计算语言学语法理论(grammatical theory of computational linguistics)

关于自然语言语法知识的系统的形式化论述。可以分为基于短语结构的语法和基于词间关系的语法两大类。前者认为句子由短语组成，短语由词组成，语法就是在区分短语和词的不同范畴基础上描述范畴间的组合规则，代表性的语法理论有**上下文无关语法**；后者认为句子由词直接组成，无需经过短语这个中间层次，语法就是描述词与词之间的支配与依存关系，代表性的语法理论有**依存语法**。这两大类语法理论都可以通过树图形式来呈现句子的结构，但具体的呈现方式有所不同。以“张三丢了一双鞋”为例，基于**上下文无关语法**对句子结构的描述如图 1 所示；基于**依存语法**对句子结构的描述如图 2 所示：

图 1

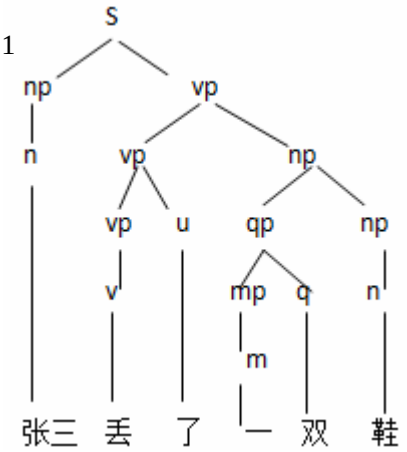

图 2

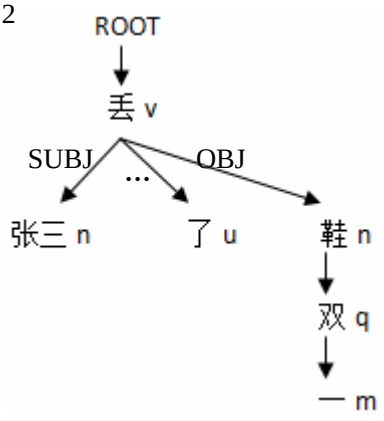

图 1 中 S 代表句子， np，vp 等分别代表名词性短语，动词性短语等短语范畴。n、v 等分别代表名词、动词等词类范畴。句子的树状结构规定了词语间的组合顺序，以及每个组合的内部语法关系，比如组成 S 的 np 与 vp 之间是主谓结构关系。图 2 中 ROOT 代表句子的“根”(或中心)，由谓语动词“丢”充当，动词支配主语“张三”、时态成分“了”、宾语“鞋”。句子的树状结构规定了句中词语间的支配与依存关系，箭头上方的节点是支配词语，下方的节点是依存词语。箭头线上还可以进一步标记词语间依存关系的不同类型。比如“SUBJ”表示主语-动词依存关系，“OBJ”表示宾语-动词依存关系，等等。

由于自然语言单位组合的复杂多样性，在计算语言学理论研究和自然语言处理的实践过程中，基于上述两大类基础语法理论框架，发展出了许多侧重点不同的语法理论模型。其中有代表性的语法理论模型包括：将短语结构区分为深层结构和表层结构来揭示不同句法结构之间关联的**转换生成语法**；引入特征结构与合一约束等形式表达机制的**词汇功能语法**、**功能合一语法**；强调短语结构内部中心成分作用的**中心语驱动短语结构语法**、**广义短语结构语法**、**X-bar 语法**；可以直接用 Prolog 程序语言来书写规则的**限定子句语法**；以树结构的替换和拼接操作代替短语两两组合方式来描述语言单位组合规律的**树邻接语法**；在短语结构语法基础上增加规则概率信息的**概率上下文无关语法**；以约分形式表达词范畴的**范畴语法**；以算子形式表达词范畴的**链语法**；侧重描述动词与名词组配关系的**格语法**、**配价语法**；侧重描述从表达功能到语法形式对应关系的**系统功能语法**，等等。

参考文献

Ruslan Mitkov, ed., 2003, The Oxford Handbook of Computational Linguistics, Oxford University Press.

Mark Baltin & Chris Collins, 2000, The Handbook of Contemporary Syntactic Theory, Blackwell Publishers Ltd.

俞士汶 主编，《计算语言学概论》.北京： 商务印书馆，2003

（詹卫东）

自然语言分析 (natural language analysis)

计算机根据给定的形式化的语言知识模型，将自然语言文本中的词语单元、句法结构、语义关系、语篇框架等信息转换为特定的计算机机内表示的技术。运用自然语言分析技术对自然语言文本加以分析，是进一步做信息检索、信息提取、文本摘要、机器翻译等各种自然语言信息处理的基础。通常可根据分析的自然语言单位大小不同以及分析的目标不同，把自然语言分析任务分为词法分析、句法分析、语义分析、语篇分析等子任务。

词法分析包括连续字符流中的词语识别，词语内部结构分析和形态分析，词类标注等具体分析任务。显然，词法分析跟所处理的具体语言有关，对于有词形屈折变化的语言，如英语、德语等，需要进行形态分析。对于没有形态变化的语言，如汉语，则不需要进行一般的形态分析。不过汉语也有自己的特点，需要做特殊的词法分析，例如汉语在书写时词和词之间没有空格。进行汉语书面文本的信息处理，首先就需要将汉语文本中的词逐个识别出来，在词和词之间加上分界标记。这个过程通常称作汉语词语切分。在这个过程中需要处理汉语的离合词、重叠词等特殊形式，也可看作是广义的词语形态分析。词法分析中的词类标注任务是确定文本中词的词类信息，这可以为后续的词义分析及句子结构分析提供帮助。自然语言中一词多类的现象比较常见。当兼类词出现在文本中时，需要根据语境来动态确定其词类。

句法分析处理的基本单位是句子，主要目标是判断输入句子是否是自然语言中合乎语法的表达式。如果是合语法的，则应分析得到能够反映句子组成情况的结构，通常用树形图表示。常见的有短语结构树和依存结构树两种形式。除对整句的句法结构进行完全句法分析外，也有所谓的浅层句法分析，即对句中特定的语言单位进行识别，比如识别出句中的命名实体、时间、处所表达式等等，因为这种分析只给出一个句子中的部分成分作为结果，所以又称部分句法分析。

语义分析主要有两类任务：一是词义分析，处理的单位是词；二是句义分析，处理的单位是句子。词义分析又称作词义标注。自然语言中的多义词现象比较常见，这些词在不同的语境中可以解释为不同的义项，词义分析的目标就是确定文本中的多义词在当前语境中所表现的义项。句义分析的目标是得到句子意义的形式化表示，通常在句法分析之后进行。典型的做法是按照一定的语义规则在句法分析的基础上推导出句子的意义表示。一般可以表示为谓词逻辑表达式，特征结构，语义网络等形式。除对于整句的完全语义分析外，也有所谓的浅层句义分析，一般又称作语义角色标注，分析目标是识别句子中心谓词的论元成分所充当的不同语义角色。典型的语义角色有施事（动作者）、受事（受动者）、与事（参与者）、工具、材料、时间、处所等。

语篇分析的对象是由句子组成的段落和完整的篇章。分析任务主要有两类：一是微观层次上的篇章分析，即确定篇章中实体成分之间的指代关系。主要是代词与其所指代的语言成分之间的共指关系。二是宏观层次上的篇章分析，即分析篇章的整体结构以及句间逻辑语义关系，如因果、假设、条件、让步、转折等关系。

上述各级语言单位的分析在具体实现时大致都可以区分为基于规则的方法和基于统计的方法两种模式。就自然语言处理的发展历史而言，早期的研究范式以基于规则的方法为主，近期则转为以基于统计的方法为主。基于规则方法的一般工作模式是，由人工提出一个语言模型，比如以上下文无关语法形式表达的语言成分组合规则，然后设计相应的搜索算法，对输入字符串进行扫描，搜索规则集中的规则，得到在模型意义下可以解释输入字符串的合理结构。在这个工作模式下，语言模型对语言知识的刻画一般是遵循布尔逻辑的原则，即语言表达式要么是符合规则的（取值为 1），要么是不符合规则的（取值为 0）。基于统计方法的一般工作模式则是，根据人工提出的对语言单位的初始分类，对大规模语言实例进行标注形成语料库，然后建立特定的数学统计模型（比如隐马尔可夫模型、最大熵模型、条件随机场

模型），从语料库中获取语言模型的参数。这个过程称为参数训练。根据训练所得参数来计算输入字符串的各种可能的结构在模型意义下的最大概率，从而选择一个最优解。这个过程一般又称为解码。在这个工作模式下，语言模型对语言知识的刻画是遵循概率分布的原则，即语言表达式之间的差异体现为在[0,1]区间内取值的分布概率，而不再仅仅是 0 跟 1 的二值对立。基于规则方法的哲学基础是所谓的理性主义，基于统计方法的哲学基础则是所谓的经验主义。尽管目前还很难说哪一种方法在自然语言处理问题上更有效，但由于自然语言客观存在的极大复杂性，以人工规则方式刻画语言知识的粒度一般来说比以统计方式刻画语言知识的粒度要粗。近年来随着统计机器学习研究的迅速发展，基于统计的方法，或者说由数据驱动的分析方法在自然语言分析技术领域得到了更为广泛的重视和应用。

除词法、句法、语义、篇章分析等传统的自然语言分析任务外，近年来还因互联网信息检索和信息抽取等应用的发展需要，产生了新的分析任务：**情感分析**和**隐喻分析**。文本情感分析通常按照处理对象的不同，分为四个层次，包括：（1）词语情感倾向性分析；（2）句子情感倾向分析；（3）篇章情感倾向性分析；（4）超大文本整体倾向性预测。其中每个层次的研究又可区分出不同的具体任务，比如词语情感倾向性分析就包括 3 个具体的任务：情感词的识别；情感词褒贬义的区分；情感词倾向义程度的度量。隐喻分析又称隐喻识别，即发现文本中的隐喻表达。目前的主要方法有基于文本线索的方法、基于语义知识的方法和基于机器学习的方法等。

参考文献

Ruslan Mitkov, ed., 2003, The Oxford Handbook of Computational Linguistics, Oxford University Press.

俞士汶 主编.《计算语言学概论》. 北京：商务印书馆. 2003

（詹卫东）

自然语言生成(natural language generation)

计算机根据给定的形式化的语言知识模型，将特定表达意图的机内形式化表示转换为自然语言形式输出的技术。自然语言生成技术可应用于机器翻译、自动文摘、多语信息发布等自然语言处理系统。自然语言生成过程一般视作一种由交际目标驱动的规划过程。通常按照生成的自然语言单位从大到小的顺序，把自然语言生成任务分解为文档规划、句子规划、表层实现等子任务。

文档规划：主要任务是确定文本的内容以及文档的整体结构，其目标是生成一个关于待生成文本的规格说明，即决定文本中应该传达哪些信息。影响文本内容确定的因素主要有文本的交际意图、文本预期读者的信息、对输出文本的要求以及文本的信息来源等。文本不是材料的任意堆积，在文本信息内容确定后，文档规划还需要就这些信息的排序和结构作出决策，完成这一任务的模块通常称作文档结构化模块。文本结构通常是一个树形结构，规定了信息如何分组以及信息分组之间的语篇关系。

句子规划：在文档规划的基础上进行，主要完成三个方面的工作：一是词汇选择，选择适于表达文本信息的动词、名词、形容词等实词性词汇；二是指称表达的生成，生成文本涉及到的人、时、地等实体性元素的具体表达形式，如限定式名词短语，代词等；三是整合，完成文档规划阶段所生成的文本结构树到篇、段、句等文本元素的映射。为了生成自然流畅的文本，文本整合模块需要对内容类似的句子进行归并，避免生成冗长刻板的表达。

表层实现：任务可分为两个方面，一是语言实现，即在句子规划阶段所确定的抽象的语篇元素的基础上，按照具体语言的形态、句法规则将其实现为合法的自然语言句子乃至完整语篇。例如对句子中使用哪些虚词作出选择；确定句子的时态、名词的性、数、格标记形式，等等；二是文本的格式化，按照具体生成系统的要求，对最终生成的文本元素进行格式标记，满足系统输出格式的需求。

参考文献

Ruslan Mitkov, ed., 2003, The Oxford Handbook of Computational Linguistics, Oxford University Press.

E. Reiter, R. Dale. 2000. Building Natural Language Generation Systems. Cambridge University Press.

俞士汶 主编.《计算语言学概论》. 北京：商务印书馆. 2003

（詹卫东）