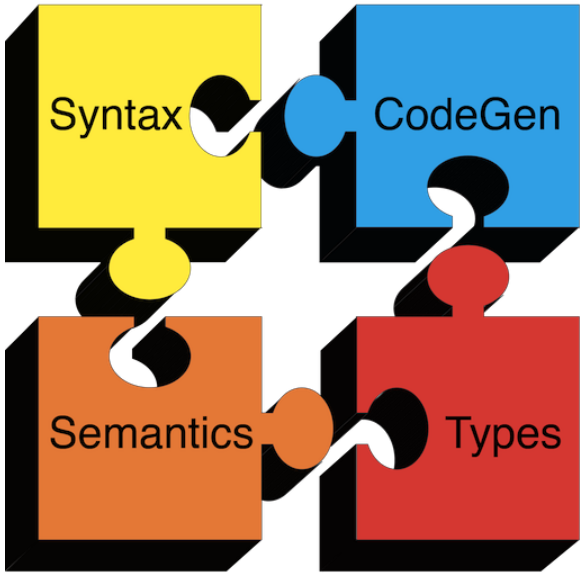

Compilers

Context-Free Grammars

CFGs

- Not all strings of tokens are programs . . .
- . . . parser must distinguish between valid and invalid strings of tokens

- We need
  - A language for describing valid strings of tokens
  - A method for distinguishing valid from invalid strings of tokens

CFGs

- Programming languages have recursive structure

- An EXPR is
  - if EXPR then EXPR else EXPR fi
  - while EXPR loop EXPR pool
  - …

- Context-free grammars are a natural notation for this recursive structure

CFGs

- A CFG consists of
  - A set of *terminals* T
  - A set of *non-terminals* N
  - A *start symbol* S $(S \in N)$
  - A set of *productions*

$$X \rightarrow Y_1 \dots Y_N$$

$$X \in N$$
$$Y_i \in N \cup T \cup \{\epsilon\}$$

CFGs

$$\left\{ \begin{array}{lll} S & \rightarrow & (\; S \;) \\ S & \rightarrow & \epsilon \end{array} \right\}$$

$N = \{ S \}$

$T = \{ (,) \}$

Productions can be read as rules.

CFGs

1. Begin with a string with only the start symbol S

2. Replace any non-terminal X in the string by the right-hand side of some production $X \rightarrow Y_1 \ldots Y_n$

3. Repeat (2) until there are no non-terminals

CFGs

$X_1 \ldots X_i \underline{X} X_{i+1} \ldots X_N \rightarrow X_1 \ldots X_i Y_1 \ldots Y_k X_{i+1} \ldots X_N$

$X \rightarrow Y_1 \ldots Y_k$ production

CFGs

$\alpha_0 \to \alpha_1 \to \alpha_2 \to \cdots \to \alpha_n$

$\alpha_0 \xrightarrow{*} \alpha_n$ (in ≥ 0 steps)

CFGs

Let *G* be a context-free grammar with start symbol S.
Then the language *L(G)* of *G* is:

$$\{ a_1 \ldots a_n \mid \forall i \; a_i \in T \wedge S \overset{*}{\Rightarrow} a_1 \ldots a_n \}$$

CFGs

- Terminals are so-called because there are no rules for replacing them
- Once generated, terminals are permanent
- Terminals ought to be tokens of the language

A fragment of COOL

EXPR → if EXPR then EXPR else EXPR fi

| while EXPR loop EXPR pool

| id

⋮

Some elements of the language:

id

EXPR → id

if id then id else id fi

while id loop id pool

if while id loop id pool then id else id

if if id then id else id fi then id else id fi

Simple arithmetic expressions

$$
\begin{aligned}
E &\rightarrow E + E \\
&\mid E * E \\
&\mid (E) \\
&\mid id
\end{aligned}
$$

id

id + id

id + id * id

(id + id) * id

⋮

CFGs

Which of the strings are in the language of the given CFG?

- ☐ abcba
- ☐ acca
- ☐ aba
- ☐ abcbcba

$$
\begin{aligned}
S &\rightarrow aXa \\
X &\rightarrow \varepsilon \\
&\mid bY \\
Y &\rightarrow \varepsilon \\
&\mid cXc
\end{aligned}
$$

CFGs

The idea of a CFG is a big step. But:

- Membership in a language is "yes" or "no"; also need parse tree of the input
- Must handle errors gracefully
- Need an implementation of CFG's (e.g., bison)

CFGs

- Form of the grammar is important
  - Many grammars generate the same language
  - Tools are sensitive to the grammar