畴，同时“打对折”跟“价格”构成受事关系（或者说，“价格”是“打对折”的受事角色），“打对折”隐含了一个“幅度”角色，该角色的取值是50%。表2.5中的“医生、医院、患者、药费、车前子”等词语，通过语义特征值的定义方式，实际上描述了词语间的语义关系，例如，“医生”是“医治”的施事（agent），“医院”是“医治”的场所（location），“患者”是“医治”的受事（patient），“药费”是购买“药物”的成本（cost），“车前子”是一种“药物”。通过这种语义特征值的定义方式，刻画了相关概念（词）间的各种语义关系，可以形成概念关系网络图。

下面以动词语义角色标注为例，介绍标注语料库的通常做法。

例如英语中表达Revenge（复仇）事件的句子通常共享以下5个语义角色，构成所谓的“复仇”语义框架。在该框架下，各元素之间的语义联系体现在表2.7所示的角色定义中。

表2.7 Revenge（复仇）语义框架元素定义表

框架元素	定义
Agent（复仇者）	惩罚**复仇对象**，使其为之前的**伤害行为**付出代价
Injured_Party（之前的受害者）	受到过**复仇对象**的伤害
Injury（之前的伤害行为）	**复仇对象**实施的动作行为
Punishment（复仇举措/惩罚方式）	**复仇者**实施的动作行为
Offender（复仇对象=之前的伤害者）	实施**伤害行为**

跟表达“复仇”事件有关的词语有16个：avenge.v, avenger.n, get back at.v, get_even.v, retaliate.v, retaliation.n, retribution.n, retributive.a, retributory.a, revenge.n, revenge.v, revengeful.a, revenger.n, vengeance.n, vengeful.a, vindictive.a。其中除一般的词语外，还包括像“get back at”“get even”这样的动词短语。这些词语中有5个动词（以.v标记），6个名词（以.n标记），5个形容词（以.a标记），这些词出现在

句子中时就激活“复仇”事件框架。对句义的理解，需要将该框架的语义角色（框架元素）跟实际句子中的成分对应起来。

下面是含有“复仇”动词avenge、get even的三个句子及其框架元素标注。

例2.35 （1）

Hook	*tries to avenge*	himself	on Peter Pan	by becoming a second and better father.
Avenger		**Injured_party**	**Offender**	**Punishment**

（2）

Yesterday	the Cowboys	*avenged*	their only defeat of the season	by beating Philadelphia Eagles 20–10.
	Avenger		**Injury**	**Punishment**

（3）

Ethel	eventually	*got even*	with Mildred	for the insult to Ethel 's family.
Avenger			**Offender**	**Injury**

通过这种标注句例，可以观察动词在实际语料中的角色分布规律，例如“复仇”框架中表达“惩罚”（Punishment）义的语言形式通常是“by”引导的介词短语，表达“复仇对象”（Offender）的语言形式一般是“on、with”引导的介词短语。英语中“伤害行为（Injury）可以直接跟在复仇动词之后，如例2.35（2）句所示。对应到汉语的复仇义动词，则没有这种句法分布形式，汉语不能说“复仇他们上个赛季的失利”。汉语中通常要采用例2.35（3）所示的句型，将伤害行为放在介词“为”之后引出，并将整个介词短语放在复仇动词之前，说成“为他们上个赛季的失利复仇”。该框架中的“复仇对象角色”，在英语中和汉语中，都需要通过介词引出，如例2.35（1）用on引出，例2.35（3）用with引出，汉语中一般用“向”引出。这是英语和汉语的相同点。可见，利用框架元素的描述方式，有利于比较语言之间在实践语义表达的形式手段方面的异同。

一般来说，语言学知识库是对相对成熟的语言学研究成果的数据化表示，是比较可靠的知识，可以反映语言系统稳定、规范的一面。标注语料库则既包括将相对成熟的语言学研究成果用于实际语料的标注，也包括将还在探索中的不太成熟的语言研究用于实际语料的标注，实际语料来源广泛，语体多样，可以体现语言系统生动、变化的一面。理想而言，这两种语言知识资源定位不同，作用互补，可以形成良性的互动，互相促进，交替提升资源的质量和规模。不过，要在实践中实现这一理想的语言知识资源生态环境，还有许多具体工作要做，包括语言知识表示的合理设计，数据交换的规范、接口等，既有理论层面的问题，也有工程技术层面，乃至知识产权方面的

现实问题。这些因素对语言知识资源总体质量和规模的提升构成了很大的挑战。

语言知识资源，无论是知识库，还是标注语料库，都是语言单位已有用法特征的记录，反映的是词语用法的部分信息（样本），而无法做到反映全部信息（总体）。对此或许可以说：当前的语言知识资源是关于语言系统的外延性的知识记录，而非内涵定义式描述。目前能够触摸到的，始终是言语，这是人们借由了解语言本身奥秘的唯一途径。

2.5 延伸阅读

对于非语言学专业背景的读者和学习者，下面三本书对了解语言学的概貌，可以起到初窥门径之效。

PARKER F, RILEY K. Linguistics for Non-linguists[M]. Boston：Taylor & Francis Ltd., 1986.

FROMKIN V, RODMAN R, HYAMS N. An Introduction to Language[M]. 7th ed. 北京：北京大学出版社，2004.

TYERS F M, BENDER E M. Linguistic fundamentals for natural language processing：100 essentials from morphology and syntax[M]. New York：Kluwer Academic Publishers, 2014.

关于如何将语言学知识、语言资源构建和机器学习任务联系起来的系统论述，可以参考：

PUSTEJOVSKY J, STUBBS A. Natural Language Annotation for Machine Learning[M]. 南京：东南大学出版社，2013.

习题

1. 请举出至少5个例子，说明语言系统产生新词和旧词产生新义的不同方式。
2. 在下面画线处填入合适的助词，并说明助词在句中的语法功

能，对句子句法结构的影响。

A.的　B.地　C.得

（1）最终他们三个头脑清醒_____完成了任务。

（2）狸跟狐相比，个头小_____多。

（3）他们打_____那个人跳下了悬崖。

（4）他高兴_____跳了起来。

3. 请解释为什么“阿Q有美元”是“阿Q拥有一定数量的美元货币”的意思，而“阿Q有钱”除了“阿Q拥有钱”，还会有“阿Q钱很多”的意思。

4. 请从语言学角度分析下面的三段论推理的逻辑错误原因是什么。

（1）**鲁迅的书**不可能在一天之内全部读完。

（2）《祝福》是**鲁迅的书**。

（3）《祝福》不可能在一天之内全部读完。

5. 下面是一句广告语，请再找一些类似这种广告语的例子，分析自然语言在词语组合方面的创造性。

某大学手语社团广告：看见你的声音

参考文献

[1] COVINGTON M. A fundamental algorithm for dependency parsing[C]//Proceedings of the 39th Annual ACM Southeast Conference, 2001.

[2] DRYER M S, HASPELMATH M. The World Atlas of Language Structures Online[M]. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013.

[3] GRICE H P. Logic and Conversation//COLE P, MORGAN J L. Speech Acts[M]. New York: Academic Press, 1975.

[4] JAKOBSON R. Linguistics and Poetics//SEBEOK T A. Style In Language[M]. Cambridge Massachusetts: MIT Press, 1960.

[5] FIRTH J R. A synopsis of linguistic theory 1930—1955// Special Volume of the Philological Society[M]. Oxford: Oxford University Press, 1957.

[6] GEOFFREY L. Principle of Pragmatics[M]. London: Longman, 1983.

[7] SPERBER D, WILSON D. 关联：交际与认知[M]. 2版. 蒋严，译. 北京：中国社会科学出版社，2008.

[8] TOMASELLO M. 人类沟通的起源[M]. 蔡雅菁，译. 北京：商务印书馆，2012.

[9] 徐烈炯. 语义学[M]. 北京：语文出版社，1990.

[10] 俞士汶，朱学锋，王惠，等. 现代汉语语法信息词典详解[M]. 2版. 北京：清华大学出版社，2003.

[11] 朱德熙. 汉语句法中的歧义现象[J]. 中国语文，1980, 2：21-27.