Abstract Syntax

Part IV

Syntax Trees

- In a multi-pass compiler the parser builds a syntax tree representing the program to compile.
- Later phases then work on this *abstract syntax tree (AST)*.
- The tree could be the concrete syntax tree (the parse tree) corresponding to the grammar, but this is often not very practical.

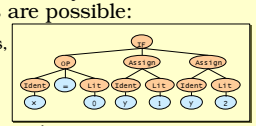
A Note on the assignment

- The skeleton for the parser that you have been given (expressive.cup) does not generate an AST.
- It just generates a string with the parsed constructs. (On success you get the input program back as a string.)
- This is useful for testing the first part of your parser, but it is not useful for the rest of the compiler.
- You will have to replace this with a parser that generates an AST.

Abstract vs. Concrete Syntax

- Compared to the concrete syntax tree, some simplifications are possible:
  - No need for parentheses, `A * (B + C)` becomes:

```
OP(MULT, A,
   OP(PLUS, B, C))
```



  - No need to maintain terminals, `if(x = 0){y:=1.} else {y:=2.}` becomes:

```
IF(OP(EQ,Ident("x"),Lit(0)),
[Assign(Ident("y"),Lit(1))],
[Assign(Ident("y"),Lit(2))])
```


Abstract Syntax Tree

- An AST is a tree with one kind of node for each alternative in the abstract syntax.
- It is simpler than a parse tree and hence easier to use.
- Still it has all the information that is necessary for the compilation.

Abstract Syntax Tree

- We represent ASTs using a set of Java classes, one for each alternative.
- We use one common abstract superclass: `Tree`.
- Each class represents subtrees as instance variables.
- Each class has a constructor to construct a node of the given kind.

Example Abstract Syntax

```
Expr ::= OP Op Expr Expr
       | Lit int
       | Ident String
Op ::= MULT | PLUS | EQ …
```


Example AST

```
public abstract class Tree {
  public static class Lit extends Tree {
    int num;
    public Lit(int num) {
      this.num = num;
    }
  }
  public static class OP extends Tree {
    Tree left, right;
    int Op;
    public OP(int Op, Tree left, Tree right) {
      this.Op = Op;
      this.right = right;
      this.left = left;
    }
  }
}
```


Repetition

- Repetition in the grammar is usually implemented by arrays.

```
Program → Statements
Statements → Statement '.' Statements
   | ε
```

```
public static abstract class Stat extends Tree {
  …
}
public static class Prog extends Tree {
  public final Stat[] body;
  …
}
```


Designing an AST

- The AST will be used by most phases of the compiler.
- Constructs in the language that can be handled in the same or in similar ways should be grouped as one node-type.
- One has to find the balance between a deep and a broad AST.
  - A deep AST has many node-types.
  - A broad AST has many instances of a few node-types.

Deep vs. Broad

Deep

```
Expr - CompOp: COp E E
     |          |- <
     |          |- =
     |
     |- ArithOp - BinA: Op E E
     |          |       |- +
     |          |       |- -
     |          |
     |          |- Una: Op E E
     |                  |- -
     |- Lit - Int: I
     |      |- Bool: Val
     |                 | - TRUE
     |                 | - FALSE
     |- ID: S
```

Broad

```
Expr: Type Op E E S I
    |     |- <
    |     |- =
    |     |- +
    |     |- -
    |
    |- BINOP
    |- UNOP
    |- ID
    |- LIT
```


Finding The Balance

- The design of the AST depends on how the AST is going to be used in later phases of the compiler.
- Getting it right takes experience and usually some experimentation.
- (For the project you will be given a complete AST.)

Constructing ASTs

- Each syntax tree recognized by the parser should somehow be turned into an AST.
- This mapping is not always one to one. Some parse trees has to be passed around in the parser before they are inserted into the AST.

Example

```
Program → Statements
Statements → Statement '.' Statements |ε
```

- In the grammar and the parser we have `Statements` as a node in the parse tree. But in the AST statements will just be an "`statement array`" in the `Prog` node.
- The parser can collect all Statements in a list and then convert the list to an array and insert it in the `Prog` node.

Constructing an AST in JavaCUP

- In JavaCUP, just as in JLex, there is an action for each rule in the grammar.
  (Actually there can be several actions for a rule, one after each symbol on the rhs)
- The action is a piece of Java code, usually used for building and returning the AST.
- The action could be used for other things, for example to implement a simple interpreter (more on this later).

Actions in JavaCup

- Actions are written in *code strings* surrounded by delimiters of the form `{:` and `:}`.
- The action for a lhs symbol S, can return values to the places where S is used on the rhs. This is done by assigning the value to `RESULT`.
- Symbols on the rhs can be labeled in order to refer to their values in the action. This is done as `sym:lab`.

Example

The Grammar:

```
Expr → Expr + Expr
     | Integer | Identifier
```

could be recognized by the rule:

```
expr ::= expr PLUS expr
       | INT
       | ID
       ;
```

Adding actions to generate the AST we get:

```
expr ::= expr:e1 PLUS expr:e2
  {: RESULT = new Op(Tokens.PLUS, e1, e2); :}
  | INT:i {: RESULT = new Lit(i.intValue()); :}
  | ID:id {: RESULT = new Ident(id); :}
  ;
```


Types

- We have to give the types of the values returned by the symbols used in the parser:

```
terminal String ID;
terminal Integer INT;
non terminal Expr expr;
```

Errors & Positions

- In order to give precise error messages we have to keep track of where in the program file a certain construct was located at.
- Since we need to be able to give error messages during the semantic analysis we need to store the positions of some tokens in the AST.

Adding pos to the AST

```
public abstract class Tree {
  public final int pos;
  public Tree(int pos) { this.pos = pos; }

  public static class Lit extends Tree {
    int num;
    public Lit(int pos, int num) {
      super(pos);
      this.num = num;
    }
  }
  …
}
```


Positions in JavaCUP

- JavaCUP uses Symbol objects for tokens and non terminals while parsing. Symbol objects has two integer fields called left and right that can be used to store the position of a token.
- In the parser these values can be accessed by appending the words *left* or *right* to the label.
- (With the given scanner and Position class, the line and column is encoded in the value for right.)

Adding tokens to the example

```
expr ::=
   expr:e1 PLUS:t expr:e2
     {:
      RESULT= new Op(tright, Tokens.PLUS, e1, e2);
     :}
   | INT:i
     {:
      RESULT = new Lit(iright, i.intValue());
     :}
   | ID:id {: RESULT = new Ident(idright, id); :}
   ;
```


A “complete” parser

```
package expressive;
import expressive.Tree.*;
terminal PLUS;
terminal String ID;
terminal Integer INT;
non terminal Expr expr;
start with expr;
expr ::=
   expr:e1 PLUS:t expr:e2
   {: RESULT= new Op(tright, Tokens.PLUS, e1, e2); :}
   | INT:i {: RESULT = new Lit(iright, i.intValue()); :}
   | ID:id {: RESULT = new Ident(idright, id); :}
   ;
```


Summary

- The abstract syntax should describe the programming language concisely without unnecessary syntactical constructs.
- The parser should create an abstract syntax tree while parsing the input.
- The AST should be designed so that the following phases can use it easily.
- The AST need information about source code positions for error handling.