CS152 Project Phase 1: Lexical Analyzer Generation Using flex

CS152 (Compiler Construction)

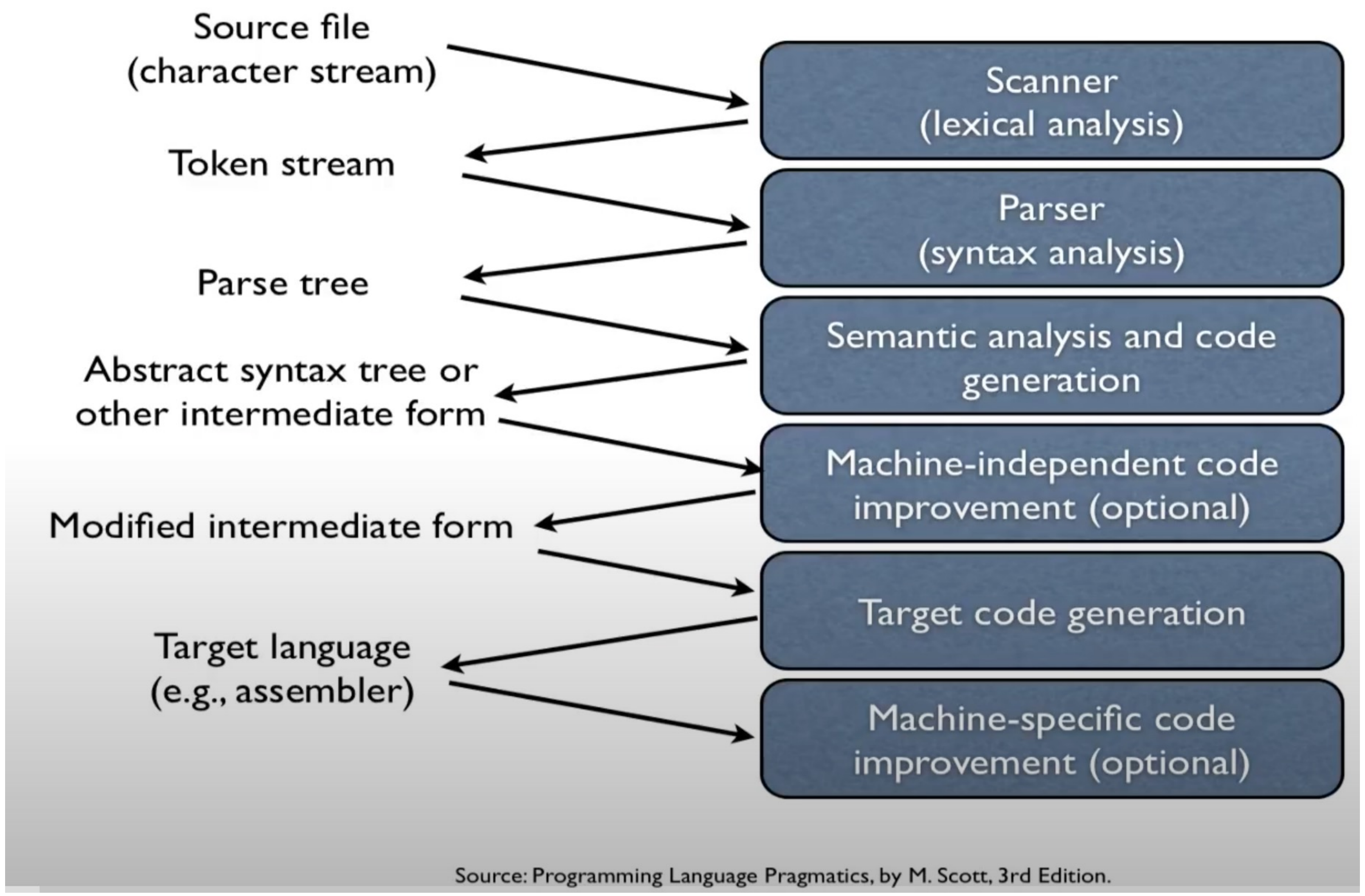
Mahbod Afarin

Spring 2021

Outline

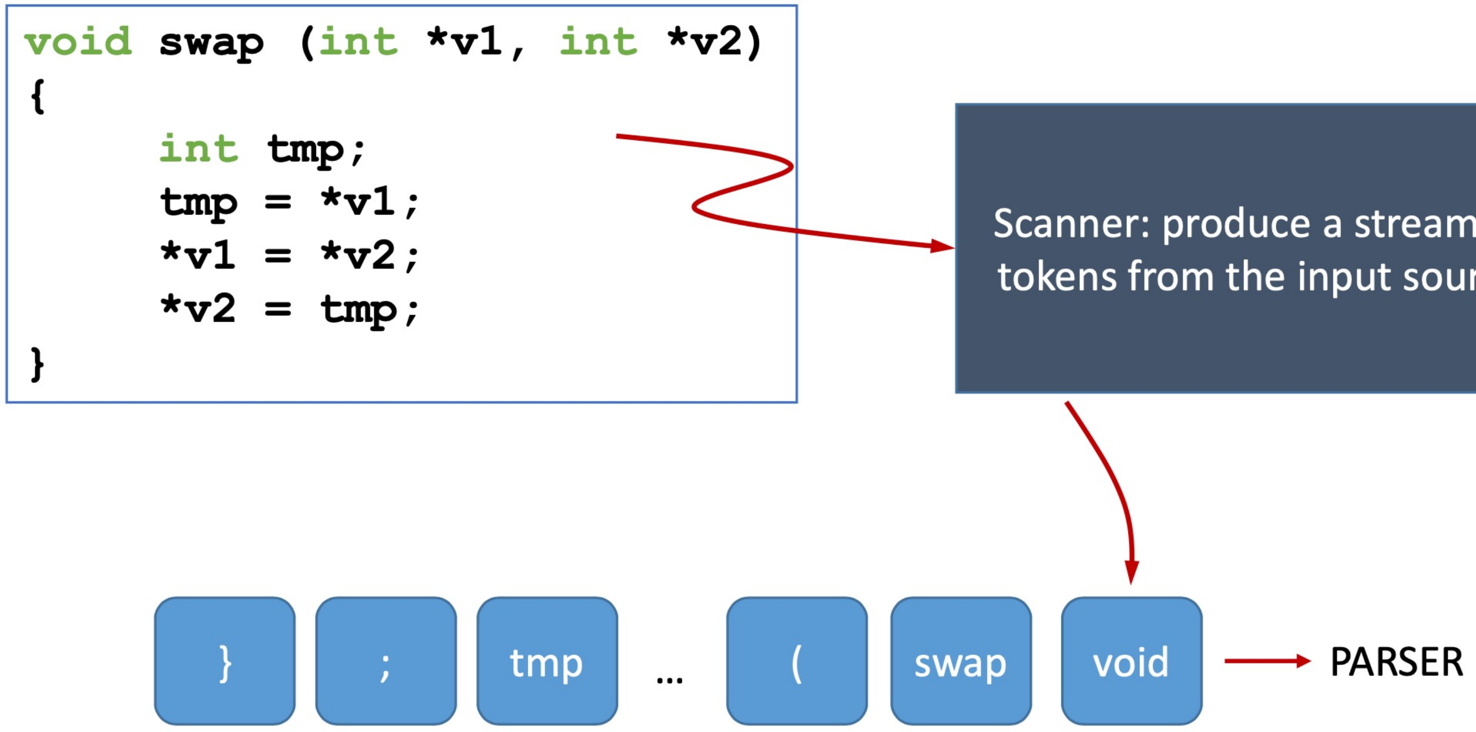
- Introduction to lexical analysis
- What is flex?
- How does it work?
- A sample program

Language Processing


Source file
(character stream)
Scanner
(lexical analysis)
Token stream
Parser
(syntax analysis)
Parse tree
Semantic analysis and code generation
Abstract syntax tree or other intermediate form
Machine-independent code improvement (optional)
Modified intermediate form
Target code generation
Target language
(e.g., assembler)
Machine-specific code improvement (optional)

Source: Programming Language Pragmatics, by M. Scott, 3rd Edition.

Lexical Analysis



```
void swap (int *v1, int *v2)
{
      int tmp;
      tmp = *v1;
      *v1 = *v2;
      *v2 = tmp;
}
```

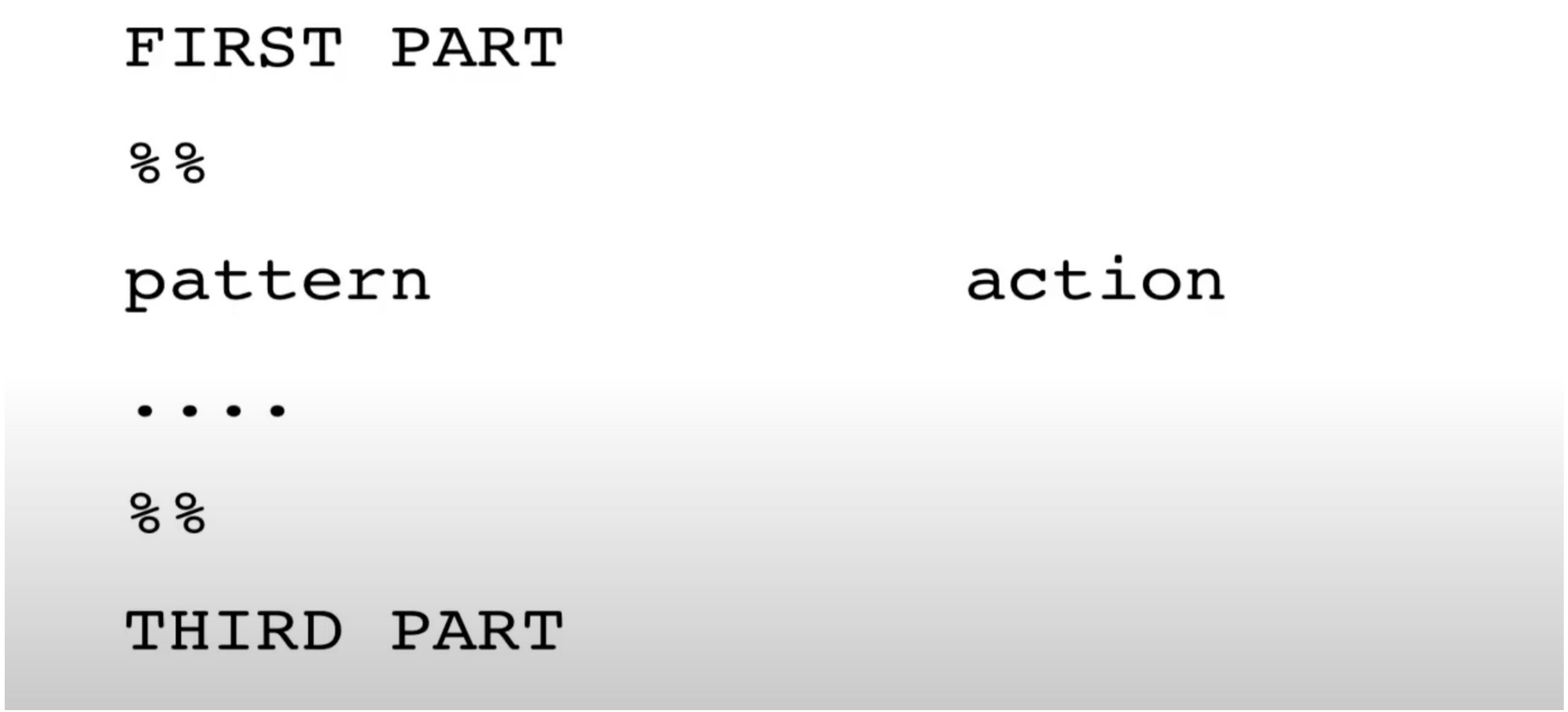
Scanner: produce a stream of tokens from the input source

Lex/flex

- lex is a scanner generator
  - Input is a set of regular expressions and associated actions (written in C)
  - Output is a table-driven scanner (lex.yy.c)

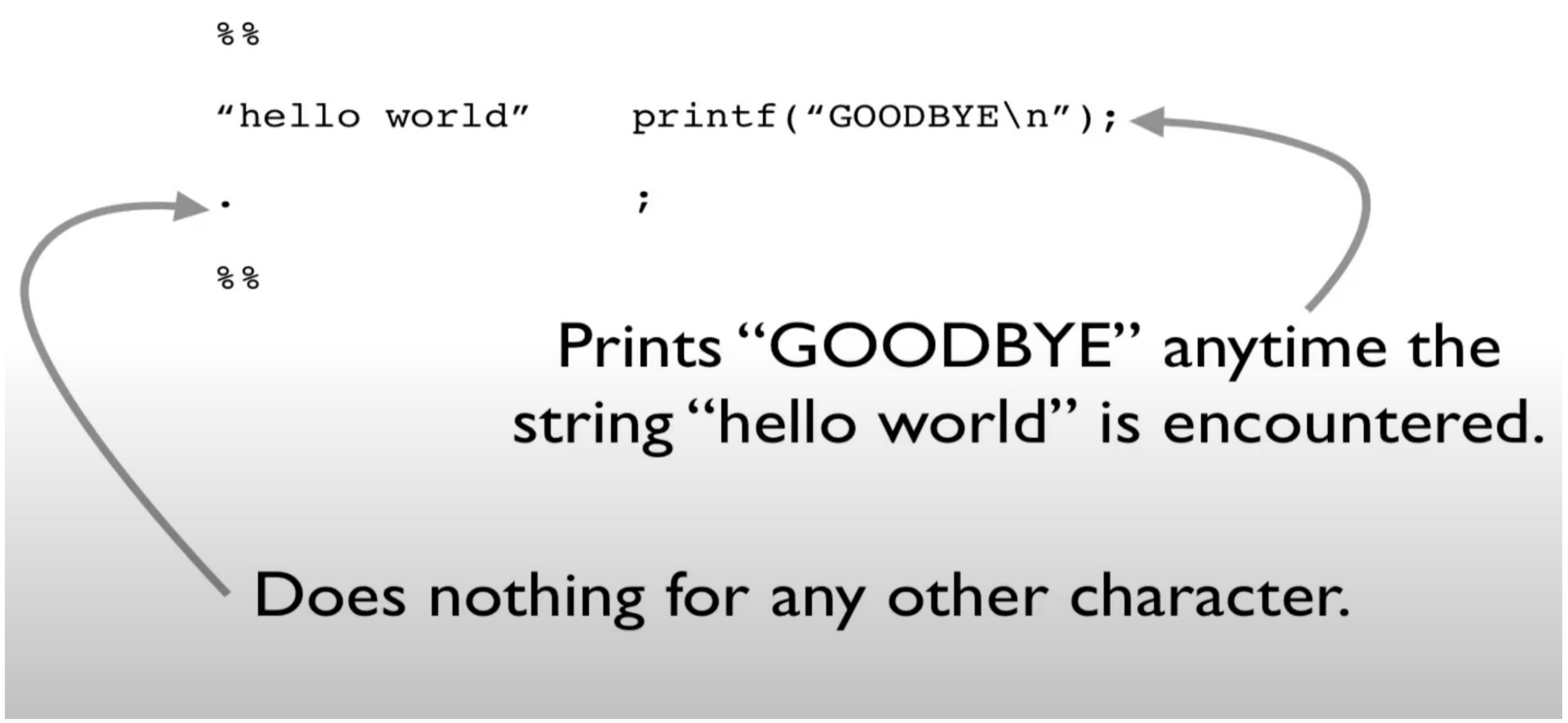
  - GNU flex: an open source implementation of the original UNIX lex utility

flex Input



```
FIRST PART
%%
pattern                     action
....
%%
THIRD PART
```

flex Input Example (I)



```
%%
“hello world”     printf(“GOODBYE\n”);
.                 ;
%%
```

Prints “GOODBYE” anytime the string “hello world” is encountered.

Does nothing for any other character.

Running flex

- $ flex exl.lex
- $ gcc lex.yy.c -lfl
- $ ./a.out
- hello world
- GOODBYE!

Process the flex file to generate a scanner (gets saved as lex.yy.c)

Compile the scanner

Run the scanner taking input from the standard input.

Flex pattern examples

Pattern	
abc	Match the string “abc”
[a-zA-Z]	Match any lower or uppercase letter
dog.*cat	Match any string starting with dog, and ending with cat
(ab)+	Match one or more occurrences of “ab” concatenated
[^a-z]+	Matches any string of one or more characters that do not include lower case a-z
[+-]?[0-9]+	Match any string of one or more digits with an optional prefix of + or -
^[a-z]+	Matches any line starting with a lower case letter

UC RIVERSIDE
UNIVERSITY OF CALIFORNIA

Thanks for your listening.