

CS 331 Programming Languages, Spring 2026  
**In-Class Lexical Specification**

*A different lexical specification will be used in the assignments.*

A **whitespace** character is a blank, tab, vertical-tab, new-line, carriage-return, or form-feed. A whitespace character is not part of any lexeme. A contiguous group of whitespace characters separates lexemes. However, adjacent lexemes are not *required* to be separate by whitespace.

A **comment** begins with `"/*` and ends with `*/` or the end of the input. The second slash in `"/*` does *not* end the comment. A comment may contain any characters. A comment is treated as whitespace and is not part of any lexeme.

**Legal** characters outside comments are whitespace and printable ASCII (values 32 [blank] to 126 [tilde]). Any other character outside comments is **illegal**.

---

There are six categories of Lexemes: **Keyword**, **Identifier**, **NumericLiteral**, **Operator**, **Punctuation**, **Malformed**. Maximal munch applies.

*Below, in a regular expression, a character preceded by a backslash means the literal character, with no special meaning.*

**Keyword**

One of the following 3: `begin` `end` `print`

**Identifier**

Any string matched by `/[a-zA-Z_][a-zA-Z_0-9]*/` that is not a **Keyword**.

**NumericLiteral**

Any string matched by `/[\+-]?([0-9]+(\.[0-9]*)?)|\.[0-9]+/`.

**Operator**

One of the following 13: `+` `-` `*` `/` `++` `--` `.` `=` `==` `+=` `--` `*=` `/=`

**Punctuation**

A single legal non-whitespace character that is not part of any of the above.

**Malformed**

A single illegal character.