**CH2 词法分析**

**2.1 scanning process扫描处理**

1.某些记号只有一个词义(lexeme)：保留字；某些记号有无限多个语义：标识符都由ID表示。

**2.2 regular expression正则表达式**

1.\*repetition>concatenation连>alternation选

2.相同的语言可以用不同的RE表示

3.R+：R(R\*)；R?：R|；[abc]：a|b|c；[a-z]：(a|b|c|..|y|z)；[^ab]除a或b；[^a-z]除a-z

4.RE匹配优先匹配保留字；最长字串优先

**2.3 finite automata有穷自动机**

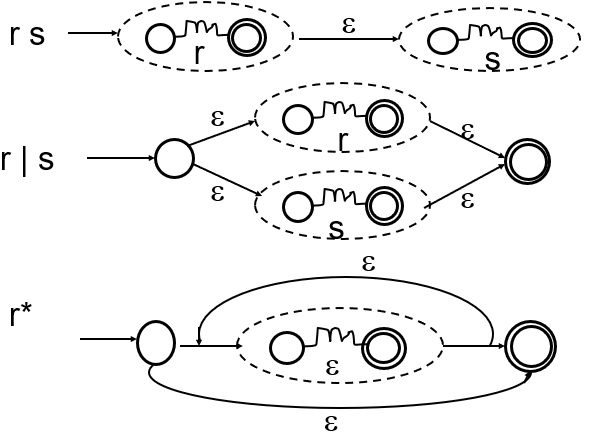
1.DFA：M由字母表、状态集S、转换函数T：S→S、初始状态S0∈S以及接受状态AS。

2.错误状态默认不画，但是存在；错误状态下的任何转移均回到自身，永远无法进入接受。

3.NFA：M由字母表、状态集S、转换函数T：S→(S)、初始状态S0以及接受状态A的集合。

**2.4 RE To DFAs正则表达式到DFA**

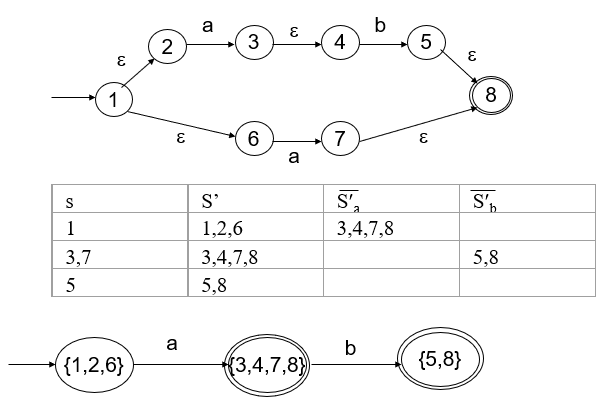
2.Thompson结构通过转移将NFAglue together

3.被合并的那个接受状态如果没有从它到其他状态的转移时，可以将该接受状态和后面的起始状态合并。

4.子集构造的过程：

首先列出所有状态的闭包；然后将初始状态的闭包作为新的初始状态；然后计算在每个新状态下在各个字符上的转移的闭包作为新的状态，转移自然成为新的转移；包含原接受状态的所有新状态都是接受状态

PS：闭包首先包含自身。下面步骤缺一个所有状态的闭包；S代表是哪几个状态的闭包得到的S’



4.DFA状态数最小化：最小状态数的DFA唯一。

步骤：创建两个集合，一个包含所有接受，另一个是剩余；考虑每个字符上的转换，如果所有的接受在a上都有到接受的转换，或是都有到非接受的转换，那么这就定义了一个从新接受到新非接受的转移；如果两个接受有转移但是不在相同集合或是一个有转换，一个没转换，那么两个接受在该字符上被区分，从而分割出了新的状态集合；重复如此。

**CH3 上下文无关文法与分析**

**3.2 CFG**

1.左递归：定义A的推导式的右边第一个出现的是A；右递归：定义A的推导式邮编最后一个出现的是A；

**3.3 Parse tree and AST分析数和抽象语法树**

1.同一个串存在多个推导即多个分析树

2.分析树(concrete sytax tree)是一个作了标记labeled的树，内部节点是非终结符，树叶是终结符；对一个内部节点运用推导时，推导结果从左到右依次成为该内部节点的子节点

3.最左推导和前序编号对应，最右推导后序

4.AST(syntax tree)去除了终结符和非终结符信息，仅保留了语义信息；一般用左孩子右兄弟

**3.4 Ambiguity二义性**

1.定义：带有两个不同的分析树的串的文法

2.解决方法①设置消歧规则disambiguating rule，在每个二义性情况下指出哪个是对的。无需对文法进行修改，但是语法结构就不是单纯依赖文法了，还需要规则②修改文法。

4.修改文法时需要同时保证优先级和结合律precedence and associativity

5.在语法树中，越接近根，越高，优先级越低；左递归导致左结合，右递归导致右结合

6.将相同优先级的运算符分组叫做precedence cascade优先级联

7.通过最近嵌套规则most closely nested rule解决else悬挂问题；另一种方案是为else语句使用一个括号关键字(end if {} fi都可)

8.inessential ambiguity是无关紧要的二义性，虽然语法树各不相同，但是语义相同，例如算术加法虽然可结合但是结合顺序无关紧要

**3.5 EBNF**

1.A→a{b}表b可重复，花括号在右是左递归

A→a[b]表b可选

**CH4 自顶向下分析**

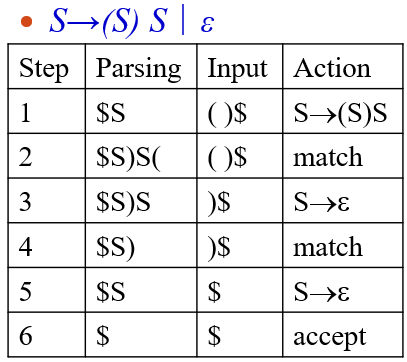
第一个L是从左到右处理，第二个L是最左推导，1代表仅使用1个符号预测分析方向

**4.1 recursive-descent递归下降**

1.将一个非终结符A的文法规则看作将识别A的一个过程的定义。递归下降需要使用EBNF；将可选[]翻译成if，将重复{}翻译成while循环

**4.2 LL(1)**

2.第一列标号；第二列为分析站内容，底座在左，栈底标注$；第三列显示了输入，从左到向右，$表示输入结束；第四列为动作

3.动作：①生成，利用文法将栈顶的N替换成串，串反向进栈②匹配：将栈顶的记号和下一个输入记号匹配③错误

4.Definition of LL(1) Grammar: A grammar is and LL(1) grammar if the associated LL(1) parsing table has at most one production in each table entry.分析表中的每个项目中至多只有一个产生式。LL(1)文法是无二义性的

6.LL(1)面对重复和选择的解决方法：消除左递归left recursion removal和提取左因子left factoring。

7.简单直接左递归：，是N，且不以A开头。，

8.普遍直接左递归：

9.一般的左递归，不能带有产生式和循环

for i:=1 to m do

for j:=1 to i-1 do

replace each grammar rule choice of the form Ai→Aj by the rule

Ai→, where Aj→ is the current rule for Aj

Remove, if necessary, immediate left recursion involving Ai 其中m是N的个数

10.提取左因子

。，

**4.3 first and follow sets**

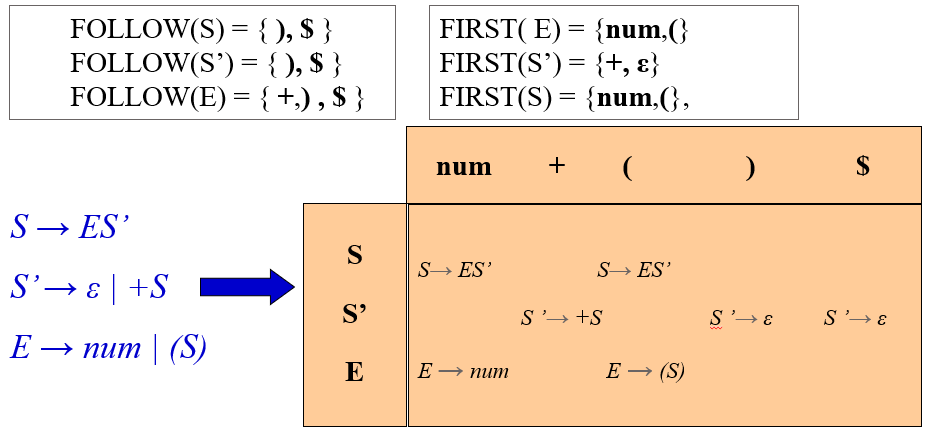
1.Fisrt定义：令X为一个T或N或，Fisrt(X)由T或组成。①若X为T或，Fisrt(X)={X}②若X为N，对于每个产生式X→X1 X2 …Xn，First(X)都包含了First(X1)-。若对于某个i<n，所有的Fisrt(X1),…,First(Xi)都含有，则First(X)也包括了First(Xi+1)-。若所有Fisrt(X1),…,First(Xn)都含有，则First(X)也包含。

2.定理：A non-terminal A is nullable if and only if First(A) contains

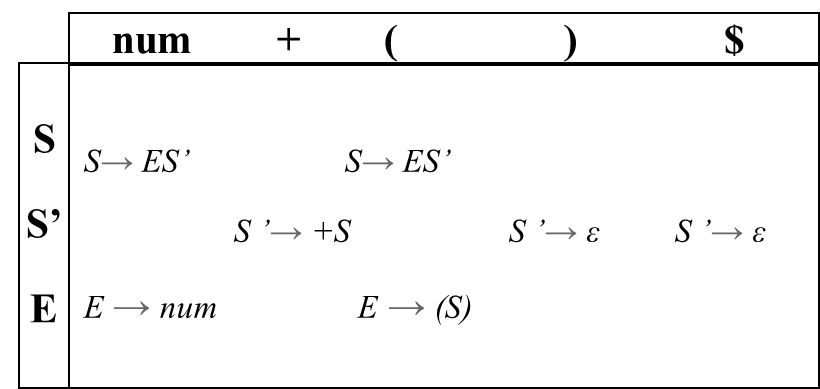
3.Follow定义：若A是一个N，那么Follow(A)由T和$组成。①若A是$，直接进入Follow(A) ②若存在产生式B→A，则First()-在Follow(A)中 ③若存在产生式B→A，且在First()中，则Follow(A)包括Follow(B)

PS：③更常见的情况是B→A，那么Follow(A)包括Follow(B)

4.First关注点在产生式左边，Follow在右边

5.LL(1)分析表M[N,T]的构造算法：为每个非终结符A和产生式A→重复以下两个步骤：①对于First()中的每个记号a，都将A→添加到项目M[A,a]中 ②若在First()中，则对于Follow(A)中的每个元素a(包括$)，都将A→添加到项目M[A,a]中

S→ES’ S→|+S E→num|(S)



6.LL(1)文法的判别：A grammar in BNF is LL(1) if the following conditions are satisfied.①For every production Ai→, First()∩First() is empty for all *i* and *j*, ②For every non-terminal A such that First(A) contains , First()∩Follow () is empty.

**4.5 error recovery**

1.遇错后的不同层次反应：给出一个错误信息后①尽可能准确定位②尝试进行错误矫正error repair③分析程序从错误程序中推断infer出正确程序

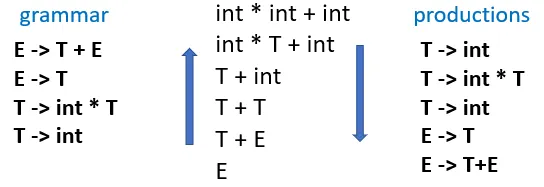
2.some important considerations:①尽快判断出错误的发生②错误发生后，必须挑选一个位置恢复resume分析，尽可能找到多的真的错误③避免出现错误级联(一个错牵出数个假错)④避免错误的无限循环

3.panic mode应急模式，递归下降中的错误矫正。基本机制为每个递归过程提供一个额外的由同步记号组成的参数。遇到错误是，就向前扫描，并且一直丢弃记号知道遇到一个同步记号，从这里恢复分析。Follow集合是同步记号中的重要一员。First集合可以避免跳过开始新的主要结构的重要记号，也可以在更早时候检测错误。同步记号随着递归不断传递并增加新值。

4.LL(1)中没有递归，因此额外增加一个栈存同步记号，算法生成每个动作前，都调用checkinput；或者在分析表中的空格中补全错误处理，共有三种可能①若当前输入为$或是在Follow(A)中，将A从栈中弹出，记作pop②当输入不是$或不在First(A)∪Follow(A)中，看到一个为了它可以重新开始分析的记号后，再弹出该记号，记作scan③特殊情况下压入一个新的N

**CH5 自底向上分析**

Yacc基于LALR(1)；使用显示栈完成分析。



**5.1 概览**

1.动作为①shift，将T从输入开头移到栈顶②reduce使用产生式A→将栈顶的规约成A③accept分析栈为开始符号，输入栈为空时的动作④error

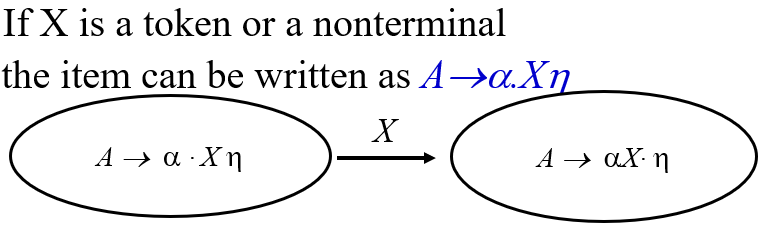
2.注意统一额外加一个S’作为新的开始符号

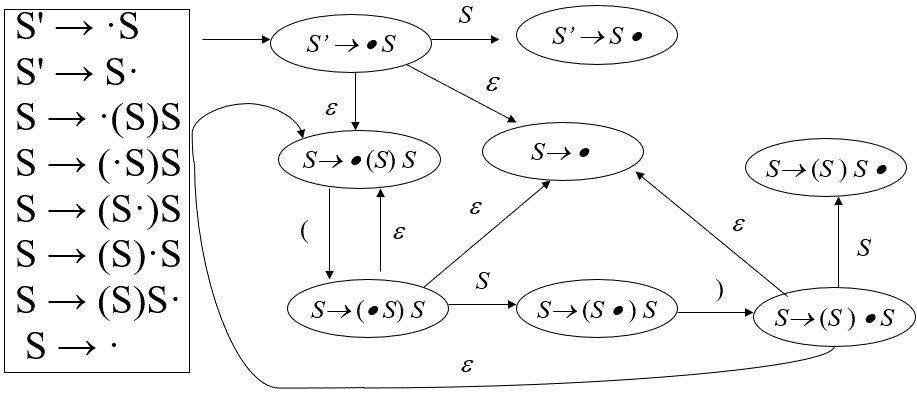
3.推导中的N和T的每个中间串都称作右句型right sentential form，这样的句型都被分析栈和输入分隔开(即使某一边空了也OK)。在每一种情况下，分析栈的符号序列都被称为右句型的可行前缀viable prefix(分析栈空时，可行前缀为)

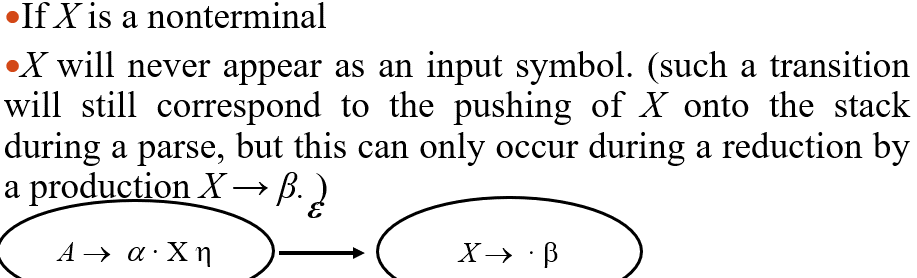
3.若语法无二义性，则句柄唯一

**5.2 FN of LR(0) items and LR(0)**

1.LR(0)的项就是在右边带有区分位置的产生式，同时就是LR(0)的FA中的一个状态

2.NFA构造算法：按照上述规则，从添加的S’→S作为起始状态进行构造。



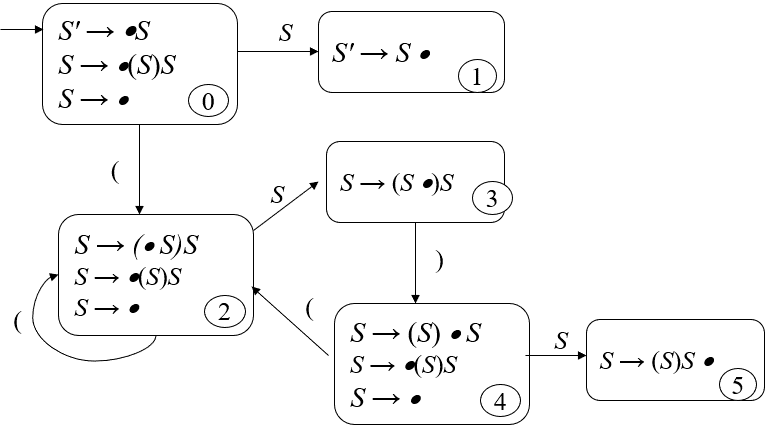


3.DFA构造算法：

每个新状态都是一个产生式的闭包。

其中在闭包步骤中通过添加到状态中的项目与引起状态的项目，前者叫闭包项closure item，后者叫做核心项kernel item。若有一个文法，核心项唯一判断出状态以及转换，那么只需要指出和心想就可以完整地表示出DFA。

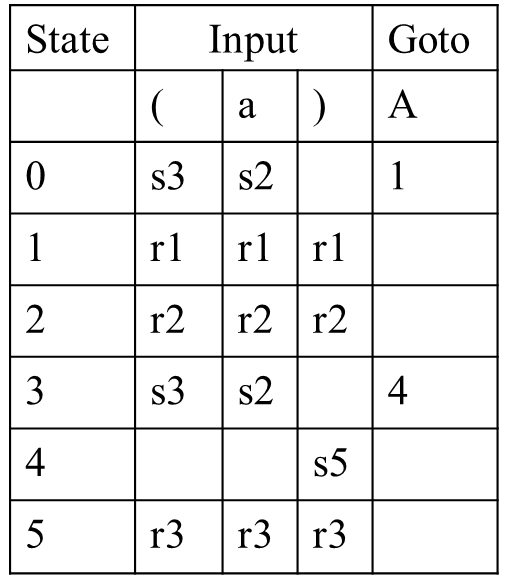
4.LR(0)分析算法的定义：Let s be the current state (at the top of the parsing stack). Then actions are defined as follow:①If state s contains any item of the form (X is a T). Then the action is to shift the current input token on to the stack.② If state s contains any complete item (), then the action is to reduce by the rule 如果为T直接移进；如果包含完整项，直接规约



5.s-r conflict：包含了完整项的状态不能包含任何其他项目，否则s-r冲突

6.r-r conflict：两个完整项共存则出现r-r冲突

7.LR(0)文法不可能是二义的

8.A grammar is LR(0) if and only if ①Each state is a shift state (a state containing only shift items) or a reduce state (containing a single complete item).

这里的r1都是用S’→S·规约，应该写成accept

**5.3 SLR(1)**

1.SLR(1)算法定义：①If state s contains any item of form (X is T), then the action is to shift the current input token onto the stack, and the new state to be pushed on the stack is the state containing the item②If state s contains the complete item and the next token in the input string is in Follow(A), then the action is to reduce by the rule 移进规则不变；规约时要求输入必须在follow中

2.SLR(1)不可能是二义性

3.A grammar is SLR(1) if and only if, for any state s, the following two conditions are satisfied:①For any item (X is T), there is no complete item B→ in s with X in Follow(B).②For any two complete item and in s, Follow(A) ∩ Follow(B) is empty.待移进的终结符不能是完整项的Follow元素；两个完整项的Follow集不相交

4.自底向上分析中右递归可能引起栈溢出，需要避免

5.SLR(1)中的两种冲突，sr冲突使用消岐规则：优先移进；rr冲突基本是设计出问题

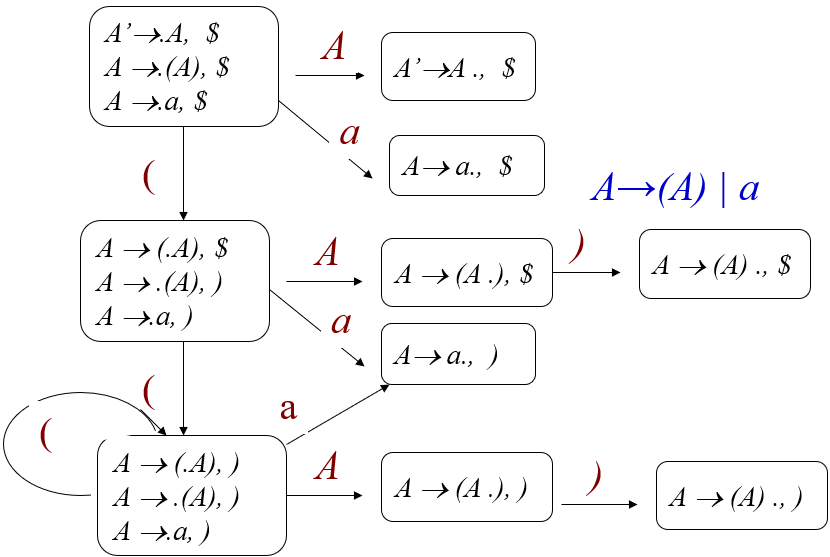
**5.4 LR(1) and LALR(1)**

1.LR(1) items：[]前面是LR(0)项，后面是lookahead token

2.LR(1)的起始状态[S’→·S,$]的闭包

3.LR(1)转移的定义：①非空转移Given an LR(1) item [], X is T or N, there is a transition on X to the item []②空转移Given an LR(1) item [], B is a N, there are -transitions to itrem [B→,b] for every production B→ and for every token b in First().第一条规则永远不会创建新的先行；

实际情况中，往往是本身就是，此时从格式[]到[B→,a]可得到转移

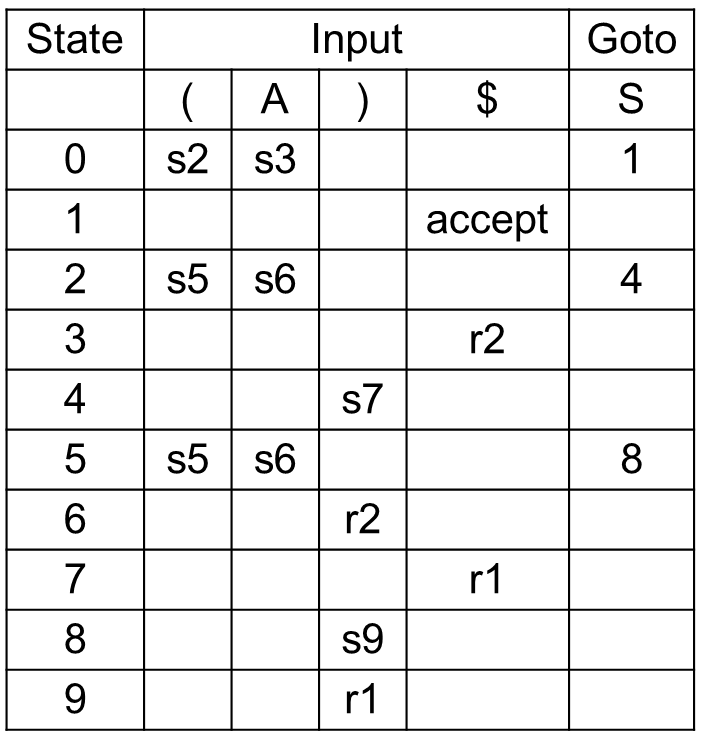


4.LR(1) definition: let s be the current state (at the top of the paring stack). Then actions are defined as follows:①If state s contains LR(1) item of the form [], X is T and X is the next token in the input string.②If state s contains LR(1) item [], the next token in the input stream is a.③If the next input token is such that neither of the above two cases applies, error.

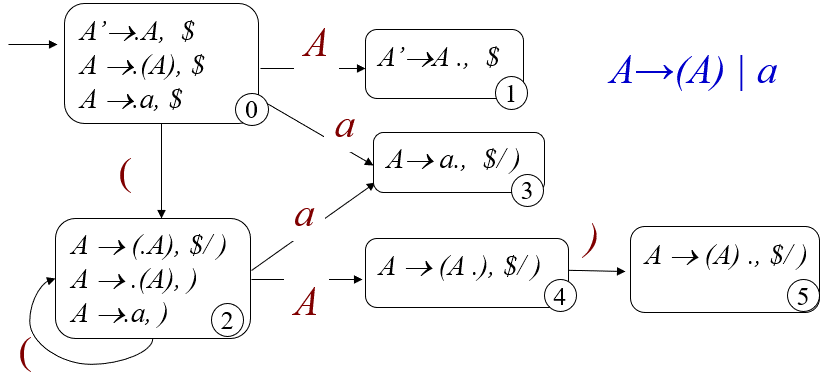
5.LR(1)文法不可能二义性

6.A grammar is LR(1) if and only if, for any state s. The following two conditions are satisfied: ①For any item [], X is T. There is no item in s of the form [B→,X]，否则sr冲突②There are no two items in s of the form [] and []，否则rr冲突

7. (1) A→(A)，(2) A→a的分析表

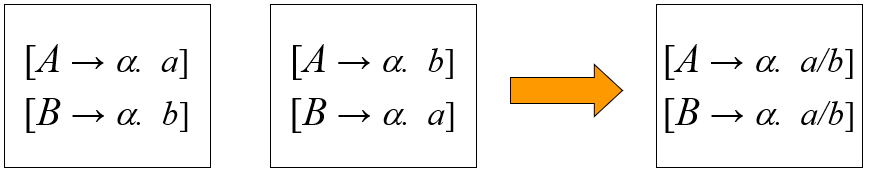


8.LALR(1)将先行合并



9.A grammar is an LALR(1) grammar if no parsing conflicts arise in the LALR(1) parsing algorithm.

10.如果文法是LR(1)，那么LALR(1)中必然没有sr冲突，但是可能有rr冲突。



11.如果文法是SLR(1),那么必然是LALR(1)。

12.通过传播先行propagating lookahead的处理从LR(0)项目的DFA直接计算出LALR(1)的DFA是可能的。

**5.7 Error recovery**

1.LR(1)比LALR(1)或SLR(1)更早检测出错误；LALR(1)和SLR(1)都比LR(0)更早

2. There are three possible alternative actions:① Pop a state from the stack.②Successively pop tokens from the input until a token is seen for which we can restart the parse.③Push a new state onto the stack.

3. When an error occurs is as follows:①Pop states from the parsing stack until a state is found with nonempty Goto entries.②If there is a legal action on the current input token from one of the Goto states, push that state onto the stack and restart the parse.③If there is no legal action on the current input token from one of the Goto states, advance the input.

**CH6 语义分析**

编译器完成的是静态语义分析

**6.1 Attributes and attribute grammars**

1.属性的计算及将计算值与语言结构联系的过程称作属性联编binding。执行之前联编的属性是静态的，期间是动态的。

2.属性计算通过属性等式attribute equation或语义规则semantics rule表示。

*Xi.aj=fij(X0.a1,…,X0.ak, X1.al, …, X1.ak,…Xn.a1, …Xn.ak)*

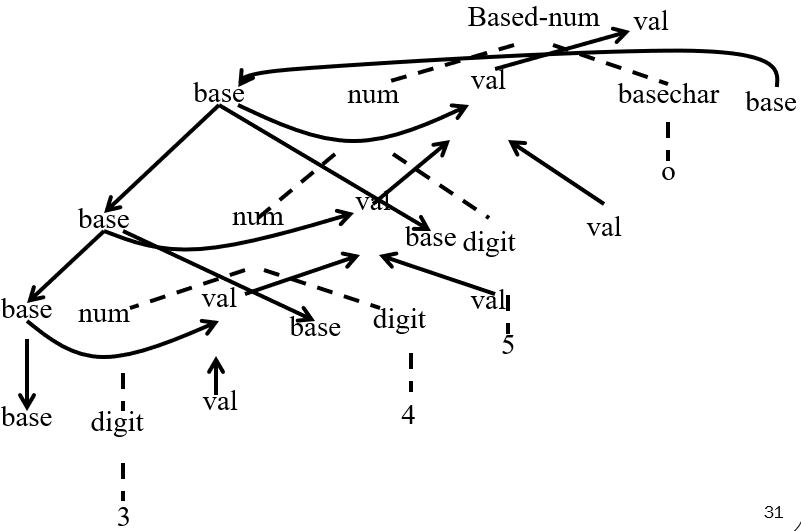
3.属性文法表，左侧列为Grammar Rule，右侧为Semantic Rules

4.属性等式中可以出现的表达式集合称作属性文法的元语言metalanguage。元语言限于算术式、逻辑式以及if switch等语句。也可以使用函数。

**6.2 Algorithms for attribute computation**

1.给定一个属性文法，每个文法规则选择有一个相关依赖图。文法规则中的每个符号在图中都有用每个属性*Xi.aj*标记的节点，对于每个属性等式*Xi.aj = fij(…, Xm.ak, …)*相关于文法规则从在右边的每个节点*Xm.ak*到节点*Xi.aj*有一条边(表示*Xi.aj*对*Xm.ak*的依赖)。

2.字符串345o(八进制)的相关图



3.存在拓扑排序的充分条件为相关图必须是肺循环的acyclic，这样的图称作确定非循环图DAGs

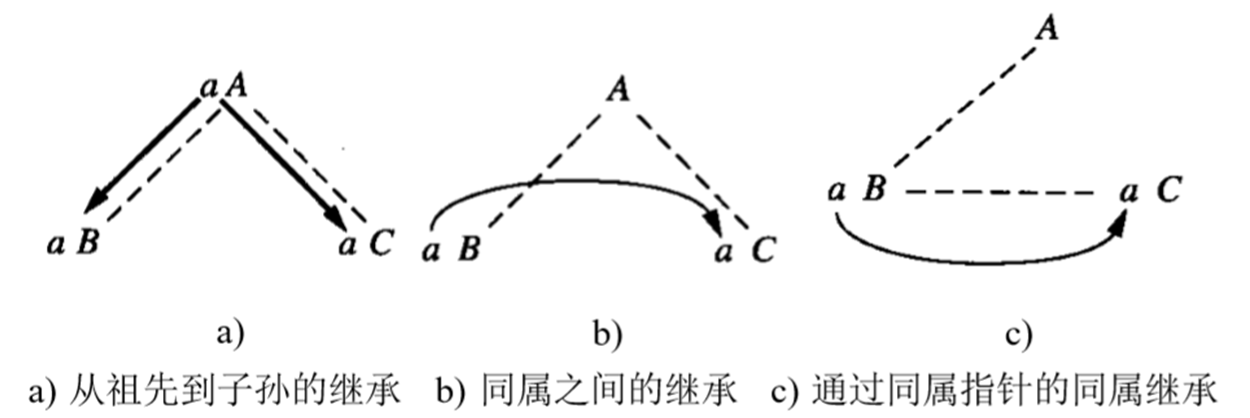
4.基于一个算法在相关图的构造上进行属性分析是可能的，因为在编译时相关图是基于特定的语法构造的，这个方法称作分析树方法。

5.依赖于对属性等式或语义规则的分析叫做基于规则的方法rule-based method

6.合成synthesized属性：一个属性是合成的，如果在语法树中它所有的相关都从子节点指向父节点。等价的，给定一个文法规则*A🡪X1X2… Xn*，左边仅有一个a的相关属性等式有以下形式*A.a = f(X1.a1,…X1.ak,…Xn.a1,… Xn.ak)*

7.一个属性文法中所有的属性都是合成的，那么就称作S属性文法S-attributed grammar。给定分析树或语法树，S属性文法的属性值可以通过对树进行简单的自底向上或后序遍历计算。

8.继承inherited属性：不是合成的属性就是继承的。



9.继承属性的计算可以通过对分析树或语法树的前序遍历或前序/中序遍历的组合来进行。与合成属性不同，子孙继承属性计算的顺序十分重要，因为在子孙的属性中继承属性可能有依赖关系。

10.递归遍历程序用前序计算继承属性，用后序计算合成属性，在子节点把继承属性作为参数传递给递归函数调用，并接收合成属性作为那些相同调用的返回值。

11.外部数据结构替代传参返回的典型是符号表

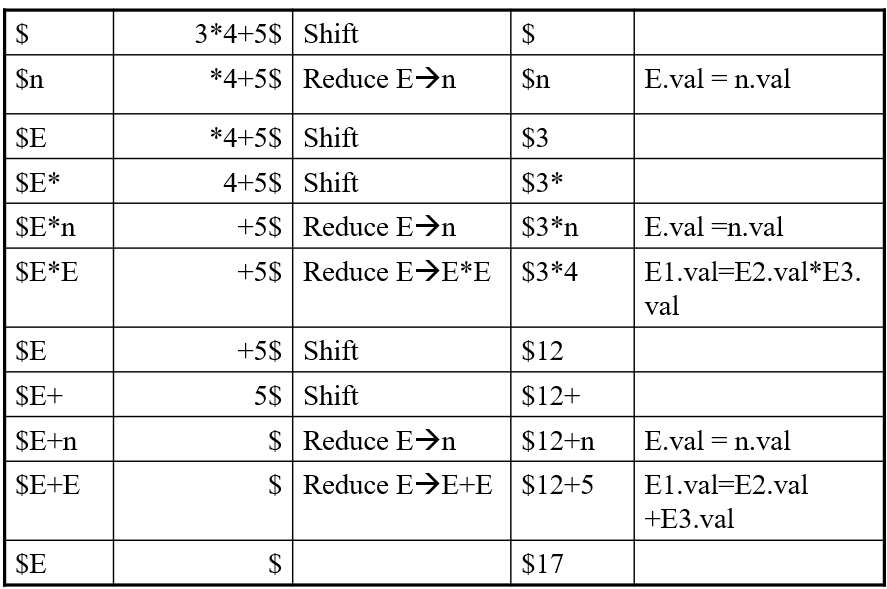
12.我们所使用的分析方法都是从左到右处理的。这等价于要求属性能通过从左向右遍历分析树进行赋值，这对于合成属性不是限制，因为子节点可以用任意顺序赋值。但是对于继承属性，这就意味着相关图中没有向后的依赖(分析树种依赖从右到左)。属性文法确保的这一特性叫L属性

13.L属性定义：属性*a1,…, ak*的一个属性文法是L属性，如果对每个继承属性aj和每个文法规则*X0 🡪 X1X2…Xn*，aj的相关等式都有以下形式*Xi.aj=fij(X0.a1,…,X0.ak, X1.al,…,X1.ak,.Xi-1.a1, …Xi-1.ak)*。也就是说，在Xi处aj的值只依赖于在文法规则种Xi左边出现的符号X0,…,Xi-1的属性。

14.S属性文法必定是L属性文法

15.给定一个L属性文法，对于Top-Down可以通过把继承属性转换成参数、把合成属性转换成返回值；对于Bottom-up，LR分析适合处理合成属性，对于继承属性很难处理。

16.值栈用来储存合成属性，在分析时并行处理



17.在LR分析中，处理继承属性的最好技术是使用外部数据结构

18.定理：给定一个属性文法，通过适当地修改文法，而无须改变文法的语言，所有的继承属性可以改变成合成属性。但是并不推荐，因为会造成文法规则晦涩难懂

**6.3 Symbol table**

1.符号表主要操作：增删查

2.符号表是一个典型的字典数据结构。典型实现包括线性表、搜索树以及hash表

3.线性表：简单直接的实现，插入操作常数时间，查找和删除是n复杂度。对于不在乎编译速度的编译器，十分合适。

4.搜索树：没什么用，删除太过复杂

5.hash表：三种操作基本都是常数时间，最常用，是最优选择

6.hash函数把索引键search key(标识符名字符串)转换成索引范围内的一个证书的hash值，对应在该索引的桶bucket中。

7.关键问题：如何处理冲突collision resolution。①开放寻址open addressing：在冲突时，将该项插入到后续相连的桶successive buckets中；性能下降很快且删除操作复杂②分离链表separate chaining：每个桶都实现为一个线性表，把新的项插入到桶对应的表中(头插入)

8.问题：桶的个数，这个个数在编译器构建期间固定不变；一般从几百到上千；大小实际应用中应为一个素数

9.hash函数三步：①将字符串中的每个字符转换成一个非负整数②将这些整数按照某种规则组合成一个整数③把整数调整到0~size-1的范围内

10.实现一：忽略大部分字符，只把开头的几个字符或者第一个、中间的和最后一个字符的值相加。对于编译器来说并不合适。

11.实现二：直接把所有字符值加起来，但是也不行，temp和ptem一样了

12.实现三：当加上下一个字符的值时，重复地使用一个常量作为乘法因子*α*。*hi+1 = αhi + ci h0 = 0*最后的hash值就是*h = hn* mod *size*等价于下列公式*h = (αn-1c1+αn-2c2 +…..+ α cn-1+cn) mod size*。惩罚因子的合理选择是2的幂，这样可以通过移位来完成乘法

13.四种基本说明declaration：常量、类型、变量以及过程/函数。①constant：把变量的值绑定到名字上②把新构建的类型绑定到名字上，或是创建已有类型的别名aliases③把名字绑定到数据类型上，变量声明同时隐含地约束了其他属性，最主要的是说明的作用域scope of declaration④procedure/function：包含显式和隐式说明

14.保存说明的策略：①使用一张符号表保存所有不同类型的说明的名字，这要求语言禁止在不同类型的说明中使用相同的名字②对每种说明使用不同的符号表③程序的不同区域(函数或过程)都有独立的符号表，并且韩赵语言的语义规则链接到一起，C Pascal Ada都是这样

15.作用域规则两条：①使用前说明declaration before use②块结构的最近嵌套规则most closely nested rule for block structure。In pascal, blocks are the main program; procedure/function declarations. In C blocks are the compilation units, procedure/function declarations; the compound statements ({})

16.同层次中不能使用相同的名字，这要求每次插入符号表前都进行一次查找。

17.顺序说明sequential：在处理时，每个说明加进符号表；并列说明collateral：替代所有要同时处理的说明，在说明部分的最后立即加入到符号表。然后说明中任意表达式的名字将引用前面的说明，不再处理新的说明；递归说明recursive：说明可以引用自身或者互相引用

18.顺序调用可以通过修改代码顺序完成；相互调用就需要在开始加一个函数原型的说明

**6.4 Data type and type checking**

1.仅有的可用于缺省类型名的，结构等价：两个类型当且仅当他们有相同的结构时才相同。等价于语法树相同时才等价。可能会允许数组大小不同、结构体成员顺序不同等；

2.基于类型名的名等价，当且仅当他们是相同的简单类型或有相同的类型名，强等价

3.说明等价，名等价的弱化版，像t2=t1，t2时作为类型别名解释的，不是新类型

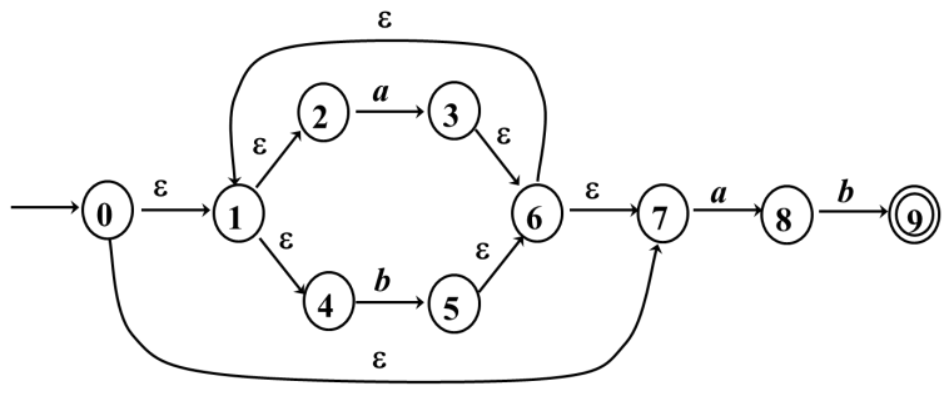
4.Pascal使用说明等价，C对结构和联合使用说明等价，对指针和数组使用结构等价。

**题目**

1. parse tree reflects the derivation steps of a string

2. left recursion is commonly used to make operations left associative

3. 



4 The topological sort of attribute computation ofis unique.(F)

5. An inherited attribute can be calculated by a postordertraversal of the parse tree.(A)

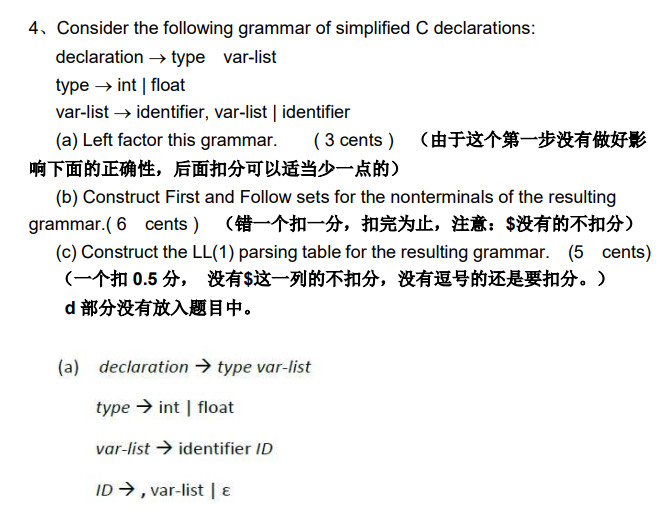
1. All inherited attributes can be changed into synthesized attributes by suitable modification of the grammar, without changing the

language of the grammar.(T)

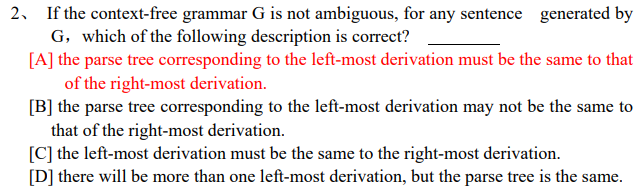
2. When inheriting a previously computed synthesized attribute during LR parsing, it is suitable to treat the computed synthesized

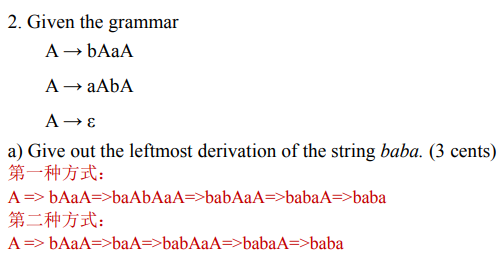
attribute as external data structure.

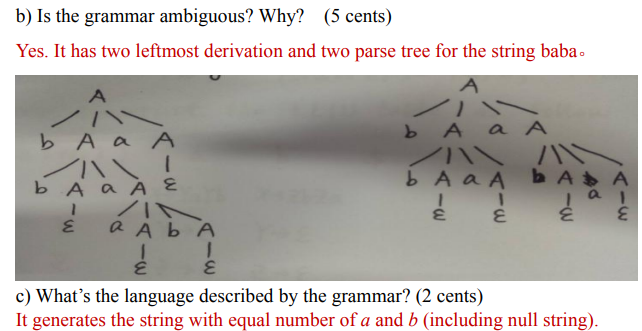
3. The symbol tables will not carry the data type information about the variables.



1. A grammar is ambiguous if it has two different derivations or two different parse trees for a sentence. (F)

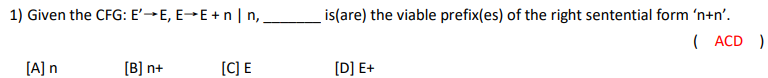
2. 

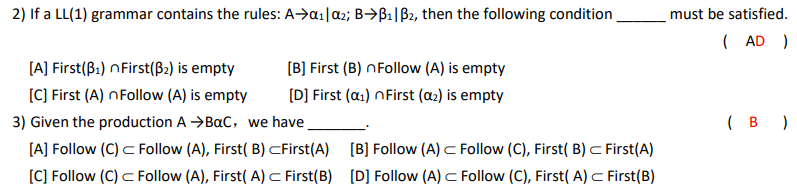
3. 

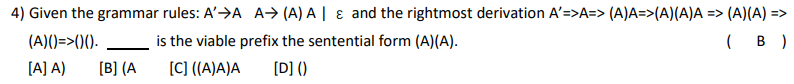


4. A LR(1) parser cannot parse any left-recursive CFG without ambiguity.（F）-

5. The parse tree will completely reflect the derivation steps for a string.（F）

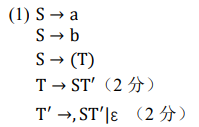
6. 

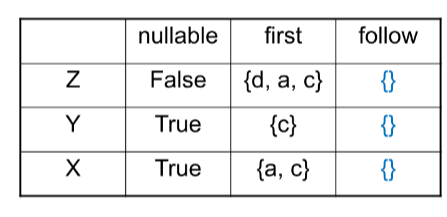
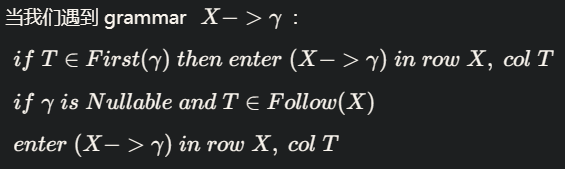
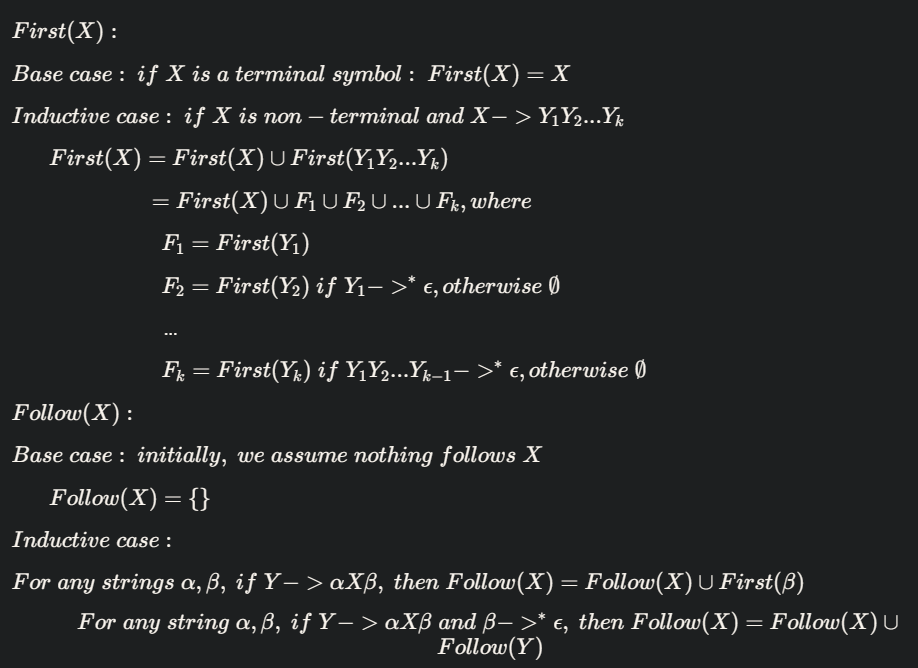
7. 

8. 

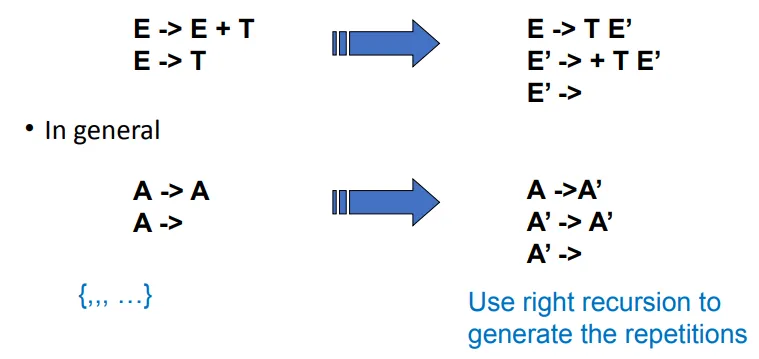
9. 

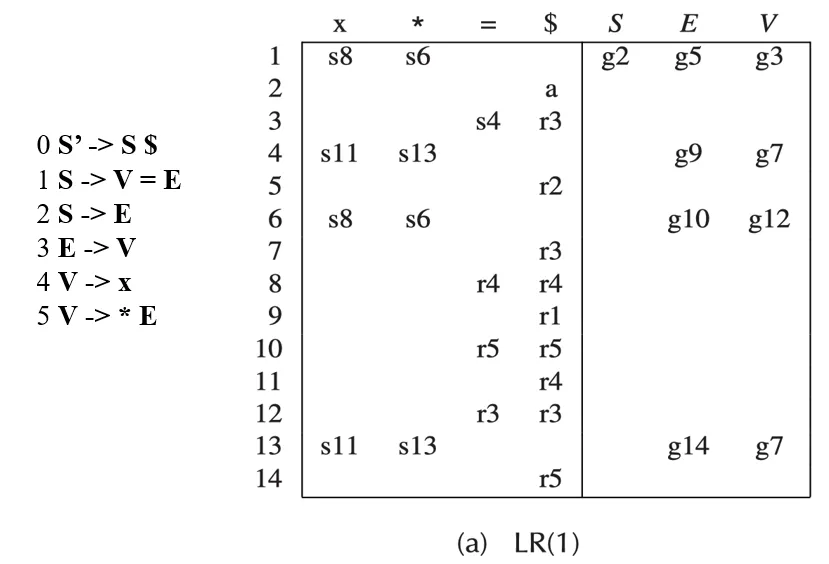
Remove the left recursion

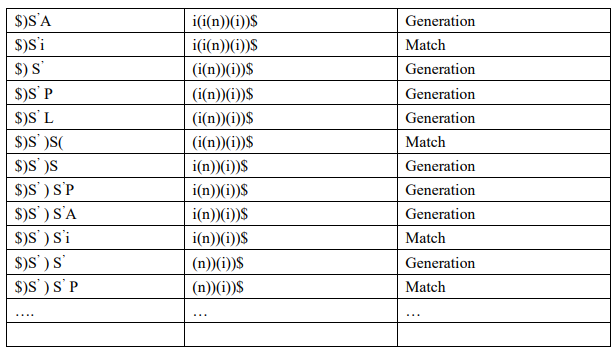
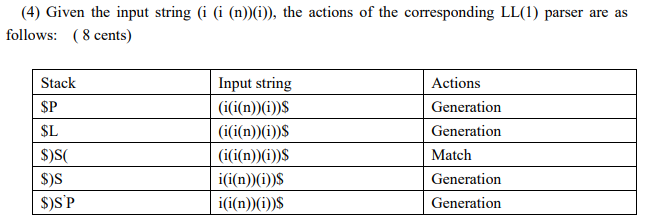
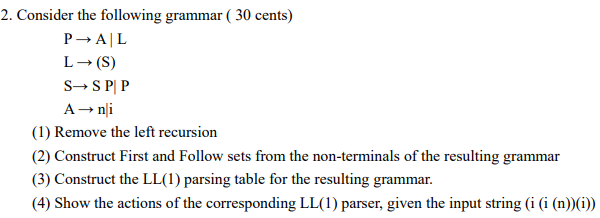


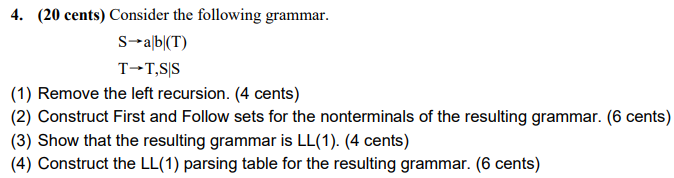


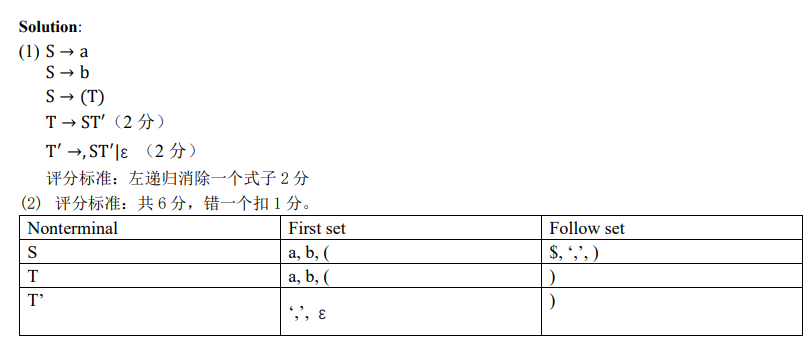
消除左递归：

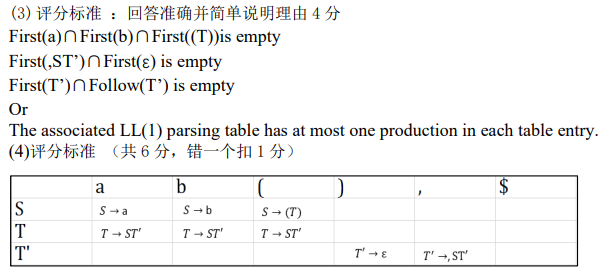






****

****

****

