三、语义分析
(9. 属性文法)

魏恒峰

hfwei@nju.edu.cn

2024 年 04 月 17 日

Regular Expression (词法分析)

Context-Free Grammar (语法分析)

用什么样的文法刻画语言的**语义**?

(类型检查、符号检查)

Donald Knuth (1938 ∼)

(Turing Award, 1974)

Semantics of Context-Free Languages

by

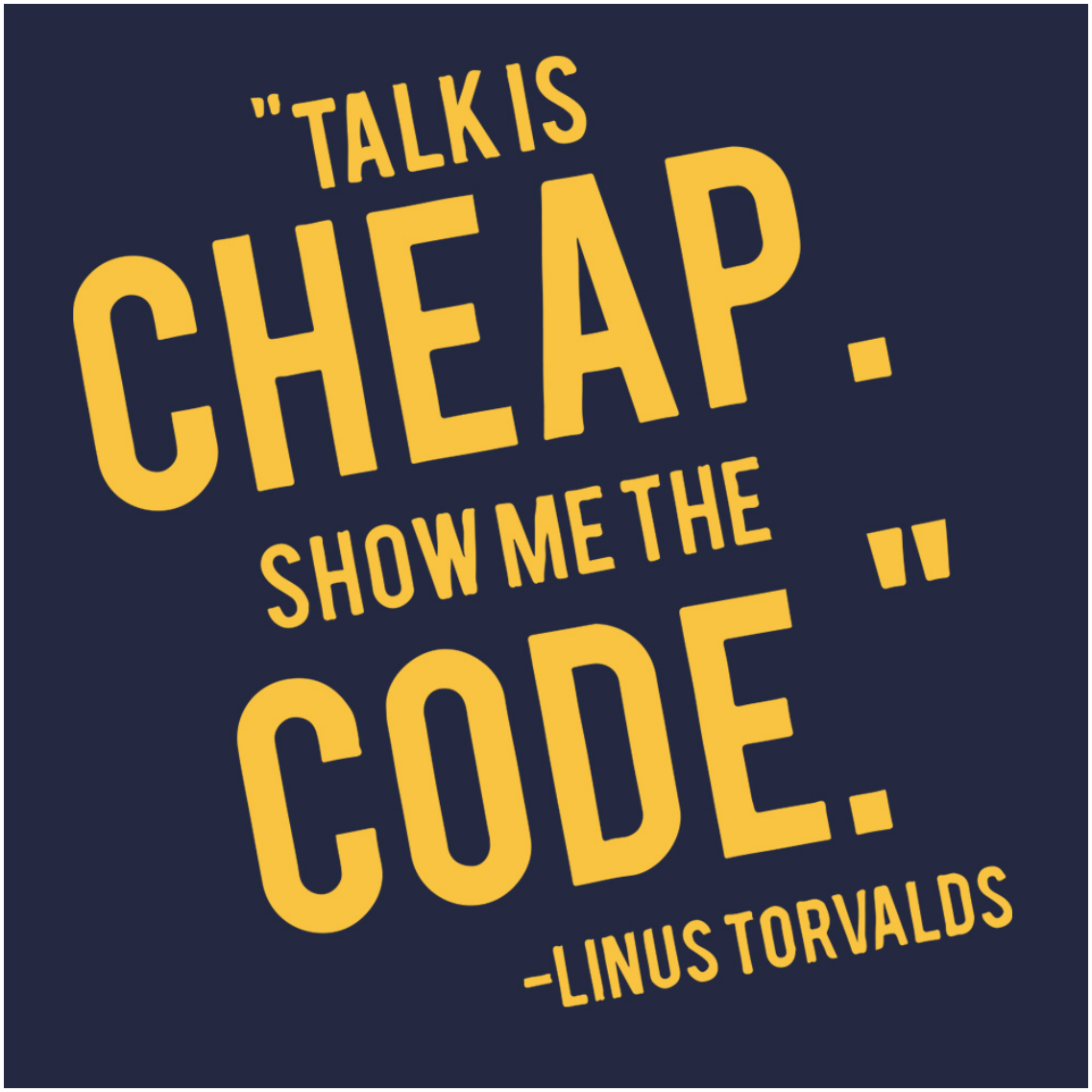
DONALD E. KNUTH

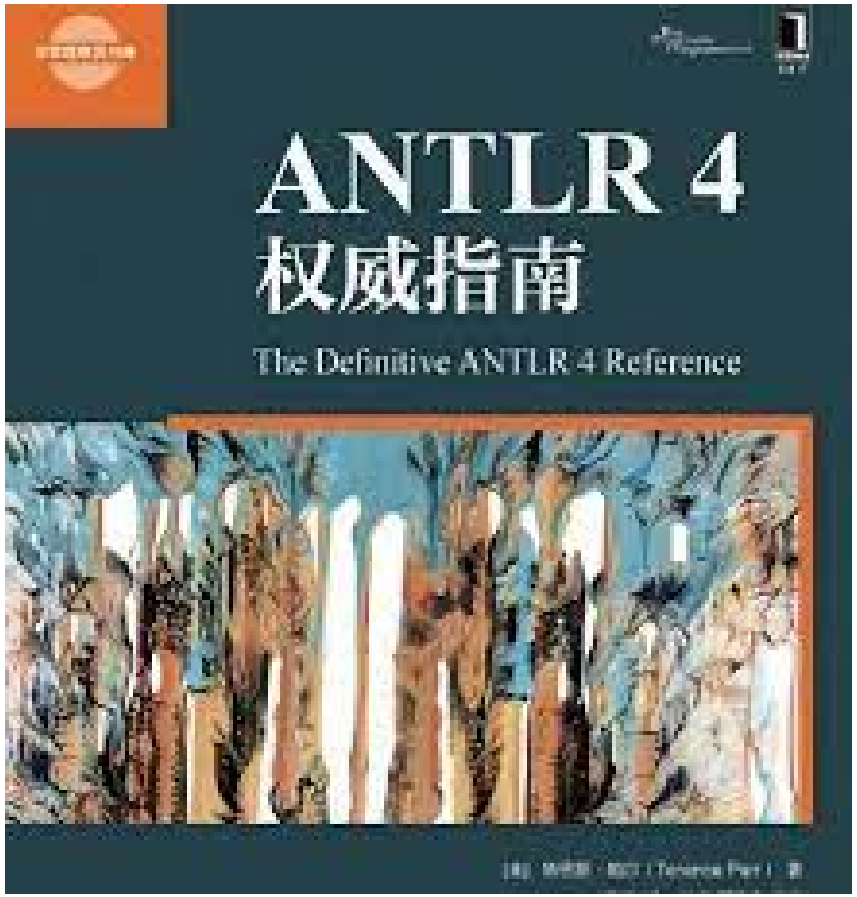
California Institute of Technology

ABSTRACT

"Meaning" may be assigned to a string in a context-free language by defining "attributes" of the symbols in a derivation tree for that string. The attributes can be defined by functions associated with each production in the grammar. This paper examines the implications of this process when some of the attributes are "synthesized", i.e., defined solely in terms of attributes of the *descendants* of the corresponding nonterminal symbol, while other attributes are "inherited", i.e., defined in terms of attributes of the *ancestors* of the nonterminal symbol. An algorithm is given which detects when such semantic rules could possibly lead to circular definition of some attributes. An example is given of a simple programming language defined with both inherited and synthesized attributes, and the method of definition is compared to other techniques for formal specification of semantics which have appeared in the literature.

属性文法 (Attribute Grammar): 为上下文无关文法赋予语义


"TALK IS
CHEAP.
SHOW ME THE
CODE."
-LINUS TORVALDS



第 10 章：属性和动作

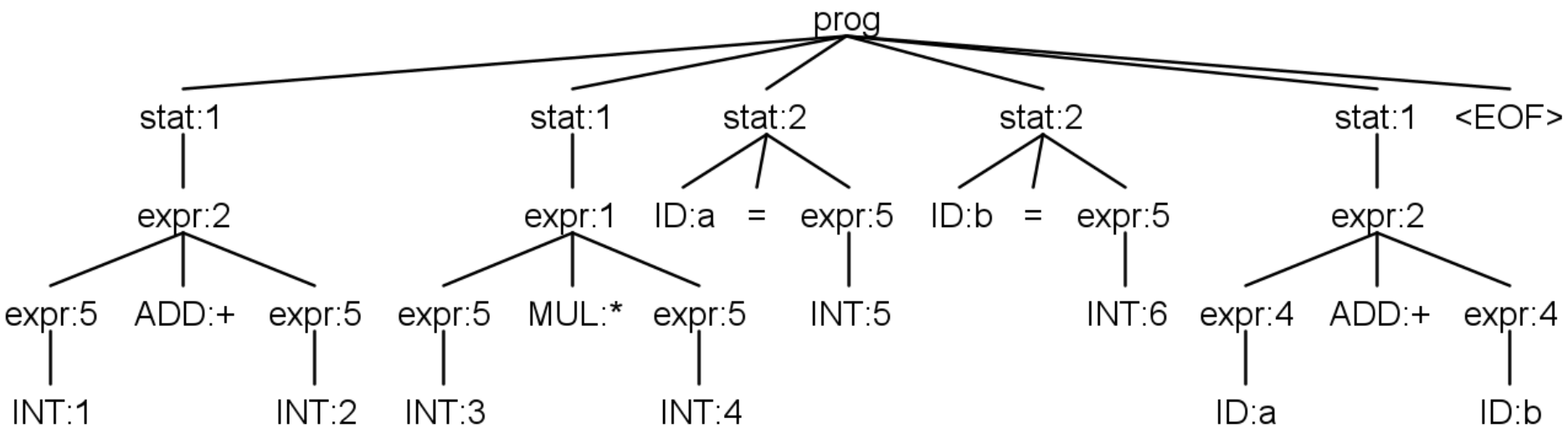
(第 15.4 节：动作和属性)

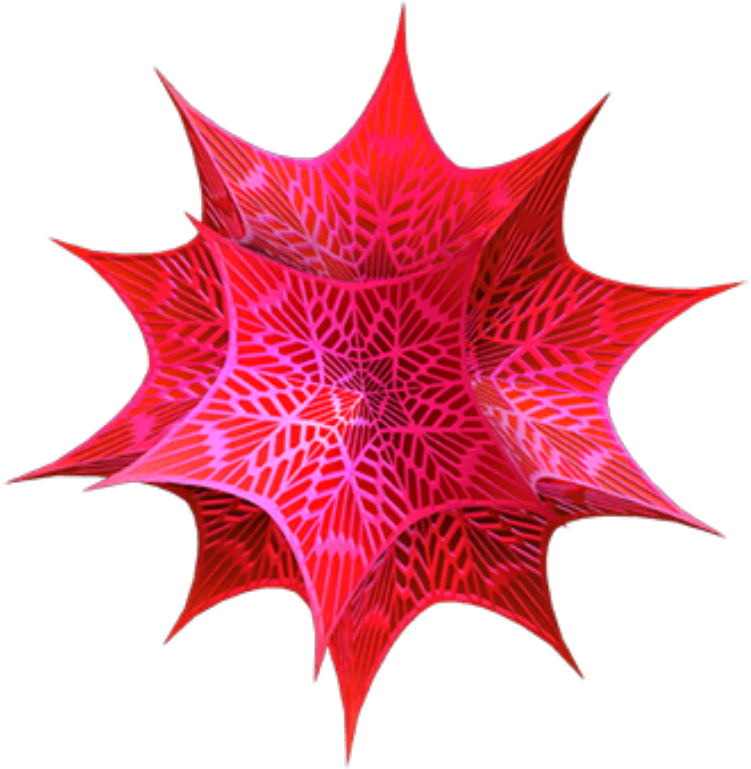
(交互式) 迷你计算器

Expr.g4

```
1 + 2
3 * 4
a = 5
b = 6
a + b
```

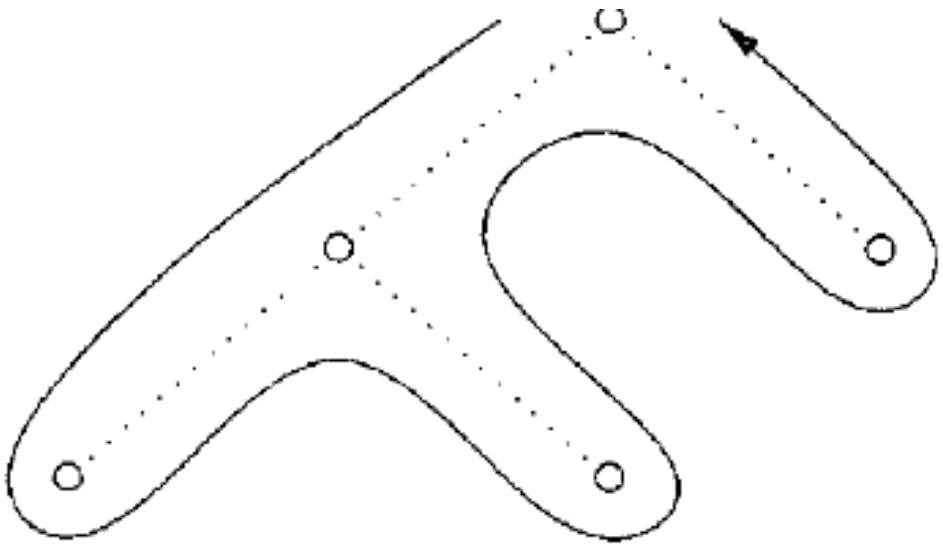
```
3
12
11
```





交互式

Offline 方式计算属性值: 已有语法分析树 (`calc`)



按照**从左到右**的**深度优先**顺序遍历语法分析树

关键: 在合适的时机执行合适的动作，计算相应的属性值

在**语法分析过程中**实现**属性文法**

$$B \to X\{a\}Y$$

语义动作嵌入的位置决定了**何时**执行该动作

基本思想: 一个动作在它**左边的**所有文法符号都**处理**过之后立刻执行

```
prog : stat+ ;

stat : expr          { System.out.println($expr.val); }
     | ID '=' expr   { memory.put($ID.text, $expr.val); }
     ;

expr returns [int val]
    : l = expr op = ('*' | '/') r = expr { $val = eval($l.val, $r.val, $op.type); }
    | l = expr op = ('+' | '-') r = expr { $val = eval($l.val, $r.val, $op.type); }
    | '(' expr ')'                       { $val = $expr.val; }
    | ID                                 { $val = memory.getOrDefault($ID.text, 0); }
    | INT                                { $val = $INT.int; }
    ;
```

ExprAG.g4

(交互式) 迷你计算器

ExprAGMain.java

“源码面前，了无秘密”

ExprAGParser.java

继承属性文法

类型声明文法举例

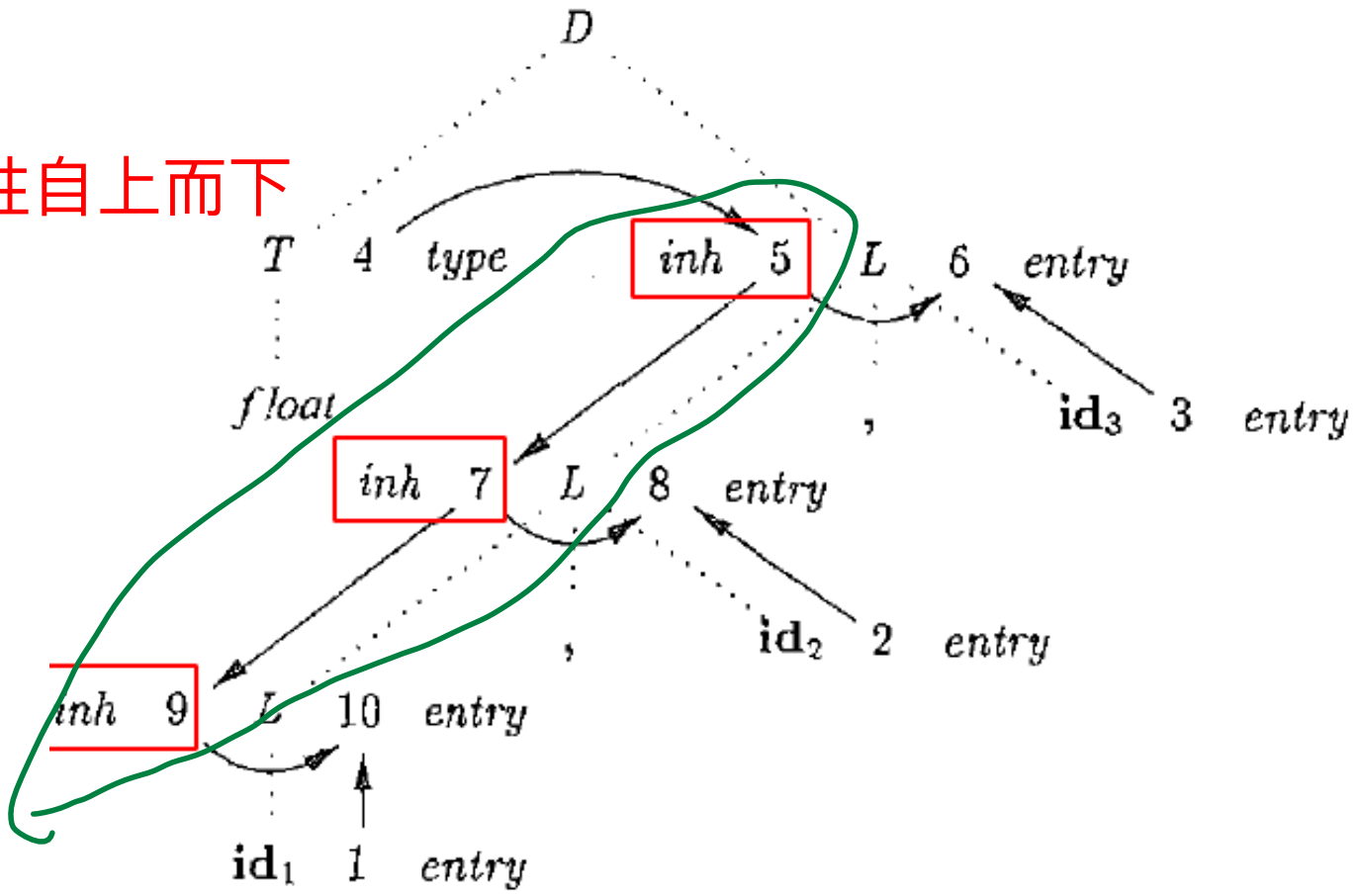
过于简单，无须多言

产生式	语义规则
1) $D \rightarrow T\ L$	$L.inh = T.type$
2) $T \rightarrow \textbf{int}$	$T.type = \text{integer}$
3) $T \rightarrow \textbf{float}$	$T.type = \text{float}$
4) $L \rightarrow L_1\ ,\ \textbf{id}$	$L_1.inh = L.inh$ $addType(\textbf{id}.entry, L.inh)$
5) $L \rightarrow \textbf{id}$	$addType(\textbf{id}.entry, L.inh)$

$\textbf{float id}_1, \textbf{id}_2, \textbf{id}_3$

$L.inh$ 将声明的类型沿着标识符列表向下传递

继承属性自上而下传递



$\textbf{float id}_1, \textbf{id}_2, \textbf{id}_3$

https://stackoverflow.com/q/76062088/1833118

```
7   decl : type vars[$type.text] ;
8   type : 'int'          # IntType
9        | 'float'        # FloatType
10       ;
11  // Unfortunately, ANTLR 4 does not support this:
12  // "rule vars is left recursive but
13  // doesn't conform to a pattern ANTLR can handle"
14  // See https://stackoverflow.com/q/76062088/1833118
15  vars[String type]
16       : left = vars[$type] ',' ID
17           { System.out.println($ID.text + " : " + $type); }
18       | ID
19           { System.out.println($ID.text + " : " + $type); }
20       ;
```

Fortunately, you can rewrite it as `vars : ID (',' ID)*`

Fortunately, you can rewrite it as `vars : ID (',' ID)*`

```
7   decl : type vars[$type.text] ;
8   type : 'int'          # IntType
9        | 'float'        # FloatType
10       ;
11  vars[String typeStr]
12       : ID { System.out.println($ID.text + " : " + $typeStr); }
13         (',' ID { System.out.println($ID.text + " : " + $typeStr); })*
14       ;
```

VarsDeclStarAG.g4

“源码面前，了无秘密”

VarsDeclStarParser.java

Semantics of Context-Free Languages

by

DONALD E. KNUTH

California Institute of Technology

ABSTRACT

"Meaning" may be assigned to a string in a context-free language by defining "attributes" of the symbols in a derivation tree for that string. The attributes can be defined by functions associated with each production in the grammar. This paper examines the implications of this process when some of the attributes are "synthesized", i.e., defined solely in terms of attributes of the *descendants* of the corresponding nonterminal symbol, while other attributes are "inherited", i.e., defined in terms of attributes of the *ancestors* of the nonterminal symbol. An algorithm is given which detects when such semantic rules could possibly lead to circular definition of some attributes. An example is given of a simple programming language defined with both inherited and synthesized attributes, and the method of definition is compared to other techniques for formal specification of semantics which have appeared in the literature.

属性文法 (Attribute Grammar): 为上下文无关文法赋予**语义**

Definition (语法制导定义 (Syntax-Directed Definition; SDD))

SDD 是一个上下文无关文法和**属性**及**规则**的结合。

每个文法符号都可以关联多个属性

产生式	语义规则
1) $L \to E\ \mathbf{n}$	$L.val = E.val$
2) $E \to E_1 + T$	$E.val = E_1.val + T.val$
3) $E \to T$	$E.val = T.val$
4) $T \to T_1 * F$	$T.val = T_1.val \times F.val$
5) $T \to F$	$T.val = F.val$
6) $F \to (\ E\ )$	$F.val = E.val$
7) $F \to \mathbf{digit}$	$F.val = \mathbf{digit}.\mathrm{lexval}$

每个产生式都可以关联一组规则

Definition (语法制导定义 (Syntax-Directed Definition; SDD))

SDD 是一个上下文无关文法和**属性**及**规则**的结合。

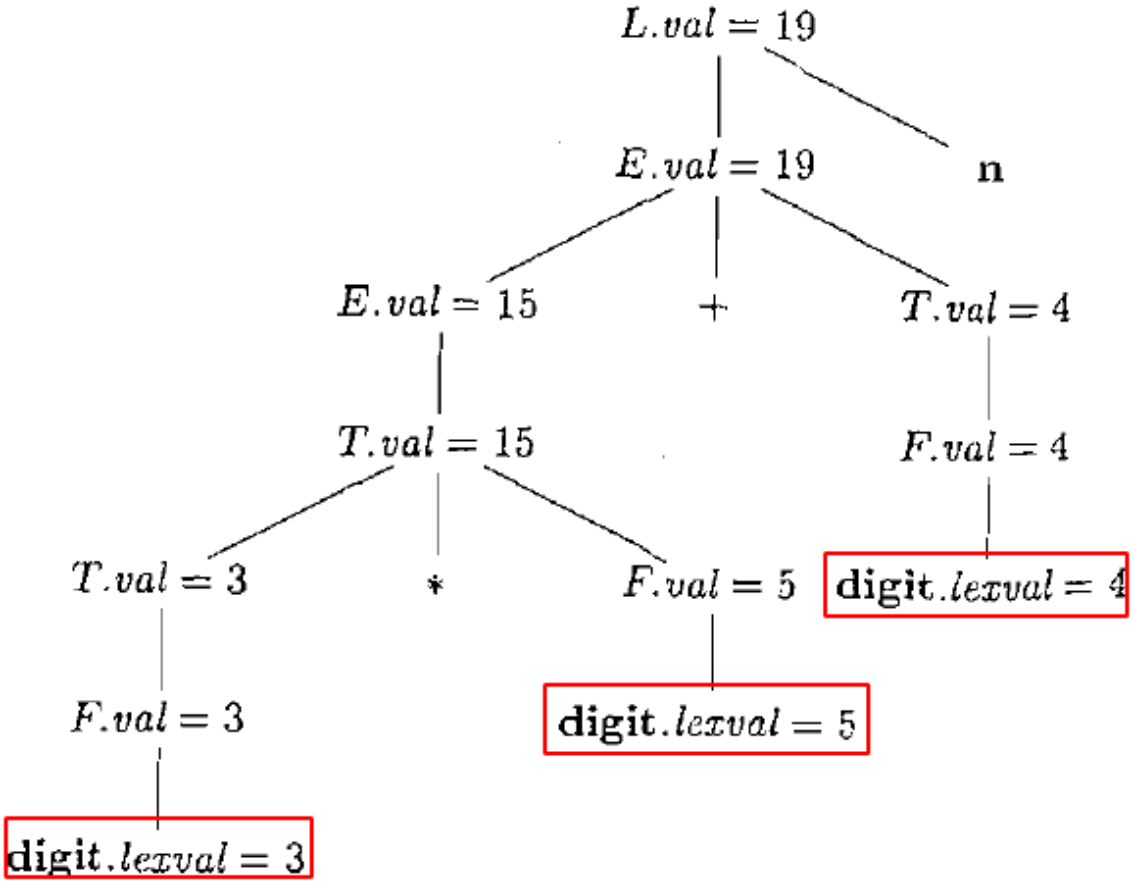
SDD **唯一确定**了语法分析树上每个非终结符节点的属性值

产生式	语义规则
1) $L \to E\ \mathbf{n}$	$L.val = E.val$
2) $E \to E_1 + T$	$E.val = E_1.val + T.val$
3) $E \to T$	$E.val = T.val$
4) $T \to T_1 * F$	$T.val = T_1.val \times F.val$
5) $T \to F$	$T.val = F.val$
6) $F \to (\ E\ )$	$F.val = E.val$
7) $F \to \mathbf{digit}$	$F.val = \mathbf{digit}.\mathrm{lexval}$

SDD **没有**规定以什么方式、什么顺序计算这些属性值

注释 (annotated) 语法分析树: 显示了各个属性值的语法分析树

`ParseTreeProperty<Integer>` `put(ctx, ...)`, `get(ctx, ...)`



$3 * 5 + 4$

Definition (综合属性 (Synthesized Attribute))

节点 N 上的**综合属性**只能通过 N 的子节点或 N 本身的属性来定义。

产生式	语义规则
1) $L \to E\ \mathbf{n}$	$L.val = E.val$
2) $E \to E_1 + T$	$E.val = E_1.val + T.val$
3) $E \to T$	$E.val = T.val$
4) $T \to T_1 * F$	$T.val = T_1.val \times F.val$
5) $T \to F$	$T.val = F.val$
6) $F \to (\ E\ )$	$F.val = E.val$
7) $F \to \mathbf{digit}$	$F.val = \mathbf{digit}.\mathrm{lexval}$

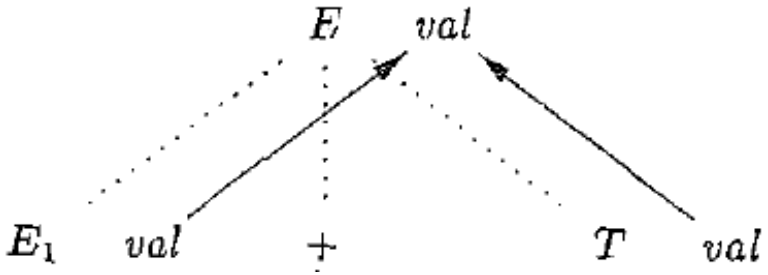
Definition (S 属性定义 (S-Attributed Definition))

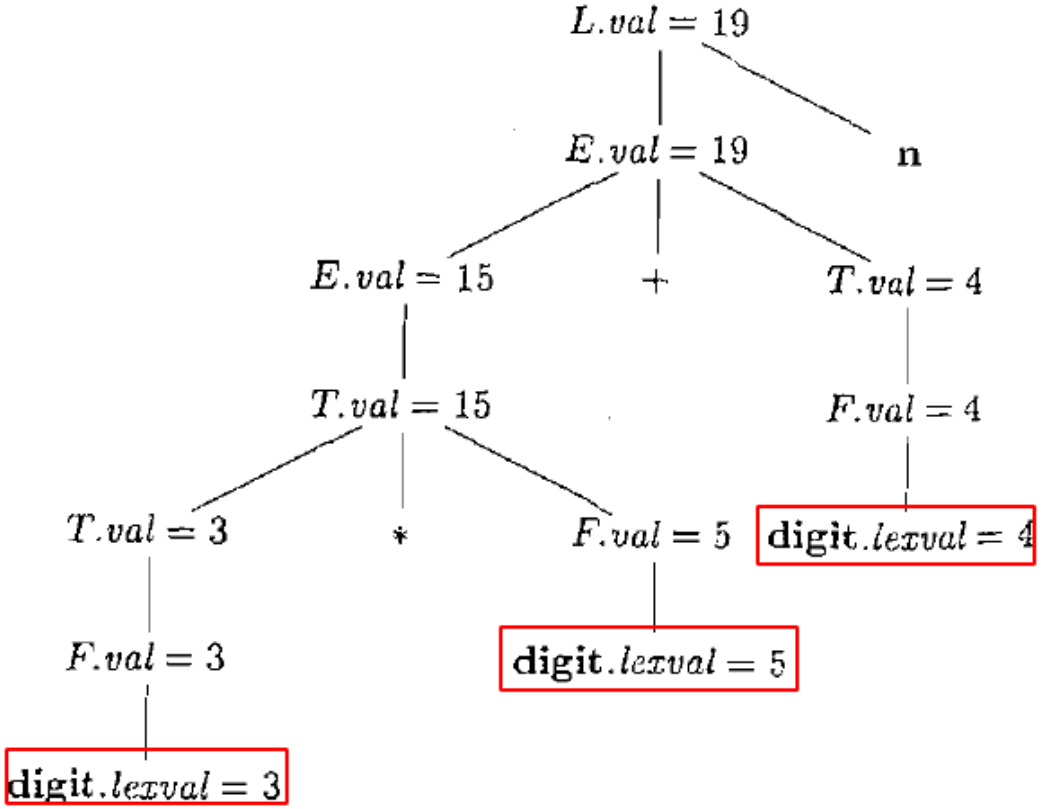
如果一个 SDD 的每个属性都是综合属性, 则它是 S 属性定义。

依赖图用于确定一棵给定的语法分析树中各个属性实例之间的依赖关系

产生式	语义规则
1) $L \to E\ \mathrm{n}$	$L.val = E.val$
2) $E \to E_1 + T$	$E.val = E_1.val + T.val$
3) $E \to T$	$E.val = T.val$
4) $T \to T_1 * F$	$T.val = T_1.val \times F.val$
5) $T \to F$	$T.val = F.val$
6) $F \to (\ E\ )$	$F.val = E.val$
7) $F \to \mathbf{digit}$	$F.val = \mathbf{digit}.lexval$

综合属性自底向上进行传递





S **属性定义**的依赖图刻画了属性实例之间**自底向上**的信息流动

S **属性定义**的依赖图描述了属性实例之间**自底向上**的信息流

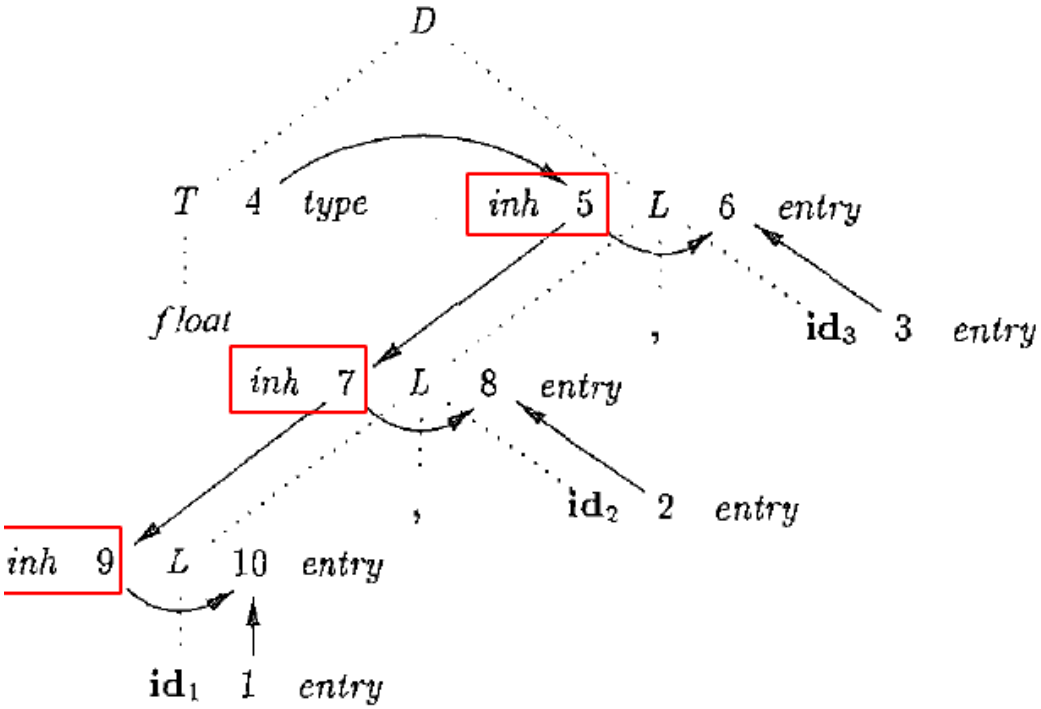
此类属性值的计算可以在自顶向下的 LL 语法分析**过程中**实现

在 LL 语法分析器中，递归下降函数 A **返回** 时，
计算相应节点 A 的综合属性值

Definition (继承属性 (Inherited Attribute))

节点 N 上的**继承属性**只能通过N 的父节点、N 本身和 N 的兄弟节点上的属性来定义。

产生式	语义规则
1) $D \to T\ L$	$L.inh = T.type$
2) $T \to \mathbf{int}$	$T.type = \text{integer}$
3) $T \to \mathbf{float}$	$T.type = \text{float}$
4) $L \to L_1\ ,\ \mathbf{id}$	$L_1.inh = L.inh$ $addType(\mathbf{id}.entry, L.inh)$
5) $L \to \mathbf{id}$	$addType(\mathbf{id}.entry, L.inh)$



float $\mathbf{id}_1, \mathbf{id}_2, \mathbf{id}_3$

继承信息流向: 先从左向右、从上到下传递信息

Definition (L 属性定义 (L-Attributed Definition))

如果一个 SDD 的每个属性

(1) 要么是综合属性,

(2) 要么是继承属性, 但是它的规则满足如下限制:
对于产生式 $A \to X_1 X_2 \ldots X_n$ 及其对应规则定义的继承属性 $X_i.a$,
则这个规则只能使用

(a) 和**产生式头 A**关联的**继承**属性;

(b) 位于**X_i 左边**的文法符号实例 X_1、X_2、...、X_{i-1} 相关的**继承**属性或**综合**属性;

(c) 和**这个 X_i 的实例本身**相关的继承属性或综合属性, 但是在由这个 X_i 的全部属性组成的依赖图中**不存在环**。

则它是 L 属性定义。

语法分析树上的有序信息流动

Learn by Examples

Definition (后缀表示 (Postfix Notation))

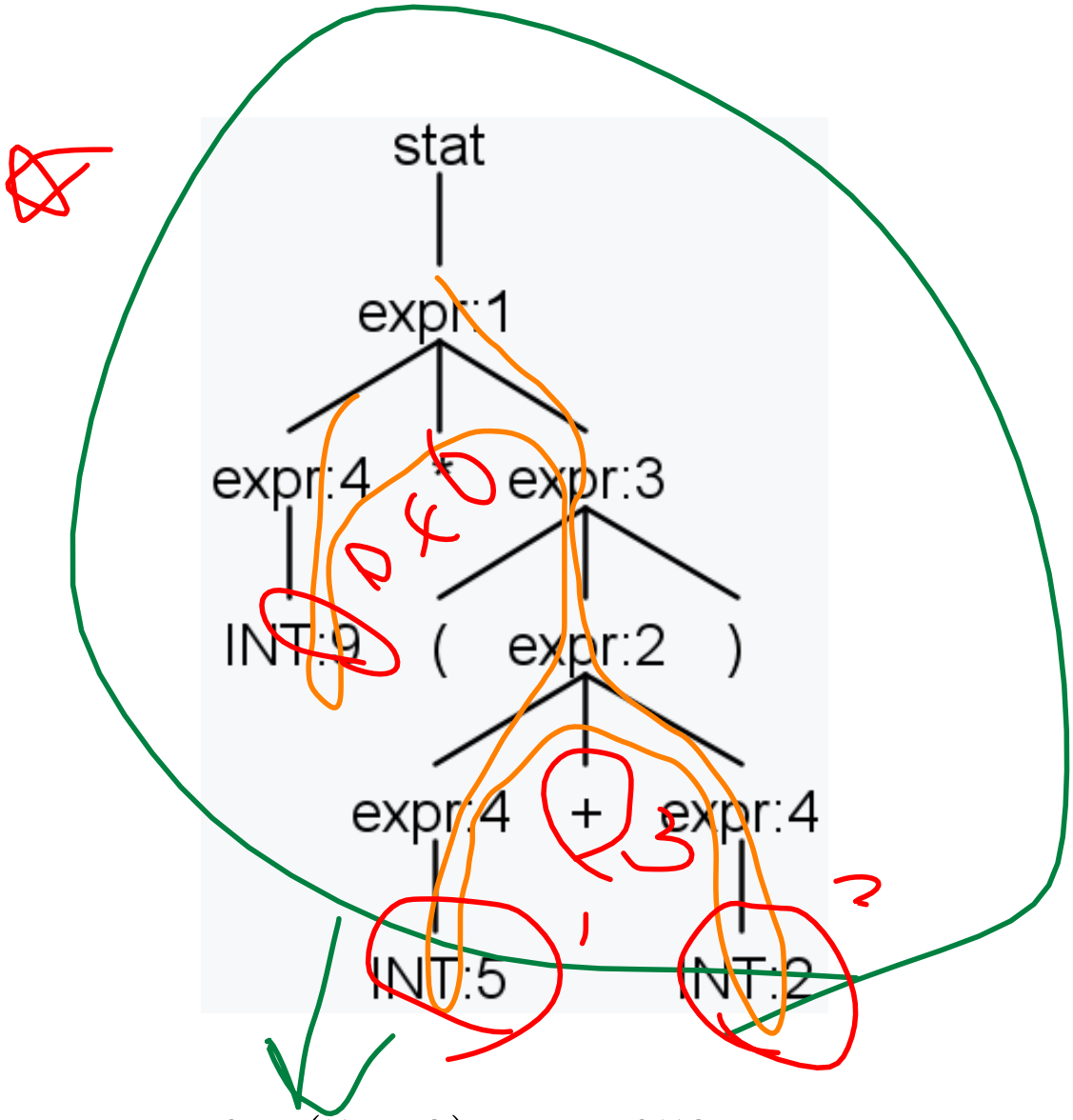
(1) 如果 E 是一个 **变量或常量**, 则 E 的后缀表示是 E 本身;

(2) 如果 E 是形如 E_1 **op** E_2 的表达式, 则 E 的后缀表示是 $E_1' E_2'$**op**, 这里 E_1' 和 E_2' 分别是 E_1 与 E_2 的后缀表达式;

(3) 如果 E 是形如 (E_1) 的表达式, 则 E 的后缀表示是 E_1 的后缀表示。

$$(9 * 5) + 2 \implies 95 * 2+$$

$$9 * (5 + 2) \implies 952 + *$$



$$9 * (5 + 2) \implies 952 + *$$

后缀表达式 S 属性定义

```
7   stat : expr { System.out.println($expr.postfix); }
8        ;
9
10  expr returns [String postfix]
11    : l = expr op = '*' r = expr { $postfix = $l.postfix + $r.postfix + $op.text; }
12    | l = expr op = '+' r = expr { $postfix = $l.postfix + $r.postfix + $op.text; }
13    | '(' expr ')'               { $postfix = $expr.postfix; }
14    | INT                        { $postfix = $INT.text; }
15    ;
```

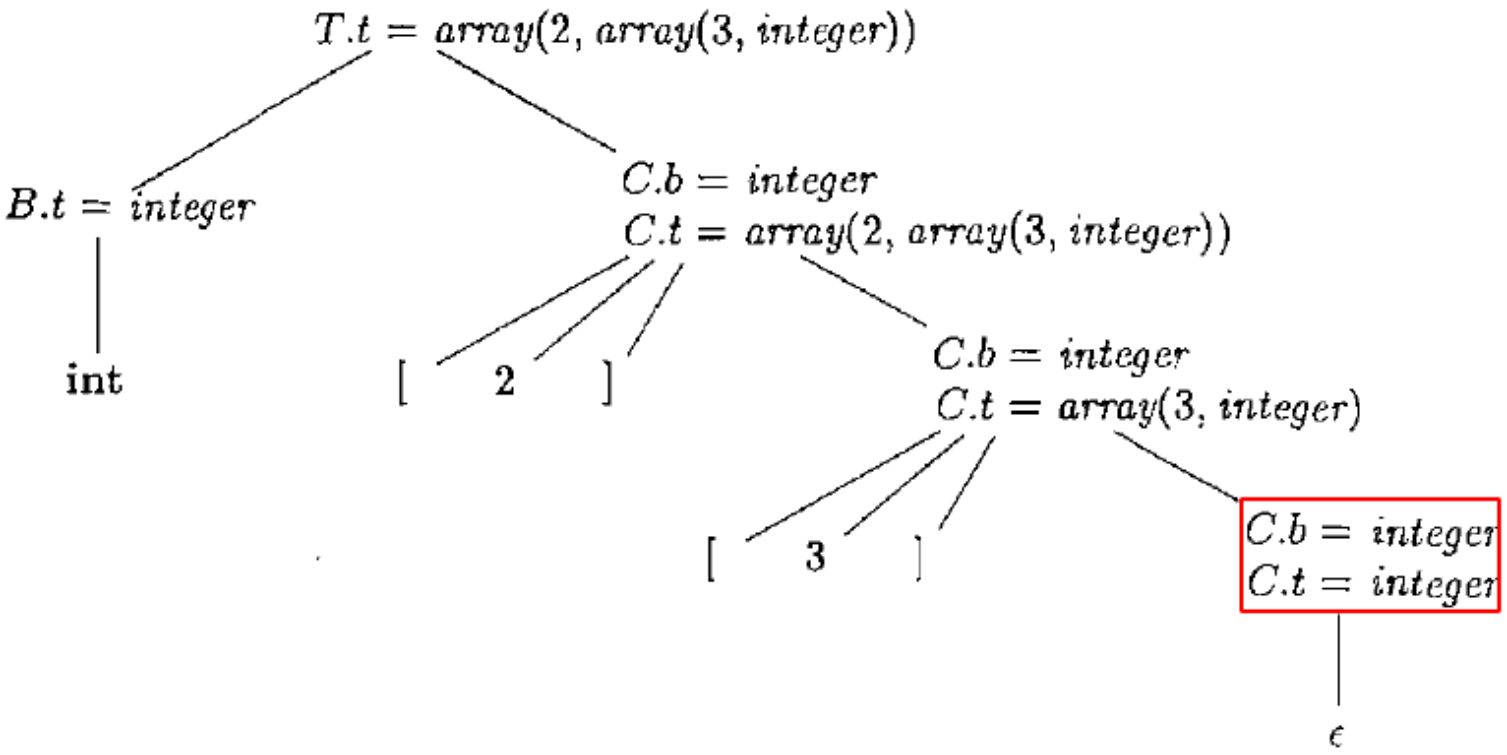
数组类型文法举例

产生式
$T \rightarrow B\ C$
$B \rightarrow \mathbf{int}$
$B \rightarrow \mathbf{float}$
$C \rightarrow [\ \mathbf{num}\ ]\ C_1$
$C \rightarrow \epsilon$

int[2][3]

类型表达式 $(2, (3, \mathrm{int}))$

继承属性 $C.b$ 将一个基本类型沿着树向下传播

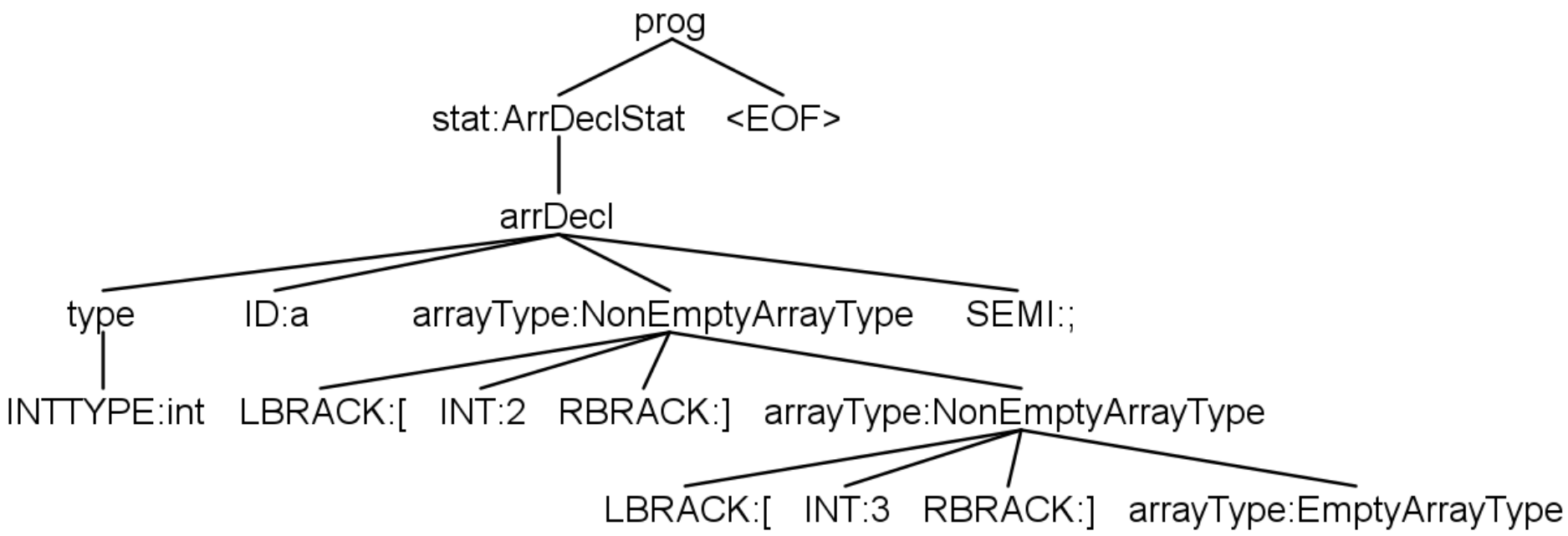


$\mathbf{int}[2][3]$ $\quad$ $(2, (3, \mathrm{int}))$

综合属性 $C.t$ 收集最终得到的类型表达式

```
// OR: type ID ('[' INT ']')* ';'
arrDecl : basicType ID arrayType ';' ;
arrayType : '[' INT ']' arrayType      # NonEmptyArrayType
          |                            # EmptyArrayType
          ;

basicType : 'int' | 'double' ;
```

ArrayType.g4


prog
stat:ArrDeclStat
<EOF>
arrDecl
type
ID:a
arrayType:NonEmptyArrayType
SEMI:;
INTTYPE:int
LBRACK:[
INT:2
RBRACK:]
arrayType:NonEmptyArrayType
LBRACK:[
INT:3
RBRACK:]
arrayType:EmptyArrayType

int[2][3]

```
// OR: type ID ('[' INT ']')* ';'
arrDecl : basicType ID arrayType[$basicType.text]
  { System.out.println($ID.text + " : " + $arrayType.array_type); } ';' ;

arrayType[String basic_type]
    returns [String array_type]
  : '[' INT ']' arrayType[$basic_type]
      { $array_type = "(" + $INT.int + ", " + $arrayType.array_type + ")"; }
  |  { $array_type = $basic_type; }
  ;
```

ArrayTypeAG.g4

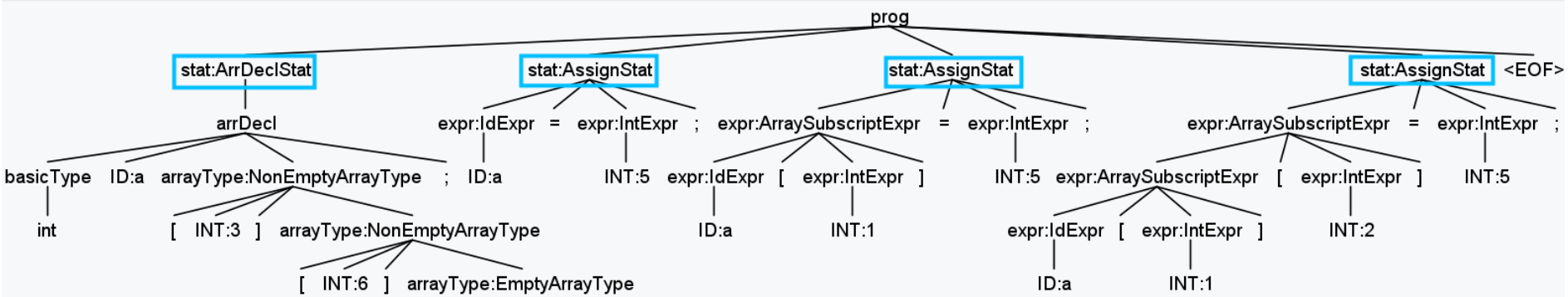
数组声明、数组引用、类型检查

```
prog : stat* EOF ;

stat : varDecl                          # VarDeclStat
     | arrDecl                          # ArrDeclStat
     | lhs = expr '=' rhs = expr ';'    # AssignStat
     ;

varDecl : basicType ID ';' ;
basicType : 'int' | 'double' ;

// OR: type ID ('[' INT ']')* ';'
arrDecl : basicType ID arrayType ';' ;
arrayType : '[' INT ']' arrayType       # NonEmptyArrayType
          |                             # EmptyArrayType
          ;

expr: primary = expr '[' subscript = expr ']'   # ArraySubscriptExpr
    | ID                                        # IdExpr
    | INT                                       # IntExpr
    ;
```

Array.g4

```
int a[3][6];
a = 5;
a[1] = 5;
a[1][2] = 5;
```

```
a : (3, (6, int))
(3, (6, int)) <=> int
 (6, int) <=> int
 int <=> int
```



```
stat : varDecl
        { String id = $varDecl.ctx.ID().getText();
          System.out.println(id + " : " + typeMap.get(id)); }
     | arrDecl
        { String id = $arrDecl.ctx.ID().getText();
          System.out.println(id + " : " + typeMap.get(id)); }
     | lhs = expr '=' rhs = expr ';'
        {  System.out.println($lhs.type + " <=> " + $rhs.type); }
     ;
```

ArrayAG.g4

```
varDecl : basicType ID ';'
    { typeMap.put($ID.text, $basicType.text); };
basicType : 'int' | 'float' ;

// OR: type ID ('[' INT ']')* ';'
arrDecl : basicType ID arrayType[$basicType.text]
  { typeMap.put($ID.text, $arrayType.array_type); }
  ';' ;
arrayType[String basic_type]
    returns [String array_type]
  : '[' INT ']' arrayType[$basic_type]
      { $array_type = "(" + $INT.int + ", " + $arrayType.array_type + ")"; }
  |  { $array_type = $basic_type; }
  ;
```

ArrayAG.g4

```
39  expr returns [String type]
40      : ID { String expr_type = typeMap.get($ID.text); }
41        ('[' INT ']'
42            {
43                int start = expr_type.indexOf(',');
44                int end = expr_type.lastIndexOf(')');
45                expr_type = expr_type.substring(start + 1, end);
46            }
47        )* { $type = expr_type; }
48      | ID { $type = typeMap.get($ID.text); }
49      | INT { $type = "int"; }
50      ;
```

ArrayAG.g4

```
TypeCheckingListener
symbolTable        Map<String, VariableSymbol>
basicTypeProperty  ParseTreeProperty<Type>
arrayTypeProperty  ParseTreeProperty<Type>
```

TypeCheckingListener.java

What is the difference between ANTLR 3 and 4?

Another big difference is that we discourage the use of actions directly within the grammar because ANTLR 4 automatically generates listeners and visitors for you to use that trigger method calls when some phrases of interest are recognized during a tree walk after parsing. See also Parse Tree Matching and XPath.

Offline (`ParseTreeWalker`)

Online (Attribute Grammar)

Online (`addParseListener`)

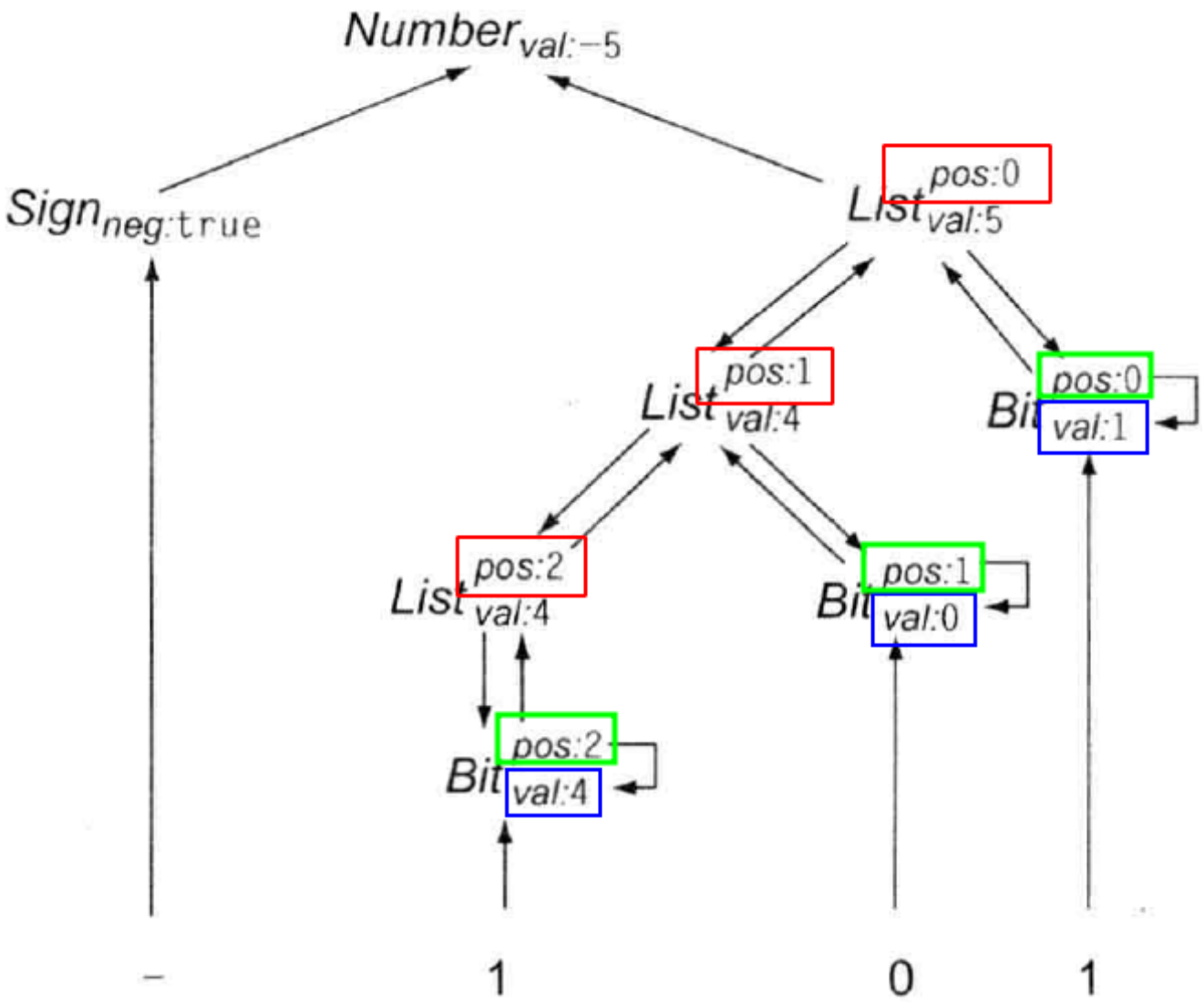
Q: What are the main design decisions in ANTLR4?

Ease-of-use over performance. I will worry about performance later. Simplicity over complexity. For example, I have taken out explicit/manual AST construction facilities and the tree grammar facilities. For 20 years I've been trying to get people to go that direction, but I've since decided that it was a mistake. It's much better to give people a parser generator that can automatically build trees and then let them use pure code to do whatever tree walking they want. People are extremely familiar and comfortable with visitors, for example.

有符号二进制数文法

$$
P = \left\{
\begin{array}{lcl}
Number & \rightarrow & Sign\ List \\
Sign & \rightarrow & + \\
 & | & - \\
List & \rightarrow & List\ Bit \\
 & | & Bit \\
Bit & \rightarrow & 0 \\
 & | & 1
\end{array}
\right\}
\qquad
\begin{array}{lcl}
T & = & \{+, -, 0, 1\} \\
NT & = & \{Number, Sign, List, Bit\} \\
S & = & \{Number\}
\end{array}
$$

$$-101_2 = -5_{10}$$


Number val:−5
Sign neg:true
List pos:0 val:5
List pos:1 val:4
Bit pos:0 val:1
List pos:2 val:4
Bit pos:1 val:0
Bit pos:2 val:4
−
1
0
1

$$-101_2 = -5_{10}$$

有符号二进制数 L 属性定义

	产生式	属性规则
1	$Number \rightarrow Sign\ List$	$List.position \leftarrow 0$ if $Sign.negative$ then $Number.value \leftarrow -List.value$ else $Number.value \leftarrow List.value$
2	$Sign \rightarrow +$	$Sign.negative \leftarrow false$
3	$Sign \rightarrow -$	$Sign.negative \leftarrow true$
4	$List \rightarrow Bit$	$Bit.position \leftarrow List.position$ $List.value \leftarrow Bit.value$
5	$List_0 \rightarrow List_1\ Bit$	$List_1.position \leftarrow List_0.position+1$ $Bit.position \leftarrow List_0.position$ $List_0.value \leftarrow List_1.value+Bit.value$
6	$Bit \rightarrow 0$	$Bit.value \leftarrow 0$
7	$Bit \rightarrow 1$	$Bit.value \leftarrow 2^{Bit.position}$

Definition (语法制导的翻译方案 (Syntax-Directed Translation Scheme; SDT))

SDT 是在其产生式体中嵌入**语义动作**的上下文无关文法。

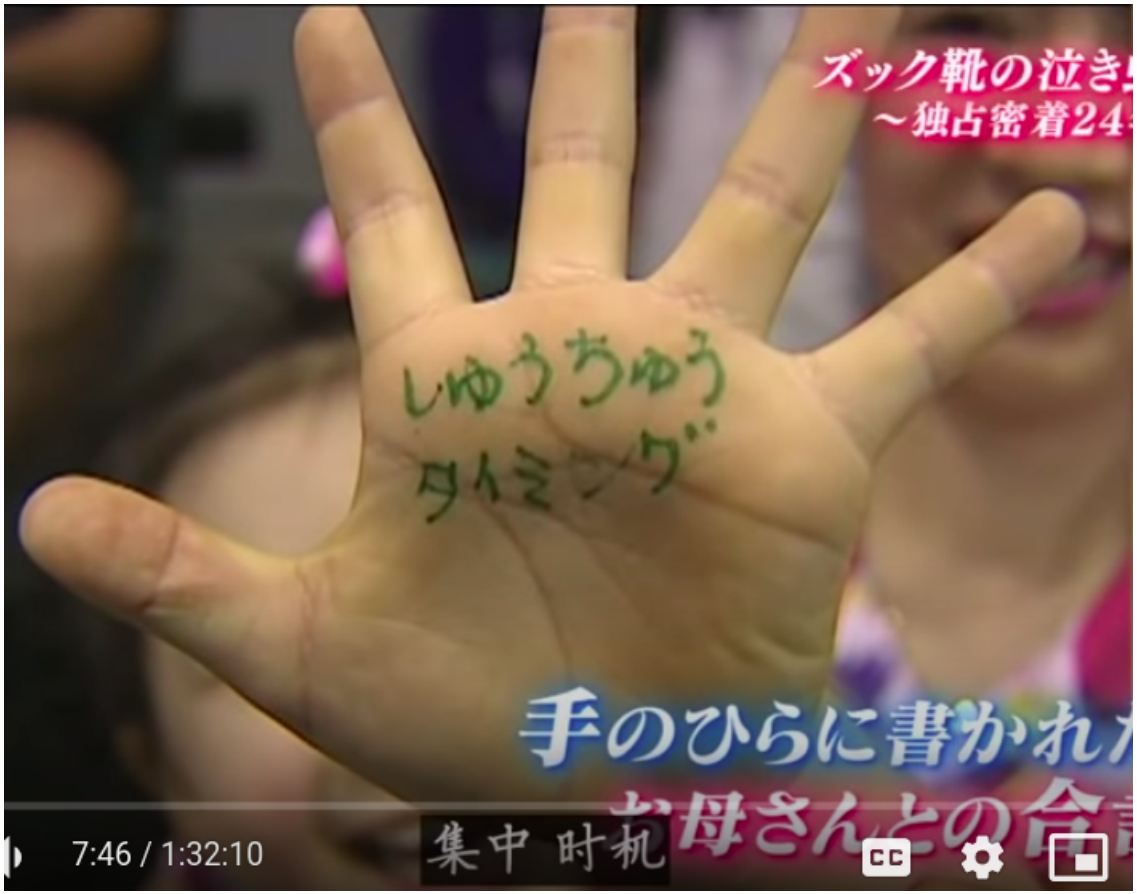
产生式	语义规则
1) $L \to E\ \mathbf{n}$	$L.val = E.val$
2) $E \to E_1 + T$	$E.val = E_1.val + T.val$
3) $E \to T$	$E.val = T.val$
4) $T \to T_1 * F$	$T.val = T_1.val \times F.val$
5) $T \to F$	$T.val = F.val$
6) $F \to (\ E\ )$	$F.val = E.val$
7) $F \to \mathbf{digit}$	$F.val = \mathbf{digit}.\mathrm{lexval}$

L	$\to$	$E\ \mathbf{n}$	$\{\ \mathrm{print}(E.val);\ \}$
E	$\to$	$E_1 + T$	$\{\ E.val = E_1.val + T.val;\ \}$
E	$\to$	T	$\{\ E.val = T.val;\ \}$
T	$\to$	$T_1 * F$	$\{\ T.val = T_1.val \times F.val;\ \}$
T	$\to$	F	$\{\ T.val = F.val;\ \}$
F	$\to$	$(\ E\ )$	$\{\ F.val = E.val;\ \}$
F	$\to$	$\mathbf{digit}$	$\{\ F.val = \mathbf{digit}.lexval;\ \}$

语义动作

Q：如何将带有**语义规则**的 SDD 转换为带有**语义动作**的 SDT

时机 (Timing; タイミング)



语义动作嵌入在什么地方？这决定了何时执行语义动作。

S 属性定义

产生式	语义规则
1) $L \to E\ \mathrm{n}$	$L.val = E.val$
2) $E \to E_1 + T$	$E.val = E_1.val + T.val$
3) $E \to T$	$E.val = T.val$
4) $T \to T_1 * F$	$T.val = T_1.val \times F.val$
5) $T \to F$	$T.val = F.val$
6) $F \to (\ E\ )$	$F.val = E.val$
7) $F \to \mathbf{digit}$	$F.val = \mathbf{digit}.\mathrm{lexval}$

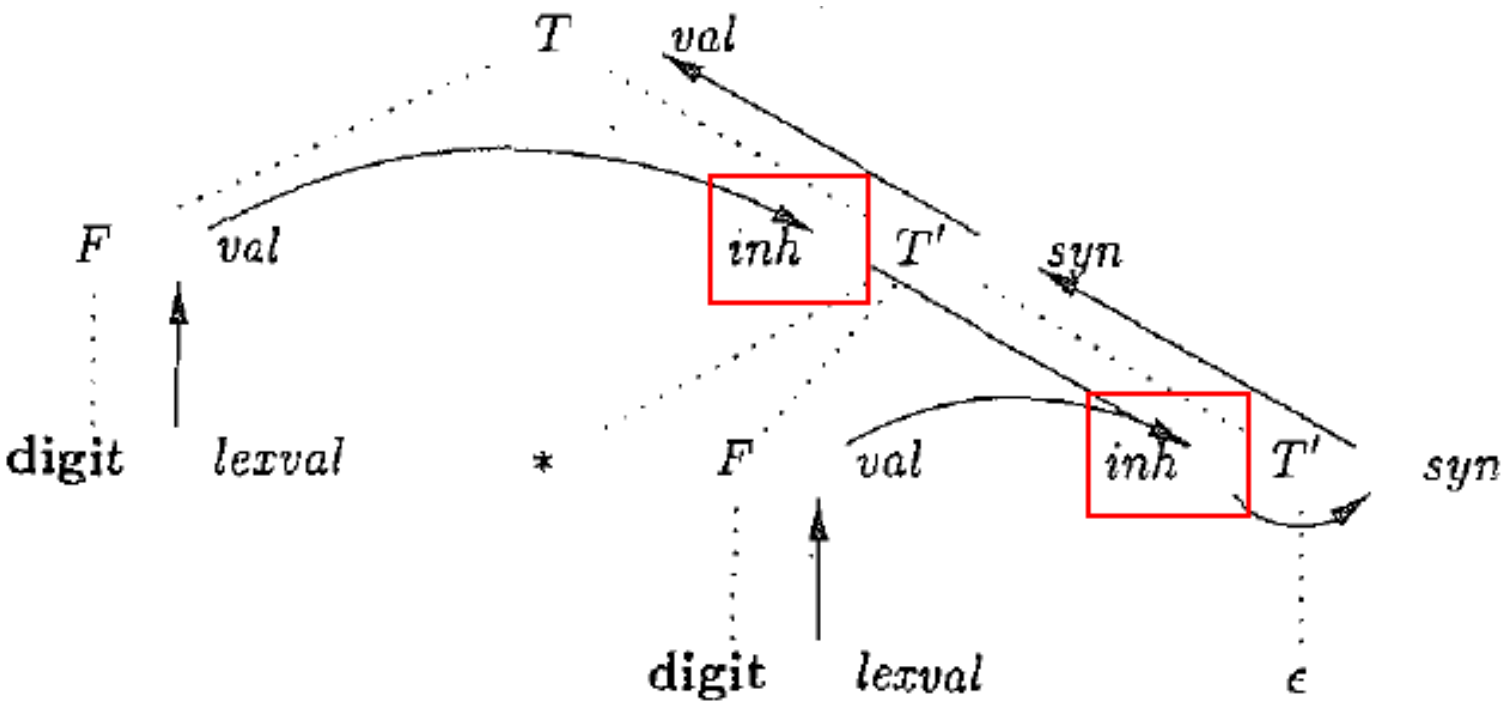
后缀翻译方案

L	$\to$	$E\ \mathrm{n}$	$\{\ \mathrm{print}(E.val);\ \}$
E	$\to$	$E_1 + T$	$\{\ E.val = E_1.val + T.val;\ \}$
E	$\to$	T	$\{\ E.val = T.val;\ \}$
T	$\to$	$T_1 * F$	$\{\ T.val = T_1.val \times F.val;\ \}$
T	$\to$	F	$\{\ T.val = F.val;\ \}$
F	$\to$	$(\ E\ )$	$\{\ F.val = E.val;\ \}$
F	$\to$	$\mathbf{digit}$	$\{\ F.val = \mathbf{digit}.lexval;\ \}$

语义动作

后缀翻译方案: 所有动作都在产生式的最后

L 属性定义 与 LL 语法分析



$3 * 5$

$$A \to X_1 \cdots X_i \cdots X_n$$

原则: **从左到右**处理各个 X_i 符号

对每个 X_i, 先计算**继承属性**, 后计算**综合属性**

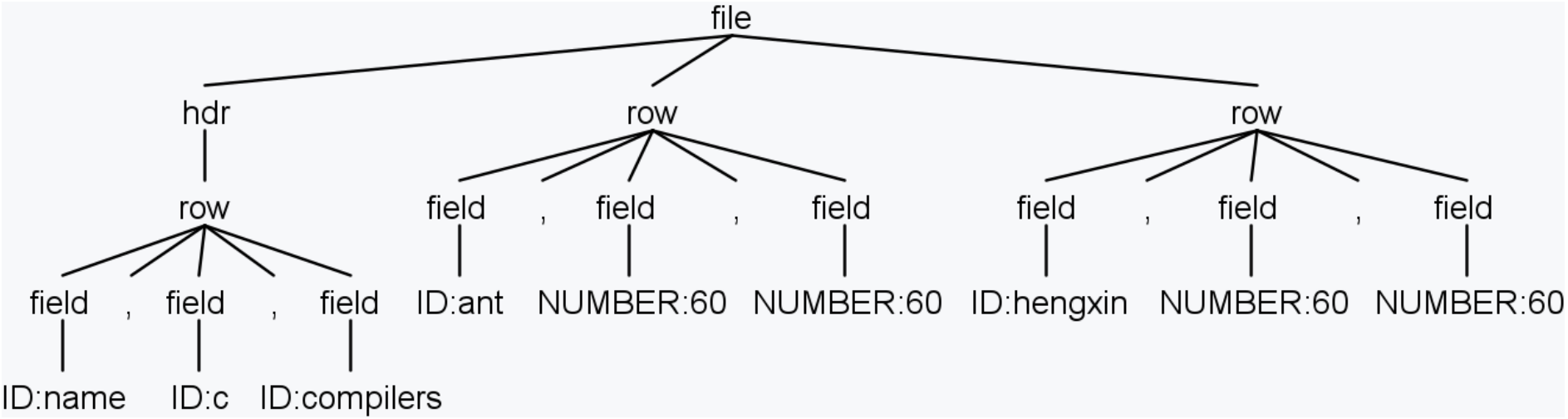
递归下降子过程 $A \to X_1 \cdots X_i \cdots X_n$

- 在调用 X_i 子过程之前，计算 X_i 的**继承属性**
  - 以 X_i 的继承属性为**参数**调用 X_i 子过程
- 在 X_i 子过程返回之前，计算 X_i 的**综合属性**
  - 在 X_i 子过程结束时**返回** X_i 的综合属性

Comma-Separated Values

```
name, c, compilers
ant, 60, 60
hengxin, 60, 60
```

```
header: name,c,compilers
{c=60, name=ant, compilers=60}
{c=60, name=hengxin, compilers=60}
Totally 2 rows
Row token interval: 5..9
Row token interval: 10..14
```

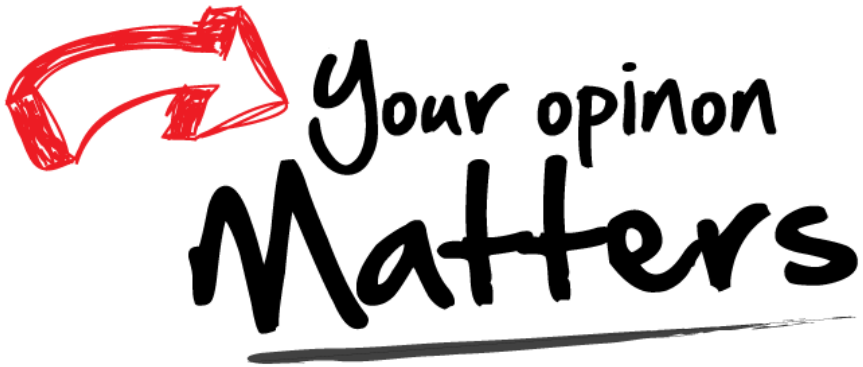



```
file : hdr row+ ;
hdr : row ;
row : field (',' field)* ;
```

```
file
locals [int i = 0]
  : hdr
    { System.out.println("header: " + $hdr.text); }
    (rs += row[$hdr.text.split(",")] { $i++; System.out.println($row.vals); } )+
    { System.out.println("Totally " + $i + " rows");
      for (RowContext r : $rs) {
        System.out.println("Row token interval: " + r.getSourceInterval());
      }
    };
hdr : row[null] ;
row[String[] columns] returns [Map<String, String> vals = new HashMap<String, String>()]
  locals [int col = 0]
    : field { if ($columns != null) $vals.put(columns[$col++], $field.text); }
      (',' field { if ($columns != null) $vals.put(columns[$col++], $field.text); } )* ;
```

CSVAG.g4

Thank You!



Office 926

hfwei@nju.edu.cn