UNIVERSITY COMPUTER SCIENCE • EXERCISE SHEET

11.4 Lexical analysis

Name: _____ Class: _____ Date: _____ **Total: 24 marks**

KEY WORDS tokens • deterministic finite automaton

Objective

Lex by hand with longest match, positions and policy.

1 Key knowledge and worked examples

■ The pipeline

Regexes compile to DFAs; the lexer emits (type, lexeme, position) with greedy longest match and priority tie-breaks ✓.

■ Worked example: the course line

`count = a1 + b2;` → IDENT ASSIGN IDENT PLUS IDENT SEMI at columns 1, 7, 9, 12, 14, 16 ✓. `ifx` is one IDENT ✓.

2 Practice

2.1 Why must the lexer keep positions? [3]

2.2 What wins between `if` and `ident` on the input `ifx`? [3]

2.3 What is discarded on the way through? [3]

3 Exam-style questions

3.1 Lex `total = ifx + 42;` under the course specification (keyword `if`; IDENT letters/digits; NUM digits; operators =, +; punctuation ;).

List the token stream with type, lexeme and column for each. Justify the `ifx` decision by longest-match-plus-priority. State the NUM policy for a following letter (`42abc`) under our spec. State which single token would change if keywords were removed from the spec. [8]

3.2 Two error policies for `12abc`: report as an error, or emit NUM(12) then IDENT(abc).

State what each policy makes legal downstream. Argue which fits a language where identifiers cannot start with a digit. State the general principle the choice illustrates about lexer specification ✓. [7]

Solutions

2.1

- Every later diagnostic (parse and semantic errors) cites them ✓.

2.2

- `ident` – the longest lexeme wins; `ifx` is one IDENT ✓.

2.3

- Whitespace and comments ✓.

3.1

- IDENT(total)@1, ASSIGN(=)@7, IDENT(ifx)@9, PLUS(+)@13, NUM(42)@15, SEMI(;)@17 ✓.
- `ifx`: the ident regex matches 3 characters, `if` matches 2 – longest wins; priority would matter only on exact ties ✓.
- Under our spec `42abc` is NUM(42) then IDENT(abc) – reported as a lexical error (adjacent value tokens with no separator) ✓.
- Without keywords, nothing changes on THIS line – `ifx` stays IDENT; `if` alone would change from KW to IDENT ✓.

3.2

- Report-as-error: the lexer refuses and no parser state is built ✓.
- Emit-both: NUM then IDENT reaches the parser, which will choke on two adjacent value tokens ✓.
- Identifiers cannot start with a digit here, so `12abc` cannot be one token of either class – the error policy matches the language ✓.
- Principle: the error policy is part of the lexer's specification – chosen, not discovered ✓.