德汉题录机译系统中基于GPSP文法的句法分析器

姚天舫　　何厚存

(上海交通大学计算机科学与工程系，上海 200030)

摘要：　广义短语结构文法(GPSG)是当今语言学家描述自然语言句法所使用的较流行的文法之一。本文首先用非形式化的表达方式介绍了GPSG的理论。在此基础上，文章进一步叙述了德语句法分析器的设计思想和主要算法，以及分析器原型的系统结构。最后讨论了分析器的实际效果。

关键词：　广义短语结构文法，句法分析，机器翻译

THE PARSER BASED ON GPSG IN A GERMAN-CHINESE TITLE MACHINE TRANSLATION SYSTEM

Yao Tianfang　He Houcun

(Shanghai Jiao-Tong University, Shanghai 200030)

Abstract: The generalized phrase structure grammar (GPSG) is currently one of the most popular formalisms used by linguists to describe the syntax of natural language. In this paper, the GPSG theory is recommended in informal expression. The consideration and main algorithms of designing german parser, based on it, are introduced. In addition, the Architecture of parser prototype is presented. At last, the practical effect for the parser is discussed.

Key Words: generalized phrase structure grammar (GPSG), parsing, machine translation

一　引言

GPSG (Generalized Phrase Structure Grammar) 文法是由G. Gazdar、E. Klein、G. Pullum 和 I. Sag共同提出的新型文法理论。它自70年代末问世以来，发展很快。至85年，四人合著的《Generalized Phrase Structure Grammar》一书全面阐述了这一理论的性质、框架和内容，标志着GPSG理论达到了成熟的阶段。GPSG理论最显著的特点是：主张句子结构只有一个平面，即表层结构。同时，认为每个句法结构都跟一个语义解释相匹配，使得句法分析和语义解释合为一体。此外，它具有严格的形式化定义，注重句法的数学性质的表达。所以，整个理论基础十分严谨。

我们考虑把GPSG理论作为分析器设计的基础，主要由于：(1) GPSG的文法对描述德语的句法范畴及句法规则是十分有利的；(2) 基于复杂特征集的合一运算及合格性条件检查对德语句法分析是十分有效的；(3) 尝试将语言学新理论直接运用于机译系统中。

在德语分析器的设计中，将GPSG作为基本文法，并与Dominance-Chart Parser算法相结合。最终实现了德语题录机译系统的分析器原型。

二　理论简介

1. 基本概念

首先要说明的是，为了通俗易懂起见，不采用GPSG的形式化定义来介绍。

图1所示是GPSP的语法框架。

树型结构是用语法规则描述的合法的自然语言句法结构，它是含有语义解释的表层结构。GPSG与传统的PSG的根本区别在于它的规则系统是通过一系列合格性条件检查之后才跟表层结构联系起来。即每条规则生成一个"候选"的局部树，它是否可成为树型结构的一部分，则要进行一系列合格性条件的检查，不合格的局部树在检查过程中被丢弃。

局部树是GPSG理论的一个基本概念，符合句法规则并经过合格性条件检查的自然语言短语或句子可以构成一棵句法树。树的结点为句法范畴。句法范畴定义为特征／值集。相邻结点构成的子树称为局部树，每一局部树对应一条句法规则。

在GPSG理论中还有两个重要的概念，即扩展和合一。

句法范畴A是另一个句法范畴B的扩展是指：

a. B中所有原子值特征说明都在A中同样被说明。

b. B中所有范畴值特征说明，A中也说明了此特征，而且A中的值是B中值的扩展。

对一组句法范畴K={C1, C2, …}，K的合一是指找出K中所有范畴的最小扩展，如果最小范畴不存在，则合一失败。

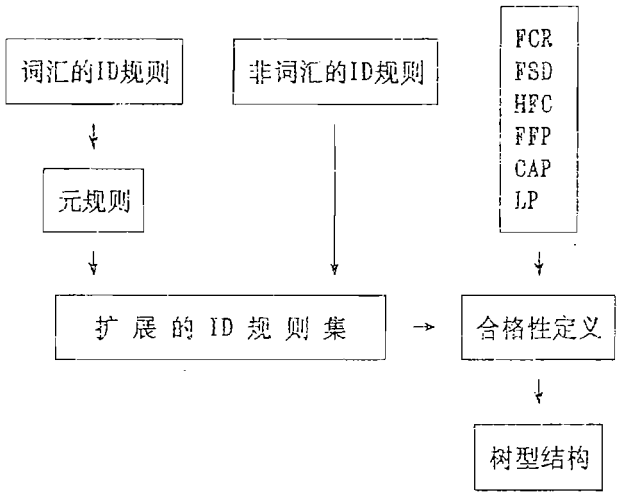

图1　GPSG的语法框架

2. 规则与元规则

GPSG的规则称为ID(Immediate Dominence)规则，它不限制规则右部各范畴的先后次序。如规则S→NP, VP和S→VP, NP两条规则。而用LP(Linear Preference)规则来规定ID规则右部范畴的顺序。如NP<VP表示在引用S→NP VP规则时，NP的位置必须优先于VP位置。ID／LP规则形式是GPSG理论的特点之一，它可使语法有更大的概括力。在具体应用GPSG文法时，LP将作为合格性条件的一部分对规则的应用加以限制。

ID规则分为两类：词汇的ID规则和非词汇的ID规则。词汇的ID 规则是指规则中范畴(中心语)具有次范畴特征如VP→V[3] VP[pas]和VP→V[2] VP[inf]。非词汇ID规则其规则范畴中不具有次范畴特征，如NP→DET N1。

在GPSG中，有一种元规则，它起到扩大短语结构语法功能的作用。如在自然语言句子中常常存在非连续现象，即某个成分在子句中不存在，而出现在子句之外。在规则中表示为范畴空缺。例如：S／NP表示在S中空缺了NP范畴。元规则可以反映出规则的变化，以及两种规则的内在联系，如

S→NP[nom] V[5]

‖ 元规则

v

S/NP[nom]→V[5]

使用元规则有一定的限制，它只能应用于词汇ID规则，这就使得元规则只能在词汇的次范畴已经明确的范围内使用，而不会任意产生与词汇中心语无关联的结构变化。

元规则的实现比较麻烦，在实际应用时，可把元规则产生的规则全部列在ID规则中。

3. 特征共现限制与隐含特征规定

FCR(Feature Cooccurrence Restrictions)：特征共现限制表达了特征之间互相依赖关系。例如：{+zweit, +ac, -rel}=>{+mc}表示在德语句子中，动词不处于句首，且此句也不是关系从句或宾语从句，那么隐含着此句一定是陈述句。从ID规则形成树型结构过程中，所有这些限制必须要满足，即必

须检查树型结构中的每一个结点。

FSD(Feature Specification Default)：隐含特征规定用来指明范畴中某些特征缺省值。例如：NP(n:+, v:-, bar:2, ..., aux:~, pf:~, ...)说明范畴NP的特征aux(助动词)和pf(完成形式)为缺省，即对于NP来说，不需要使用aux和pf特征。

4.特征例化原则

特征例化原则FIP(Feature Instantiation Principle)是GPSG理论的核心部分，它由三个独立的原则HFC、FFP和CAP构成。

我们先看一下投影(Projection)的概念。GPSC把对应于某条规则的局部语法树称为此规则的投影。一个句子是否合法即这个句子所对应的句法树是否合法，也即句法树中的所有局部树是否合法。

投影的合法性：设有ID规则r: C→C1, C2, … Cn及局部树t，t是r的合法投影当且仅当

a. C' 是C的扩展

b. 对任一Ci'，是其在ID规则中对应的某Cj的扩展

局部树如图2所示。

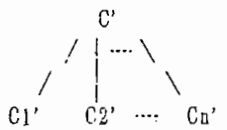

图2 局部树

ID规则合法性：当且仅当局部树是某一ID规则的合法投影。

LP规则合法性：当且仅当在局部树中，设Ci'和Cj'分别是Ci和Cj的扩展，若范畴Ci出现在Cj左边，那么在LP规则中不存在Cj<Ci的关系。

FCR的合法性：局部树中的所有范畴都符合FCR限制。

HFC(Head Feature Convention)：中心语特征规约

基本内容：对一局部树来说，任何自由的中心子结点范畴的Head特征必须与父结点范畴一致。

HFC与X阶标理论相联系。X阶标理论在修改的扩充式标准理论中提出。在这个理论中，短语被定义为其内部中心语(Head)的投影。中心语在很大程度上决定了短语内部的句法特征。在德语中，定义S范畴的阶为3，NP、VP、AP、PP、PD范畴的阶为2，介于NP与N之间的N1范畴的阶为1，终结范畴N、V、ADJ、P等的阶为0。基本的Head特征集为：Head={n, v, per, plu, gen, cas, flex, ac, mc, aux, vf}。如ID规则NP→DET N1和N1→N[1] PP[1]，这里N是NP的中心，所以，N的中心特征值，如per(人称) plu(数) gen(性) cas(格)等必须与NP中心特征一致。

在HFC中提出了“自由”的概念。有些特征，如bar(阶)，为了处理某些特殊现象，也被作为Head特征。在规则N1->N[1] PP[1]中，N1与N的bar不一致，因此，N[1]就不是自由的中心子结点。

FFP(Foot Feature Principle)：基础特征原则

基本内容：对一局部树来说，在父结点中例化的Foot特征必须是所有子结点中例化的Foot特征的合一。

什么叫例化的特征？如果特征由特征例化原则(FIP)说明的，即称此特征的赋值为例化的。

在德语中，基本的Foot特征集为：Foot={slash}，slash特征表示空缺的范畴，它的值是范畴。上述内容包含两方面含义：

a. 局部树中所有在子结点中例化的Foot特征，必须在父结点中例化为同一值。

例：图3的局部树满足FFP。

b. 局部树中，父结点中某特征不可能例化，除非在某一子结点中，此特征被例化为同一值。

例：图4的局部树是非法的。

图3 局部树满足FFP　　图4 局部树不满足FFP

在这里要强调的是：从某局部树所对应的ID规则中指定特征说明称为是“继承”，而从规则中复制并且不由FSD指定特征说明称为是“例化”。例化是FFP的基础，必须注意区分。

CAP(Control Agreement Principle)：控制一致原则

基本内容：对一局部树来说，函子范畴(受控者)中说明的Control特征值必须与参数范畴(控制者)一致。

Control为控制特征集。在德语中，基本的Control特征集为：Control={per, plu, gen, cas, flex}。上述内容包含两方面含义：

a. 如果控制对象C在局部树中有控制者C'，则在C中Control特征值必须与C'相同。

b. 如果在局部树中不存在控制者，那么C中Control特征值必须与其父结点中的Control特征值相同。

举例来说，典型的函子范畴就是VP，它要与参数范畴NP(主语)的Control特征值一致。实际上是把NP的有关特征值复制到VP中。

GPSG理论简单介绍至此，有兴趣的读者可参阅文献[1]。

三 分析算法

分析器采用了Dominance-Chart Parser算法作为实现的基础，它是一种改进的Earley算法。由于DC-Parser是一种自底向上的耗尽搜索算法，所以，对于共同的结构，可能会被重复分析。

DC-Parser算法如下：

输入：a. IDLPG＝〈N，T，ID，LP，B〉

N为非终结范畴集合，T为终结范畴集合，ID为直接支配规则，LP 为线性优先表述规则，B为开始范畴。在直接支配规则中无$A\rightarrow\{\ \}$或$A\overset{*}{\rightarrow}\{A\}$的产生式。

b. DOM关系集合，DOM ⊆ ((N∪T)×N)，即<S，k>∈ DOM，符号S是在第k条产生式中出现在右部的范畴。

c. 终结范畴链w＝a1a2...an∈ T^*

输出：边集E0，E1，...En

处理过程：

```
E0:={ };
FOR i:=0 TO (n-1) DO BEGIN
    Ei+1:={ };
    process(ai+1, i, (i+1))
END;
IF <R, 0, _> ∈ En
THEN RETURN('输入链合法')
ELSE RETURN('输入链非法');

PROCEDURE reduce(k, A, β, w, i, j)
BEGIN
    IF β={ }
    THEN
        IF (w=a1a2...am) and (ax≤ay) and (1≤x<y≤m)
        THEN process(A, i, j)
        ELSE
    FOR all <S, h, M>∈ Ei where k∈ M DO
        reduce(k, A, β\{S}, Sw, h, j)
END;

PROCEDURE process(S, i, j)
BEGIN
    M:={ };
    FOR all <S, k> ∈ DOM DO M:=M∪{k}
    add <S, i, M> to Ej
    FOR all k∈ M where <k, A→α>∈ ID DO
        reduce(k, A, α\{S}, S, i, j)
END;
```

在DC-Parser算法中，增加了GPSG的成分，在这里主要介绍一下FIP的算法。

FFP算法：

```
PROCEDURE ffp(n, m, dl)
BEGIN
    FOR all f∈ foot DO BEGIN
        vl:={ };
        FOR all k∈ dl DO BEGIN
            IF not(inherited(k(f)) in id(n)))
            THEN vl:=vl∪{k(f)}
        END;
        IF not(inherited(m(f) in id(n)))
        THEN
            IF vl={'~'}
            THEN m(f):='~'
            ELSE
                FOR all k1, k2∈ (vl∪{m(f)}) DO
                    unify(k1(f), k2(f))
```

CAP算法：

```
PROCEDURE cap(m, dl)
BEGIN
    FOR all f∈ agr DO BEGIN
        vl:={ };
        FOR all k∈ dl
            where k(agrm)='+' DO
            vl:=vl∪{k(f)};
        vl:=(vl∪{m(f)})\{'~'};
        FOR all k1, k2∈ vl DO
            unify(k1(f), k2(f))
    END
END;
```

注：这里的agr就是控制一致特征集，agrm为需要进行控制一致特征值检查的标志，\表示集合的差运算。

```
    END
END;
```
注：在这里，n是ID规则编号，在这条规则中，m是父范畴，dl是子范畴集。'~'为特征缺省值。

HFC算法：

```
PROCEDURE hfc(m, dl)
BEGIN
    FOR all f∈ head DO BEGIN
        v:=w(f);
        FOR all k∈ dl where k(headm)='+' DO
            IF specified(k(f)) THEN v:=v∩{k(f)} ELSE unify(m(f), k(f));
        IF (v≠{ }) and not(specified(m(f))) THEN unify(m(f), k(f))
    END
END;
```
注：这里w(f)表示特征f的值范围。

四　实现

句法分析器的系统结构如图5所示。

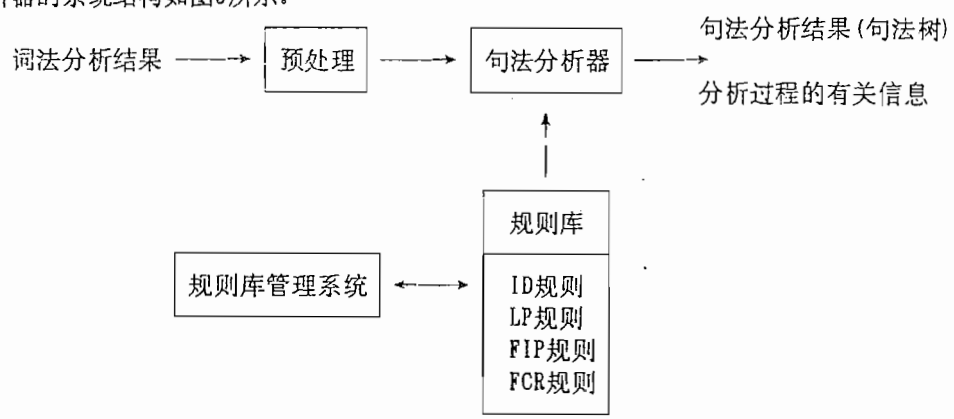

图5　句法分析器系统结构

为了提高分析器的执行效率，我们对词法分析结果进行了一次预处理，以去除输入链前后范畴之间在某些特征(如性、数、格等)不可能匹配的候选范畴。如介词后的冠词、形容词和名词的格应与介词所要求的格一致，不一致的候选范畴被去掉，以减少句法分析器的时空开销。

在句法分析过程中，产生每一棵合法的局部树进行合格性规则检查的顺序是：ID→FCR→FFP→FCR→CAP→FCR→HFC→FCR→LP，在检查中，只要局部树不满足其中一条规则，则放弃对它继续分析，意味着这棵局部树非法。如果分析结果多于一棵句法树，则全部句法树都作为结果输出。

规则库管理系统用来维护规则库。它用于定义新范畴、定义范畴的特征名、特征值范围，定义Head、Foot和Agr的集合成员。还可以定义范畴的别名(外部名)。此外，还具有增加、删除、修改规则等功能。

句法分析器用PDC PROLOG 3.2语言编码，在PC 386机上实现。该系统共有57条ID规则、22条LP规则、26条FCR规则。

我们对几百条德语题录进行了分析实验，分析结果基本上是正确的，句法信息是丰富的。对后续的转换生成工作提供了较好的基础。

下面给出一条德语题录的分析结果：

Verfahren und Vorrichtung zum kontinuierlichen Verpressen von Werkstoffbahnen bei erhoehten Temperaturen.　　在较高的温度条件下连续挤压带料的方法和设备。

分析结果如下所示:

```
np(3, -, 3, nom, Flex, -, Top, -)
- - np(3, -, 3, nom, Flex, -, Top, -)
- - - - np(3, -, 3, nom, Flex, -, Top, -)
- - - - - - - n1(-, 3, nom, Flex)
- - - - - - - - - n1(-, 3, nom, Flex)
- - - - - - - - - - - n(1, -, 3, nom, Flex)   Verfahren
- - - - - - - - - conp(nom)
- - - - - - - - - - - conj(conj)   und
- - - - - - - - - - - n1(-, 2, nom, Flex)
- - - - - - - - - - - - - n(1, -, 2, nom, Flex)   Vorrichtung
- - - - pp(2, -)
- - - - - - p(2, dat)   zu
- - - - - - np(3, -, 3, dat, 1, -, -, Rel)
- - - - - - - - det(-, 3, dat, 1, Rel)   dem
- - - - - - - - n1(-, 3, dat, 1)
- - - - - - - - - - ap(-, 3, dat, 1)
- - - - - - - - - - - - adj(-, 3, dat, 1)   kontinuierlichen
- - - - - - - - - - n1(-, 3, dat, 1)
- - - - - - - - - - - - n(1, -, 3, dat, Flex)   Verpressen
- - - - - - - - - - - - pp(1, -)
- - - - - - - - - - - - - - p(1, dat)   von
- - - - - - - - - - - - - - np(3, -, 3, dat, Flex, -, Top, Rel)
- - - - - - - - - - - - - - - - n1(-, 3, dat, Flex)
- - - - - - - - - - - - - - - - - - - n(1, -, 3, dat, Flex)   Werkstoffbahnen
- - pp(2, -)
- - - - p(2, dat)   bei
- - - - np(3, +, 2, dat, 3, -, -, -)
- - - - - - n1(+, 2, dat, 3)
- - - - - - - - ap(+, 2, dat, 3)
- - - - - - - - - - adj(+, 2, dat, 3)   erhoehten
- - - - - - - - n1(+, 2, dat, 3)
- - - - - - - - - - n(1, +, 2, dat, Flex)   Temperaturen
```

五 结论

我系开发的德汉题录机译系统原型已经实现，其中句法分析部分采用了上述分析算法。实验证明: GPSG理论应用于德语句法分析是行之有效的。

句法分析器还存在的问题: (1) 由于采用的是DC-Parser算法，尽管对词法分析的结果进行了预处理，但执行效率较低，运行速度慢; (2) GPSG 理论强调每个句法结构与一个语义解释相匹配，但这部分工作实现起来较困难，所以，未在分析器中应用，对消除题录语义的歧义带来了不利的影响。

参考文献

[1] Gazdar, Gerald; Klein, Ewan; Pullum, Geoffreg K.; and Sag, Ivan
'Generalized Phrase Structure Grammar'
Basil Blackwell, Oxford, U.K. 1985

[2] Preuss, Susanne
'GPSG-Syntax fuer ein Fragment des Deutschen',
TU Berlin, KIT-IA20, 1987