Today

• Quiz 1
• Compilation (overview)
• Lexical analysis (intro)
• Formal grammars (intro)

Assignments

• HW1 due
• HW2 out (due Tues)
Compilation (typically) moves in increasingly abstract steps

1. identify language “tokens” in source (more later)
2. ensure syntax is correct (“parse”)
3. ensure source code is “correct” (nothing used before defined, no type errors)
4. generate an intermediate representation (for optimization)
5. improve performance of code (optimization)
6. generate executable code (machine code)

Example of “separation of concerns” (engineering design strategy)

• too complex to do “all at once” (similarly, single vs multi-pass)
• easier to manage, maintain, update/improve
Abstract Syntax Tree (AST) Interpreters
- execute the program directly from the AST ... what we'll do for MyPL

Bytecode Interpreters
- intermediate representation is bytecode
- usually includes optimization (back end) steps
- interpreter runs bytecode directly ... “write once run anywhere”

Just-in-time Compiler
- intermediate representation is bytecode (with optimization)
- instead of interpreting bytecode, converts to and runs machine code
- still “write once run anywhere” ... Java, JS, C#
Lexical Analysis

Tokens are the smallest meaningful units of a program

Some examples:

- Special words ("reserved" words)
  
  `int`, `if`, `while`, `new`, `class`, `public`, and so on

- Operators and Punctuation
  
  `+`, `=`, `==`, `<=`, `(`, `;`, and so on

- Identifiers
  
  variable names, function names, class names, etc.

- Constant values ("Literal" values)
  
  `42`, `3.14`, `true`, "abc", and so on

White space and comments are (usually) not tokens

- some exceptions such as Python

Tokens include a type and a "lexeme" (a value)

- the lexeme is just the token’s value in the source file

- e.g., in the statement: `x = 42;`
  
  - the token types might be `ID`, `ASSIGN`, `INTVAL`, `SEMICOLON`
  
  - and the corresponding lexemes “x”, “=”, “42”, and “;”

- for some tokens, the lexemes are needed for program execution
  
  - e.g., the variable name of the identifier ("x") and the int value “42"
Lexical Analysis Basics

- Goal is to simplify syntax analysis (parsing) ... and detect errors early
- The basic idea:

  Source Code: Token Sequence (as TYPE(lexeme)):

  int fun() INTTYPE("int"), ID("fun"), LPAREN(""), RPAREN(""),
  { LBRACKET("{"),
  int x = 0; INTTYPE("int"), ID("x"), ASSIGN("="), INTVAL("0"), SEMICOLON(";"),
  return x; RETURN("return"), ID("x"), SEMICOLON(";"),
  } RBRACKET("{")

- Source code is converted to a sequence (or a stream) of tokens
- Removes non-tokens (white space, comments)
- Tokens also typically stored with corresponding line and column numbers