SAARLAND UNIVERSITY
COMPUTER SCIENCE

Prof. Dr. Reinhard Wilhelm
Prof. Dr. Sebastian Hack
Dipl.-Inform. Daniel Grund
Michael Jacobs, M.Sc.

Compiler Construction WS11/12

Exercise Sheet 1

Please hand in the solutions to the theoretical exercises until the beginning of the lecture next Friday 2011-10-28, 12:00. Please write the number of your tutorial group or the name of your tutor on the first sheet of your solution.

Exercise 1.1. Regular Expressions and Finite Automata (Points: 3+3+3)

1. For the regular expression $(a|b)^*bac(a|bc)^*$ over the alphabet $\Sigma = \{a, b, c\}$ construct:
   a) a nondeterministic finite automaton,
   b) a deterministic finite automaton, and
   c) a minimal deterministic finite automaton

   following the proof sketches from the lecture.

Exercise 1.2. Greed (Points: 3+3)

1. Provide an automaton and a corresponding input word, such that the maximal munch strategy has runtime in $\Omega(n^2)$. Prove all your claims!

2. Is there a sequence of regular expressions $\alpha_1, \ldots, \alpha_n$ over Σ and a word $w \in \Sigma^*$, such that
   - there exist $w_1, \ldots, w_k \in \Sigma^*$ and $t_1, \ldots, t_k \in \{1, \ldots, n\}$, such that $w = w_1 \ldots w_k$ and $w_i \in \alpha_{t_i}$ (that is, there is a match of the complete word), and
   - for each such dissection there exist $j \in \{1, \ldots, k-1\}$, $x \in \Sigma^+$, $y \in \Sigma^*$, and $s \in \{1, \ldots, n\}$, such that $w_{j+1} \ldots w_k = xy$, $w_j x \in \alpha_s$, and there is no dissection $y = w'_1 \ldots w'_m$ and $t'_1, \ldots, t'_m \in \{1, \ldots, n\}$, such that $w'_i \in \alpha_{t'_i}$ (that is, greedy scanning does not take us to a match)?

Exercise 1.3. Minimizing Automata (Points: 5)

Consider the following alternative algorithm for minimizing automata: Let $\langle \Sigma, Q, \Delta, q_0, F \rangle$ be a DFA. We merge two states $q_1, q_2 \in Q$ iff $q_1 \in F \Leftrightarrow q_2 \in F$ and for all $a \in \Sigma$, for all $p \in Q$ holds: $(q_1, a, p) \in \Delta \Leftrightarrow (q_2, a, p) \in \Delta$. We apply this rule until no further states can be merged.

Does this algorithm correctly minimize arbitrary deterministic finite automata? Find a proof or a counter example to back up your claim!

Project task A. General Information

In the practical project you will implement a compiler for a small subset of Java. The milestones of the project roughly reflect the structure of a compiler, which also structures the lecture. Although the milestones will not be graded, you will receive feedback that hints at existing shortcomings in your compiler. The final submission will be graded on the basis of percentage of tests passed and a code review where the group members have to justify their work.

Technical requirements and restrictions:

- You are not allowed to use existing tools that more or less directly solve the assignments, e.g. you must not use lexer or parser generators. If you feel like it, however, you are free to implement your own tools.
- The compiler itself must be written in C or C++.
- Executing the command `make` in your top-level directory must produce an executable file `mjavac` in that same directory.[1]
- Upon acceptance of a source program, `mjavac` must terminate with return code 0. Rejection of a source program must be indicated by return code 1 and should print an error message.
- The individual assignments will refine the requirements.

Project task B. Lexer

To get started:

- Set up a version control system of your choice.
- Go to the web page of the lecture and download the project materials.
- Examine the contents of the starter kit. It provides a Makefile and the public tests for the lexer. Each `.ref` file contains the expected output for the respective `.java` file.
- Read the language specification of MiniJava and identify its symbols/tokens.

Implement a lexical analysis for MiniJava. You do not necessarily have to follow the automata approach presented in the lecture.

- Files that contain only valid tokens, comments, and whitespace must be accepted and `mjavac --tokens [file]` must print the corresponding token stream to the standard output. For example, the line `int i = 423;` must produce the following token stream:

  ```
  int
  IDENT i
  =
  INTEGER_LITERAL 423
  ;
  ```

  Remember that whitespace and comments are not part of the token stream.
- Files that contain lexical errors must be rejected.
- The running time of your parser should be proportional to the input size.
- Write more test cases, especially some that test the rejection capabilities of your lexer.
- The next project task will be to implement a parser. You therefore might want to structure your source code in a way that separates lexing and printing of tokens. . . .

Please check in your solution into your repository until 2011-11-03, 12:00. Solutions submitted later may not receive feedback.

[1] Do not even think about trying to include malicious or "funny" code.