

LIMBAJE FORMALE ȘI AUTOMATE

CURSUL 6

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.

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

- N este alfabetul neterminalilor;
- Σ este alfabetul terminalilor, $\Sigma \cap N = \emptyset$;
- $S \in N$ este simbolul de start al gramaticii;
- P este o mulțime finită, $P \subseteq (N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$, ale cărei elemente se numesc producții. Pentru fiecare pereche $(\alpha, \beta) \in P$ vom scrie $\alpha \rightarrow \beta$. Observăm că membrul stâng al acestei producții, α , este un șir de simboluri din $(N \cup \Sigma)^* N (N \cup \Sigma)^*$, în care apare cel puțin un simbol neterminal, iar membrul drept al producției, β , este un șir de simboluri din $(N \cup \Sigma)^*$, posibil λ .

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

- Spunem că din α se derivă direct β în G (sau că β este derivat direct din α în G) și scriem $\alpha \Rightarrow_G \beta$ (sau $\alpha \Rightarrow \beta$ când G este subînțeles) dacă $\alpha = uxv$, $\beta = uyv$ și $x \rightarrow y \in P$. Cu alte cuvinte, subșirul x din α este înlocuit cu y .

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

Definiție 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

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

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

GRAMATICI ȘI LIMBAJE INDEPENDENTE DE CONTEXT

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

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

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

Observații.

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

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

Exemple.

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

Observație. 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:

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

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

Fie $w \in L(G_5), S \xRightarrow{n} w, w \in \{a, b\}^*$. 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 $S \xRightarrow{p} w$ avem $w \in L_5$.

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

a) $S \Rightarrow aSb \xRightarrow{n} z$. Rezultă că $z = az'b$ și $S \xRightarrow{n} z'$. Din ipoteza de inducție rezulta că $|z'|_a = |z'|_b$. Dar atunci rezulta că $|z|_a = |z|_b$.

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

c) $S \Rightarrow SS \xRightarrow{n} z$. Rezulta că $z = z_1 z_2$ și $S \xRightarrow{\leq n} z_1$, $S \xRightarrow{\leq n} z_2$. Din ipoteza de inducție avem $|z_1|_a = |z_1|_b$, $|z_2|_a = |z_2|_b$, de unde rezulta că $|z|_a = |z|_b$.

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

Baza inducției. Pentru $n = 0$, rezulta că $w = \lambda$ și $\lambda \in L(G_5)$.

Ipoteza inductivă. Presupunem că 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 inducție $w \in L(G_5)$, deci $S \xRightarrow{*} w$.

Atunci în G avem derivarea $S \Rightarrow aSb \xRightarrow{*} awb = z$, 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 $z = a_1 a_2 \dots a_{2n}$, $z_i = a_1 \dots a_i$, $a_1 = a_{2n} = a$ și funcția definită prin $f(i) = |z_i|_a - |z_i|_b$. Avem $f(1) = 1$, $f(2n - 1) = -1$. Rezulta că există j , $2 \leq j \leq 2n - 2$ astfel încât $f(j) = 0$.

Luăm $u = a_1 a_2 \dots a_j$, $v = a_{j+1} \dots a_{2n}$, $z = uv$. Avem $|u|_a = |u|_b$, deci $|v|_a = |v|_b$.

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

GRAMATICI ȘI LIMBAJE REGULATE

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

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

$$A \rightarrow b, \text{ unde } b \in \Sigma \cup \{\lambda\}$$

Propoziția 2. Limbajul generat de o gramatică regulată este un limbaj regulat.

Demonstratie. Fie $G = (N, \Sigma, S, P)$ o gramatică regulată. Construim $AFN_\lambda A = (N \cup \{f\}, \Sigma, \delta, S, \{f\})$, unde $\delta(A, a) = \{B \mid A \rightarrow aB \in P\}$ și $\delta(A, a) = \{f\}$ dacă și numai dacă $A \rightarrow a \in P$. Aratam ca $L(G) = L(A)$.

Fie $w \in \Sigma^*, A \in N$ și $A \xRightarrow{n} w, n \geq 1$. Aratam prin inducție după n ca $(A, w) \vdash^* (f, \lambda)$.

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

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

Saltul inductiv. Fie $A \xRightarrow{n+1} w$. Punem în evidență primul pas al acestei derivări, $A \xRightarrow{n} aB \xRightarrow{1} w$, unde $A \rightarrow aB \in P$. Rezulta ca $w = az, B \xRightarrow{n} z$. Din definiția lui δ avem ca $\delta(A, a) = B$, iar din ipoteza de inducție rezulta $(B, z) \vdash^* (f, \lambda)$. Atunci $(A, w) = (A, az) \vdash (B, z) \vdash^* (f, \lambda)$.

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

Reciproc, aratam ca $L(A) \subseteq L(G)$. Pentru aceasta se arată că dacă $(A, w) \vdash^n (f, \lambda)$ atunci $A \xRightarrow{n} w$ prin inducție după n . **EXERCITIU**

Propoziția 3. Orice limbaj regulat este acceptat de o gramatică regulată.

Demonstratie. Fie $L \subseteq \Sigma^*$ un limbaj regulat acceptat de AFN $A = (Q, \Sigma, \delta, s, F)$. Construim gramatică regulată $G = (N, \Sigma, S, P)$ cu $N = Q, S = s, P$ definită prin

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

Aratam ca $L(A) = L(G)$. Fie $w \in \Sigma^*$. Avem următoarele echivalente:

$$\begin{aligned} w \in L(A) &\Leftrightarrow s \xRightarrow{n} w \Leftrightarrow s \Rightarrow a_1 q_1 \Rightarrow \dots \Rightarrow a_1 \dots a_n q_n \Rightarrow a_1 \dots a_n = w \Leftrightarrow s \rightarrow \\ &a_1 q_1 \in P, \dots, q_{n-1} \rightarrow a_n q_n, q_n \rightarrow \lambda, q_n \in F \Leftrightarrow q_1 \in \delta(s, a_1), \dots, q_n \in \\ &\delta(q_{n-1}, a_n), q_n \in F \Leftrightarrow (s, a_1 \dots a_n) \vdash (q_1, a_2 \dots a_n) \vdash \dots \vdash (q_n, \lambda) \Leftrightarrow \\ &a_1 \dots a_n = w \in L(A). \end{aligned}$$

Rezulta ca $L(A) = L(G)$.

Teoremă. Familia limbajelor generate de gramaticile regulate, notată cu \mathfrak{L}_{REG} , este egală cu familia limbajelor regulate.

Demonstratie. Rezulta din Propozițiile 1 și 2

Exemple de gramatici regulate.

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