4 Syntax and Regular Expressions

Friday, September 11, 2026 12:04 PM

- Borrow from Linguistics & Mathematics.

– Orthography: ¿How to write words?

apple manzana pómme [illegible]

A complete list of words.: Dictionary

In a programming language: keywords.

– Syntax: ¿How to write sentences?

"has jump apple dog green"

Syntax has rules.

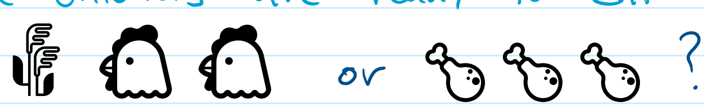
also programming languages.

– Semantics: ¿what does sentences mean?

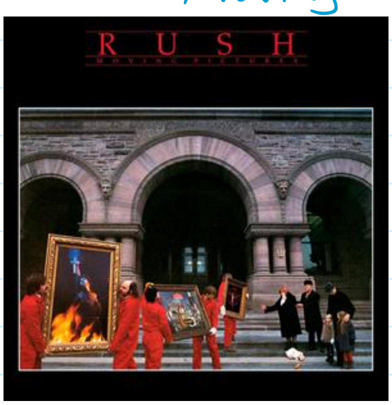
"colorless green ideas sleep fouriously"

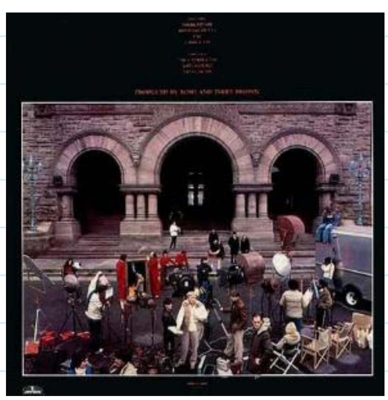
Adj Adj noun verb Adv.

"the chickens are ready to eat"



"Moving Pictures"





In programming languages we should avoid
- non-sensical sentences.
- ambiguous sentences.

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

- PROGRAMMING LANGUAGE ORTHOGRAPHY
What are the words of a programming language?

Symbols :- ASCII
- keywords.
- literals : constant values.
- identifiers: named entities
  - variables
  - classes
  - constants
  - functions
  - structs
  - modules.

- keywords can be listed
- How many identifiers are there? infinite
- How many literals are there? infinite.
Python: 5 6 7 "hello" "banana"

- We need Rules
We borrow from mathematics.
"Mathematical Languages" E. post S. kleene.

- An alphabet is a finite set of symbols:
$\Sigma = \{a, b\}$ $\Sigma = \{1, 0\}$

- A language is a set of sentences / strings / words made of the symbols in the alphabet
e.g. { a, ab, ba, bab, baab }

- Problem #1 : The specification problem.
How to precisely describe a language?

e.g.: $L = \{ a, ab, abb, abbb, \ldots \}$

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

- Problem #2: the Recognition problem.
  given a word w and a language L

  $w \in L$? is w a word in language L

- Answer #1: Regular Expressions
  Meta-Language to precisely describe a Language.

DEF: build from Σ plus | * ( )

- a word w is a RegEx
- if r_1, r_2 are RegEx then
  - $r_1 r_2$
  - $r_1 | r_2$
  - r_1^*

  } are also RegEx

- A RegEx specifies a Language
  - w $\{w\}$
  - $r1\ r2$ $\{w_1 w_2$ s.t. $w_1 \in r_1 \wedge w_2 \in r_2\}$
  - $r1 | r2$ $\{w$ s.t. $w \in r_1 \vee w \in r_2\}$
  - r^* $\{w$ s.t. w is zero or more repetitions of a word in $r\}$
    ↑ Kleene star.

e.g. $\Sigma = \{0, 1\}$

$((01)|((10)^*))10 = \{0110, 10, 1010, 101010, 10101010, \ldots\}$

$01 = \{01\}$
$10 = \{10\}$
$(10)^* = \{\lambda, 10, 1010, 101010, 10101010, 1010101010, \ldots\}$
$01 | (10)^* = \{01, \lambda, 10, 1010, 101010, \ldots\}$

Note: an Infinite Set is precisely defined by finite means

Note: an Infinite set is precisely defined by finite means

$\Sigma = \{a, b\}$

e.g:

- $(a|b)b = \{ab, bb\}$

  $a = \{a\}$

  $b = \{b\}$

  $(a|b) = \{a, b\}$

- $(aa|b)^* = \{\lambda, b, aa, bb, aab, bbbaa, aabaa, \ldots\}$

  $aa = \{aa\}$

  $b = \{b\}$

  $aa|b = \{aa, b\}$

- Shorthands
  - $r? \equiv (r|\lambda)$ "maybe r"
  - $r+ \equiv r(r^*)$ "one or more repetitions of a word in r"

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

$a = \{a\}$

$b? = \{b, \lambda\}$

$a^+ = \{a, aa, aaa, aaaa, aaaaa, \ldots\}$

- Regular Expressions and Programming Languages

Σ = ASCII keyboard symbols.

$\mathcal{L}_{Python}$ = The set of all python programs. RegEx is not enough.

Note: RegEx does not have the power to specify such set

Why?

e.g. $\{a^n b^n \mid n \in \mathbb{N}\}$