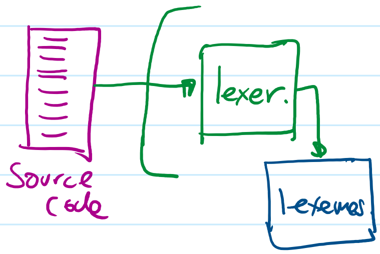
4 Syntax and Regular Expressions

Wednesday, September 10, 2025 11:51 AM



- Borrow from Linguistics and Mathematics.

- Orthography.- ¿How to write words?

  apple manzana pómme [illegible]

  A list of words gives you a Dictionary

  In a programming language.- Keywords.

- Syntax.- ¿How to write sentences?

  "has jump apple dog green"

  Syntax has rules.

  In programming languages also

- Semantics.- ¿What do sentences mean?

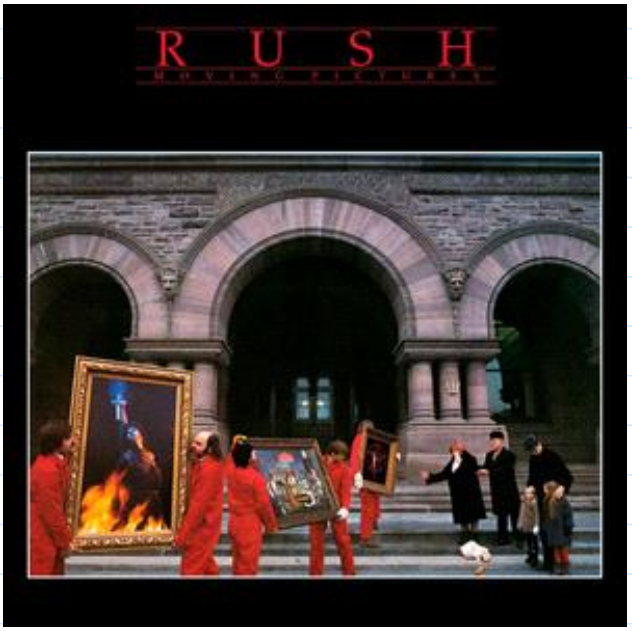
  "colorless green ideas sleep furiously" N. Chomsky

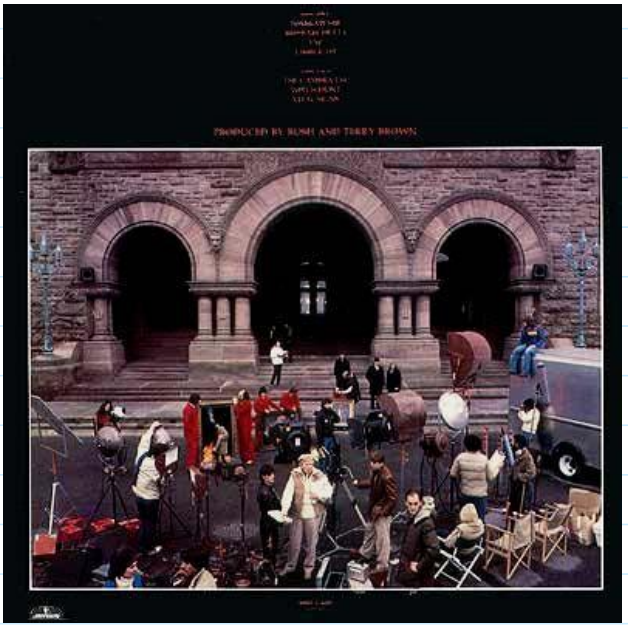
  Adj Adj noun verb adverb

  "The chickens are ready to eat"

  or ?

  "Moving Pictures"





- Pragmatics:- How meaning changes over time
  "the container has to handle an unreliable cloud"

- Programming Language Orthography and Syntax

What are the words of a programming language?

Symbols :- ASCII

- Keywords
  - Operators
- Literals. constant values 3 7.5 True 'apple'
- identifiers :- named entities.
  - variables
  - functions
  - classes
  - modules
  - Namespaces
  - Templates

How many literals or identifiers are there?
potentially infinite !!!

- We need rules!

We Borrow from Mathematics
"Mathematical Languages" S. Kleene E. Post

DEF:- An alphabet is a finite set of symbols

e.g. $\Sigma = \{a, b\}$

A language is a set of sequences / strings / words made of symbols in an alphabet

e.g $L = \{ a, aaa, aba, bba, bab, bbb \}$

- Problem #1 - the Specification problem
  - how to precisely describe a language?

  e.g $L = \{ a, ab, aba, abb, \ldots \}$???

- Problem #2 - the Recognition Problem
  given a word w and a language L
  is w in L? $w \in L$?

- Answer to #1: Regular Expressions
  built from Σ plus | * ( )

DEF: Regular Expressions

- a word w is a reg.ex.
- if r_1, r_2 are regex then
  - $r_1 r_2$
  - $r_1 \mid r_2$
  - r_1^*

  } are also regex

- A regular expression specifies a language

• w	$\{w\}$
• $r_1 r_2$	$\{ w_1 w_2 \mid w_1 \in r_1 \wedge w_2 \in r_2 \}$
• $r_1 \mid r_2$	$\{ w \mid w \in r_1 \vee w \in r_2 \}$
• r^* (Kleene Star)	$\{ w \mid w$ is zero or more repetitions of a word in r $\}$

E.g. $\Sigma = \{a, b\}$

- $(a \mid b)\, b = \{ab, bb\}$
  - $a = \{a\}$
  - $b = \{b\}$
  - $(a \mid b) = \{a, b\}$
- $(ab)^* = \{\lambda, ab, abab, ababab, \ldots\}$ (λ: empty string)

Note: an infinite set is clearly specified by finite means.

Note: an infinite set is clearly specified by finite means.

- $(aa \mid b)^* = \{\lambda, b, aa, bb, bbb, aab, baa, aaaa, bbbb, baab, \ldots\}$
  - $aa = \{aa\}$
  - $b = \{b\}$
  - $aa \mid b = \{aa, b\}$

Shorthands:
- $r? \equiv (r \mid \lambda)$ "to r, or not to r"
- $r+ \equiv r(r^*)$ "one or more repetitions of r"

e.g. $a(b?)(a^+) = \{aa, aba, aaa, abaa, aaaa, \ldots\}$
- $a = \{a\}$
- $b? = \{b, \lambda\}$
- $a+ = \{a, aa, aaa, \ldots\}$

- Regular expressions and Programming Languages

Σ = ASCII Keyboard characters

L_{C++} = The set of all c++ programs. RegEx not enough.

Note: RegEx do not have the power to specify such set.

Why?

e.g. $\{a^n b^n \mid n \in \mathbb{Z}^+\} = \{ab, aabb, aaabbb, aaaabbbb, \ldots\}$

consider matching brackets

{ ( ) [ ] } < > { ([ )] }

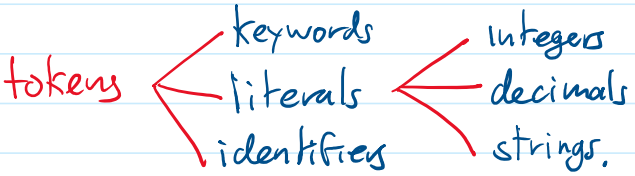
- Famous Stack Overflow question
  what is a regex to validate HTML? closed

 .....

<h3> ..... </h3>

- RegEx are good enough to describe tokens
  token - a category of atomic strings
  lexeme - an instance of a token

token :- a category of atomic strings
lexeme :- an instance of a token



lexemes while foo 3.7

- Describing Tokens using RegEx

  Σ = ASCII

  - Regex for keywords:
    C++ (if | while | for | swich | do | .....)
  - Regex for operators:
    (+ | * | - | % | / | == | <= | .....)