# HW #3 (Parser Code Generation)

COP3402: System Software

Fall 2025

Instructor: Jie Lin, Ph.D.

Due date: **Friday, October 31, 2025 at 11:59 PM ET**

Last updated: October 20, 2025

**Disclaimer:** This document specifies all deliverables and constraints for HW3. If anything is unclear, contact the instructor before implementing assumptions.

**Updates:** All official updates and clarifications will be posted as **Webcourses announcements**. Check Webcourses regularly for critical updates.

**Due Date: Friday, October 31, 2025 at 11:59 PM ET**

**Submission Method:** Submit only via Webcourses. Submissions by email, chat/DM, cloud links, or any other channel are not accepted.

**Timestamp:** The Webcourses submission timestamp is authoritative. All deadlines use U.S. Eastern Time.

**Group Work:** This is a group assignment. Submit a single solution per group that clearly lists every member in the header comment and in the Webcourses group submission.

**Language Restriction:** All code must be written in C and compiled with GCC on Eustis. Code in any other language will receive a score of zero.

**Required Filename:** Name your program file `parsercodegen.c`.

# Contents

# 1 Academic Integrity, AI Usage, and Late Policy

## 1.1 Academic Integrity Expectations

If you plan to use AI tools while preparing your assignment, you must disclose this usage. Complete the **AI Usage Disclosure Form** provided with this assignment. If you used AI, include a separate markdown file describing:

- The name and version of the AI tool used.
- The dates used and specific parts of the assignment where the AI assisted.
- The prompts you provided and a summary of the AI output.
- How you verified the AI output against other sources and your own understanding.
- Reflections on what you learned from using the AI.

If you did not use any AI, check the appropriate box on the form. **Each team member must submit their own signed AI disclosure form individually in separate submissions (not as a group submission, since group submissions only allow one member to submit and will override other submissions).** Submit the signed form and the markdown file (if applicable) along with your assignment. Failure to disclose AI usage will be treated as academic dishonesty.

## 1.2 Plagiarism Detection

**All submissions will be processed through JPlag**, which detects the similarity score between you and the other students' submitted code. If the similarity score is above certain threshold, your code will be considered as plagiarism.

While AI tools may assist with brainstorming or initial code draft (if properly disclosed), the final submission must represent your own work and understanding. It is important to notice that if the similarity score is above the threshold, it does not matter if you have the AI disclosure or not, your program will be considered plagiarism. Also importantly, AI tends to draft the code in a similar way, if you just copy and paste, the similarity score will be very high.

## 1.3 Late Policy

**Due vs. late.** Anything submitted after the posted due date/time is late.

**Late window.** Late submissions are accepted for up to 48 hours after the due date/time; after that, the assignment is missed (score 0).

**Penalty (points, not percentages).** 5 points are deducted for each started 12-hour block after the due date/time (any fraction counts as a full block):

- 0:00:01–12:00:00 late → -5 points
- 12:00:01–24:00:00 late → -10 points

- 24:00:01–36:00:00 late → -15 points
- 36:00:01–48:00:00 late → -20 points
- After 48:00:00 late → Not accepted; recorded as missed (0 points)

Example: If the assignment is scored out of 100 points, the penalties above are deducted from your earned score.

**Technical issues.** Individual device/network problems do not justify exceptions—submit early and verify your upload.

# 2 Assignment Overview

HW3 builds directly upon Homework 2 by adding a deterministic recursive-descent parser and PM/0 code generator to your existing scanner. Your HW2 scanner (`lex.c`) will produce a token stream file, which your new `parsercodegen.c` program will read, parse according to the PL/0 grammar, and—only when no errors occur—generate executable PM/0 assembly code in the ISA format specified in Appendix A.

# 3 Implementation Procedure

This assignment requires two separate C source files that work together in a pipeline:

## 3.1 Scanner (`lex.c`) from HW2

- Keep your `lex.c` file as an independent source code file from Homework 2.
- Modify `lex.c` to write output to a file instead of only printing to the terminal.
- The output file should contain either the lexeme table or token list (your choice based on your HW2 implementation).
- This output file will serve as input for the new `parsercodegen.c` file.
- `lex.c` must accept exactly **ONE** command-line argument: the input PL/0 source file.

## 3.2 Parser/Code Generator (`parsercodegen.c`) for HW3

- Create a new file `parsercodegen.c` that implements the parser and code generator.
- **Important distinction:** `parsercodegen.c` requires **NO** command-line arguments.
- The input filename should be hard-coded in `parsercodegen.c` (e.g., `tokens.txt` or `token_list.txt`—whatever `lex.c` outputs).
- The parser reads the token file produced by `lex.c`, validates the grammar, and generates PM/0 assembly code (see Appendix A for the ISA specification).

- The output file should contain PM/0 instructions in the format: `OP L M` (one instruction per line).

# 4 Grammar Specification

Your parser must accept exactly the productions in Figure 1. In EBNF notation, nonterminals are enclosed in angle brackets (e.g., `<program>`, `<statement>`), while terminals are keywords, symbols, or literals shown in quotes (e.g., `"const"`, `":="`) or unquoted special symbols (e.g., `empty` for epsilon).

PL/0 Grammar (EBNF)

```
<program> ::= <block> "."

<block> ::= <const-declaration> <var-declaration> <statement>

<const-declaration> ::= [ "const" <ident> "=" <number>
                        {"," <ident> "=" <number>} ";"]

<var-declaration>  ::= [ "var" <ident> {"," <ident>} ";"]

<statement> ::=
    [
      <ident> ":=" <expression>
    | "begin" <statement> { ";" <statement> } "end"
    | "if" <condition> "then" <statement> "fi"
    | "while" <condition> "do" <statement>
    | "read" <ident>
    | "write" <expression>
    | empty
    ]

<condition> ::= "even" <expression>
            | <expression> <rel-op> <expression>

<expression> ::= <term> { ("+" | "-") <term> }

<term> ::= <factor> { ("*" | "/" ) <factor> }

<factor> ::=
      <ident>
    | <number>
    | "(" <expression> ")"

<rel-op>  ::= "=" | "<>" | "<" | "<=" | ">" | ">="

<number> ::= <digit> { <digit> }

<ident>  ::= <letter> { <letter> | <digit> }

<digit>  ::= "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"

<letter> ::= "a" | "b" | ... | "z" | "A" | "B" | ... | "Z"
```

Figure 1: Grammar to be implemented by the HW3 parser

**Important notes about this grammar:**

- There is NO "procedure" declaration in this grammar.
- Statement can be empty (epsilon production).

Reserved words, symbols, and token type names must exactly match the table provided in HW2. Identifiers and numbers retain the lexemes emitted by the scanner.

# 5 Compilation Instructions

Compile and run on **Eustis**. The only permitted language is C.

Compilation Commands

```
// Compile the scanner to produce executable
gcc -O2 -std=c11 -o lex lex.c

// Compile the parser/code generator to produce executable
gcc -O2 -std=c11 -o parsercodegen parsercodegen.c
```

**Note:** The only permitted language is C. The main logic for HW3 is in the `parsercodegen.c` file.

Do not require additional libraries beyond the C standard library. Any makefiles or helper scripts must preserve the same compiler flags and target names.

# 6 Testing Instructions

Invoke programs from the terminal/command line on **Eustis** (the university grading server).

## 6.1 Scanner Requirements

- `lex.c` must accept exactly **ONE** command-line parameter: the path to the PL/0 source program to be scanned.
- Do not prompt for input.
- Reject incorrect argument counts with a helpful usage message and exit.
- The scanner should write all output to a file (not standard output for the token list).
- Never request additional input from the user.

## 6.2 Parser/Code Generator Requirements

- `parsercodegen.c` requires **NO** command-line arguments.
- The input filename should be hard-coded (e.g., `tokens.txt` or `token_list.txt`).
- The parser reads the token file and generates assembly code.

## 6.3 Command-Line Usage Examples

Usage Examples

```
./lex inputfile.txt
./parsercodegen
```

**Explanation:** This works because `lex` produces the token list file, and `parsercodegen` reads that token list file (with hard-coded filename) for parsing and code generation.

# 7 Input and Output Specifications

This section describes the exact input and output requirements for both the scanner and parser/code generator, as well as comprehensive error handling requirements. For complete working examples, see Appendix B (correct input), Appendix C (scanning error), and Appendix D (syntax error).

## 7.1 Scanner (lex.c) Input and Output

### 7.1.1 Input

- **Input file:** A PL/0 source program provided as a command-line argument
- **Format:** Plain text file containing PL/0 source code
- **Example:** `./lex input.txt`
- **See:** Appendix B for a complete example of correct PL/0 input

### 7.1.2 Output

- **Output file:** The scanner must write tokens to a file with a hard-coded filename
  - **Filename flexibility:** You may use `tokens.txt`, `token_list.txt`, or similar variations
  - **Important:** Your parser/code generator must read from the same filename that your scanner writes to

  - The exact filename is your choice, but it must be consistent between scanner and parser

- **Token format flexibility:** You may choose your own token format
  - **Example format 1:** Two lines per token (token type on line 1, token value on line 2 if applicable)
  - **Example format 2:** One line per token (token type and value on same line)
  - **Example format 3:** Any other format that your parser can read
  - **Important:** The format shown in Appendix B is just ONE example—your implementation may use a different format
  - Your scanner and parser must work together with whatever format you choose

- **Error handling:** If lexical errors are detected (identifier too long, number too long, or invalid symbols), replace the erroneous lexeme with token type **1** (skipsym) and continue processing the entire file
  - **See:** Appendix C for a complete example of scanning error handling

- **No terminal output required** for the scanner (all output goes to file)

## 7.2 Parser/Code Generator (parsercodegen.c) Input and Output

### 7.2.1 Input

- **Input file:** Token list file with hard-coded filename (must match the output filename your scanner uses)

- **Format:** Token list produced by your scanner (must be able to read whatever format your scanner produces)

- **Important:** Your parser and scanner work together—they must agree on both filename and format

### 7.2.2 Output

- **Terminal output:**
  - If no errors: Display generated assembly code with line numbers and symbol table
    - **See:** Appendix B for an example of correct terminal output
  - If errors detected: Display appropriate error message (see Section 7.4)
    - **See:** Appendix C for scanning error output
    - **See:** Appendix D for syntax error output

- **File output (`elf.txt`):**

- If no errors: Write assembly code in numeric format (one instruction per line: `OP L M`)
  - **See:** Appendix B for an example of correct `elf.txt` output
- If errors detected: Write the same error message as displayed on terminal
  - **See:** Appendix C and Appendix D for error output examples

## 7.3 Error Handling Requirements

Your parser/code generator must detect and report errors appropriately. When an error is detected, the program must:

1. Stop processing immediately (do not attempt to continue parsing or code generation)
2. Output the appropriate error message to the terminal
3. Write the same error message to `elf.txt`
4. Exit gracefully

## 7.4 Required Error Messages

Your `parsercodegen.c` implementation **must** support the following error messages. Each error message must be output **exactly** as specified below (case-sensitive, exact wording):

### 7.4.1 Scanning Error (Lexical Error)

Scanning Error Message - REQUIRED

**Error Message:**
`Error: Scanning error detected by lexer (skipsym present)`

**When to output:** Token type 1 (skipsym) is found in the token list
**Meaning:** The scanner detected a lexical error (identifier too long, number too long, or invalid symbol)
**Complete Example:** See Appendix C for a detailed walkthrough showing:

- Input file with lexical error (identifier too long)
- Scanner output with token type 1
- Parser detection and error message output
- Terminal and `elf.txt` output

### 7.4.2 Syntax Errors (Grammar Violations)

The following error messages correspond to violations of the PL/0 grammar rules. See Appendix D for a complete example of syntax error handling.

**Syntax Error Messages - ALL REQUIRED**

**Your parser must support ALL of the following error messages:**

1. `Error: program must end with period`
   - **Complete Example:** See Appendix D for detailed walkthrough
2. `Error: const, var, and read keywords must be followed by identifier`
3. `Error: symbol name has already been declared`
4. `Error: constants must be assigned with =`
5. `Error: constants must be assigned an integer value`
6. `Error: constant and variable declarations must be followed by a semicolon`
7. `Error: undeclared identifier`
8. `Error: only variable values may be altered`
9. `Error: assignment statements must use :=`
10. `Error: begin must be followed by end`
11. `Error: if must be followed by then`
12. `Error: while must be followed by do`
13. `Error: condition must contain comparison operator`
14. `Error: right parenthesis must follow left parenthesis`
15. `Error: arithmetic equations must contain operands, parentheses, numbers, or symbols`

**Important:** Each error message must be output **exactly** as shown above (case-sensitive, exact wording). The autograder will check for exact string matches.

### 7.4.3 Important Notes on Error Messages

- All error messages must begin with `Error:` followed by a space
- Error messages are case-sensitive and must match exactly as specified
- When an error is detected, output the error message to both terminal and `elf.txt`
- Do not output multiple error messages—stop at the first error encountered

# 8 Submission Instructions

Submit on **Webcourses**. Programs are compiled and tested on **Eustis**. Follow these to avoid deductions.

## 8.1 Code Requirements

- **Program names.** Submit exactly two source files: `lex.c` and `parsercodegen.c`.

- **Header comment (copy/paste).** Place this header comment at the top of your source file (both `lex.c` and `parsercodegen.c`).

REQUIRED Header Comment - Copy and Paste Into Your Code

**Instructions:** Copy the entire comment block below and paste it at the very top of both `lex.c` and `parsercodegen.c`. Replace `<Full Name 1>, <Full Name 2>` with the actual names of all group members.

```
/*
Assignment:
HW3 - Parser and Code Generator for PL/0

Author(s): <Full Name 1>, <Full Name 2>

Language: C (only)

To Compile:
  Scanner:
    gcc -O2 -std=c11 -o lex lex.c
  Parser/Code Generator:
    gcc -O2 -std=c11 -o parsercodegen parsercodegen.c

To Execute (on Eustis):
  ./lex <input_file.txt>
  ./parsercodegen

where:
  <input_file.txt> is the path to the PL/0 source program

Notes:
  - lex.c accepts ONE command-line argument (input PL/0 source file)
  - parsercodegen.c accepts NO command-line arguments
  - Input filename is hard-coded in parsercodegen.c
  - Implements recursive-descent parser for PL/0 grammar
  - Generates PM/0 assembly code (see Appendix A for ISA)
  - All development and testing performed on Eustis

Class: COP3402 - System Software - Fall 2025

Instructor: Dr. Jie Lin

Due Date: Friday, October 31, 2025 at 11:59 PM ET
*/
```

## 8.2 What to Submit

**IMPORTANT:** All items marked as **REQUIRED** MUST be submitted. Failure to submit any required items will result in automatic penalties as specified in the grading section.

### 8.2.1 Category A - Always Required:

- Your source code files (`lex.c` and `parsercodegen.c`) - **REQUIRED**
- The AI Usage Disclosure Form with your signature for each group member - **REQUIRED**

### 8.2.2 Category B - Conditionally Required:

- **If you used AI:** A separate markdown file describing your AI usage - **REQUIRED when AI was used**
- **If you did not use AI:** No additional files needed beyond Category A items

**WARNING:** The submission requirements above are unambiguous. Any missing required items will result in automatic point deductions. There are no exceptions for misunderstanding these requirements.

## 8.3 Submission Guidelines

- Submit your source and any required build files on Webcourses before the due date. Late submissions may incur penalties.
- Ensure both `lex.c` and `parsercodegen.c` compile without warnings using the commands in Section 5.
- `lex.c` must accept exactly ONE command-line argument (the input PL/0 source file path). Print a usage message and exit for any other argument count.
- `parsercodegen.c` must accept NO command-line arguments and use a hard-coded input filename (e.g., `tokens.txt`).

# 9 Grading

This homework assignment is graded based on correctness and completeness through detailed manual examination and review of your implementation. Grading follows a deduction-based model: all students start with a maximum score of 100 points, and points are deducted for errors, missing requirements, compilation failures, academic integrity violations, and other issues discovered during the grading process.

## 9.1 Submission Files

The following deductions apply to missing or incorrect submission files:

- **-100 points:** Either `lex.c` or `parsercodegen.c` is missing from the submission.
- **-100 points:** Program names are not `lex.c` and `parsercodegen.c` exactly as specified.
- **-100 points:** Missing the header usage instructions/comment block in either file (see Section 8.1 for detailed requirements). This is considered an academic integrity violation.
- **-100 points:** Author section in the header comment is not modified to include the actual names of all group members. This is considered an academic integrity violation.

## 9.2 Compilation and Execution

The following deductions apply to compilation and execution issues:

- **-100 points:** Either program cannot compile on Eustis or does not compile from command line using the prescribed commands (`gcc -O2 -std=c11 -o lex lex.c` and `gcc -O2 -std=c11 -o parsercodegen parsercodegen.c`).
- **-100 points:** `lex.c` does not support exactly 1 command line parameter (the input PL/0 source file path).
- **-100 points:** `parsercodegen.c` requires command line arguments instead of using a hard-coded input filename.
- **-100 points:** Programs cannot read the required input files.
- **-100 points:** Code is written in a language other than C.

## 9.3 Academic Integrity

Academic integrity violations result in severe penalties:

- **-100 points:** Any instance of plagiarism, including direct copying from other students or sources, fabricated symbol table entries, fabricated implementations, or use of AI-generated code without substantial modification and understanding.
- **-100 points:** Missing or incomplete header comment blocks (see Section 8.1).
- **-100 points:** Failing to modify the author section in the header comment to include the actual names of all group members.
- **-100 points:** Implementation that does not follow the specified PL/0 grammar.

- **-100 points:** `lex.c` does not write output to a file (fails to create the token list file in the directory), OR `parsercodegen.c` does not read any input file for the token list. This indicates failure to implement the core file I/O functionality required for the assignment.
- **-5 points:** Missing or incomplete AI Usage Disclosure Form.
- Plagiarism detection is performed using automated similarity analysis (see Section 9.7 at the end of this section for detailed information).

## 9.4 Code Generation and Output Format

The following deductions apply to code generation and output formatting issues:

- **-100 points:** The first instruction in the generated code file (`elf.txt`) is not `JMP 0 3` (encoded as `7 0 3`). This instruction appears as `JMP 0 3` when displayed to the terminal but must be encoded as `7 0 3` in `elf.txt` (where 7 is the opcode for JMP).
- **-10 points:** Errors in LOD (load) and STO (store) instruction generation.
- **-5 points per instance:** Implementation does not correctly follow the grammar specification.
- **-5 points:** Incorrect PM/0 assembly output compared to expected results.
- **-5 points:** Output not formatted according to the ISA specification in Appendix A.

## 9.5 Symbol Table Requirements

The following deductions apply to symbol table implementation errors:

- **-5 points:** Incorrect value field for variable symbols.
- **-5 points:** Level not set to 0 for all symbols (since there are no procedures in this grammar).
- **-5 points:** Incorrect address calculation for variables.
- **-5 points:** Incorrect marking field (not initialized to 0 or not updated to 1 after use).
- **-5 points:** Any other symbol table implementation errors.

## 9.6 Error Handling

The following deductions apply to error handling issues:

- **-5 points:** Incorrect or missing error messages for invalid input.
- **-5 points:** Not detecting all required errors from HW2 scanner.

**Note on Late Submissions:** Late submissions are penalized according to Section 1.3.
**Note on Additional Deductions:** Graders reserve the right to apply additional 5-point deductions for significant errors discovered during manual review that do not fit into the predefined categories above. This grading rubric is comprehensive but not exhaustive.

## 9.7 Plagiarism Detection and Manual Verification

**Important Notice:** We take academic integrity seriously. All submissions undergo automated similarity analysis through JPlag. If similarity scores exceed established thresholds, the following process will be initiated:

1. **Manual Verification:** Submissions with high similarity scores will be manually reviewed by instructional staff.
2. **Zero Tolerance Policy:** If manual verification reveals direct copying from other students or direct usage of AI-generated code without substantial modification and understanding, the entire assignment will receive a score of zero (**-100 points**).
3. **AI Disclosure Irrelevant:** Having an AI disclosure form does not exempt submissions from plagiarism penalties if direct copying is detected.
4. **Similarity Patterns:** AI tools often generate similar code patterns. Simply copying and pasting AI-generated code will likely result in high similarity scores and trigger manual review.

The primary grading criterion is correctness and completeness combined with adherence to academic integrity standards. Follow all instructions precisely; deviations may result in additional deductions at the graders' discretion.
**FINAL GRADE CALCULATION:** Your final grade is calculated by starting with 100 points and subtracting all applicable deductions from the categories above. Academic integrity violations result in automatic assignment scores of zero (-100 points), regardless of technical performance.

# A Instruction Set Architecture (ISA)

The PM/0 virtual machine supports nine opcodes. Each instruction is encoded by a three-number tuple $\langle OP, L, M \rangle$. The tables below summarize each opcode along with a brief description and pseudocode. See Table 2 for `OPR` sub-operations.

**Note:** Your parser and code generator must emit instructions in this ISA format. The generated assembly code file should contain one instruction per line in the format: `OP L M`.

Table 1: PM/0 Instruction Set (Core)

| Opcode | OP Mnemonic | L | M | Description & Pseudocode |
|---|---|---|---|---|
| 01 | LIT | 0 | $n$ | **Literal push.**<br>`sp` $\leftarrow$ `sp` $-1$<br>`pas[sp]` $\leftarrow$ $n$ |
| 02 | OPR | 0 | $m$ | **Operation code;** see Table 2 for specific operations.<br>*See OPR table for operation details* |
| 03 | LOD | $n$ | $a$ | **Load value to top of stack** from offset $a$ in the AR $n$ static levels down.<br>`sp` $\leftarrow$ `sp` $-1$<br>`pas[sp]` $\leftarrow$ `pas[base(bp,`$n$`)` $-a$`]` |
| 04 | STO | $n$ | $o$ | **Store top of stack** into offset $o$ in the AR $n$ static levels down.<br>`pas[base(bp,`$n$`)` $-o$`]` $\leftarrow$ `pas[sp]`<br>`sp` $\leftarrow$ `sp` $+1$ |
| 05 | CAL | $n$ | $a$ | **Call procedure** at code address $a$; create activation record.<br>`pas[sp`$-1$`]` $\leftarrow$ `base(bp,`$n$`)`<br>`pas[sp`$-2$`]` $\leftarrow$ `bp`<br>`pas[sp`$-3$`]` $\leftarrow$ `pc`<br>`bp` $\leftarrow$ `sp`$-1$<br>`pc` $\leftarrow$ $499 - a$ |
| 06 | INC | 0 | $n$ | **Allocate $n$ locals** on the stack.<br>`sp` $\leftarrow$ `sp` $-n$ |
| 07 | JMP | 0 | $a$ | **Unconditional jump** to address $a$.<br>`pc` $\leftarrow$ $499 - a$ |
| 08 | JPC | 0 | $a$ | **Conditional jump:** if value at top of stack is 0, jump to $a$; pop the stack.<br>`if pas[sp]` $=$ `0 then pc` $\leftarrow$ $499 - a$<br>`sp` $\leftarrow$ `sp` $+1$ |
| 09 | SYS | 0 | 1 | **Output integer** value at top of stack; then pop.<br>`print(pas[sp])`<br>`sp` $\leftarrow$ `sp` $+1$ |
| 09 | SYS | 0 | 2 | **Read an integer** from stdin and push it.<br>`sp` $\leftarrow$ `sp` $-1$<br>`pas[sp]` $\leftarrow$ `readInt()` |
| 09 | SYS | 0 | 3 | **Halt the program.**<br>`halt` |

Table 2: PM/0 Arithmetic and Relational Operations (OPR, opcode 02, L=0)

| Opcode | OP Mnemonic | L | M | Description & Pseudocode |
|---|---|---|---|---|
| 02 | RTN | 0 | 0 | **Return from subroutine** and restore caller's AR.<br>`sp` ← `bp` +1<br>`bp` ← `pas[sp−2]`<br>`pc` ← `pas[sp−3]` |
| 02 | ADD | 0 | 1 | **Addition.**<br>`pas[sp+1]` ← `pas[sp+1]` + `pas[sp]`<br>`sp` ← `sp` +1 |
| 02 | SUB | 0 | 2 | **Subtraction.**<br>`pas[sp+1]` ← `pas[sp+1]` − `pas[sp]`<br>`sp` ← `sp` +1 |
| 02 | MUL | 0 | 3 | **Multiplication.**<br>`pas[sp+1]` ← `pas[sp+1]` ∗ `pas[sp]`<br>`sp` ← `sp` +1 |
| 02 | DIV | 0 | 4 | **Integer division.**<br>`pas[sp+1]` ← `pas[sp+1]` / `pas[sp]`<br>`sp` ← `sp` +1 |
| 02 | EQL | 0 | 5 | **Equality comparison** (result 0/1).<br>`pas[sp+1]` ← `(pas[sp+1]` == `pas[sp])`<br>`sp` ← `sp` +1 |
| 02 | NEQ | 0 | 6 | **Inequality comparison** (result 0/1).<br>`pas[sp+1]` ← `(pas[sp+1]` ≠ `pas[sp])`<br>`sp` ← `sp` +1 |
| 02 | LSS | 0 | 7 | **Less-than comparison** (result 0/1).<br>`pas[sp+1]` ← `(pas[sp+1]` < `pas[sp])`<br>`sp` ← `sp` +1 |
| 02 | LEQ | 0 | 8 | **Less-or-equal comparison** (result 0/1).<br>`pas[sp+1]` ← `(pas[sp+1]` ≤ `pas[sp])`<br>`sp` ← `sp` +1 |
| 02 | GTR | 0 | 9 | **Greater-than comparison** (result 0/1).<br>`pas[sp+1]` ← `(pas[sp+1]` > `pas[sp])`<br>`sp` ← `sp` +1 |
| 02 | GEQ | 0 | 10 | **Greater-or-equal comparison** (result 0/1).<br>`pas[sp+1]` ← `(pas[sp+1]` ≥ `pas[sp])`<br>`sp` ← `sp` +1 |
| 02 | EVEN | 0 | 11 | **Even check** (result 0/1).<br>`pas[sp+1]` ← `(pas[sp] % 2 == 0)`<br>`sp` ← `sp` +1 |

**Important:** For this assignment (HW3), since there are no procedures in the grammar, you will **not** use the `CAL` or `RTN` instructions. All generated code should use `L=0` for `LOD` and `STO` instructions.

# B Appendix B: Complete Example for Correct Input

This appendix provides a complete, correct example showing the entire pipeline from PL/0 source code through lexical analysis to parser/code generation output. This example demonstrates what your implementation should produce.

## B.1 Input: PL/0 Source Code

The following PL/0 program is the input file that `lex.c` reads:

Input File

```
var x, y;
begin
    if y <> x then
        write y
    fi;
    write x+ 1;
end.
```

## B.2 Scanner Output: tokens.txt

After running `./lex input.txt`, your `lex.c` may produce a `tokens.txt` file similar to the following. **Note:** This is one possible implementation. Your token file format may differ based on your HW2 design, but it must contain equivalent information.

Example tokens.txt File

```
29
2 x
16
2 y
17
20
22
2 y
9
2 x
24
31
2 y
23
17
31
2 x
4
3 1
17
21
18
```

**Hint:** This shows one way to represent tokens. Token type 29 represents `var`, token type 2 represents identifiers (followed by the identifier name), token type 16 represents comma, etc. Your implementation may use a different format (e.g., a lexeme table), but it must convey the same token sequence to the parser.

## B.3 Parser/Code Generator Terminal Output

After running `./parsercodegen`, your program reads the token file and outputs the following to the terminal. This output shows both the generated assembly code and the symbol table:

Terminal Output

```
Assembly Code:

Line     OP   L   M
  0     JMP   0   3
  1     INC   0   5
  2     LOD   0   4
  3     LOD   0   3
  4     OPR   0   6
  5     JPC   0  24
  6     LOD   0   4
  7     SYS   0   1
  8     LOD   0   3
  9     LIT   0   1
 10     OPR   0   1
 11     SYS   0   1
 12     SYS   0   3

Symbol Table:

Kind | Name        | Value | Level | Address | Mark
---------------------------------------------------
   2 |           x |     0 |     0 |       3 |    1
   2 |           y |     0 |     0 |       4 |    1
```

**Explanation:**

- The assembly code shows the PM/0 instructions in human-readable format with line numbers.
- The symbol table shows all declared variables with their properties.
- Kind 2 indicates a variable (as opposed to a constant).
- Level is 0 for all symbols (no procedures in this grammar).
- Address shows the memory offset for each variable.
- Mark indicates whether the variable has been used (1 = used, 0 = unused).

## B.4 Parser/Code Generator File Output: elf.txt

In addition to terminal output, `parsercodegen` must write the executable code to a file (e.g., `elf.txt`). This file contains the same instructions but in numeric format (opcode, L, M) without line numbers or mnemonics:

File Output: elf.txt

```
7 0 3
6 0 5
3 0 4
3 0 3
2 0 6
8 0 24
3 0 4
9 0 1
3 0 3
1 0 1
2 0 1
9 0 1
9 0 3
```

**Important Notes:**

- The first instruction must be `7 0 3` (JMP 0 3), which jumps to the main program code.
- Each line contains three space-separated integers: opcode, L-value, and M-value.
- This file format is required for the PM/0 virtual machine to execute your code.
- Refer to Appendix A for the complete ISA specification and opcode mappings.

## B.5 Summary

This complete example demonstrates:

1. How `lex.c` transforms PL/0 source code into a token stream.
2. How `parsercodegen.c` reads the token stream and generates PM/0 assembly code.
3. The required output formats for both terminal display and file output.
4. The symbol table structure and content.

Use this example to verify your implementation produces correct output for valid PL/0 programs.

# C Appendix C: Scanning Error Example

This appendix demonstrates how your implementation should handle input files containing **lexical/scanning errors**. The key principle is that **the scanner (lex.c) continues processing the entire input file even when errors are detected**, but represents errors as a special error token (token type 1). The parser (parsercodegen.c) then detects this error token and halts immediately.

## C.1 Input: PL/0 Source Code with Error

The following PL/0 program contains a lexical error (identifier name too long):

Input File with Error

```
const z = 5;
var xcjioasunihujacioj, y;
begin
    xcjioasunihujacioj := y * 2;
end.
```

**Error:** The identifier `xcjioasunihujacioj` is 18 characters long, which exceeds the maximum allowed length of 11 characters.

## C.2 Scanner Behavior and Output

**Important:** Your `lex.c` scanner must **NOT** stop when it encounters an error. Instead, it should:

1. Continue processing the entire input file
2. Replace any erroneous lexeme with token type **1** (error token/skipsym)
3. Generate a complete token list including the error token(s)

After running `./lex input.txt`, your `lex.c` produces the following `tokens.txt` file:

tokens.txt with Error Token

```
28
2 z
8
3 5
17
29
1
16
2 y
17
20
1
19
2 y
6
3 2
17
21
18
```

**Explanation:**

- Token type **1** represents an error token (skipsym)
- The first occurrence of token **1** (line 7) represents the identifier that was too long in the variable declaration
- The second occurrence of token **1** (line 11) represents the same identifier in the assignment statement
- All three scanner errors (name too long, number too long, invalid symbols) are represented by token type **1**
- The scanner continues processing and generates tokens for all valid lexemes

## C.3 Parser/Code Generator Behavior

When `parsercodegen.c` reads the `tokens.txt` file and encounters token type **1**, it must:

1. Immediately recognize that the scanner detected an error
2. Stop parsing and code generation
3. Output a unified error message
4. Write the same error message to both terminal and file

## C.4 Terminal Output

After running `./parsercodegen`, the terminal displays:

Terminal Output for Error

```
Error: Scanning error detected by lexer (skipsym present)
```

## C.5 File Output: elf.txt

The `elf.txt` file contains the same error message:

elf.txt for Error

```
Error: Scanning error detected by lexer (skipsym present)
```

## C.6 Error Handling Summary

### C.6.1 Scanner (lex.c) Responsibilities

- Detect all three types of lexical errors:
  1. Identifier name too long (exceeds 11 characters)
  2. Number too long (exceeds maximum allowed digits)
  3. Invalid symbols (characters not in the PL/0 alphabet)
- Replace each error with token type **1** (skipsym)
- **Continue processing** the entire input file
- Generate a complete token list including error tokens

### C.6.2 Parser/Code Generator (parsercodegen.c) Responsibilities

- Read the token list from the file
- Check for the presence of token type **1**
- If token type **1** is found:
  - Stop immediately (do not attempt to parse)
  - Output the unified error message to terminal
  - Write the same error message to `elf.txt`
- The exact error message must be: `Error: Scanning error detected by lexer (skipsym present)`

### C.6.3 Key Principle

All three scanner errors produce the same unified error message from the parser. The parser does not need to distinguish between different types of lexical errors—it only needs to detect that an error occurred (presence of token type 1) and halt with the appropriate message.

# D Appendix D: Syntax Error Example

This appendix demonstrates how your implementation should handle **syntax errors** (grammar violations). Unlike Appendix C which dealt with **scanning errors** (lexical errors detected by `lex.c`), this appendix focuses on **syntax errors** (grammar violations detected by `parsercodegen.c`).

## D.1 Key Distinction: Scanning vs. Syntax Errors

- **Scanning Errors (Appendix C):** Lexical errors detected by the scanner (e.g., identifier too long, number too long, invalid symbols). The scanner represents these as token type 1 (skipsym), and the parser detects the presence of this error token.
- **Syntax Errors (This Appendix):** Grammar violations detected by the parser during parsing (e.g., missing period, missing semicolon, incorrect statement structure). The scanner produces valid tokens, but the token sequence does not match the grammar rules.

## D.2 Input: PL/0 Source Code with Syntax Error

The following PL/0 program contains a syntax error (missing period at the end):

Input File with Missing Period

```
var x, y;
begin
    read y;
    x := y * 2;
end
```

**Error:** According to the grammar specification (see Figure 1), a program must follow the production:

```
<program> ::= <block> "."
```

The period symbol after `end` is **required** but missing in this input.

## D.3 Scanner Output: tokens.txt

The scanner (`lex.c`) processes this input and produces valid tokens. **Note:** The scanner does not detect this error because all lexemes are valid—the error is in the grammar structure, not in individual tokens.

tokens.txt for Syntax Error Example

```
29
2 x
16
2 y
17
20
32
2 y
17
2 x
19
2 y
6
3 2
17
21
```

**Explanation:**

- All tokens are valid (no token type 1 present)
- Token type 21 represents `end`
- Token type 18 (period) is missing at the end
- The scanner successfully tokenized all lexemes
- The error will be detected by the parser, not the scanner

## D.4 Parser Behavior

When `parsercodegen.c` parses the token list:

1. It successfully parses the variable declaration (`var x, y;`)
2. It successfully parses the `begin...end` block
3. After parsing `end` (token 21), it expects a period (token 18) according to the grammar rule: `<program> ::= <block> "."`
4. It encounters end-of-file instead of the required period
5. It detects a syntax error and halts

## D.5 Terminal Output

After running `./parsercodegen`, the terminal displays:

Terminal Output for Syntax Error

```
Error: program must end with period
```

## D.6 File Output: elf.txt

The `elf.txt` file contains the same error message:

elf.txt for Syntax Error

```
Error: program must end with period
```

## D.7 Syntax Error Handling Summary

### D.7.1 Scanner (lex.c) Behavior

- Processes all lexemes successfully
- Produces valid tokens (no token type 1)
- Does **NOT** detect syntax errors
- Syntax error detection is the parser's responsibility

### D.7.2 Parser (parsercodegen.c) Responsibilities

- Parse tokens according to the grammar rules
- Detect when the token sequence violates grammar rules
- For missing period error specifically:
  - After parsing the block, expect token type 18 (period)
  - If period is missing, output the specific error message
  - Halt parsing and code generation
- Output error message to both terminal and `elf.txt`
- The exact error message must be: `Error: program must end with period`

### D.7.3 Key Principle

The scanner handles lexical errors by emitting error tokens (token type 1). The parser handles both scanning errors (by detecting token type 1) and syntax errors (by detecting grammar violations during parsing). Each type of error produces a specific, appropriate error message.

# E Appendix E: Symbol Table Structure

This appendix provides the recommended data structure for implementing the symbol table in your `parsercodegen.c` implementation.

## E.1 Recommended Symbol Table Structure

**Symbol Table Data Structure - RECOMMENDED**

**Important:** While you may use alternative data structures (e.g., linked lists, dynamically allocated arrays), the **structure fields must remain the same** as shown below. The autograder expects these specific field names and types.

```
#define MAX_SYMBOL_TABLE_SIZE 500

typedef struct {
    int kind;       // const = 1, var = 2, proc = 3
    char name[12];  // name up to 11 chars
    int val;        // number (ASCII value)
    int level;      // L level
    int addr;       // M address
    int mark;       // to indicate unavailable or deleted
} symbol;

symbol symbol_table[MAX_SYMBOL_TABLE_SIZE];
```

**Field Descriptions:**

- `kind`: Type of symbol (1 = constant, 2 = variable, 3 = procedure)
- `name[10]`: Symbol name (up to 11 characters including null terminator)
- `val`: For constants, stores the constant value
- `level`: Lexicographical level (L in PM/0 instructions)
- `addr`: Memory address (M in PM/0 instructions)
- `mark`: Flag to indicate if symbol is available (0) or deleted/unavailable (1)

## E.2 Symbol Table Size

- **Maximum size:** 500 symbols (`MAX_SYMBOL_TABLE_SIZE = 500`)
- **Sufficiency:** A table size of 500 is sufficient for all test cases in this assignment

- **Recommendation:** Use the simple array-based structure shown above for ease of implementation
- **Alternative structures:** You may use linked lists or dynamically allocated memory if you prefer, but the structure fields must remain the same

## E.3 Required Storage for Different Symbol Types

### E.3.1 For Constants (kind = 1)

When storing a constant symbol, you **must** populate the following fields:

**Required Fields for Constants**

- `kind = 1`
- `name`: The constant's identifier name
- `val`: The constant's integer value
- `level = 0` (all symbols are at level 0 in this grammar)
- `mark = 0` (initially available)

**Note:** The `addr` field is not used for constants.

### E.3.2 For Variables (kind = 2)

When storing a variable symbol, you **must** populate the following fields:

**Required Fields for Variables**

- `kind = 2`
- `name`: The variable's identifier name
- `level = 0` (all symbols are at level 0 in this grammar)
- `addr`: The variable's memory address (M value for LOD/STO instructions)
- `mark = 0` (initially available)

**Note:** The `val` field is not used for variables.

## E.4 Symbol Table Operations

Your implementation should support the following operations:

- **Insert:** Add a new symbol to the table (check for duplicate names first)

- **Lookup:** Search for a symbol by name
- **Mark:** Set the `mark` field to 1 to indicate the symbol can be no longer used

# F Appendix F: Pseudocode

This appendix provides pseudocode for implementing the recursive-descent parser and code generator. The pseudocode closely follows the grammar structure and demonstrates the general approach for parsing and code generation.

**CRITICAL WARNING: This pseudocode is NOT an exact implementation of the grammar!**

**Important points:**

- The pseudocode provides a **starting point** but is **NOT 100% accurate**
- You **MUST** carefully study the grammar specification (Figure 1) and understand how it differs from this pseudocode
- **Blindly translating this pseudocode to C will result in deductions**
- You are responsible for modifying and adapting the pseudocode to correctly match the grammar
- Use this as a **guide**, not as a complete solution
- The pseudocode is a good starting point, but you must verify every detail against the grammar

Each function below is presented in a separate box for clarity. Study each function carefully and compare it with the corresponding grammar production.

## F.1 Symbol Table Helper Function

```
SYMBOLTABLECHECK (string)
    linear search through symbol table looking at name
    return index if found, -1 if not
```

## F.2 PROGRAM Function

```
PROGRAM
    BLOCK
    if token != periodsym
        error
    emit HALT
```

## F.3 BLOCK Function

```
BLOCK
    CONST-DECLARATION
    numVars = VAR-DECLARATION
    emit INC (M = 3 + numVars)
    STATEMENT
```

## F.4 CONST-DECLARATION Function

```
CONST-DECLARATION
    if token == const
        do
            get next token
            if token != identsym
                error
            if SYMBOLTABLECHECK (token) != -1
                error
            save ident name
            get next token
            if token != eqlsym
                error
            get next token
            if token != numbersym
                error
            add to symbol table (kind 1, saved name, number, 0, 0)
            get next token
        while token == commasym
        if token != semicolonsym
            error
        get next token
```

## F.5 VAR-DECLARATION Function

```
VAR-DECLARATION - returns number of variables
    numVars = 0
    if token == varsym
        do
            numVars++
            get next token
            if token != identsym
                error
            if SYMBOLTABLECHECK (token) != -1
                error
            add to symbol table (kind 2, ident, 0, 0, var# + 2)
            get next token
        while token == commasym
        if token != semicolonsym
            error
        get next token
    return numVars
```

## F.6 STATEMENT Function

```
STATEMENT
    if token == identsym
        symIdx = SYMBOLTABLECHECK (token)
        if symIdx == -1
            error
        if table[symIdx].kind != 2 (not a var)
            error
        get next token
        if token != becomessym
            error
        get next token
        EXPRESSION
        emit STO (M = table[symIdx].addr)
        return
    if token == beginsym
        do
            get next token
            STATEMENT
        while token == semicolonsym
        if token != endsym
            error
        get next token
        return
    if token == ifsym
        get next token
        CONDITION
        jpcIdx = current code index
        emit JPC
        if token != thensym
            error
        get next token
        STATEMENT
        code[jpcIdx].M = current code index
        return
    if token == whilesym
        get next token
        loopIdx = current code index
        CONDITION
        if token != dosym
            error
        get next token
        jpcIdx = current code index
```

```
        emit JPC
        STATEMENT
        emit JMP (M = loopIdx)
        code[jpcIdx].M = current code index
        return
    if token == readsym
        get next token
        if token != identsym
            error
        symIdx = SYMBOLTABLECHECK (token)
        if symIdx == -1
            error
        if table[symIdx].kind != 2 (not a var)
            error
        get next token
        emit READ
        emit STO (M = table[symIdx].addr)
        return
    if token == writesym
        get next token
        EXPRESSION
        emit WRITE
        return
```

## F.7 CONDITION Function

```
CONDITION
    if token == oddsym
        get next token
        EXPRESSION
        emit ODD
    else
        EXPRESSION
        if token == eqlsym
            get next token
            EXPRESSION
            emit EQL
        else if token == neqsym
            get next token
            EXPRESSION
            emit NEQ
        else if token == lessym
            get next token
            EXPRESSION
            emit LSS
        else if token == leqsym
            get next token
            EXPRESSION
            emit LEQ
        else if token == gtrsym
            get next token
            EXPRESSION
            emit GTR
        else if token == geqsym
            get next token
            EXPRESSION
            emit GEQ
        else
            error
```

## F.8 EXPRESSION Function

```
EXPRESSION
    if token == minussym
        get next token
        TERM
        emit NEG
        while token == plussym || token == minussym
            if token == plussym
                get next token
                TERM
                emit ADD
            else
                get next token
                TERM
                emit SUB
    else
        if token == plussym
            get next token
        TERM
        while token == plussym || token == minussym
            if token == plussym
                get next token
                TERM
                emit ADD
            else
                get next token
                TERM
                emit SUB
```

## F.9 TERM Function

```
TERM
    FACTOR
    while token == multsym || token == slashsym || token == modsym
        if token == multsym
            get next token
            FACTOR
            emit MUL
        else if token == slashsym
            get next token
            FACTOR
            emit DIV
        else
            get next token
            FACTOR
            emit MOD
```

## F.10 FACTOR Function

```
FACTOR
    if token == identsym
        symIdx = SYMBOLTABLECHECK (token)
        if symIdx == -1
            error
        if table[symIdx].kind == 1 (const)
            emit LIT (M = table[symIdx].value)
        else (var)
            emit LOD (M = table[symIdx].addr)
        get next token
    else if token == numbersym
        emit LIT
        get next token
    else if token == lparentsym
        get next token
        EXPRESSION
        if token != rparentsym
            error
        get next token
    else
        error
```