从属关系语法对机器翻译研究的作用

冯志伟

从属关系语法(grammaire de dependance)又称依存语法,最早是法国语言学家特思尼耶尔(L. Tesnière)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》一书中,但是,他于1934 年在《怎样建立一种句法》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。

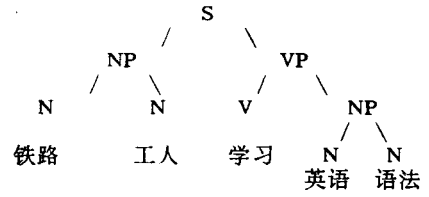
从属关系语法认为,一切结构句法现象可以概括为关联(connexion)、组合(jonction)和转位(translation)三大核心。关联赋予句子以严谨的组织和生命的气息,它是句子的生命线。句法关联建立起词与词之间的从属关系,这种从属关系是由支配词和从属词联结而成的。动词是句子的中心,它支配着别的成分,而它本身则不受其它任何成分支配。直接受动词支配的有名词词组和副词词组,名词词组形成"行动元"(actant),副词词组形成"状态元"(circonstant)。从理论上说,状态元是无限的,而行动元不得超过三个:主语、宾语 1、宾语 2。行动元的数目决定动词的价(valence)的数目。一个动词,如果不支配任何的行动元,则为零价动词,如果支配一个行动元,则为一价动词,如果支配两个行动元,则为二价动词,如果支配三个行动元,则为三价动词。

从属关系语法中的"价",是从化学中借用来的一个概念,在化学中,一个元素的"价"是指这个元素的一个原子与氢原子化合或者被氢原子置换时氢原子的数目,特思尼耶尔把这个术语引入语法研究,用以说明动词支配的行动元数目的多少;一个动词能支配多少行动元,这个动词的价的数目就是多少。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。因此,从属关系语法又叫做"配价语法"(valence grammar)。

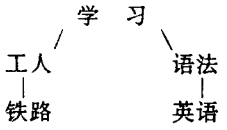
特思尼耶尔的从属关系语法受到了机器翻译研究者的欢迎,1980 年,我在法国格勒诺布尔理科医科大学应用数学研究所(IMAG)自动翻译中心研究机器翻译时,曾经利用从属关系语法设计了汉—法/英/日/德/俄多语言自动翻译系统。我把特思尼耶尔关于"价"的概念引入机器翻译的研究中,把动词和形容词的行动元分为主体者、对象者、受益者三个,把状态元分为时刻、时段、时间起点、空间点、空间段、空间起点、空间终点、初态、末态、原因、结果、目的、工具、方式、范围、条件、作用、内容、论题、比较、伴随、程度、判断、陈述、附加、修饰等 27 个,以此来建立多语言的自动句法分析系统;对于一些表示观念、感情的名词,也分别给出了它们的价。我还把从属关系语法和短语结构语法结合起来,在表示结构关系的树形图中,明确指出中心词的位置,并用核心(GOV)、枢轴(PIVOT)等结点来表示中心词。这可能是我国学者最早利用从属关系语法来进行自然语言计算机处理的一次成功的尝试。90 年代以来,我国清华大学以黄昌宁教授为首的自然语言处理研究者们,又利用从属关系语法来进行汉语的自动处理,取得很好的成果。事实证明,从属关系语法确实是自然语言处理的一种较好的理论语法。

与短语结构语法比较起来,从属关系语法没有词组这个层次,每一个结点都与句子中的单词相对应,它能直接处理句子中词与词之间的关系,而结点数目大大减少了,便于直接标注词性,具有简明清晰的长处。特别在语料库文本的自动标注中,使用起来比短语结构语法方便。

例如,"铁路工人学习英语语法"这个句子,如果用短语结构语法来表示,其结构是:



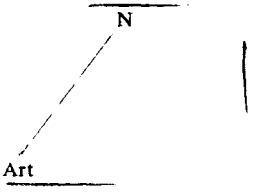
如果用从属关系语法来表示，其结构是：



显而易见，这样的结构比基于短语结构的树形图简洁得多，层次和结点数都减少了。因此，从属关系语法受到了自然语言处理研究者的欢迎。

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从属分析法(dependency analysis)，尽管海斯的从属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从属关系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。

例如，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，冠词位于名词的右侧，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，$X_{j1}, X_{j2} \cdots, X_{jk}$ 为 X_i 的左侧从属词(X_{j1} 位于最左侧)，$X_{jk+1}, X_{jk+2}, \cdots X_{jn}$ 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：

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

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，达到自动句法分析的目的。

1970 年，美国计算语言学家罗宾孙(J. Robinson)提出了从属关系语法的 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从属于某一成分；

3. 任何一个成分都不能从属于两个或两个以上的成分；

4. 如果成分 A 直接从属于成分 B，而成分 C 在句子中位于 A 和 B 之间，那么，成分 C 或者从属于 A，或者从属于 B，或者从属于 A 和 B 之间的某一成分。

1987 年，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从属关系语法 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. 句子中的单词通过从属关系而相互关联；

4. 从属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从属关系来描述；

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从属关系结构中出现一次；

11. 从属关系结构是一种真正的树结构；

12. 在从属关系结构中应该避免出现空结点。

不难看出，舒贝尔特的这 12 条原则包含了罗宾孙的 4 条公理，并且把从属关系扩展到了语素和语篇的领域，可计算性和可操作性更好，更加适合于自

然语言处理的要求。

从属关系可以用树形图来表示。表示从属关系的树形图，叫做“从属树”(dependency tree)。这种从属树是机器翻译中句子结构的一种形式描述方式，因此，我们有必要进一步研究从属树中结点之间的各种关系。

从属树中的结点之间的关系，主要有支配关系和前于关系两种。

如果从结点 X 到结点 Y 有一系列的树枝把它们连接起来，系列中所有的树枝从 X 到 Y 自上而下都有同一个方向，那么，我们就说结点 X 支配结点 Y。例如，在表示“铁路工人学习英语语法”这个句子的从属树中，标有“学习”的结点支配标有“工人”和“铁路”的结点，标有“工人”的结点支配标有“铁路”的结点；标有“学习”的结点还支配标有“语法”和“英语”的结点，标有“语法”的结点支配标有“英语”的结点。

从属树中的两个结点，只有当它们之间没有支配关系的时候，才能够在从左到右的方向上排序，这时，这两个结点之间就存在着前于关系。例如，在前面的从属树中，标有“工人”的结点前于标有“语法”和“英语”这两个结点之间，也不存在支配关系；同样地，标有“铁路”的结点前于标有“语法”和“英语”的结点，“铁路”与“语法”这两个结点之间，不存在支配关系，“铁路”与“英语”这两个结点之间也不存在支配关系。

根据机器翻译研究的实践，我们认为，从属树应该满足如下 5 个条件：

1. 单纯结点条件：在从属树中，只有终极结点，没有非终极结点，也就是说，从属树中的所有结点所代表的都是句子中实际出现的具体的单词。

2. 单一父结点条件：在从属树中，除了根结点没有父结点之外，所有的结点都只有一个父结点。

3. 独根结点条件：一个从属树只能有一个根结点，这个根结点，也就是从属树中唯一没有父结点的结点，这个根结点支配着其他的所有的结点。

4. 非交条件：从属树中的树枝不能彼此相交。

5. 互斥条件：从属树中的结点之间，从上到下的支配关系和从左到右的前于关系是互相排斥的，也就是说，如果两个结点之间存在着支配关系，那么，它们之间就不能存在前于关系。

我们这里提出的从属树的 5 个条件，更加形象地描述了从属树中各个结点之间的联系，显然比罗宾孙的 4 条公理和舒贝尔特的 12 条原则更加直观，更加便于在机器翻译中使用。

用从属关系语法来进行自动分析是很好的，因为分析得到的从属树层次不多，结点数目少，清晰地表示了句子中各个单词之间的从属关系。但是，用从属树来进行自动生成时，必须把表示句子层次结构的从属树转变成线性的自然语言的句子，根据从属树的第 5 个条件(互斥条件)，从属树中的结点之间的支配关系和前于关系是互相排斥的，从结点之间的支配关系，不能直接地推导出它们之间的前于关系，所以，还应该按照具体自然语言中词序的特点，提出适当的生成规则，把表示结构关系的从属树，转变成表示线性关系的句子。在这方面，各种自然语言的生成规则是不尽相同的。例如，汉语的修饰成分一般应置于中心成分之前，而法语的某些修饰成分则置于中心成分之后；汉语主动句的宾语一般应置于谓语之后，而日语的宾语则置于谓语之前。

与短语结构语法相比，从属树也有它的不足之处。在短语结构语法的成分结构树中，由于终极结点之间的前于关系直接地反映了单词顺序，只要顺次取终极结点上的单词，就能够直接生成句子。所以，在自动生成方面，从属树不如短语结构语法的成分结构树方便。为了弥补从属树的这种不足，我在机器翻译研究中，把短语结构语法和从属关系语法结合起来，较好地解决了句子的自动生成问题。

60 年代初期，德国学者把特思尼耶尔的从属关系语法引进了德语研究。从属关系语法在德国一般叫“配价语法”(Valenzgrammatik)。赫尔比希(G. Herbig)提出了补足语(Erganzungen)和说明语(Angaben)的概念，补足语大致相当于特思尼耶尔的行动元，说明语大致相当于特思尼耶尔的状态元，赫尔比希指出某些状语也是动词要求的配价成分，并把补足语分为必有补足语和可有补足语两种。他还与申克尔(W. Schenkel)合编了《德语动词配价与分布词典》，于 1969 年出版。恩格(U. Engel)在《现代德语语法》一书中，建立了完善的德语配价语法体系，他把补足语定义为动词在次范畴化形成一个句子时所特有的被支配成分的集合，对补足语和说明语进行了详尽的分类和论述。舒马赫(H. Schumacher)主编了《动词配价分类词典》，对补足语的种类进行了调整，该词典于 1986 年出版，是一部研究德语动词配价的专著。托依拜特(W. Teubert)把“价”的概念扩展到名词，深入地研究了德语名词的价，于 1979 年出

版了专著《名词的配价》，这是关于名词配价的最早著作，开名词配价研究的先河。

配价可以从逻辑、句法和语义三个不同的层次来认识：

1. 逻辑配价：德国学者邦茨欧（W. Bondzio）认为，在句法结构的组合过程中，词汇的意义提供了决定性的前提，词汇本身具有联结的可能，其联结的能力来源于词汇的语义特点，词义的概念核心反映了语言之外的现实中各种现象之间的关系。例如，德语的 verbinden（联结）这个词的词义表示了联结者、联结的对象、同联结的对象相连的成分三者的关系，德语的 besuchen（访问）这个词的词义表示了访问者和被访者两者之间的关系。配价学者用“空位”这个谓词逻辑的术语来表示词义所具有的关系。动词 verbinden 的词义含有三个空位，动词 besuchen 的词义含有两个空位。空位的数量是完全由单词的词义决定的，在词义的基础上产生的空位就是“价”，某个单词的词义含有空位数就是该词的价数。这种由于词义的逻辑关系所决定的配价，叫做逻辑配价。在不同的语言中，同一个概念所表示的逻辑配价的价数是相同的，在汉语中，“联结”这个动词也是三价的，“访问”这个动词也是两价的。不过，在某一具体的语言中，逻辑关系如何实现，则要借助于该语言特殊的表现方法。

2. 句法配价：逻辑配价在某一具体语言中的表现形式是不尽相同的，这种不同的表现形式，是由具体语言的特有的形式决定的，逻辑配价在具体语言中的表现形式就是句法配价。例如，“帮助”这个动词的逻辑配价为三价：帮助者、被帮助者、所提供帮助的内容，这种逻辑配价在德语中的表现是：谓语动词需要变位，帮助者用主格表示，被帮助者用给予格表示，所提供的帮助用 bei 构成介词结构表示。“他帮助我工作”的德语是“Er hilft mir bei der Arbeit”。同一语言中的同义词的逻辑配价是相同的，但却往往具有不同的句法配价。例如，德语的 warten 和 erwarten 都表示“等待”，逻辑配价是一样的，它们都是二价动词，有两个空位：等待者、被等待者。但是，warten 的被等待者要用 auf 构成介词结构表示，而 erwarten 的被等待者则用宾格表示。比较：

Er wartet auf seine Freundin

Er erwartet seine Freundin

这两个句子的含义都是“他等待他的女朋友”。

3. 语义配价：语义配价是指充当补足语的词语在语义上是否与动词相容。语义配价在不同语言中往往有不同的特点。例如，汉语中可以说“喝汤”，补足语“汤”在语义上与动词“喝”是相容的，但是，在德语中，Suppe（汤）与 trinken（喝）是不相容的，德语中不说“eine Suppe trinken”（喝汤），而要说“eine Suppe essen”（吃汤），而在汉语普通话中是不能说“吃汤”的。这种语义配价也同样反映了不同的语言的特性。

德国学者对于配价理论的研究也同样受到了自然语言处理研究者的欢迎。我在 1983 年研制的德汉机器翻译系统 GCAT 中，就应用了德国学者当时已经取得的成果。在这个德汉机器翻译系统中，语言成分之间的关系不仅有句法关系，还有逻辑关系和语义关系，自觉地接受了德国配价学者关于逻辑配价、句法配价和语义配价相区分的理论。这个机器翻译系统还确定了以词的词汇意义为基础来建立句子成分支配关系的原则，也明显地接受了德国配价学者关于词义决定动词的空位的思想。

配价理论对于自然语言的计算机处理是很有价值的。在我国的机器翻译研究中，有必要重视配价理论的研究，这是机器翻译基础理论研究的一个重要方面。

参考文献

1. L. Tesniere, *Elements de Syntaxe Structurale*, Paris, 1965.

2. K. Schubert, Metataxis: *Contrastive Dependency Syntax for MT*, Dordrecht: Foris, 1987.

3. W. Teubert, Valenz des Substantives, Verlag Schwann, Duesseldorf. 1979.

4. 冯志伟，自然语言机器翻译新论，语文出版社，1995 年。

5. 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996 年。