计算语言学

第12讲 形式语法理论III

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

内容提要

- （几乎）不使用范畴的语法
  - 依存语法
  - 配价语法
  - 词语法
  - 范畴语法
  - 链语法
- 工程性语法

不使用范畴的语法与词汇主义

- 有一大类语法形式，几乎不使用词性和短语类等句法语义范畴，直接刻划词与词之间的搭配关系，这一类语法我们称为（几乎）不使用范畴的语法；
- 由于这一类语法不使用句法语义范畴，因此也无法使用Chomsky形式的重写式规则，几乎所有的语言知识都体现在词典中，因此这一类语法又称为基于词的语法；
- 词汇主义（即所谓“小规则，大词典”）是现代语法理论研究的一个趋势，即使在使用复杂范畴的语法理论中，词典的作用也越来越大，例如HPSG就公开宣称自己是基于“词汇主义”的语法理论；在GB理论中，所有的语法规则被抽象为一条原则（即标杆理论），而词典的内容则越来越丰富。

依存语法－来源

- 依存语法
  Dependency Grammar
- 最早是法国语言学家特思尼耶尔(L. Tesniere, 1893-1954)提出的。特思尼耶尔的主要思想反映在他1959年出版的《结构句法基础》(Element de Syntaxe Structurale)一书中，但是，他于1934年在《怎样建立一种句法》(Comment construire une syntaxe)这篇论文中，就提出从属关系语法的基本论点。特思尼耶尔是从属关系语法的创始人。
- 参考：冯志伟，1983，《特思尼耶尔从属关系语法》，载《国外语言学》1983年第1期

依存语法－特点

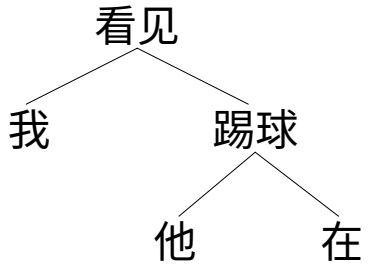
- 依存语法的核心是依存树结构；
- 依存树和短语结构树一样，也是句子结构的一种表示方式，区别在于依存树上的结点都是词语，而不是句法范畴，依存语法不使用词类和短语类等语法标记；
- 依存语法对于什么样的树才是合法的依存树有明确的规定；
- 依存语法适合于中心词分析法，而短语结构语法适合于层次分析法；

特思尼耶尔依存语法的基本框架

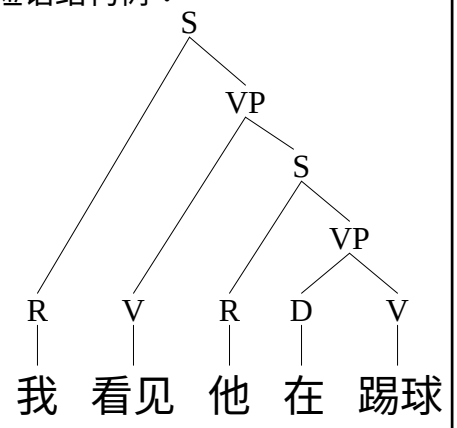
- 一切结构句法现象都可以概括为以下三类：
  - 关联（Connexion）：句子成分之间的从属关系
  - 组合（Jonction）：句子的并列扩展
  - 转位（Translation）：词语功能的转移
- 句法关联建立起词与词之间的依存关系，这种依存关系是由支配词和从属词联结而成的：
  - 动词是句子的中心，动词支配其他成分，它本身不受支配；
  - 直接受动词支配的有名词词组和副词词组；
  - 名词词组是动词的行动元（actant）；
  - 副词词组是动词的状态元（circonstant）；
  - 行动元的数目就是动词的“价”数（valence）；

依存树 vs. 短语结构树

依存树：



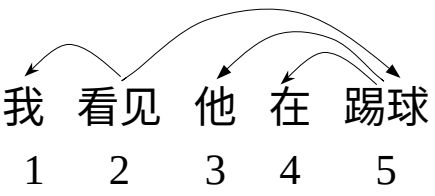
短语结构树：



依存树的线性表示

序号	词语	支配词
1	我	2
2	看见	0
3	他	5
4	在	5
5	踢球	2

支配词序号为0
表示根结点



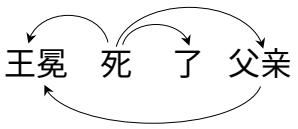
依存语法的四条公理

1970年，美国计算语言学家J. 罗宾孙(J. Robinson)提出了依存语法的4条公理

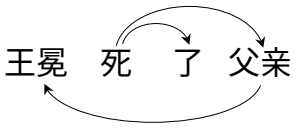
1. 一个句子只有一个成分是独立的；
2. 句子中的其它成分直接依存于某一成分；
3. 任何一个成分都不能依存于两个或两个以上的成分；
4. 如果成分A直接依存于成分B，而成分C在句子中位于A和B之间，那么，成分C或者依存于A，或者从依存于B，或者依存于A和B之间的某一成分。

这4条公理显然也是依存语法不可忽视的形式特性。

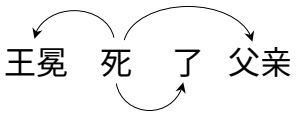
依存树：合法与非法



不合法：违反公理3，“王冕”有两个父结点



不合法：违反公理4，“死”位于“王冕”和“父亲”之间，但不依存于其间任何一个结点



合法

依存语法的12条原则

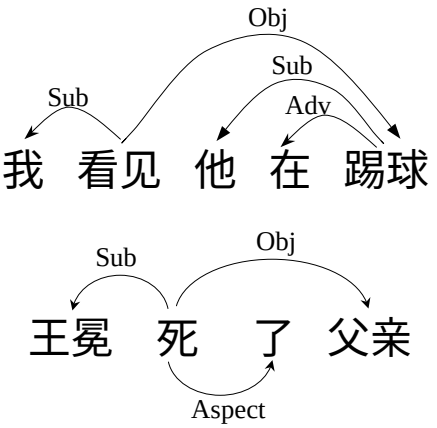
1987年，K. 舒贝尔特(K. Schubert)在研制多语言机器翻译系统DLT的工作中，从计算语言学的角度出发，提出了用于计算语言学的依存语法12条原则：

1. 句法只与语言符号的形式有关;
2. 句法研究从语素到语篇各个层次的形式特征；
3. 句子中的单词通过依存关系而相互关联；
4. 依存关系是一种有向的同现关系；
5. 单词的句法形式通过词法、构词法和词序来体现；
6. 一个单词对于其它单词的句法功能通过依存关系来描述；
7. 词组是作为一个整体与其它词和词组产生聚合关系的语言单位，而词组内部的各个单词之间存在着句法关系，形成语言组合体；
8. 一个语言组合体内部只有一个支配词，这个支配词代表该语言组合体与句子中的其它成分发生联系；
9. 句子的主支配词支配着句子中的其它词而不受任何词的支配，除了主支配词之外，句子中的其它词只能有一个直接支配它的词；
10. 句子中的每一个词只在依存关系结构中出现一次；
11. 依存关系结构是一种真正的树结构；
12. 在依存关系结构中应该避免出现空结点。

依存关系

- 虽然依存语法本身并没有规定要对依存关系进行分类，但在实际应用中，通常都会对依存关系进行分类，也就是说，给依存树上的边加上不同的标记，否则依存树传达的句法信息过少
- 依存关系可以是句法关系，如Subject，Object等等，也可以是语义关系，如Agent，Patient等等

标注了依存关系的依存树



依存树和短语结构树的转换

- 短语结构树所含有的信息比依存树更丰富，只要对于任何一个短语结构规则规定一个唯一的中心成分结点，就可以从短语结构树可以唯一地生成依存树。
- 从依存树生成短语结构树存在不确定性：
  - 必须对依存关系进行标记，才有可能生成短语结构标记
  - 依存关系的结合次序不同，可能导致不同的短语结构树。如依存树B←A→C就可以产生两种短语结构树：(B (A C))和((B A) C)

依存分析法 1

- 美国语言学家 D.G海斯(D.G.Hays)于1960年根据机器翻译的特点提出了依存分析法(dependency analysis)，尽管海斯的依存分析法是独立提出的，但是，这种分析法在基本原则方面与特思尼耶尔的依存语法有许多共同之处。
- 这种分析法力图从形式上建立句子中词与词之间的从属关系，比特思尼耶尔的理论更加形式化，因此，可以看成是对依存语法的形式特性的重要描述。

依存分析法 2

- 在英语中，冠词(Art)与名词(N)之间的关系是：名词是中心词，冠词是从属词，冠词位于名词的左侧，这种从属关系图示如下：

从属词写于中心词的下方，如从属词位于中心词的右侧，就写在右下方。

- 这种从属关系还可以用符号来表示。假定X_i为中心词，$X_{j1},X_{j2}\ldots,X_{jk}$ 为X_i的左侧从属词(X_{j1}位于最左侧)，$X_{jk+1},X_{jk+2},\ldots X_{jn}$为$X_i$的右侧从属词($X_{jn}$位于最右侧)，那么，表示$X_i$与其从属词之间的语法规则可写为：

$$X_i(X_{j1},X_{j2}\ldots X_{jk},*,X_{jk+1},X_{jk+2},\ldots,X_{jn})$$

式中*代表中心词相对于从属词的位置。这个规则记为规则①。

- 除了这种形式的规则之外，还有两种形式的规则，分别记为②和③：
  - ② $X_i(*)$:表示X_i在句子中没有从属性，这是终极型规则；
  - ③ $*(X_i)$:表示X_i不是任何词的从属词，即X_i为全句的中心词，这是初始型规则
- 采用这3种形式的规则，可以从形式上表示句子的中心词及其从属词之间的关系，以造出句子的从属关系树形图从而表示出句子的句法结构，实现自动句法分析。

依存语法的优缺点

- 依存语法直接刻划词与词之间的关系，不使用词性和短语类型标记，形式简洁、精炼，冗余信息少，不过也使得语法的表达能力受到限制。
- 单纯的依存语法由于没有描述依存树结点之间的顺序关系，因而不利于语言的生成。

配价语法 1

- 配价语法是特思尼耶尔早期依存语法理论中关于“价”的理论的进一步发展；
- 配价语法通过对词语配价的描述来刻划一种语言；
- “价”是从化学中借用过来的一个概念，在配价理论中表示一个动词所能够支配的名词性成分（即特思尼耶尔依存语法理论中的行动元）的数目。
- 语言学的进一步发展发现，不仅动词有价，形容词和名词也有价。因此，价可以理解为语言中的动词、形容词或某些名词在其周围开辟一定数量的空位，并要求用特定的成分来加以填补的特性，有多少空位就有多少价。
- 配价语法后来又发展出句法配价、语义配价、逻辑配价等理论。

配价语法 2

- 动词的配价
  - 一价动词：工作
  - 二价动词：购买
  - 三价动词：送给
- 形容词的配价
  - 一价形容词：高兴
  - 二价形容词：愤怒
- 名词的配价
  - 零价名词：桌子
  - 一价名词：姐姐
  - 二价名词：意见

词语法 1

- 词语法认识到，规则是由词汇制约的，也就是说，每一条规则的使用仅局限于某些词汇。
- 词语法试图通过详尽的描写，给出每一个词语在特定的结构下可能和不可能出现的句子形式。
- 下面的例子给出了汉语离合词“穿鞋”在把字句中的用法
- 例子摘自：郑定欧，词汇语法理论与汉语句法研究，北京语言文化大学出版社，1999

词语法 2

N_0		词汇	W_0			N_1				W_1		W_2		T		用例
生命	非生命		助动词	副词	兼语结构	代词	带修饰语	不带修饰语	熟语性	副词	副词短语	补语	带宾语	施事主语句	受事主语句	
+		穿鞋						+				+		−	+	(01)张三把鞋穿反了
+							+					+		−	+	(02)张三把我的鞋穿跑了
	+					+						+		−	+	(03)高跟鞋把我穿怕了
	+						+					+	+	−	+	(04)高跟鞋把我脚穿痛了
+								+					+	−	−	(05)张三把鞋穿了个窟窿
+								+				+		−	−	(06)张三把鞋穿掉了后跟
+			+					+			+			−	−	(07)张三喜欢把鞋趿拉着穿
+					+			+		+				+	−	(08)*张三讨厌人家把休闲鞋到处穿
+				+				+				+		−	+	(09)张三可把雨鞋穿够了
+				+					+			+		−	+	(10)张三可把小鞋穿够了

范畴语法

- 范畴语法试图用两个最基本的范畴S和N来刻划所有的句法成分，这里S表示句子(Sentence)，N表示定指的名词短语(Noun Phrase)。
- 在范畴语法中，所有的其他范畴都被认为是由基本范畴导出的，导出运算只有两种：/ 和 \（读作“右缺”和“左缺”），可以用括号来改变运算的顺序。
- 例如动词短语VP可以认为是由一个S “左缺”一个N构成的，记为：N\S

范畴语法：句法类型

(1) 如果有某个词B，其后面的词C的句法类型为 γ，而它们所构成的序列BC的功能与 β 相同，则这个词B的句法类型记为 β / γ。

(2) 如果有某个词B，其前面的词A的句法类型为 α，而它们所构成的序列AB的功能与 β 相同，则这个词B的句法类型记为 α \ β。

(3) 如果有某个词B，其前面的词A的句法类型为 α，其后面的词C的句法类型为 γ，而它们所构成的序列ABC的功能与 β 相同，则这个词B的句法类型记为 α \ β / γ。

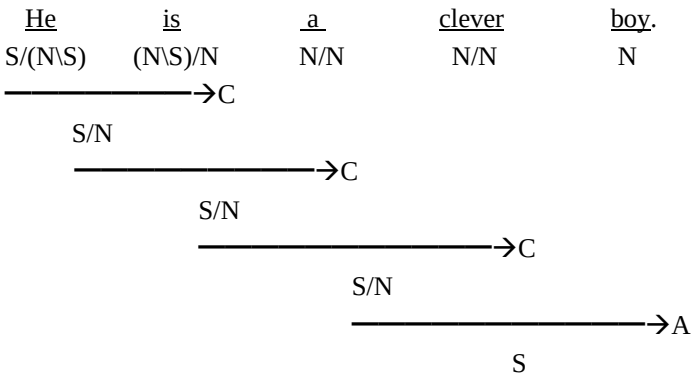
常见英语词类的范畴标注

词类	范畴标注	说明
句子	S	基本范畴
名词	N	基本范畴
不及物动词	N\S	左方缺少主语
及物动词	(N\S)/N 或者 N\(S/N)	左方少主语，右方少宾语
形容词（做定语）	N/N	右方少中心语
形容词（做表语）	(S/N)\S	左方少“缺宾语句子”
副词（做前置状语）	(N\S)/(N\S)	右方少中心语
副词（做后置状语）	(N\S)\(N\S)	左方少中心语
介词（做后置状语）	((N\S)/(N\S))/N	右方少介词宾语
介词（做后置定语）	(N\N)/N	右方少介词宾语
冠词	N/N	右方少名词
代词（主格）	S/(N\S)	右方少不及物动词
代词（宾格）	(S/N)\S	左方少“缺宾语句子”

范畴演算

- 范畴演算的具体操作分为两种：
  - “应用"(Application)，简记为A
    - (1) 如果有形如α，α\β的符号序列，那么就用β来替换它。
    - (2) 如果有形如β/γ，γ的符号序列，那么就用β来替换它。
  - "合成"(Composition)，简记为C
    - (1) 如果有形如α\β，β\γ的符号序列，那么就用α\γ来替换它。
    - (2) 如果有形如α/β，β/γ的符号序列，那么就用α/γ来替换它。
- 对于任何有限的的词语序列，如果通过有限的演算步骤，可以把该词语序列转化为s，那么这个词语序列便是语言中合法的句子，否则为不合法的句子。

范畴演算示例



范畴语法的优缺点

- 优点
  - 形式化程度高，理论优美
  - 词负载结构：完全的词汇主义
  - 范畴语法由于可以在句法结构和语义结构之间建立起同构关系，在形式语义理论中受到了广泛的重视
- 缺点
  - 范畴标记可读性差
  - 对于一个具体的语言单位（如一个词），在不同的语言结构中使用时往往对应着完全不同的范畴，使得语言的歧义现象变得非常严重
- 结论：理论意义大，实用价值有限

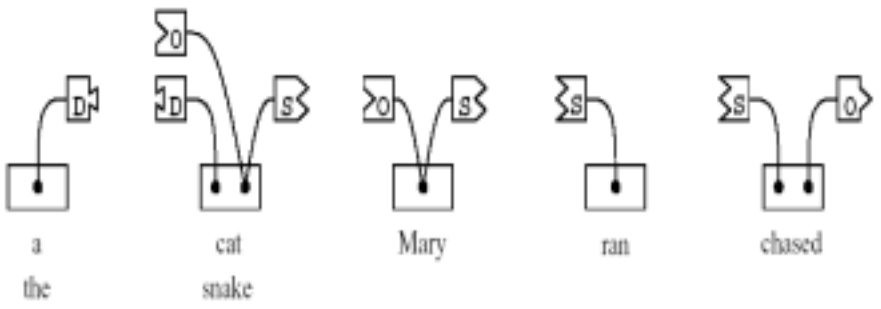
链语法

- 原始文献：
  Daniel Sleator and Davy Temperley, *Parsing English with a Link Grammar*, Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.
- 主页：http://www.link.cs.cmu.edu/link/
  主页上不仅提供了相关的文献，还包括一套完整的英语链语法和相应的分析器及源代码，以及一个用链语法实现的简单的英语到德语的机器翻译系统

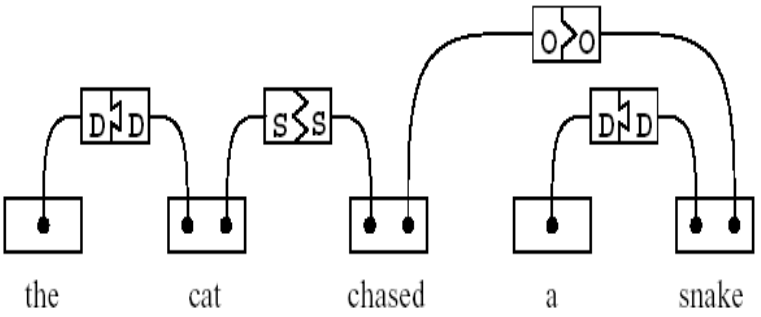
链语法－特点

- 链语法是一种“词汇主义”的语法体系，不使用规则，所有语法信息都由词语来承载；
- 对于每一个词，链语法规定了一些要满族的链接条件（link requirement）；
- 当一些词语组成句子时，词与词之间就建立起一些链接，使得这些词语的链接条件被满足；
- 对于一个合法的句子，词与词之间的链接必须符合一些元规则；
- 链语法的表达能力等价于上下文无关语法。

链语法－词典示例



链语法－一个合法的句子



链接条件

- 词语的链接条件（link requirement）可以通过两种形式来表示：
  - 链接表达式（formula）
    便于理解，适合于人类
  - 析取形式（disjunctive form）
    便于计算，适合于计算机
- 这两种形式是等价的，可以互相转化。

链接表达式

- 链接表达式由一些链接子（connector）通过“与(&)”、“或(or)”运算组成，可以用括号“( )”改变运算的顺序；
- 一个链接子由两部分组成：
  - 名称：由若干个大写字母后接若干个小写字母组成，小写字母部分称为下标（subscript），链接子可以没有下标；
    - 名称的下标部分可以出现通配符 *，表示可以和任意小写字符串匹配；
    - 没有下标等价于任意下标，也就是说下标只有一个通配符 *；
    - 名称前面可以引入符号@，表示该链接子在句子中可以出现零次到多次；
  - 方向：用符号“+”和“-”表示向右和向左。
- 链接表达式中链接子的顺序是有意义的，不能改变。

链接表达式举例

- a, the: D+
- chased: S- & O+
- Mary: S+ or O-
- run: S-
- green, black: A+
- cat, snake: { @A- } & D- & (O- or S+)

析取形式

- 析取形式（disjunctive form）由一系列析取式（disjunct）组成，在句子中这些析取式只要有一个满足即可；
- 一个析取式由左右两个列表组成，左列表中的链接子都是左链接的，右列表中的链接子都是右链接的，因此析取式中链接子的后缀（+和-）可以省略。

链接表达式和析取形式的转换

- 链接表达式：

  (A- or ()) & D- & (B+ or ()) & (O- or S+)

- 析取形式（与上述连接表达式等价）：

((A, D) (S, B))
((A,D,O) (B))
((A,D) (S))
((A,D,O) ( ))
((D) (S,B))
((D,O) (B))
((D) (S))
((D,O) ( ))

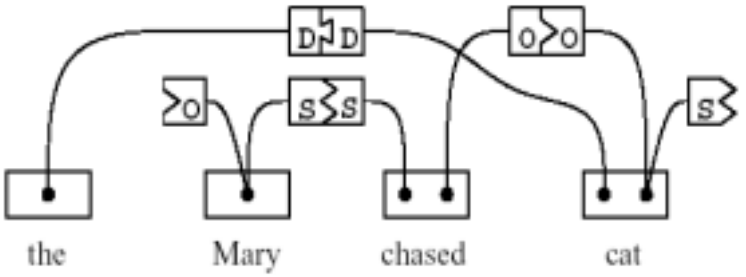
链接条件的满足

- 单词串中某个单词如果有一个向右的链接子，例如X+，而另一个单词有一个向左的链接子X-，那么这两个链接子相互匹配(match)，在这两个单词之间就可以画一条X链。这时，我们说链接子X+或X-得到了满足(Satisfication)或说存在一个链接，满足了链接子X+或X-。
- 链接表达式 X & Y 要被满足，则链接必须同时满足链接子X和Y。
- 链接表达式 X or Y 要被满足，则链接必须至少满足链接子X和Y中的一个。

链接集与元规则

- 对于一个合法的句子，要求句子中所有的单词的链接条件都被满足，并且所有的链接符合下面4条元规则（Meta Rule）的要求：
  - ✓ 平面性（Planarity），链与链之间互相不交叉；
  - ✓ 连通性（Connectivity），所有的单词应该链在一起，形成连通图。
  - ✓ 顺序性（Ordering），链接表达式中靠前的链接子跟距离该单词较近的单词链接，链接表达式中靠后的链接子跟距离该单词较远的单词链接。
  - ✓ 排它性（Exclusion），一对单词之间不能同时有两个链接。
- 一个合法的句子中所有的链接称为一个链接集（linkage），链接集就是链语法分析句子的结果。

链语法－不合法的句子



· 链接有交叉：不满足平面性

复习思考题

- 试比较依存语法和链语法的相同之处和不同之处；
- 试比较配价和链接语法中的链接子有何相同和不同之处；
- 给出用依存语法、范畴语法和链语法分析以下句子的结果：
  - 我是县长；
  - 我是县长派来的；
  - 衣服洗干净了；
  - 小王上街买菜，看见一个人，穿着军大衣，打了一个人一拳，血都流出来了。