# LIMBAJE FORMALE ŞI AUTOMATE

## CURSUL 5

## GRAMATICI FORMALE CHOMSKY

Gramaticile ne furnizează mecanisme de generare a limbajelor, spre deosebire de automate care sunt mecanisme de acceptare sau de recunoaștere de limbaje. De aceea, automatele se mai numesc și acceptoare.

Vom da pentru început cea mai generală definiție pentru gramatici, introdusă de Naum Chomsky în 1950, pentru a modela limbajele naturale. Vom particulariza această definiție pentru gramaticile independente de context, liniare și regulate.

**Definiţie 1.** O gramatică formală are o structură de forma $G = (N, \Sigma, S, P)$, unde:

- $N$ este alfabetul neterminalilor;
- $\Sigma$ este alfabetul terminalilor, $\Sigma \cap N = \emptyset$;
- $S \in N$ este simbolul de start al gramaticii;
- $P$ este o multime finita, $P \subseteq (N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$, ale cărei elemente se numesc producții. Pentru fiecare pereche $(\alpha, \beta) \in P$ vom scrie $\alpha \to \beta$. Observăm că membrul stâng al acestei producții, $\alpha$, este un șir de simboluri din $(N \cup \Sigma)^* N (N \cup \Sigma)^*$, in care apare cel putin un simbol neterminal, iar membrul drept al productiei, $\beta$, este un șir de simboluri din $(N \cup \Sigma)^*$, posibil $\lambda$.

**Definiţie 2.** Fie $G = (N, \Sigma, S, P)$, o gramatică definită ca mai sus, $\alpha, \beta \in (N \cup \Sigma)^*$.

- Spunem că din $\alpha$ se derivă direct $\beta$ în $G$ (sau că $\beta$ este derivat direct din α in G) și scriem $\alpha \underset{G}{\Rightarrow} \beta$ (sau $\alpha \Rightarrow \beta$ cand $G$ este subânțeles) dacă $\alpha = uxv, \beta = uyv$ și $x \to y \in P$. Cu alte cuvinte, subșirul $x$ din $\alpha$ este înlocuit cu $y$.

- Spunem că din $\alpha$ se derivă $\beta$ în $G$ (sau că $\beta$ este derivat din α în G) și scriem $\alpha \overset{*}{\Rightarrow}_G \beta$ (sau $\alpha \overset{*}{\Rightarrow} \beta$ cand $G$ este subânțeles) dacă: fie $\alpha = \beta$, fie există $\alpha_1, \dots, \alpha_n = \beta, \alpha_1, \dots, \alpha_n \in (N \cup \Sigma)^*$ și $\alpha \underset{G}{\Rightarrow} \alpha_1, \alpha_1 \underset{G}{\Rightarrow} \alpha_2, \dots, \alpha_{n-1} \underset{G}{\Rightarrow} \alpha_n, n \geq 1$. În acest caz spunem că $\alpha \overset{*}{\Rightarrow}_G \beta$ este o derivare în $n$ pași, pe care o putem nota $\alpha \overset{n}{\Rightarrow}_G \beta$ sau $\alpha \underset{G}{\Rightarrow} \alpha_1 \underset{G}{\Rightarrow} \alpha_2 \underset{G}{\Rightarrow} \dots \underset{G}{\Rightarrow} \alpha_{n-1} \underset{G}{\Rightarrow} \alpha_n$. Pentru cazul $\alpha = \beta$ spunem că $\alpha \overset{*}{\Rightarrow}_G \beta$ este o derivare în 0 pași.
- În cazul în care $\alpha \overset{n}{\Rightarrow}_G \beta, n \geq 1$, putem scrie $\alpha \overset{+}{\Rightarrow}_G \beta$
- În cazul în care $S \overset{*}{\Rightarrow}_G \alpha,\ \alpha \in (N \cup \Sigma)^*$, spunem că $\alpha$ este o <u>formă sentențială</u> pentru $G$.

**Definiţie 3.** Fie $G = (N, \Sigma, S, P)$, o gramatică definită ca mai sus. Limbajul generat de gramatica $G$ se notează cu $L(G)$ și este definit prin

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

Limbajul generat de $G$ constă din mulțimea șirurilor terminale care sunt derivate din simbolul de start al gramaticii.

## GRAMATICI ŞI LIMBAJE INDEPENDENTE DE CONTEXT

Gramaticile independente de context sunt cel mai mult utilizate in specificarea sintaxei unui limbaj de programare. Formal:

**Definiţie 4.** Fie $G = (N, \Sigma, S, P)$ o gramatică Chomsky. Spunem că $G$ este independentă de context dacă orice producţie din $P$ este de forma

$$A \to \alpha, \text{unde } A \in N, \alpha \in (N \cup \Sigma)^*$$

**Observaţii.**

1. $P \subseteq N \times (N \cup \Sigma)^*$;
2. Multimile $N, \Sigma, P$ sunt finite;
3. Producţia $A \to \alpha$ mai este numită și $A$-producţie.

4. Dacă toate producţiile lui $A$ sunt $A \to \alpha_1, \dots, A \to \alpha_n$, pentru simplificare putem folosi notaţia $A \to \alpha_1 | \dots | \alpha_n$
5. În general, notăm cu litere mari simbolurile neterminale

**Exemple.**

1. Gramatică care genereaza $\{a^n b^n | n \geq 0\}$
   $G_1 = (\{S\}, \{a, b\}, S, \{S \to aSb, S \to \lambda\}$
   O derivare in $G$ este de forma
   $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow \cdots \Rightarrow a^n S b^n \Rightarrow a^n b^n$
2. Gramatică care genereaza $\{a^m b^n | m \geq n \geq 0\}$
   $G_2 = (\{S\}, \{a, b\}, S, \{S \to aSb, S \to aS, S \to \lambda\}$
3. Gramatică care genereaza expresiile formate cu ajutorul operatorilor $+, -, *, ()$ şi a operandului $a$
   $G_3 = (\{S\}, \{a, +, -, *, (, )\}, S, \{S \to S + S, S \to S - S, S \to S * S, S \to a,$
   $S \to (S)\}$
4. Gramatica $G_4$ care generează şirurile palindroame peste $\{a, b\}$ ($w$ este palindrom dacă şirul obţinut prin citirea de la dreapta la stânga a literelor lui $w$ este identic cu $w$; exemple: $\lambda, a, b, aba, abba$)
   $G_4$ are producţiile
   $S \to aSa, S \to bSb, S \to a, S \to b, S \to \lambda$
5. Gramatică care generează şirurile $w$ peste $\{a, b\}$ pentru care numărul apariţiilor lui $a$ în $w$, notat cu$|w|_a$, este egal cu numărul apariţiilor lui $b$ în $w$, notat cu $|w|_b$.
   Fie $G_5$ cu producţiile $S \to aSb \mid bSa \mid SS \mid S \to \lambda$

**Observaţie.** Pentru a arăta că pentru $G_i$ limbajul generat, $L(G_i)$, este cel indicat în exemplul $i, i = 1, \dots, 5$, trebuie folosită dubla incluziune. Spre exemplu:

**Propoziţia 1.** Limbajul generat de gramatica $G_5$ cu producţiile $S \to aSb \mid bSa \mid SS \mid S \to \lambda$ este $L_5 = \{w \in \{a, b\}^* | \, |w|_a = |w|_b\}$

**Demonstratie.** Aratam mai intai ca $L(G_5) \subseteq L_5$.

Fie $w \in L(G_5), S \overset{n}{\Rightarrow} w, w \in \{a, b\}^*$. Aratam prin inductie dupa $n$ ca $w \in L_5$

**Baza inductiei.** Pentru $n = 1$, rezulta ca $w = \lambda$, iar $\lambda \in L_5$.

**Ipoteza inductiva.** Presupunem ca pentru orice $p \leq n$ si pentru orice $w$ cu $S \overset{p}{\Rightarrow} w$ avem $w \in L_5$.

**Saltul inductiv.** Fie $S \overset{n+1}{\Longrightarrow} z$. Vom pune în evidenţă primul pas al derivării. Avem 3 cazuri:

a) $S \Rightarrow aSb \overset{n}{\Rightarrow} z$. Rezultă că $z = az'b$ si $S \overset{n}{\Rightarrow} z'$. Din ipoteza de inductie rezulta ca $|z'|_a = |z'|_b$. Dar atunci rezulta ca $|z|_a = |z|_b$.

b) $S \Rightarrow bSa \overset{n}{\Rightarrow} z$. Analog cazului a).

c) $S \Rightarrow SS \overset{n}{\Rightarrow} z$. Rezulta ca $z = z_1 z_2$ si $S \overset{\leq n}{\Rightarrow} z_1$, $S \overset{\leq n}{\Rightarrow} z_2$. Din ipoteza de inducţie avem $|z_1|_a = |z_1|_b$ , $|z_2|_a = |z_2|_b$, de unde rezulta ca $|z|_a = |z|_b$.

**Reciproc**, arătăm că $L_5 \subseteq L(G_5)$. Fie $w \in L_5, |w|_a = |w|_b$. Aratam prin inductie dupa $n = |w|_a$ ca $w \in L(G_5)$.

**Baza inductiei.** Pentru n $= 0$, rezulta ca w $= \lambda$ si $\lambda \in L(G_5)$.

**Ipoteza inductiva.** Presupunem ca pentru orice $w \in L_5, |w|_a = |w|_b = p$, p $\leq n$, rezulta $w \in L(G_5)$.

**Saltul inductiv.** Fie $z \in L_5, |z|_a = |z|_b = n + 1$. Avem cazurile:

a) $z = awb, |w|_a = |w|_b = n$. Din ipoteza de inductie $w \in L(G_5)$, deci $S \overset{*}{\Rightarrow} w$. Atunci in $G$ avem derivarea $S \Rightarrow aSb \overset{*}{\Rightarrow} awb = z$, deci $z \in L(G_5)$.

b) $z = bwa, |w|_a = |w|_b = n$. Analog cazului a).

c) $z = awa, |w|_a = n - 1, |w|_b = n + 1$. Fie $z = a_1 a_2 \dots a_{2n}, z_i = a_1 \dots a_i$, $a_1 = a_{2n} = a$ si functia definita prin $f(i) = |z_i|_a - |z_i|_b$. Avem $f(1) = 1, f(2n-1) = -1$. Rezulta ca exista $j, 2 \leq j \leq 2n-2$ astfel incat $f(j) = 0$. Luam $u = a_1 a_2 \dots a_j, v = a_{j+1} \dots a_{2n}, z = uv$ . Avem $|u|_a = |u|_b$, deci $|v|_a = |v|_b$.

Conform ipotezei de inductie $u, v \in L(G_5)$, deci $S \overset{*}{\Rightarrow} u$ si $S \overset{*}{\Rightarrow} v$. Atunci in $G$ avem derivarea $S \Rightarrow SS \overset{*}{\Rightarrow} uv = z$, deci $z \in L(G_5)$.

## GRAMATICI ŞI LIMBAJE REGULATE

**Definiţie 5.** Fie $G = (N, \Sigma, S, P)$ o gramatică independentă de context. Spunem că $G$ este o gramatică regulată dacă orice producţie din $P$ are una din formele

$$A \to aB, \text{unde } B \in N, a \in \Sigma \cup \{\lambda\}$$
$$A \to b, \text{unde } b \in \Sigma \cup \{\lambda\}$$

**Propoziţia 2.** Limbajul generat de o gramatică regulată este un limbaj regulat.
**Demonstratie.** Fie $G = (N, \Sigma, S, P)$ o gramatică regulata. Construim $AFN_\lambda$ $A = (N \cup \{f\}, \Sigma, \delta, S, \{f\})$, unde $\delta(A, a) = \{B | A \to aB \in P\}$ si $\delta(A, a) = \{f\}$ daca si numai daca $A \to a \in P$. Aratam ca $L(G) = L(A)$.

Fie $w \in \Sigma^*, A \in N$ si $A \overset{n}{\Rightarrow} w, n \geq 1$. Aratam prin inductie dupa $n$ ca $(A, w) \vdash^* (f, \lambda)$.

**Baza.** $n = 1$. Rezulta ca $A \to w \in P, w \in \Sigma \cup \{\lambda\}$, deci $\delta(A, w) = \{f\}$, adica $(A, w) \vdash (f, \lambda)$.

**Ipoteza de inductie.** Presupunem ca pentru orice $w \in \Sigma^*$ si orice $A \in N$ cu $A \overset{n}{\Rightarrow} w, n \geq 1$, avem $(A, w) \vdash^* (f, \lambda)$.

**Saltul inductiv.** Fie $A \overset{n+1}{\Longrightarrow} w$. Punem in evidenta primul pas al acestei derivari, $A \Rightarrow aB \overset{n}{\Rightarrow} w$, unde $A \to aB \in P$. Rezulta ca $w = az, B \overset{n}{\Rightarrow} z$. Din definitia lui $\delta$ avem ca $\delta(A, a) = B$, iar din ipoteza de inductie rezulta $(B, z) \vdash^* (f, \lambda)$. Atunci $(A, w) = (A, az) \vdash (B, z) \vdash^* (f, \lambda)$.

Pentru $A = S$, rezulta ca $S \overset{n}{\Rightarrow} w$ implica $(S, w) \vdash^* (f, \lambda)$, adica $w \in L(G)$ implica $w \in L(A)$, cu alte cuvinte $L(G) \subseteq L(A)$.

Reciproc, aratam ca $L(A) \subseteq L(G)$. Pentru aceasta se arata ca daca $(A, w) \vdash^n (f, \lambda)$ atunci $A \overset{n}{\Rightarrow} w$ prin inductie dupa $n$. **EXERCITIU**

**Propoziţia 3.** Orice limbaj regulat este acceptat de o gramatica regulata.
**Demonstratie.** Fie $L \subseteq \Sigma^*$ un limbaj regulat acceptat de AFN $A = (Q, \Sigma, \delta, s, F)$. Construim gramatica regulata $G = (N, \Sigma, S, P)$ cu $N = Q, S = s, P$ definita prin

$$P = \{p \to aq | q \in \delta(p, a)\} \cup \{p \to \lambda | p \in F\}$$

Aratam ca $L(A) = L(G)$. Fie $w \in \Sigma^*$. Avem urmatoarele echivalente:
$w \in L(A) \Leftrightarrow s \overset{n}{\Rightarrow} w \Leftrightarrow s \Rightarrow a_1 q_1 \Rightarrow \dots \Rightarrow a_1 \dots a_n q_n \Rightarrow a_1 \dots a_n = w \Leftrightarrow$
$s \to a_1 q_1 \in P, \dots, q_{n-1} \to a_n q_n, q_n \to \lambda, q_n \in F \Leftrightarrow q_1 \in \delta(s, a_1), \dots, q_n \in \delta(q_{n-1}, a_n), q_n \in F \Leftrightarrow (s, a_1 \dots a_n) \vdash (q_1, a_2 \dots a_n) \vdash \dots \vdash (q_n, \lambda)$
$\Leftrightarrow a_1 \dots a_n = w \in L(A)$. Rezulta ca $L(A) = L(G)$.

**Teoremă.** Familia limbajelor generate de gramaticile regulate, notată cu $\mathfrak{L}_{REG}$, este egala cu familia limbajelor regulate.
**Demonstratie.** Rezulta din Propozitiile 1 si 2

## Exemple de gramatici regulate.

1. $S \to aS \mid bS \mid \lambda$. Aceasta gramatica genereaza $\{a, b\}^*$.
2. $S \to aA,\ A \to bB, A \to aS, S \to \lambda$. Aceasta gramatica genereaza $\{aba\}^*$.