Abstract Syntax Trees

1

Abstract Syntax Trees

- We will modify our parser one more time so that, as it parses the source code, it will also generate an intermediate representation of the program known as abstract syntax trees.
- An abstract syntax tree is similar to a parse tree but without extraneous nonterminal and terminal symbols.
- Abstract syntax trees provide an explicit representation of the structure of the source code that can be used for
  - additional constraint analysis (e.g., for type constraints)
  - some optimization (tree transformations)
  - code generation

2

Representing Abstract Syntax Trees

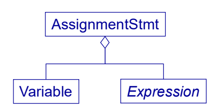
- We will use different classes to represent different node types in our abstract syntax trees. Examples include
  - `Program`
  - `AssignmentStmt`
  - `Variable`
  - `ProcedureDecl`
  - `LoopStmt`
  - `Expression`
- Each AST class has named instance variables (fields) to reference its children. These instance variables provide the “tree” structure.
- Occasionally we also include additional fields to support error handling (e.g., position) and code generation.

Terence Parr refers to this type of AST structure as an irregular (named child fields) heterogeneous (different node types) AST.

3

Abstract Syntax Trees: Example 1

- Consider the grammar for an assignment statement.
  `assignmentStmt = variable ":=" expression ";" .`
- The important parts of an assignment statement are
  - variable (the left side of the assignment)
  - expression (the right side of the assignment)
- We create an AST node for an assignment statement with the following structure:



4

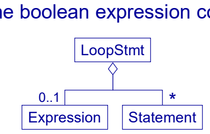
Class `AssignmentStmt`

```
public class AssignmentStmt extends Statement
  {
    private Variable   variable;
    private Expression expr;

    // position of assignment operator (for error reporting)
    private Position assignPosition;

    public AssignmentStmt(Variable   variable,
                          Expression expr,
                          Position   assignPosition)
      {
        this.variable = variable;
        this.expr = expr;
        this.assignPosition = assignPosition;
      }
    ...
  }
```


5

Abstract Syntax Trees: Example 2

- Consider the following grammar for a loop statement:
```
  loopStmt = ( "while" booleanExpr )?
             "loop" statements "end" "loop" ";" .
```
- Once a loop statement has been parsed, we don’t need to retain the nonterminal symbols. The AST for a loop statement would contain only the statements in the body of the loop and the optional boolean expression (e.g., the reference to the boolean expression could be null).



6

Class `LoopStmt`

```
public class LoopStmt extends Statement
  {
    private Expression whileExpr;
    private List<Statement> statements;

    public LoopStmt(Expression whileExpr, List<Statement> statements)
      {
        this.whileExpr  = whileExpr;
        this.statements = statements;
        ...
      }
    ...
  }
```

Note that `whileExpr` can be null.

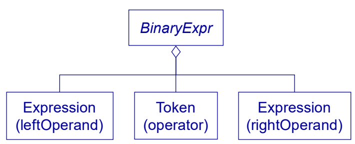
7

Abstract Syntax Trees: Example 3

- For binary expressions, part of the grammar exists simply to define operator precedence.
- Once an expression has been parsed, we do not need to preserve additional information about nonterminals that were introduced to define precedence (`relation`, `simpleExpr`, `term`, `factor`, etc.).
- A binary expression AST would contain only the operator and the left and right operands. The parsing algorithm would build the AST so as to preserve operator precedence.

8

Abstract Syntax Trees: Example 3
(continued)



9

Class `BinaryExpr`

```
public abstract class BinaryExpr extends Expression
  {
    private Expression leftOperand;
    private Token operator;
    private Expression rightOperand;

    public BinaryExpr(Expression leftOperand, Token operator,
                      Expression rightOperand)
      {
        this.leftOperand  = leftOperand;
        this.operator     = operator;
        this.rightOperand = rightOperand;
        ...
      }
    ...
  }
```


10

Structure of Abstract Syntax Trees

- There is an abstract class `AST` that serves as the superclass for all other abstract syntax tree classes.
- Class `AST` contains implementations of methods common to all subclasses plus declarations of abstract methods required by all concrete subclasses.
- All AST classes will be defined in an "…`ast`" subpackage.

Note the use of `AST` (in monospaced font) for the specific class and AST (in normal font) as an abbreviation for "abstract syntax tree".

11

Outline of Class `AST`

```
public abstract class AST
  {
    ...

    /** Check semantic/contextual constraints. */
    public abstract void checkConstraints();

    /** Emit the object code for the AST. */
    public abstract void emit()
        throws CodeGenException, IOException;
  }
```

Methods `checkConstraints()` and `emit()` provide a mechanism to "walk" the tree structure using recursive calls to subordinate tree nodes.

12

Subclasses of AST

- We will create a hierarchy of classes, some of which are abstract, that are all direct or indirect subclasses of `AST`.
- Each node in the abstract syntax tree constructed by the parser will be an object of a class in the `AST` hierarchy.
- Most classes in the hierarchy will correspond to and have names similar to the nonterminal symbols in the grammar, but not all abstract syntax trees have this property. See, for example, the earlier discussion about binary expressions. We do not need abstract syntax tree classes corresponding to nonterminals `simpleExpr`, `term`, `factor`, etc.

13

Using Collection Classes

- Some parsing methods simply return lists of AST objects.
- Examples

```
public List<InitialDecl> parseInitialDecls()
    throws IOException
public List<SubprogramDecl> parseSubprogramDecls()
    throws IOException
public List<Token> parseIdentifiers() throws IOException
public List<Statement> parseStatements()
    throws IOException
public List<ParameterDecl> parseFormalParameters()
    throws IOException
public List<Expression> parseActualParameters()
    throws IOException
```


14

Naming Conventions for AST

- Most AST classes have names similar to nonterminals in the grammar.
  - `Program`
  - `FunctionDecl`
  - `AssignmentStmt`
  - `LoopStmt`
- The parsing method for that nonterminal will create the corresponding AST object.
  - `parseProgram` returns a `Program` object
  - `parseLoopStmt` returns a `LoopStmt` object
  - etc.
- Parsing methods with plural names will return lists of AST objects.
  - the grammar was written to have this property.

15

Naming Conventions for AST (continued)

- Example

```
public abstract class Statement extends AST ...
public class LoopStmt extends Statement ...
```

- The parsing method `parseLoopStmt()` would be responsible for creating the AST node for `LoopStmt`. Instead of returning `void`, method `parseLoopStmt()` will return an object of class `LoopStmt`.
- Similarly, the parsing method `parseStatements()` will return a list of `Statement` objects, where each `Statement` object is either an `AssignmentStmt`, a `LoopStmt`, an `IfStmt`, etc.

16

Method `parseLiteral()`

- Method `parseLiteral()` is a special case.
- Since literals are tokens returned from the scanner, method `parseLiteral()` simply returns a `Token`. There is no AST class named `Literal`.
- Relevant Grammar Rules

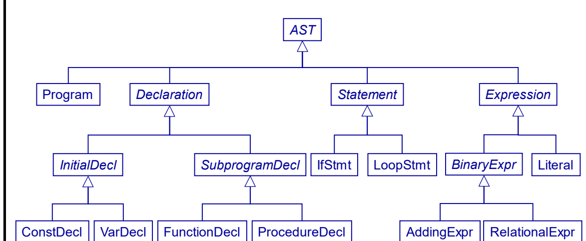
```
literal = intLiteral | charLiteral | stringLiteral
        | booleanLiteral .
booleanLiteral = "true" | "false" .
```

- Method

```
public Token parseLiteral() throws IOException
  {
    ...
  }
```


17

Partial AST Inheritance Diagram for the Language CPRL



(names for abstract classes are shown in italics)

18

Language Constraints Associated With Identifiers

- A parser built using only the set of parsing rules will not reject programs that violate certain language constraints such as “an identifier must be declared exactly once”.
- Examples: Valid syntax but not valid with respect to contextual constraints

```
var x : Integer;            var c : Char;
begin                       begin
   y := 5;                     c := -3;
end.                        end.
```

©SoftMoore Consulting Slide 19

19

Class `IdTable`

- We will extend class `IdTable` to help track not only of the types of identifiers that have been declared, but also of their declarations.
- Class `Declaration` is part of the AST hierarchy. A declaration object contains a reference to the identifier token and information about its type. We will use different subclasses of `Declaration` for kinds of declarations; e.g., ConstDecl, VarDecl, ProcedureDecl, etc.

©SoftMoore Consulting Slide 20

20

Selected Methods in the Modified Version of `IdTable`

```
/**
 * Opens a new scope for identifiers.
 */
public void openScope()

/**
 * Closes the outermost scope.
 */
public void closeScope()

/**
 * Add a declaration at the current scope level.
 * @throws ParserException if the identifier token associated
 *                         with the declaration is already
 *                         defined in the current scope.
 */
public void add(Declaration decl) throws ParserException
```

©SoftMoore Consulting Slide 21

21

Selected Methods in Class `IdTable` (continued)

```
/**
 * Returns the Declaration associated with the identifier
 * token's text.  Returns null if the identifier is not found.
 * Searches enclosing scopes if necessary.
 */
public Declaration get(Token idToken)

/**
 * Returns the current scope level.
 */
public ScopeLevel getCurrentLevel()
```

Note: `ScopeLevel` is an enum class with only two values, `PROGRAM` and `SUBPROGRAM`.

©SoftMoore Consulting Slide 22

22

Adding Declarations to `IdTable`

- When an identifier is declared, the parser will attempt to add the declaration to the table within the current scope. (The declaration already contains the identifier token.)
  - throws an exception if a declaration with the same name (same token text) has been previously declared in the current scope.
- Example (in method parseConstDecl())

```
Token constId = scanner.getToken();
...
constDecl = new ConstDecl(constId, constType, literal);
idTable.add(constDecl);
```

Throws a `ParserException` if the identifier token `constId` is already defined in the current scope

©SoftMoore Consulting Slide 23

23

Interface `NamedDecl`

- Identifiers declared using `VarDecl` (which we convert to a list of `SingleVarDecl` as described later) or `ParameterDecl` have similar uses within CPRL; e.g.,
  `x := y;`
- Variable x could have been declared in a variable declaration or a parameter declaration.
  - similarly for the named value y
- There is a need to treat both types of declarations uniformly at several points during parsing, which we achieve by creating interface `NamedDecl` and specifying that `SingleVarDecl` and `ParameterDecl` implement this interface.

©SoftMoore Consulting Slide 24

24

Interface `NamedDecl` (continued)

- Five methods in interface `NamedDecl`

```
public Type getType();
public int getSize();
public ScopeLevel getScopeLevel();
public void setRelAddr(int relAddr);
public int getRelAddr();
```

 Slide 25

25

Example: Using Interface `NamedDecl`

```
// excerpt from parseStatement()

if (symbol == Symbol.identifier)
  {
    Declaration decl = idTable.get(scanner.getToken());

    if (decl != null)
      {
        if (decl instanceof NamedDecl)
            stmt = parseAssignmentStmt();
        ...
      }
  }
...
```

 Slide 26

26

Using `IdTable` to Check Applied Occurrences of Identifiers

- When an identifier is encountered in the statement part of the program or a subprogram (e.g., as part of an expression or subprogram call), the parser will
  - check that the identifier has been declared
  - use the information about how the identifier was declared to facilitate correct parsing (e.g., you can't assign a value to an identifier that was declared as a constant.)

 Slide 27

27

Using `IdTable` to Check Applied Occurrences of Identifiers (continued)

- Example (in method `parseVariableExpr()`)

```
Token idToken = scanner.getToken();
match(Symbol.identifier);
Declaration decl = idTable.get(idToken);

if (decl == null)
    throw error("Identifier \"" + idToken
              + "\" has not been declared.");
else if (!(decl instanceof NamedDecl))
    throw error("Identifier \"" + idToken
              + "\" is not a variable.");
```

 Slide 28

28

Types in CPRL

- The compiler uses two classes to provide support for CPRL types.
- Class `Type` encapsulates the language types and their sizes.
  - Predefined types are declared as static constants.
  - Class `Type` also contains a static method that returns the type of a literal symbol.

    `public static Type getTypeOf(Symbol literal)`
- Class `ArrayType` extends `Type` to provide additional support for arrays.

 Slide 29

29

Class Type

- Class Type encapsulates the language types and sizes (number of bytes) for the programming language CPRL.
- Type sizes are initialized to values appropriate for the CPRL virtual machine.
  - 4 for Integer
  - 1 for Boolean
  - 2 for Character
  - etc.
- Predefined types are declared as static constants.

```
public static final Type Boolean = new Type(...);
public static final Type Integer = new Type(...);
public static final Type Char    = new Type(...);
public static final Type String  = new Type(...);
public static final Type Address = new Type(...);
public static final Type UNKNOWN = new Type(...);
```

 Slide 30

30

Class `ArrayType`

- Class `ArrayType` extends class `Type`.
  - therefore array types are also types
- In addition to the total size of the array, class `ArrayType` also keeps track of the number and type of elements.

```
/**
 * Construct an array type with the specified name, number of
 * elements, and the type of elements contained in the array.
 */
public ArrayType(String typeName, int numElements,
                 Type elementType)
```

- When the parser parses an array type declaration, the constructor for AST class `ArrayTypeDecl` creates an `ArrayType` object.

31

Example: Parsing a `ConstDecl`

```
/**
 * Parse the following grammar rule:<br>
 * <code>constDecl = "const" constId ":=" literal ";" .</code>
 *
 * @return the parsed constant declaration.
 *         Returns null declaration if parsing fails.
 */
public InitialDecl parseConstDecl() throws IOException
  {
    try
      {
        match(Symbol.constRW);
        Token constId = scanner.getToken();
        match(Symbol.identifier);
```

(continued on next slide)

32

Example: Parsing a `ConstDecl` (continued)

```
        match(Symbol.assign);
        Token literal = parseLiteral();
        match(Symbol.semicolon);

        Type constType = Type.UNKNOWN;
        if (literal != null)
            constType = Type.getTypeOf(literal.getSymbol());

        ConstDecl constDecl
                  = new ConstDecl(constId, constType, literal);
        idTable.add(constDecl);
        return constDecl;
      }
```

(continued on next slide)

33

Example: Parsing a `ConstDecl` (continued)

```
    catch (ParserException e)
      {
        ErrorHandler.getInstance().reportError(e);
        recover(initialDeclFollowers);
        return null;
      }
  }
```


34

The Scope Level of a Variable Declaration

- During code generation, when a variable or named value is referenced in the statement part of a program or subprogram, we need to be able to determine where the variable was declared.
- Class `IdTable` contains a method `getCurrentLevel()` that returns the block nesting level for the current scope.
  - `PROGRAM` for objects declared at the outermost (program) scope.
  - `SUBPROGRAM` for objects declared within a subprogram.
- When a variable is **declared**, the declaration is initialized with the current level.

```
ScopeLevel scopeLevel = idTable.getCurrentLevel();
varDecl = new VarDecl(identifiers, varType, scopeLevel);
```


35

Example: Scope Levels

```
var x : Integer;   // scope level of declaration is PROGRAM
var y : Integer;   // scope level of declaration is PROGRAM

procedure p is       // scope level of declaration is PROGRAM
   var x : Integer;  // scope level of declaration is SUBPROGRAM
   var b : Integer;  // scope level of declaration is SUBPROGRAM
begin
   ... x ...    // x was declared at SUBPROGRAM scope
   ... b ...    // b was declared at SUBPROGRAM scope
   ... y ...    // y was declared at PROGRAM scope
end p;

begin
    ... x ...     // x was declared at PROGRAM scope
    ... y ...     // y was declared at PROGRAM scope
    ... p ...     // p was declared at PROGRAM scope
end.
```


36

`VarDecl` versus `SingleVarDecl`

- A variable declaration can declare several identifiers all with the same type, as in
```
var x, y, z : Integer;
```
- This declaration is logically equivalent to declaring each variable separately, as in
```
var x : Integer;
var y : Integer;
var z : Integer;
```
- To simplify constraint checking and code generation, within the AST we will view a variable declaration as a collection of single variable declarations.

Class `SingleVarDecl`

```
public class SingleVarDecl extends InitialDecl
                           implements NamedDecl
  {
    private ScopeLevel scopeLevel;
    ...

    public SingleVarDecl(Token identifier, Type varType,
                         ScopeLevel scopeLevel)
      {
        super(identifier, varType);
        this.scopeLevel = scopeLevel;
      }

    ...
  }
```


Class `VarDecl`

```
public class VarDecl extends InitialDecl
  {
    // the list of SingleVarDecls for the variable declaration
    private List<SingleVarDecl> singleVarDecls;

    public VarDecl(List<Token> identifiers, Type varType,
                   ScopeLevel scopeLevel)
      {
        super(null, varType);
        singleVarDecls = new
            ArrayList<SingleVarDecl>(identifiers.size());
        for (Token id : identifiers)
            singleVarDecls.add(new SingleVarDecl(id,
                               varType, scopeLevel));
      }
    ...
  }
```

A `VarDecl` is simply a list of `SingleVarDecls`.

Method `parseInitialDecls()`

- Method `parseInitialDecls()` constructs/returns a list of initial declarations.
- For constant and array type declarations, this method simply adds them to the list.
- For variable declarations (`VarDecls`), this method extracts the list of single variable declarations (`SingleVarDecls`) and adds them to the list. The original `VarDecl` is no longer used after this point.

Method `parseInitialDecls()` (continued)

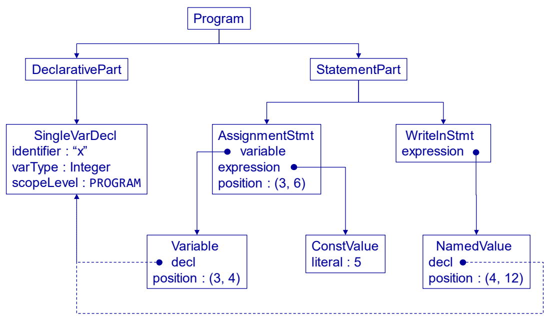
```
...

InitialDecl decl = parseInitialDecl();

if (decl instanceof VarDecl)
  {
    // add the single variable declarations
    VarDecl varDecl = (VarDecl) decl;
    for (SingleVarDecl singleVarDecl : varDecl.getSingleVarDecls())
        initialDecls.add(singleVarDecl);
  }
else
    initialDecls.add(decl);
```


Example: Abstract Syntax Tree

```
var x : Integer;
begin
   x := 5;
   writeln x;
end.
```

(AST for this example is on the next slide.)

Example: Abstract Syntax Tree (continued)



43

Determining Types of Expressions

- Since CPRL is statically typed, it is possible to determine the type of every expression at compile time, and AST class Expression has a property for the expression type that is inherited by all expression subclasses.
- Where within the compiler should type determination take place? In general, we will determine the type of an expression in the constructor for the expression's AST class.

44

Example: `RelationalExpr`

- A relational expression is a binary expression where the operator is a relational operator such as "<=" or ">".
- Regardless of the types of the two operands, a relational expression always has type `Boolean`.
- Constructor for `RelationalExpr`

```
public RelationalExpr(Expression leftOperand,
                      Token      operator,
                      Expression rightOperand)
  {
    super(leftOperand, operator, rightOperand);
    setType(Type.Boolean);
    ...
  }
```


45

Example: `AddingExpr`

- For most "real" programming languages, determining the type of an adding expression can be somewhat complicated.
  - C and Java have multiple numeric types with rules about automatic conversions (coercions) when an operator has different operand types.
- In CPRL, an adding expression always has type Integer. (Similarly for a multiplying expression in CPRL.)

46

Example: `AddingExpr` (continued)

- Constructor for `AddingExpr`

```
public AddingExpr(Expression leftOperand, Token operator,
                  Expression rightOperand)
  {
    super(leftOperand, operator, rightOperand);
    setType(Type.Integer);
    ...
  }
```


47

Example: `Variable`

- The type for a variable (and therefore also for a named value) is initialized to the type specified in the variable's declaration.
- Constructor for Variable

```
public Variable(NamedDecl decl, Position position,
                List<Expression> indexExprs)
  {
    super(decl.getType(), position);
    this.decl       = decl;
    this.indexExprs = indexExprs;
  }
```


48

Example: `Variable` (continued)

- The initialized type for a variable is correct for predefined types such as `Integer` or `Char`, but additional work is required for arrays.
- Consider the following declarations:

```
type T1 is array(10) of Integer;
type T2 is array(10) of T1;
var a, b : T2;
```

- While the declared (initialized) type of both a and b is `T2`, we could have a variable or named value with zero, one, or two index expressions, as in the following:

```
a := b;             // type of var and named val is T2
a[0] := b[0];       // type of var and named val is T1
a[1][6] := b[5][7]; // type of var and named val is Integer
```

©SoftMoore Consulting Slide 49

49

Example: `Variable` (continued)

- For arrays, we determine the actual type of a variable or named value in method `checkConstraints()`.

```
for (Expression expr : indexExprs)
  {
    expr.checkConstraints();

    if (expr.getType() != Type.Integer)
        throw error(...);

    if (getType() instanceof ArrayType)
      {
        ArrayType type = (ArrayType) getType();
        setType(type.getElementType());
      }
    else
        throw error(...);
  }
```

©SoftMoore Consulting Slide 50

50

Maintaining Context During Parsing

- Certain CPRL statements need access to an enclosing context for constraint checking and code generation.
- Example: `exit when n > 10;`
  An exit statement has meaning only when nested inside a loop., and code generation for an exit statement requires knowledge of which loop encloses it.
- Similarly, a return statement needs to know which subprogram it is returning from.
- Classes `LoopContext` and `SubprogramContext` will be used to maintain contextual information in these cases.

©SoftMoore Consulting Slide 51

51

Class `LoopContext`

```
/**
 * Returns the loop statement currently being parsed;
 * returns null if no such loop statement exists.
 */
public LoopStmt getLoopStmt()

/**
 * Called when starting to parse a loop statement.
 */
public void beginLoop(LoopStmt stmt)

/**
 * Called when finished parsing a loop statement.
 */
public void endLoop()
```

©SoftMoore Consulting Slide 52

52

Class `SubprogramContext`

```
/**
 * Returns the subprogram declaration currently being
 * parsed.  Returns null if no such procedure exists.
 */
public SubprogramDecl getSubprogramDecl()

/**
 * Called when starting to parse a subprogram declaration.
 */
public void beginSubprogramDecl(SubprogramDecl subprogDecl)

/**
 * Called when finished parsing a subprogram declaration.
 */
public void endSubprogramDecl()
```

©SoftMoore Consulting Slide 53

53

Example: Using Context During Parsing

- When parsing a loop statement:

```
LoopStmt stmt = new LoopStmt();
...
loopContext.beginLoop(stmt);
stmt.setStatements(parseStatements());
loopContext.endLoop();
```

- When parsing an exit statement:

```
LoopStmt loopStmt = loopContext.getLoopStmt();
if (loopStmt == null)
    throw error(exitPosition,
        "Exit statement is not nested within a loop");
return new ExitStmt(expr, loopStmt);
```

©SoftMoore Consulting Slide 54

54

Version 3 of the Parser
(Abstract Syntax Trees)

- Create AST classes in package "…`ast`"
- Add generation of AST structure; i.e., parsing methods should return AST objects or lists of AST objects.
- Use empty bodies when overriding abstract methods `checkConstraints()` and `emit()`.
- Use complete version of `IdTable` to check for scope errors.
- Use class `Context` to check exit and return statements.

At this point your compiler should accept all legal programs and reject most illegal programs. Some programs with type or other miscellaneous errors will not yet be rejected.

 Slide 55

55