Лабораторная работа №4
«Объектно-ориентированный лексический анализатор»

Скоробогатов С. Ю., Коновалов А. В.

22 марта 2016

1 Цель работы

Целью данной работы является приобретение навыка реализации лексического анализатора на объектно-ориентированном языке без применения каких-либо средств автоматизации решения задачи лексического анализа.

2 Задание

В лабораторной работе предлагается реализовать на языке Java две первые фазы стадии анализа: чтение входного потока и лексический анализ. При этом следует придерживаться схемы реализации объектно-ориентированного лексического анализатора, рассмотренной на лекции.

Входной поток должен загружаться из файла (в UTF-8). В результате работы программы в стандартный поток вывода должны выдаваться описания распознанных лексем в формате

```
Тег (координаты_фрагмента): атрибут лексемы
```

При этом для лексем, не имеющих атрибутов, нужно выводить только тег и координаты. Например,

```
IDENT (1, 2)-(1, 4): str
ASSIGN (1, 8)-(1, 9):
STRING (1, 11)-(1, 16): qwerty
```

Лексемы во входном файле могут разделяться пробельными символами (пробел, горизонтальная табуляция, маркеры окончания строки), а могут быть записаны слитно (если это не приводит к противоречиям).

Входной файл может содержать ошибки, при обнаружении которых лексический анализатор должен выдавать сообщение с указанием координаты, восстанавливаться и продолжать работу.

Варианты языков для лексического анализа приведены в таблицах 1 и 2.

Таблица 1: Краткое описание лексики вариантов языков

1	Идентификаторы: последовательности латинских букв, начинающиеся с гласной буквы. Числовые литералы: последовательности десятичных цифр, перед которыми может стоять знак «минус». Операции: «−−», «<», «<=».
2	Комментарии: начинаются с «/*», заканчиваются на «*/» и могут пересекать границы строк текста. Идентификаторы: последовательности латинских букв и десятичных цифр, в которых буквы и цифры чередуются. Ключевые слова: «for», «if», «m1».
3	Строковые литералы: ограничены апострофами, для включения апострофа в литерал он удваивается, не пересекают границы строк текста. Числовые литералы: последовательности десятичных цифр, которые могут включать точку и предваряться знаком «минус». Идентификаторы: последовательности буквенных символов Unicode, точек и цифр, начинающиеся с буквы.
4	Идентификаторы: либо последовательности латинских букв, либо непустые последовательности десятичных цифр, ограниченные круглыми скобками. Числовые литералы: либо последовательности десятичных цифр, не начинающиеся с нуля, либо «0». Операции: «()», «:», «:=».
5	Комментарии: начинаются с «//» и продолжаются до окончания строки текста. Идентификаторы: любой текст, не содержащий «/» и ограниченный символами «/». Ключевые слова: «/while/», «/do/», «/end/».
6	Строковые литералы: ограничены двойными кавычками, могут содержать Escape-последовательности «\"», «\n» и «\t», не пересекают границы строк текста. Числовые литералы: последовательности десятичных цифр, разбитые точками на группы по три цифры («100», «1.000», «1.000.000»). Идентификаторы: последовательности буквенных символов Unicode, знаков подчёркивания и цифр, начинающиеся с буквы или подчёркивания.
7	Идентификаторы: последовательности латинских букв и десятичных цифр, оканчивающиеся на цифру. Числовые литералы: последовательности десятичных цифр, органиченные знаками «<» и «>». Операции: «<=», «=», «==».
8	Комментарии: целиком строка текста, начинающаяся с «*». Идентификаторы: либо последовательности латинских букв нечётной длины, либо последовательности символов «*». Ключевые слова: «with», «end», «**».
9	Строковые литералы: органичены двойными кавычками, для включения двойной кавычки она удваивается, для продолжения литерала на следующей строчке текста в конце текущей строчки ставится знак «\». Числовые литералы: либо последовательности десятичных цифр, либо последовательности шестнадцатеричных цифр, начинающиеся с «$». Идентификаторы: последовательности буквенных символов Unicode, цифр и знаков «$», начинающиеся с буквы.
10	Идентификаторы: последовательности заглавных латинских букв, за которыми могут располагаться последовательности знаков «+», «−» и «*». Числовые литералы: знак «*» или последовательности, состоящие целиком либо из знаков «+», либо из знаков «−». Ключевые слова: «ON», «OFF», «**».

Таблица 2: Краткое описание лексики вариантов языков (продолжение)

11	Комментарии: начинаются с «(*» или «{», заканчиваются на «*)» или «}» и могут пересекать границы строк текста. Идентификаторы: последовательности латинских букв, представляющие собой конкатенации двух одинаковых слов («zz», «abab»). Ключевые слова: « ifif », «do», «dodo».
12	Строковые литералы: ограничены обратными кавычками, могут занимать несколько строчек текста, для включения обратной кавычки она удваивается. Числовые литералы: десятичные литералы представляют собой последовательности десятичных цифр, двоичные – последовательности нулей и единиц, оканчивающиеся буквой «b». Идентификаторы: последовательности десятичных цифр и знаков «?», «*» и «|», не начинающиеся с цифры.
13	Идентификаторы: последовательности буквенных символов Unicode и десятичных цифр, начинающиеся и заканчивающиеся на одну и ту же букву. Числовые литералы: последовательности шестнадцатеричных цифр (чтобы литерал не был похож на идентификатор, его можно предварять нулём). Ключевые слова: «qeq», «xx», «xxx».
14	Символьные литералы: ограничены апострофами, могут содержать Escape-последовательности «\'», «\n» и «\xxxx» (в последней Escape-последовательности буквы «x» обозначают шестнадцатеричные цифры). Идентификаторы: последовательности символов Unicode длиной от 2 до 10 символов, начинающиеся и заканчивающиеся буквой. Ключевые слова: «z», «for», «forward».
15	Идентификаторы: последовательности буквенных символов Unicode и цифр, начинающиеся с буквы. Знаки операций: либо последовательности, состоящие из знаков !, #, $, %, &, ∗, +, ., /, <, =, >, ?, @, \, ^, |, − и ~, либо идентификаторы, записанные в обратных кавычках (например, ‘plus‘). Ключевые слова where, −>, =>.
16	Комментарии: начинаются с −− и продолжаются до конца строки, либо начинаются с {− и заканчиваются на −}, могут занимать несколько строк текста. Целочисленные литералы: последовательности десятичных цифр. Вещественные литералы: последовательности десятичных цифр, за которой следует либо дробная часть (десятичная точка и последовательность десятичных цифр, возможно, пустая), либо показатель степени (буква e или E, за которой следует непустая последовательность десятичных цифр), либо дробная часть и показатель степени.
17	Идентификаторы: последовательности буквенных символов Unicode и цифр, начинающиеся с буквы.Числовые литералы: десятичные литералы представляют собой последовательности десятичных цифр, шестнадцатиричные — начинаются на десятичную цифру и заканчиваются символом h. Ключевые слова mov, eax.
18	Числовые литералы: знак 0 либо последовательности знаков 1. Строковые литералы: регулярные строки — ограничены двойными кавычками, могут содержать escape-последовательности \", \t, \n, не пересекают границы строк текста; буквальные строки — начинаются на @", заканчиваются на двойную кавычку, пересекают границы строк текста, для включения двойной кавычки она удваивается.
19	Идентификаторы: последовательности десятичных цифр. Числовые литералы: римские цифры или ключевое слово NIHIL, не чувствительны к регистру. Принцип вычитания (числа вида IV, IX, XL и др.) допустимо не реализовывать.