Lecture 10 — Context-Free Grammars

CSE205: Theory of Computation

Minseok Jeon

2026 Fall

Context-Free Languages

The *context-free languages* are an extension of the regular languages, with many applications in computer science:

- Most programming languages (e.g., C, Java, ML, etc.).
- Markup languages (e.g., HTML, XML, etc.).
- Essential to the design of programming languages and the construction of compilers.

Context-free languages are described by *context-free grammars*.

Example: Palindromes

A string is a *palindrome* if it reads the same forward and backward:

$$\boldsymbol{L} = \{\, \boldsymbol{w} \in \{0,1\}^* \mid \boldsymbol{w} = \boldsymbol{w}^R \,\}.$$

$\boldsymbol{L}$ is not regular, but it is context-free. It has a recursive definition:

- **Basis:** $\boldsymbol{\varepsilon}$, 0, and 1 are palindromes.
- **Induction:** If $\boldsymbol{w}$ is a palindrome, so are $0\boldsymbol{w}0$ and $1\boldsymbol{w}1$.

This recursive definition is expressed by a context-free grammar:

$$\begin{aligned} \boldsymbol{P} &\to \boldsymbol{\varepsilon} \\ \boldsymbol{P} &\to 0 \\ \boldsymbol{P} &\to 1 \\ \boldsymbol{P} &\to 0\,\boldsymbol{P}\,0 \\ \boldsymbol{P} &\to 1\,\boldsymbol{P}\,1 \end{aligned}$$

Context-Free Grammars

Definition (Context-Free Grammar)

A context-free grammar G is a quadruple

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

where

- V: a finite set of *variables* (nonterminals);
- T: a finite set of *terminals* (terminal symbols);
- $S \in V$: the *start variable*;
- P: a finite set of *productions*. Each production has the form

$$x \to y$$

  where $x \in V$ and $y \in (V \cup T)^*$.

Example: Palindromes

The palindrome grammar is

$$\boldsymbol{G} = (\{\boldsymbol{P}\}, \{0, 1\}, \boldsymbol{P}, \boldsymbol{A})$$

where $\boldsymbol{A}$ is the set of five productions:

$$\begin{array}{lcl} \boldsymbol{P} & \to & \boldsymbol{\varepsilon} \\ \boldsymbol{P} & \to & 0 \\ \boldsymbol{P} & \to & 1 \\ \boldsymbol{P} & \to & 0\,\boldsymbol{P}\,0 \\ \boldsymbol{P} & \to & 1\,\boldsymbol{P}\,1 \end{array}$$

Using BNF alternation, these can be written compactly as

$$\boldsymbol{P} \to \boldsymbol{\varepsilon} \mid 0 \mid 1 \mid 0\boldsymbol{P}0 \mid 1\boldsymbol{P}1.$$

Example: Simple Arithmetic Expressions

$$G = (\{E, I\},\ \{+, *, (,), a, b, 0, 1\},\ E,\ P)$$

where P is the set of productions:

$$\begin{array}{rcl} E & \to & I \\ E & \to & E + E \\ E & \to & E * E \\ E & \to & (E) \\ I & \to & a \mid b \\ I & \to & Ia \mid Ib \mid I0 \mid I1 \end{array}$$

Here E generates expressions and I generates identifiers.

Derivation

Definition (Derivation Relation, $\Rightarrow$)

Let $G = (V, T, S, P)$ be a context-free grammar. Let $\alpha A \beta$ be a string of terminals and variables, where $A \in V$ and $\alpha, \beta \in (V \cup T)^*$. If $A \to \gamma$ is a production in G, then we say $\alpha A \beta$ *derives* $\alpha \gamma \beta$, and write

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

Definition ($\Rightarrow^*$: Closure of $\Rightarrow$)

$\Rightarrow^*$ represents zero or more derivation steps:

- **Basis:** for any string α of terminals and variables, $\alpha \Rightarrow^* \alpha$.
- **Induction:** if $\alpha \Rightarrow^* \beta$ and $\beta \Rightarrow \gamma$, then $\alpha \Rightarrow^* \gamma$.

Example: A Derivation

A derivation for $\boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{b}00)$ in the expression grammar:

$$\boldsymbol{E} \Rightarrow \boldsymbol{E} * \boldsymbol{E} \Rightarrow \boldsymbol{I} * \boldsymbol{E} \Rightarrow \boldsymbol{a} * \boldsymbol{E} \Rightarrow \boldsymbol{a} * (\boldsymbol{E}) \Rightarrow$$
$$\boldsymbol{a} * (\boldsymbol{E} + \boldsymbol{E}) \Rightarrow \boldsymbol{a} * (\boldsymbol{I} + \boldsymbol{E}) \Rightarrow \boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{E}) \Rightarrow \boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{I}) \Rightarrow$$
$$\boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{I}0) \Rightarrow \boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{I}00) \Rightarrow \boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{b}00).$$

Thus, $\boldsymbol{E} \Rightarrow^* \boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{b}00)$.

Leftmost and Rightmost Derivations

- **Leftmost derivation** ($\Rightarrow_{lm}$): replace the leftmost variable at each derivation step.
- **Rightmost derivation** ($\Rightarrow_{rm}$): replace the rightmost variable at each derivation step.

A *rightmost* derivation for $\boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{b}00)$:

$$\begin{aligned}&\boldsymbol{E} \Rightarrow \boldsymbol{E} * \boldsymbol{E} \Rightarrow \boldsymbol{E} * (\boldsymbol{E}) \Rightarrow \boldsymbol{E} * (\boldsymbol{E} + \boldsymbol{E}) \Rightarrow \boldsymbol{E} * (\boldsymbol{E} + \boldsymbol{I}) \Rightarrow \\ &\boldsymbol{E} * (\boldsymbol{E} + \boldsymbol{I}0) \Rightarrow \boldsymbol{E} * (\boldsymbol{E} + \boldsymbol{I}00) \Rightarrow \boldsymbol{E} * (\boldsymbol{E} + \boldsymbol{b}00) \Rightarrow \boldsymbol{E} * (\boldsymbol{I} + \boldsymbol{b}00) \Rightarrow \\ &\boldsymbol{E} * (\boldsymbol{a} + \boldsymbol{b}00) \Rightarrow \boldsymbol{I} * (\boldsymbol{a} + \boldsymbol{b}00) \Rightarrow \boldsymbol{a} * (\boldsymbol{a} + \boldsymbol{b}00).\end{aligned}$$

Language of a Grammar

Definition (Language of a Grammar)

Let $G = (V, T, S, P)$ be a context-free grammar. The *language* of G, denoted $L(G)$, is the set of terminal strings that have derivations from the start symbol:

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

Definition (Context-Free Language)

If a language L is the language of some context-free grammar G, i.e., $L = L(G)$, then L is a *context-free language* (CFL).

Sentential Forms

Definition (Sentential Form)

If $G = (V, T, S, P)$ is a context-free grammar, then any string $\alpha \in (V \cup T)^*$ such that $S \Rightarrow^* \alpha$ is a *sentential form*.

- If $S \Rightarrow^* \alpha$ is a leftmost derivation, α is a *left-sentential form*.
- If $S \Rightarrow^* \alpha$ is a rightmost derivation, α is a *right-sentential form*.

A sentential form may contain both variables and terminals; a string in $L(G)$ is a sentential form containing only terminals.

Example: Sentential Forms

Leftmost derivation:

$$E \Rightarrow E * E \Rightarrow I * E \Rightarrow a * E \Rightarrow a * (E) \Rightarrow$$
$$a * (E + E) \Rightarrow a * (I + E) \Rightarrow a * (a + E) \Rightarrow a * (a + I) \Rightarrow$$
$$a * (a + I0) \Rightarrow a * (a + I00) \Rightarrow a * (a + b00).$$

Neither leftmost nor rightmost:

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

Exercises

Give context-free grammars for the following languages:

- $L = \{ ww^R \mid w \in \{a, b\}^* \}$
- $L = \{ a^n b^n \mid n \geq 0 \}$
- $L = \{ a^n b^m \mid n \neq m \}$
- The language of balanced parentheses.
  - E.g., ε, (), ()(), (()), (()()).