Context-Free Languages

Sipser: Section 2.1 pages 101 - 111

Extending Our Reach

Finite automata and people are *language recognition devices*.

Regular expressions and people are *language generating devices*.

Finite automata *recognize* and regular expressions *generate* an important but limited class of languages.

People Languages

A grammar for the English language tells us whether a particular sentence is well formed or not.

A typically English grammar is "a sentence can consist of a noun phrase followed by a predicate."

More concisely, we write

<sentence> → *<noun_phrase>* *<predicate>*

So What's a Noun Phrase?

A sentence is

<sentence> → *<noun_phrase>* *<predicate>*

We must also provide definitions for the newly introduced constructs *<noun_phrase>* and *<predicate>*.

<noun_phrase> → <article> <noun>

<predicate> → *<verb>*

Generating Well Formed Sentences

Grammar rules so far:

<sentence> → *<noun_phrase>* *<predicate>*

<noun_phrase> → <article> <noun>

<predicate> → *<verb>*

To complete our simple grammar, we associate actual words with the terms *<article>*, *<noun>*, and *<verb>*.

<article> → a

<article> → the

<noun> → student

<verb> → relaxes

<verb> → studies

Context-Free Grammars

A *context-free grammar* G is a quadruple (V, Σ, R, S), where

V is a finite set called ***variables***,

Σ is a finite set, disjoint from V, called the ***terminals***

R is a finite subset of $V \times (V \cup \Sigma)^*$ called ***rules***, and

S (the ***start symbol***) is an element of V.

For any $A \in V$ and $u \in (V \cup \Sigma)^*$, we write $A \to u$ whenever $(A, u) \in R$.

The Language of a Grammar

If $u, v, w \in (V \cup \Sigma)^*$ and $A \to w$ is a rule, then we say uAv ***yields*** uwv and write $uAv \Rightarrow uwv$.

Write $u \overset{*}{\Rightarrow} v$ if

$$u \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \cdots \Rightarrow u_k \Rightarrow v.$$

The ***language of the grammar*** G is

$$L(G) = \{ w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w \}.$$

For Example

Consider $G = (V, \Sigma, R, S)$, where

$V = \{S\}$,

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

$R = \{ S \to aSa \mid bSb \mid aSb \mid bSa \mid \varepsilon \}$.

Is there a grammar whose language is

$$PAL = \{ w \in \Sigma^* \mid w = reverse(w) \}?$$

Arithmetic Expressions & Parse Trees

Consider $G = (V, \Sigma, R, S)$, where

$V = \{\langle EXPR\rangle, \langle TERM\rangle, \langle FACTOR\rangle\}$,

$\Sigma = \{ a, +, \times, (,) \}$,

$R = \{ \langle EXPR\rangle \to \langle EXPR\rangle + \langle TERM\rangle \mid \langle TERM\rangle,$
$\langle TERM\rangle \to \langle TERM\rangle \times \langle FACTOR\rangle \mid \langle FACTOR\rangle,$
$\langle FACTOR\rangle \to (\langle EXPR\rangle) \mid a \},$

$S = \langle EXPR\rangle$.

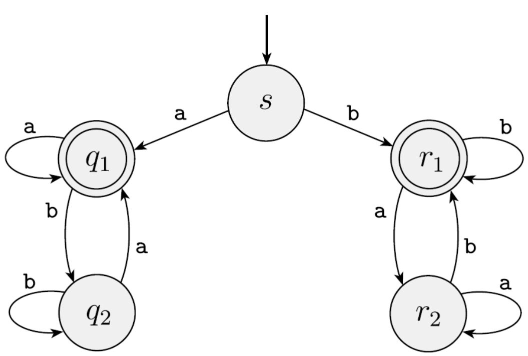
Needlessly Complicated?

How about just

$$\begin{aligned} \langle EXPR\rangle \to\ & \langle EXPR\rangle + \langle EXPR\rangle \mid \\ & \langle EXPR\rangle \times \langle EXPR\rangle \mid \\ & (\langle EXPR\rangle) \mid \\ & a \end{aligned}$$

A grammar G is ***ambiguous*** if some string w has two or more different leftmost derivations.

Regular Languages are Context-Free



Chomsky Normal Form

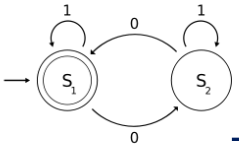
A context-free grammar G is in ***Chomsky normal form*** if every rule is of the form

$S \to \varepsilon$

$A \to BC$

$A \to a$

where $A, B, C \in V$, $B \neq S \neq C$, and $a \in \Sigma$.



Chomsky Normal Form

Theorem Any context-free language is generated by a context-free grammar in Chomsky normal form.

Proof

1. Make sure S appears only on the left.
2. Remove empty rules: $A \to \varepsilon$.
3. Handle unit rules: $A \to B$.
4. Fix all the rest.

$S \to ASA \mid aA$

$A \to b \mid \varepsilon$