Lexical Conventions and Grammars of Languages: C vs. Python vs. Javascript

Ananya Mantravadi - CS19B1004

Lexical Aspects of C, Javascript, and Python

Line Structure and Whitespace

C and Javascript contain lines of statements that are terminated by semicolons whereas they are terminated with a newline in Python. Braces { } are used to group declarations and statements in C and Javascript, and indentation (with spaces preferably) is used in Python. Whitespace in C and Javascript includes comments starting with // or between /* and */ and # in Python. But since in Javascript /* or */ can occur in its regular expression literals, it is avoided.

Identifiers and Keywords

In C, Javascript, and Python, identifiers or names consist of letters (upper and lower cases), digits, and underscore(_) and must begin with a letter. They cannot be any of the keywords or reserved words of their language. Further, identifiers in Python that start or end with underscores may have special meanings.

Number Literals

Numbers are preceded by an optional + or - sign. In C, the integer type is determined by how the number begins (0x→hexadecimal, 0→octal, decimal otherwise) and is suffixed by (u/U→unsigned, l/L→long). Floating constants contain decimal part, fractional part, optional (e/E), and optional suffix (f/F/l/L). Javascript contains only a single number type (internally represented as a floating-point). In both Javascript and Python floating-point numbers contain decimal, fractional, and exponent parts. Python additionally has the numeric types Booleans (False→0, True→1), Integers (0x→hexadecimal, 0→octal, decimal otherwise), Long integers (suffixed with l/L), and Complex Numbers (suffixed with j/J).

String Literals

In Javascript, string literals are defined by enclosing text in single or double quotes, each character 16 bits wide (Unicode). A character is a string with just one character in it. Python supports both types of string literals 8-bits wide (ASCII) - defined by enclosing in

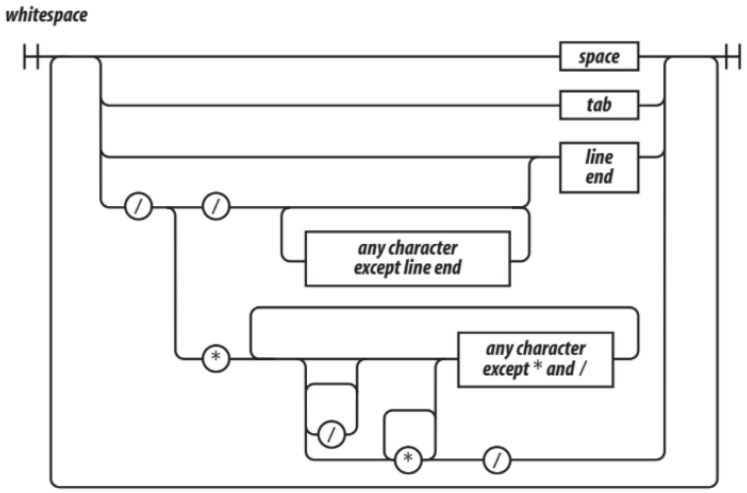
single, double, or triple quotes and Unicode - preceding with (u/U). However, in C, character and string literals (array of characters) are defined separately and are enclosed in single and double quotes respectively. In all three languages, escape sequences using backslash (\) allow for inserting special characters such as newlines, quotes, backslashes, nonprinting characters, etc. into these string/character literals. In Python, one may also precede string literals with (r/R) to keep all backslashes intact.

Operators and Special Symbols

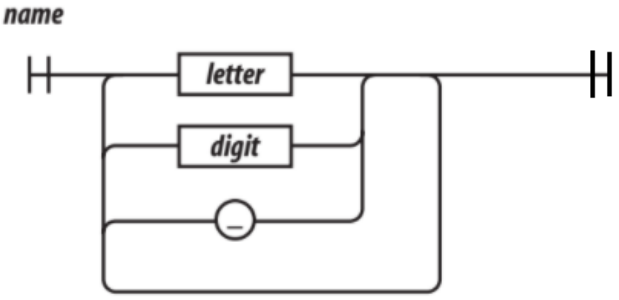
Operators are symbols used in expressions to evaluate arithmetic, relational, logical operations. They follow a precedence order defined by the language. Special symbols have special meaning and hence cannot be used for other purposes. Python uses ( ), [ ], { }, to define tuples, lists, and dictionaries respectively. C uses the same for indicating function calls, array element references, and marking blocks of code. Other special symbols include *, … , : , ; #, etc.

Transition Diagrams

Whitespace:
This diagram tells us what all are considered to be a part of whitespace in the code. This includes spaces, tabs, line ends, // followed by text till line end, and any text (except characters in the order */) between /* and */.



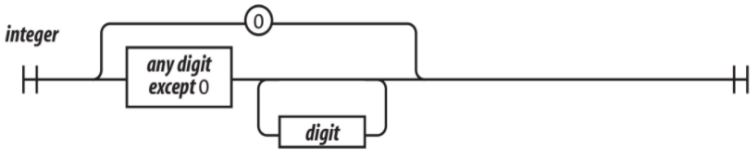
Names:
Here a name must definitely begin with a letter and can be followed by one or more letters, digits or underscores.



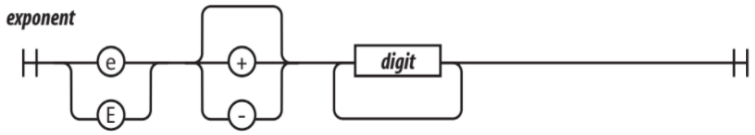
Numbers:
Each number literal consists of integer part with optional fraction and exponent parts.

Integer:
If integer starts with 0, it cannot be followed by more digits. If it begins with any digit except 0, it can be followed by more digits.

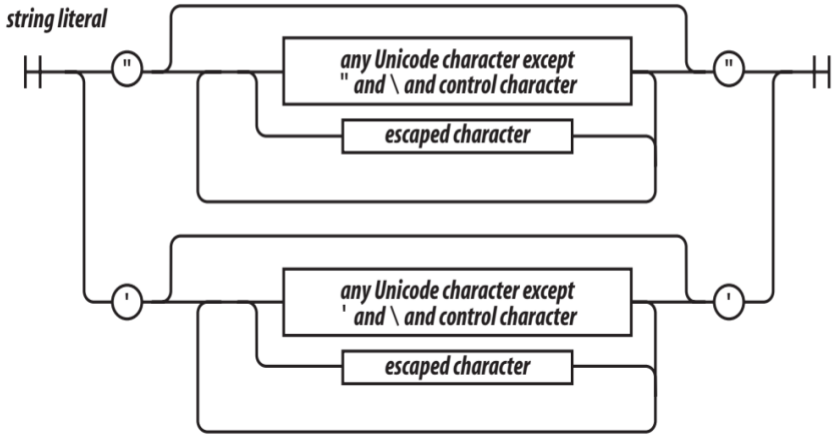
Fraction:
It always begins with . (decimal point), and is followed by zero or more digits.



Exponent:
It begins with either e or E, followed by optional sign symbols +/- , followed by one or more digits.

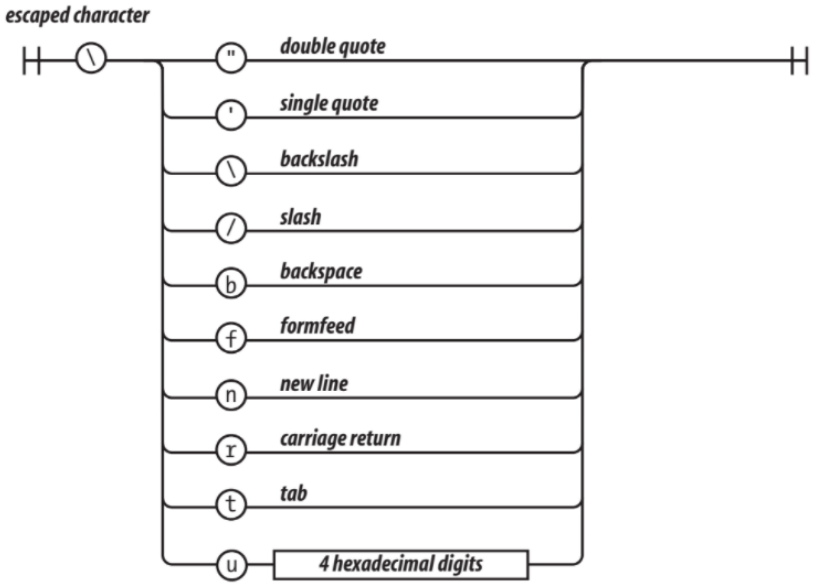


Strings:
It starts and ends with either double quotes (“) or single quotes (‘) and contains zero or more characters. Enclosed within quotes can be one or more unicode characters except for the quotes the string started with, backslash and control character or one or more escaped characters



interchangeably. If an escaped character is encountered, the transition digram below is followed.

Escaped Character:
It always begins with a backslash and is followed either by ”, ’, \, /, b, f, n, r, t, u and 4 hexadecimal digits.



Statements:
var statements are used to declare and optionally initialize a variable. It could be an empty line, or multiple var statements. If it begins with var keyword and name, either an expression can be assigned to the name or the statement is terminated by semicolon. For declaring multiple variables in one var statement, comma is used after name or expression to separate variables.