System Software Crash Couse

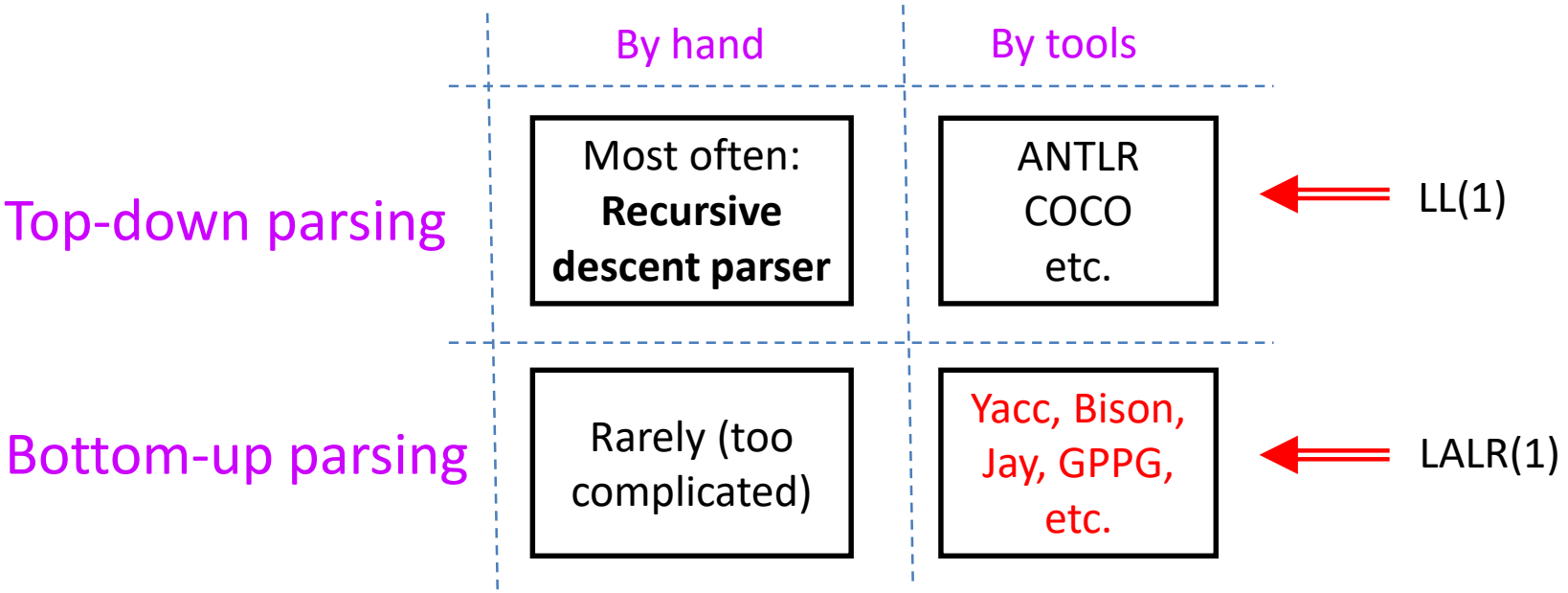
Samsung Research Russia
Moscow 2019

Block C Compiler Construction
6-7. YACC & Bootstrapping

Eugene Zouev

Automatic parser generation

- Top-down or bottom-up parsing?
- «Hand-made» or automated development?



Yacc/Bison & clones

- **YACC** – Yet another compiler compiler 1970: based on C.
- **Bison** – Yacc version for для GNU: based on C.
- **GPPG** – Gardens Point Parser Generator: Yacc version for C# and .NET.
- **Jay** – Yacc version for Java.
- …A lot of YACC clones for almost all popular languages.

All YACCs have **identical** parsing algorithm.

Yacc/Bison: references (Russian)

YACC - Yet Another Compiler Compiler

http://yacc.solotony.com/yacc_rus/index.html

Перевод оригинальной статьи (так себе, но понятно)

Компилятор компиляторов Bison – первое знакомство

http://trpl.narod.ru/CC_Bison.htm

Bison - Генератор синтаксических анализаторов, совместимый с YACC

http://www.opennet.ru/docs/RUS/bison_yacc/bison_1.html

Перевод официального руководства GNU

Lex и YACC в примерах

http://rus-linux.net/lib.php?name=/MyLDP/algol/lex-yacc-howto.html

Gardens Point Parser Generator

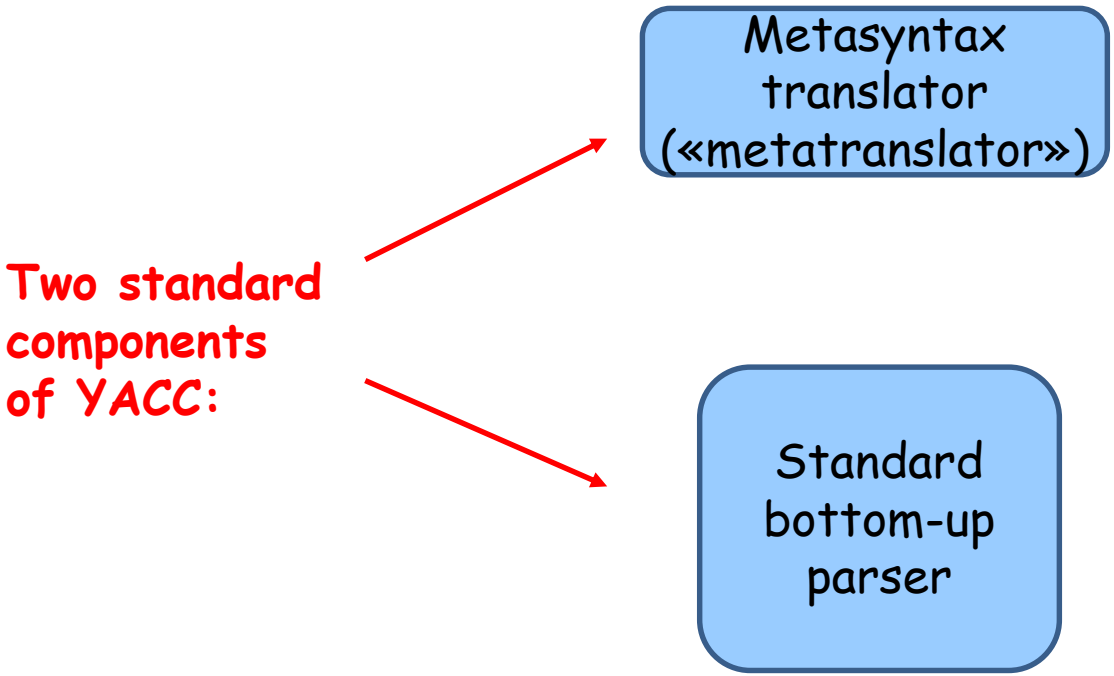
YACC-совместимый генератор для C#; http://gppg.codeplex.com/

Yacc/Bison & clones: features

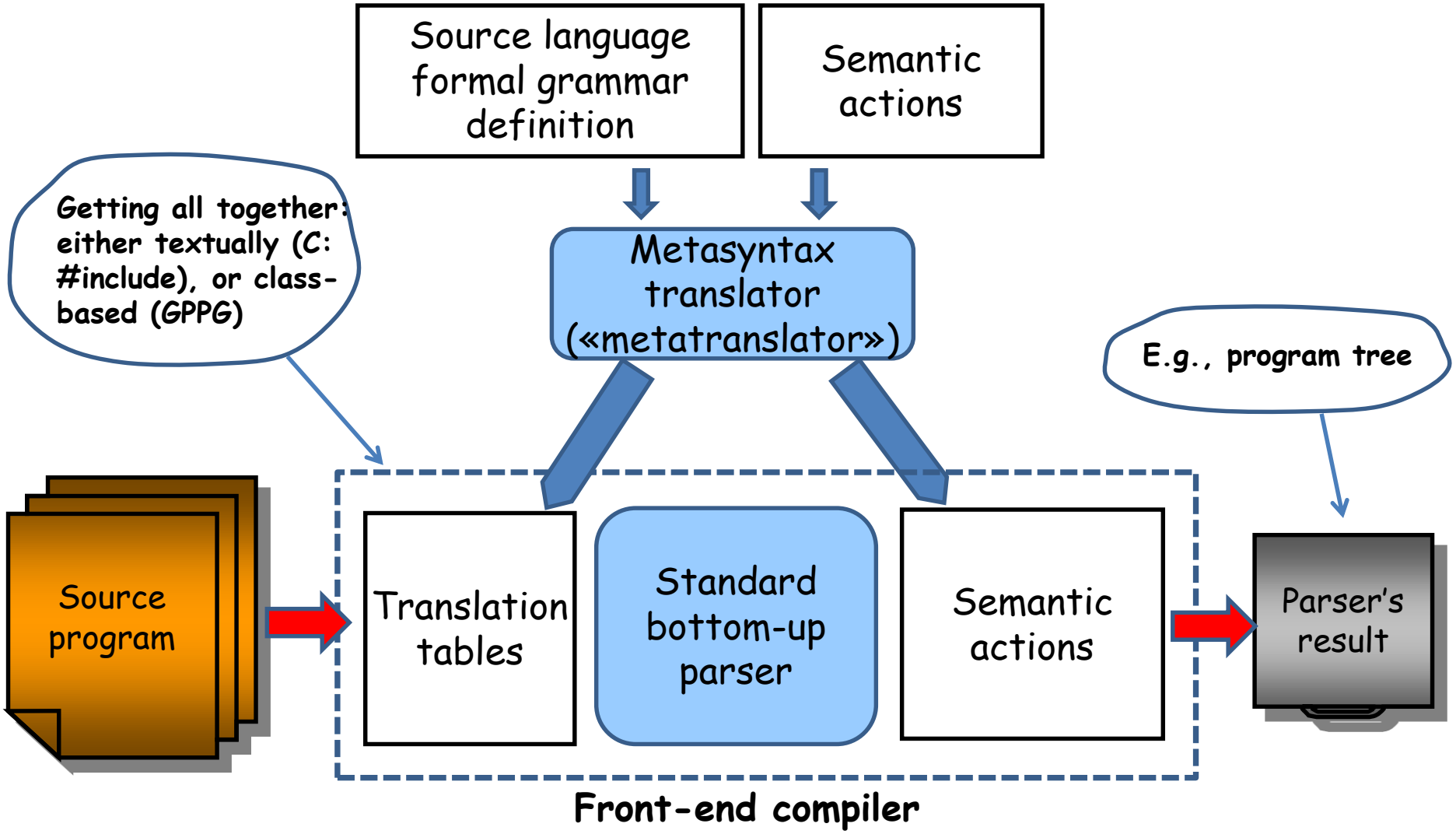
- Generates **bottom-up** syntax parsers.
- Has its **own notation** (formalism) for grammar specification.
- Internally, the grammar is represented in a **table form**; the generated parser is table-driven.
- Source tokens should be generated by a separate lexical analyzer: either by a hand made analyzer or by Lex/Flex or compatible (Yacc uses integer token codes).

- Very good grammar readability.
- Separation the grammar from semantic actions.
- Rules with left recursion are allowed.
- Good standard support for error recovery.
- Hard to debug the grammar and to find ambiguities.

Yacc based technology



Yacc based technology



Yacc: Grammar structure

Common declarations (implementation language)
%%
Declarations of token, types, associativity,...
Declaration of the **main rule**
%%
Grammar rules (together with semantic actions)
%%
Common declarations (implementation language)

Grammar examples

- **ANSI C & Oberon grammars.**

  Are in DropBox: `Lecture 7-9 Addendum`

- **C++ & C# grammars** (a bit out of date).

  In the Appendix of the Russian translation of the "Dragon Book" (1st edition).

- Following slides: **Toy language grammar.**

  Are in DropBox: `Lecture 7-9 Addendum`

Example: Toy language grammar

```
// Identifiers & numbers
%token IDENTIFIER
%token NUMBER

// Keywords
%token IMPORT CLASS EXTENDS PRIVATE PUBLIC STATIC VOID IF ELSE
%token WHILE LOOP RETURN PRINT NULL NEW INT REAL

// Delimiters
%token LBRACE       //  {
%token RBRACE       //  }
%token LPAREN       //  (
%token RPAREN       //  )
%token LBRACKET     //  [
%token RBRACKET     //  ]
%token COMMA        //  ,
%token DOT          //  .
%token SEMICOLON    //  ;

// Operator signs
%token ASSIGN       //  =
%token LESS         //  <
%token GREATER      //  >
%token EQUAL        //  ==
%token NOT_EQUAL    //  !=
%token PLUS         //  +
%token MINUS        //  -
%token MULTIPLY     //  *
%token DIVIDE       //  /

%start CompilationUnit
```

Language alphabet

Grammar main rule

Gets converted to

```
enum Tokens
{
    IDENTIFIER,
    NUMBER,
    ...
};
```

Example: Toy language grammar

```
CompilationUnit
        : Imports ClassDeclarations
        ;
Imports
        :  /* empty */
        | Import Imports
        ;
Import
        : IMPORT IDENTIFIER SEMICOLON
        ;
ClassDeclarations
        : /* empty */
        | ClassDeclaration ClassDeclarations
        ;
ClassDeclaration
        :        CLASS IDENTIFIER SEMICOLON Extension ClassBody
        | PUBLIC CLASS IDENTIFIER SEMICOLON Extension ClassBody
        ;
Extension
        : /* empty */
        | EXTENDS Identifier
        ;
ClassBody
        : LBRACE               RBRACE
        | LBRACE ClassMembers RBRACE
        ;
 ClassMembers
        :              ClassMember
        | ClassMembers ClassMember
        ;
```

Grammar: program & classes

Example: Toy language grammar

```
ClassMember
        : FieldDeclaration
        | MethodDeclaration
        ;
FieldDeclaration
        : Visibility Staticness Type IDENTIFIER SEMICOLON
        ;
Visibility
        : /* empty */
        | PRIVATE
        | PUBLIC
        ;
Staticness
        : /* empty */
        | STATIC
        ;
MethodDeclaration
        : Visibility Staticness MethodType IDENTIFIER Parameters Body
        ;
Parameters
        : LPAREN                RPAREN
        | LPAREN ParameterList RPAREN
        ;
ParameterList
        :                      Parameter
        | ParameterList COMMA Parameter
        ;
Parameter
        : Type IDENTIFIER ;
```

Grammar: declarations

Example: Toy language grammar

```
MethodType
        : Type
        | VOID
        ;
Body
        : LBRACE LocalDeclarations Statements RBRACE
        ;
LocalDeclarations
        :                       LocalDeclaration
        | LocalDeclarations LocalDeclaration
        ;
LocalDeclaration
        : Type IDENTIFIER SEMICOLON
        ;
```

Grammar: declarations

Example: Toy language grammar

```
Statements
        :             Statement
        | Statements Statement
        ;
Statement
        : Assignment | IfStatement | WhileStatement | ReturnStatement
        | CallStatement | PrintStatement | Block
        ;
Assignment
        : LeftPart ASSIGN Expression SEMICOLON
        ;
LeftPart
        : CompoundName
        | CompoundName LBRACKET Expression RBRACKET
        ;
CompoundName
        :                   IDENTIFIER
        | CompoundName DOT IDENTIFIER
        ;
IfStatement
        : IF LPAREN Relation RPAREN Statement
        | IF LPAREN Relation RPAREN Statement ELSE Statement
        ;
WhileStatement
        : WHILE Relation LOOP Statement SEMICOLON
        ;
ReturnStatement
        : RETURN             SEMICOLON
        | RETURN Expression SEMICOLON
        ;
```

Grammar: statements

Example: Toy language grammar

```
CallStatement
        : CompoundName LPAREN                RPAREN SEMICOLON
        | CompoundName LPAREN ArgumentList RPAREN SEMICOLON
        ;
ArgumentList
        :                      Expression
        | ArgumentList COMMA Expression
        ;
PrintStatement
        : PRINT Expression SEMICOLON
        ;
Block
        : LBRACE              RBRACE
        | LBRACE Statements RBRACE
        ;
```

Example: Toy language grammar

```
Relation
        : Expression
        | Expression RelationalOperator Expression
        ;
RelationalOperator
        : LESS | GREATER | EQUAL | NOT_EQUAL
        ;
Expression
        :          Term Terms
        | AddSign Term Terms
        ;
AddSign
        : PLUS | MINUS
        ;
Terms
        : /* empty */
        | AddSign Term Terms
        ;
Term
        : Factor Factors
        ;
Factors
        : /* empty */
        | MultSign Factor Factors
        ;
MultSign
        : MULTIPLY | DIVIDE
        ;
```

Grammar: expressions

Example: Toy language grammar

```
Factor
        : NUMBER
        | LeftPart
        | NULL
        | NEW NewType
        | NEW NewType LBRACKET Expression RBRACKET
        ;
NewType
        : INT
        | REAL
        | IDENTIFIER
        ;
Type
        : INT        ArrayTail
        | REAL       ArrayTail
        | IDENTIFIER ArrayTail
        ;
ArrayTail
        : /* empty */
        | LBRACKET RBRACKET
        ;
```

Grammar: types

Toy grammar: comments

1. No means for expression repetitions (like in BNF format) in YACC notation; we have to use recursion instead.

```
ParameterList
        :                           Parameter
        | ParameterList COMMA Parameter
        ;
...
ArgumentList
        :                         Expression
        | ArgumentList COMMA Expression
        ;
...
Statements
        :              Statement
        | Statements Statement
        ;
```

Toy grammar: comments

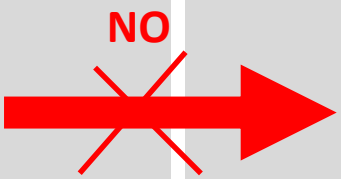
2. Both right and left recursions are allowed and supported.

```
Expression
        :          Term Terms
        | AddSign Term Terms
        ;
AddSign
        : PLUS | MINUS
        ;
Terms
        : /* empty */
        | AddSign Term Terms
        ;
Term
        : Factor Factors
        ;
Factors
        : /* empty */
        | MultSign Factor Factors
        ;
MultSign
        : MULTIPLY | DIVIDE
        ;
```

Toy grammar: comments

3. Grouping is not supported; we have to add extra rules for grouping

```
AddSign
        : PLUS | MINUS
        ;
Terms
        : /* empty */
        | AddSign Term Terms
        ;
Term
        : Factor Factors
        ;
Factors
        : /* empty */
        | MultSign Factor Factors
        ;
MultSign
        : MULTIPLY | DIVIDE
        ;
```




```
Terms
        : /* empty */
        | (PLUS|MINUS) Term Terms
        ;

Factors
        : /* empty */
        | (MULTIPLY | DIVIDE)
                  Factor Factors
        ;
```

Toy grammar translation

```
C:\Lectures\GPG 1.5.0\binaries>
                gppg /conflicts "C:\Lectures\Lecture 8\Toy.yacc

Shift/Reduce conflict
 Shift "IDENTIFIER":    State-20 -> State-21
 Reduce 30:     MethodType -> Type

Shift/Reduce conflict
 Shift "ELSE":  State-87 -> State-88
 Reduce 50:     IfStatement -> IF, LPAREN, Relation, RPAREN, Statement

Shift/Reduce conflict
 Shift "LBRACKET":      State-120 -> State-122
 Reduce 48:     CompoundName -> IDENTIFIER
```

1

2

3

Toy grammar translation

```
C:\Lectures\GPG 1.5.0\binaries>
              gppg /conflicts "C:\Lectures\Lecture 8\Toy.yacc

Shift/Reduce conflict
 Shift "IDENTIFIER":    State-20 -> State-21
 Reduce 30:     MethodType -> Type

FieldDeclaration: Visibility Staticness Type . IDENTIFIER SEMICOLON
MethodType: Type .
```


```
FieldDeclaration
        : Visibility Staticness Type IDENTIFIER SEMICOLON
        ;
...
MethodDeclaration
        : Visibility Staticness MethodType IDENTIFIER Parameters Body
        ;
...
Type
        : ...
        | IDENTIFIER ArrayTail
        ;
MethodType
        : Type
        | ...
        ;
```

Shift/Reduce conflicts are resolved in favor of Shift

Toy grammar translation

```
C:\Lectures\GPG 1.5.0\binaries>
                gppg /conflicts "C:\Lectures\Lecture 8\Toy.yacc


Shift/Reduce conflict
 Shift "ELSE":  State-87 -> State-88
 Reduce 50:     IfStatement -> IF, LPAREN, Relation, RPAREN, Statement

IfStatement: IF LPAREN Relation RPAREN Statement .
IfStatement: IF LPAREN Relation RPAREN Statement . ELSE Statement
```

2

```
IfStatement
        : IF LPAREN Relation RPAREN Statement
        | IF LPAREN Relation RPAREN Statement ELSE Statement
        ;
```

```
IfStatement
        : IF LPAREN Relation RPAREN Statement ElseTail
        ;
ElseTail
        : /* empty */
        | ELSE Statement
        ;
```

Toy grammar translation

```
Shift/Reduce conflict
 Shift "LBRACKET":      State-120 -> State-122
 Reduce 48:     CompoundName -> IDENTIFIER

CompoundName: IDENTIFIER .
Type: IDENTIFIER . ArrayTail
```

3

```
C [ 10 ] = 7 ;   // assignment
C [ ] a ;        // declaration
```

```
Assignment
        : LeftPart ASSIGN Expression SEMICOLON
        ;
LeftPart
        : CompoundName
        | CompoundName LBRACKET Expression RBRACKET
        ;
Body
        : LBRACE LocalDeclarations Statements RBRACE
        ;
LocalDeclaration
        : Type IDENTIFIER SEMICOLON
        ;
```

```
Type
        : ...
        | IDENTIFIER ArrayTail
        ;
ArrayTail
        : ...
        | LBRACKET RBRACKET
        ;
```

Toy grammar translation

Let's introduce an error to the grammar:

```
Body
        : LBRACE LocalDeclarations Statements RBRACE
        ;
...
Statement
        : LocalDeclaration
        | Assignment
        | IfStatement
        | WhileStatement
        | ReturnStatement
        | CallStatement
        | PrintStatement
        | Block
        ;
```

```
Reduce/Reduce conflict in state 131 on symbol INT
 Reduce 34:     LocalDeclarations -> LocalDeclarations, LocalDeclaration
 Reduce 38:     Statement -> LocalDeclaration
```

Reduce/Reduce should be resolved by developer (by transforming the grammar)

Toy grammar: semantic actions

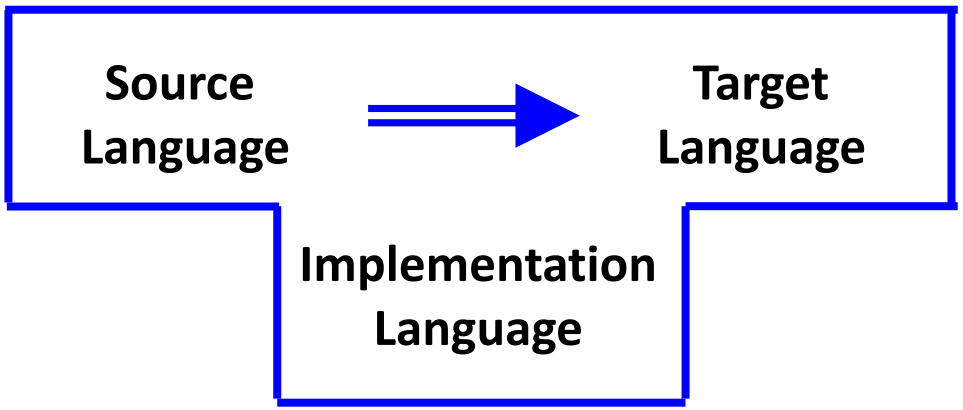
```
Statements
       :            Statement { $$ = createStmtList($1); }
       | Statements Statement { $$ = addStmtToList($1,$2); }
       ;
Statement
       : Assignment | IfStatement | WhileStatement | ReturnStatement
       | ...
       ;
Assignment
       : LeftPart ASSIGN Expression SEMICOLON  { $$ = createAssign($1,$3); }
       ;
IfStatement
       : IF LPAREN Relation RPAREN Statement
                                      { $$ = createIf($3,$5,NULL); }
       | IF LPAREN Relation RPAREN Statement ELSE Statement
                                      { $$ = createIf($3,$5,$7); }
       ;
WhileStatement
       : WHILE Relation LOOP Statement SEMICOLON { $$ = createWhile($2,$4); }
       ;
ReturnStatement
       : RETURN            SEMICOLON { $$ = createReturn(NULL); }
       | RETURN Expression SEMICOLON { $$ = createReturn($2); }
       ;
```

Bootstrapping Technology

- When implementation language & source language are **the same**. Examples: Ada; Eiffel; Scala.
- **Advantages**:
  More stable technology; supports graduate language & compiler improvement; no dependency on any third-party tools; the code of the compiler is **an excellent test** for both language and compiler itself.
- **Disadvantages**:
  A bit awkward technology; requires non-trivial management & powerful management tools (e.g., ***ant***).

Bootstrapping Technology

- Reference: Terence Pratt.
- Graphical notation ("T Notation"):



Bootstrapping the Compiler

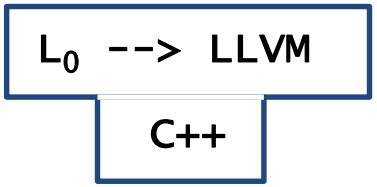
Initial development step

L_0

- **Define a very simple <u>subset</u> of the target L language: L_0**

Bootstrapping the Compiler

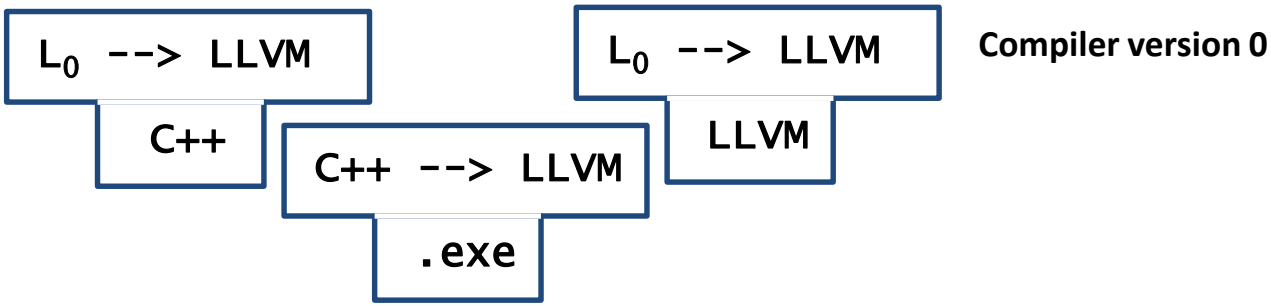
Initial development step



- **Define a very simple <u>subset</u> of the target L language: L_0**
- **Write the prototype compiler for L_0 using a third party language with an existing compiler**

Bootstrapping the Compiler

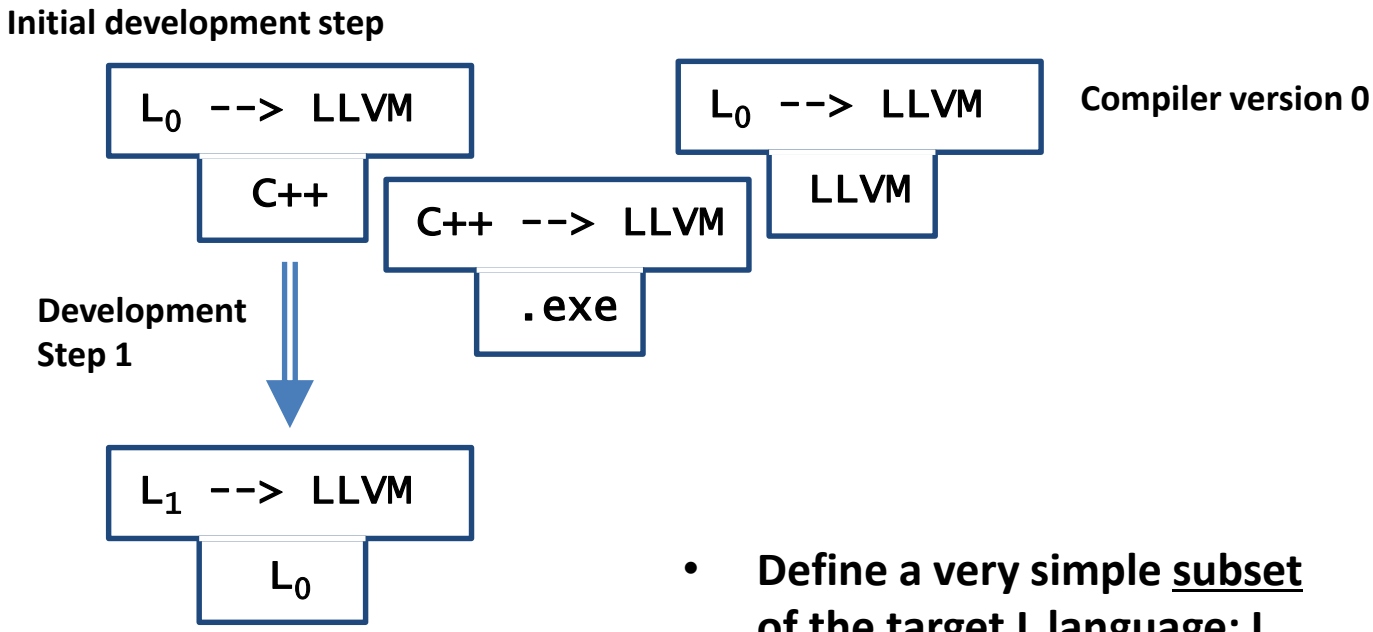
Initial development step



Compiler version 0

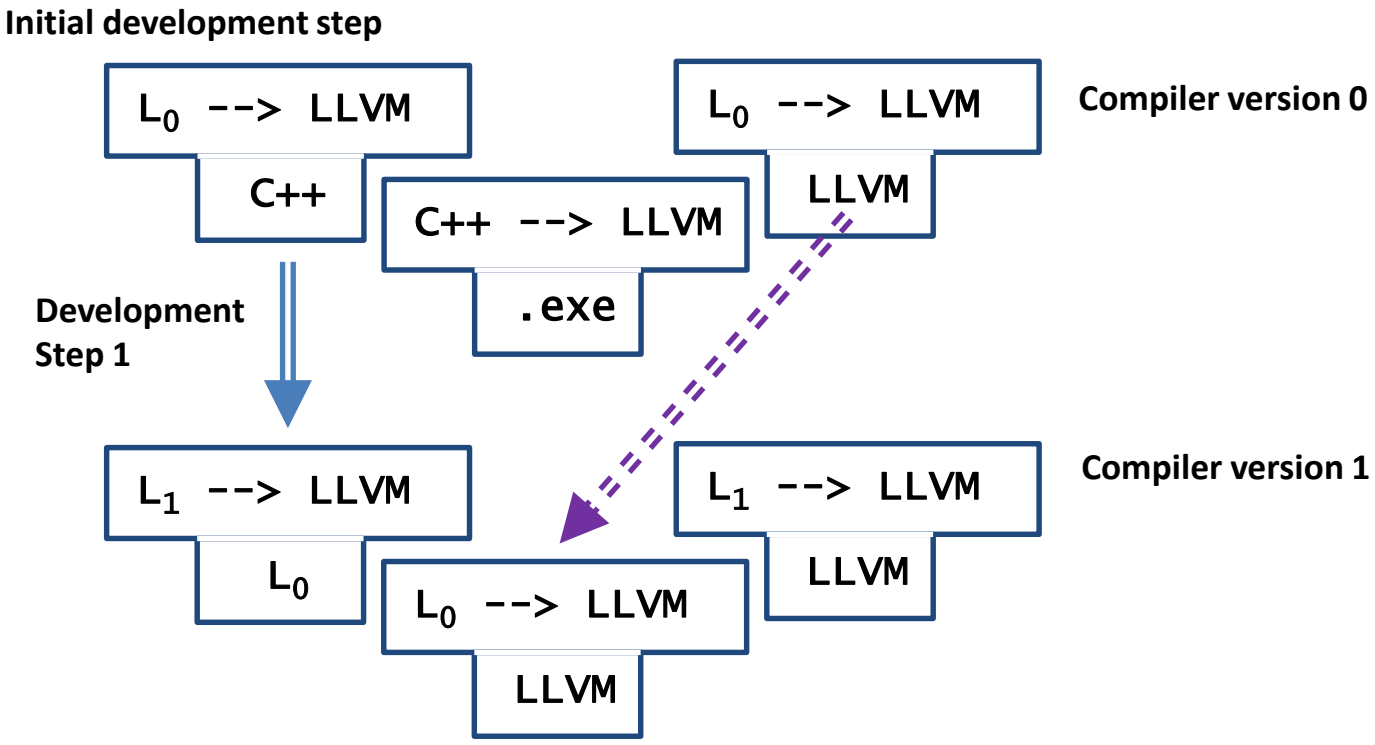
- **Define a very simple <u>subset</u> of the target L language: L_0**
- **Write the prototype compiler for L_0 using a third party language with an existing compiler**
- **Compile the L_0 compiler getting the 0th version of the target compiler.**

Bootstrapping the Compiler



- Define a very simple <u>subset</u> of the target L language: L_0
- Write the prototype compiler for L_0 using a third party language with an existing compiler
- Compile the L_0 compiler getting the 0th version of the target compiler.
- DevStep1: Create the next version of the language and (re)write the compiler in L_0 language…

Bootstrapping the Compiler



- **Dev Step1:**
  …and compile it using the previous compiler version!

Bootstrapping the Compiler