CS 331 Programming Languages, Spring 2009

Lexeme Description

Wednesday, February 4, 2009

We wish to write a lexical analyzer for a language with the following lexemes.

No lexeme contains a *whitespace* character (blank, tab, vertical-tab, new-line, carriage-return, form-feed). A whitespace character, or any contiguous group of whitespace characters, is considered as a separator between lexemes. However, pairs of lexemes are not *required* to be separated by whitespace.

Comments, which begin with two slashes ("//") and end with the next new-line, or the end of the input string if there is no following new-line, are treated as whitespace.

There are four kinds of lexemes: Identifier, Operator, Number, and Illegal.

Identifier

> Begins with letter (upper- or lower-case) or underscore ("_"). Contains only letters, digits, or underscores.

Operator

> Any single printable ASCII character other than letters, digits, underscores, and space characters, or "++", "--".

Number

> Where *D* represents a nonempty string of digits, a number has the form "*D*", "*D*. ", ".*D*", "*D*.*D*" (the *D*s may represent different strings), or any of these with a single "+" or "-" prepended.

Illegal

> Any non-ASCII or non-printable ASCII character, other than space characters. *Note: The printable ASCII set consists of those characters whose values range from 32 (blank) to 126 (tilde).*

Once a lexeme has begun, the complete lexeme is considered to be the longest legal substring beginning from the starting point.