Context-Free Grammars and Languages

• We have seen that many languages cannot be regular. Thus we need to consider larger classes of langs.

• *Contex-Free Languages* (CFL's) played a central role natural languages since the 1950's, and in compilers since the 1960's.

• *Context-Free Grammars* (CFG's) are the basis of BNF-syntax.

• Today CFL's are increasingly important for XML and their DTD's.

We'll look at: CFG's, the languages they generate, parse trees, pushdown automata, and closure properties of CFL's.

Pumping Lemma, decision property.

Informal example of CFG's

Consider $L_{pal} = \{w \in \Sigma^* : w = w^R\}$

For example $\texttt{otto} \in L_{pal}$, $\texttt{madamimadam} \in L_{pal}$.

In Finnish language e.g. $\texttt{saippuakauppias} \in L_{pal}$ ("soap-merchant")

DoGeeseSeeGod
NoMelonNoLemon

Let $\Sigma = \{0, 1\}$ and suppose L_{pal} were regular.

Let n be given by the pumping lemma. Then $0^n10^n \in L_{pal}$. In reading 0^n the FA must make a loop. Omit the loop; contradiction.

Let's define L_{pal} inductively:

Basis: $\epsilon, 0$, and 1 are palindromes.

Induction: If w is a palindrome, so are $0w0$ and $1w1$.

Circumscription: Nothing else is a palindrome.

CFG's is a formal mechanism for definitions such as the one for L_{pal}.

1. $P \rightarrow \epsilon$
2. $P \rightarrow 0$
3. $P \rightarrow 1$
4. $P \rightarrow 0P0$
5. $P \rightarrow 1P1$

0 and 1 are *terminals*

P is a *variable* (or *nonterminal*, or *syntactic category*)

P is in this grammar also the *start symbol*.

1–5 are *productions* (or *rules*)

$\boxed{\text{Some real examples from Sipser.}}$

Formal definition of CFG's

A *context-free grammar* is a quadruple

$$G = (V, T, P, S)$$

where

V is a finite set of *variables*.

T is a finite set of *terminals*.

P is a finite set of *productions* of the form $A \rightarrow \alpha$, where A is a variable and $\alpha \in (V \cup T)^*$

S is a designated variable called the *start symbol*.

Example: $G_{pal} = (\{P\}, \{0, 1\}, A, P)$, where $A = \{P \to \epsilon, P \to 0, P \to 1, P \to 0P0, P \to 1P1\}$.

Sometimes we group productions with the same head, e.g. $A = \{P \to \epsilon|0|1|0P0|1P1\}$.

Example: Regular expressions over $\{0, 1\}$ can be defined by the grammar

$$G_{regex} = (\{E\}, \{0, 1\}, A, E)$$

where $A =$ $+,.,\phi,\varepsilon,^{\star},(,)$

$\{E \to \mathbf{0}, E \to \mathbf{1}, E \to E.E, E \to E{+}E, E \to E^{\star}, E \to (E)\}$

E –> ε, E –> φ

Example: (simple) expressions in a typical prog lang. Operators are $+$ and *, and arguments are identfiers, i.e. strings in $L((\boldsymbol{a}+\boldsymbol{b})(\boldsymbol{a}+\boldsymbol{b}+\mathbf{0}+\mathbf{1})^*)$

The expressions are defined by the grammar

$$G = (\{E, I\}, T, P, E)$$

where $T = \{+, *, (,), a, b, 0, 1\}$ and P is the following set of productions:

1. $E \to I$
2. $E \to E + E$
3. $E \to E * E$
4. $E \to (E)$
5. $I \to a$
6. $I \to b$
7. $I \to Ia$
8. $I \to Ib$
9. $I \to I0$
10. $I \to I1$

Derivations using grammars

- *Recursive inference*, using productions from body to head

- *Derivations*, using productions from head to body.

Example of recursive inference:

	String	Lang	Prod	String(s) used
(i)	a	I	5	-
(ii)	b	I	6	-
(iii)	$b0$	I	9	*(ii)*
(iv)	$b00$	I	9	*(iii)*
(v)	a	E	1	*(i)*
(vi)	$b00$	E	1	*(iv)*
(vii)	$a + b00$	E	2	*(v)*, *(vi)*
(viii)	$(a + b00)$	E	4	*(vii)*
(ix)	$a * (a + b00)$	E	3	*(v)*, *(viii)*

Let $G = (V, T, P, S)$ be a CFG, $A \in V$, $\{\alpha, \beta\} \subset (V \cup T)^*$, and $A \rightarrow \gamma \in P$.

Then we write

$$\alpha A \beta \underset{G}{\Rightarrow} \alpha \gamma \beta$$

or, if G is understood

$$\alpha A \beta \Rightarrow \alpha \gamma \beta$$

and say that $\alpha A \beta$ *derives* $\alpha \gamma \beta$.

We define $\stackrel{*}{\Rightarrow}$ to be the reflexive and transitive closure of $\Rightarrow$, IOW:

Basis: Let $\alpha \in (V \cup T)^*$. Then $\alpha \stackrel{*}{\Rightarrow} \alpha$.

Induction: If $\alpha \stackrel{*}{\Rightarrow} \beta$, and $\beta \Rightarrow \gamma$, then $\alpha \stackrel{*}{\Rightarrow} \gamma$.

Example: Derivation of $a * (a + b00)$ from E in the grammar of slide 138:

$$E \Rightarrow E * E \Rightarrow I * E \Rightarrow a * E \Rightarrow a * (E) \Rightarrow$$

$$a*(E+E) \Rightarrow a*(I+E) \Rightarrow a*(a+E) \Rightarrow a*(a+I) \Rightarrow$$

$$a * (a + I0) \Rightarrow a * (a + I00) \Rightarrow a * (a + b00)$$

Note: At each step we might have several rules to choose from, e.g.
$I * E \Rightarrow a * E \Rightarrow a * (E)$, versus
$I * E \Rightarrow I * (E) \Rightarrow a * (E)$.

Note2: Not all choices lead to successful derivations of a particular string, for instance

$$E \Rightarrow E + E$$

won't lead to a derivation of $a * (a + b00)$.

Leftmost and Rightmost Derivations

Leftmost derivation $\underset{lm}{\Rightarrow}$: Always replace the leftmost variable by one of its rule-bodies.

Rightmost derivation $\underset{rm}{\Rightarrow}$: Always replace the rightmost variable by one of its rule-bodies.

Leftmost: The derivation on the previous slide.

Rightmost:

$$E \underset{rm}{\Rightarrow} E * E \underset{rm}{\Rightarrow}$$

$$E*(E) \underset{rm}{\Rightarrow} E*(E+E) \underset{rm}{\Rightarrow} E*(E+I) \underset{rm}{\Rightarrow} E*(E+I0)$$

$$\underset{rm}{\Rightarrow} E*(E+I00) \underset{rm}{\Rightarrow} E*(E+b00) \underset{rm}{\Rightarrow} E*(I+b00)$$

$$\underset{rm}{\Rightarrow} E*(a+b00) \underset{rm}{\Rightarrow} I*(a+b00) \underset{rm}{\Rightarrow} a*(a+b00)$$

We can conclude that $E \overset{*}{\underset{rm}{\Rightarrow}} a*(a+b00)$

The Language of a Grammar

If $G(V, T, P, S)$ is a CFG, then the *language of* G is

$$L(G) = \{w \in T^* : S \overset{*}{\underset{G}{\Rightarrow}} w\}$$

i.e. the set of strings over T^* derivable from the start symbol.

If G is a CFG, we call $L(G)$ a *context-free language*.

Example: $L(G_{pal})$ is a context-free language.

Theorem 5.7:

$$L(G_{pal}) = \{w \in \{0, 1\}^* : w = w^R\}$$

Proof: ($\supseteq$-direction.) Suppose $w = w^R$. We show by induction on $|w|$ that $w \in L(G_{pal})$

Basis: $|w| = 0$, or $|w| = 1$. Then w is $\epsilon, 0$, or 1. Since $P \to \epsilon, P \to 0$, and $P \to 1$ are productions, we conclude that $P \overset{*}{\underset{G}{\Rightarrow}} w$ in all base cases.

Induction: Suppose $|w| \geq 2$. Since $w = w^R$, we have $w = 0x0$, or $w = 1x1$, and $x = x^R$.

If $w = 0x0$ we know from the IH that $P \overset{*}{\Rightarrow} x$. Then

$$P \Rightarrow 0P0 \overset{*}{\Rightarrow} 0x0 = w$$

Thus $w \in L(G_{pal})$.

The case for $w = 1x1$ is similar.

($\subseteq$-direction.) We assume that $w \in L(G_{pal})$ and must show that $w = w^R$.

Since $w \in L(G_{pal})$, we have $P \stackrel{*}{\Rightarrow} w$.

We do an induction on the length of $\stackrel{*}{\Rightarrow}$.

Basis: The derivation $P \stackrel{*}{\Rightarrow} w$ is done in one step.

Then w must be $\epsilon, 0$, or 1, all palindromes.

Induction: Let $n \geq 1$, and suppose the derivation takes $n + 1$ steps. Then we must have

$$w = 0x0 \stackrel{*}{\Leftarrow} 0P0 \Leftarrow P$$

or

$$w = 1x1 \stackrel{*}{\Leftarrow} 1P1 \Leftarrow P$$

where the second derivation is done in n steps.

By the IH x is a palindrome, and the inductive proof is complete.

Ex. Design CFGs for the following languages:
L_1 = {balanced parentheses} = {ε,(),(()),()(),((())),()(()),(())(), ...} (the Dyck language)
$L_2 = \{0^m 1^n 2^p \mid m,n, p >= 0, m+n = p\}$
$L_3 = \{w \mid w \, \varepsilon \, \{0,1\}^*, w <> w^R\}$

Sentential Forms

Let $G = (V, T, P, S)$ be a CFG, and $\alpha \in (V \cup T)^*$. If

$$S \stackrel{*}{\Rightarrow} \alpha$$

we say that α is a *sentential form*.

If $S \underset{lm}{\Rightarrow} \alpha$ we say that α is a *left-sentential form*, and if $S \underset{rm}{\Rightarrow} \alpha$ we say that α is a *right-sentential form*

Note: $L(G)$ is those sentential forms that are in T^*.

Example: Take G from slide 138. Then $E * (I + E)$ is a sentential form since

$$E \Rightarrow E*E \Rightarrow E*(E) \Rightarrow E*(E+E) \Rightarrow E*(I+E)$$

This derivation is neither leftmost, nor rightmost

Example: $a * E$ is a left-sentential form, since

$$E \underset{lm}{\Rightarrow} E * E \underset{lm}{\Rightarrow} I * E \underset{lm}{\Rightarrow} a * E$$

Example: $E*(E+E)$ is a right-sentential form, since

$$E \underset{rm}{\Rightarrow} E * E \underset{rm}{\Rightarrow} E * (E) \underset{rm}{\Rightarrow} E * (E + E)$$