

Simplificarea gramaticilor independente de context. Gramatici regulate. Automate finite deterministe

1. Simplificarea gramaticilor independente de context

Gramaticile independente de context, deci și gramaticile regulate, pot fi aduse la o formă mai simplă prin eliminarea a două tipuri de simboluri: **neobservabile** și **inaccesibile**.

- **Neterminale neobservabile**

Definiția 1.1. Fie $G = (N, T, S, P)$ o gramatică independentă de context și $X \in N$. Spunem că X este **neterminal observabil** dacă există cel puțin o derivare de forma $X \xRightarrow{*} w$, unde $w \in T^*$. În caz contrar spunem că X este **neterminal neobservabil**.

Este evident faptul că numai neterminalele observabile sunt utile în generarea cuvintelor din $L(G)$, deci producțiile care conțin în membrul stâng sau în cel drept neterminale neobservabile pot fi eliminate din P .

Propoziția 1.1. Neterminalele observabile ale unei gramatici independente de context $G = (N, T, S, P)$ se pot obține prin metoda șirului crescător de mulțimi, astfel:

$$\begin{cases} M_0 = \{X \in N \mid \exists X \rightarrow \alpha \in P \text{ cu } \alpha \in T^*\} \\ M_{n+1} = M_n \cup \{X \in N \mid \exists X \rightarrow \alpha \in P \text{ cu } \alpha \in (T \cup M_n)^*\} \end{cases}$$

Se observă că

$$M_0 \subset M_1 \subset \dots \subset M_n \subset M_{n+1} \subset N.$$

Deoarece mulțimea N este finită, rezultă că există un indice $k \geq 0$ pentru care șirul de mulțimi se stabilizează, respectiv

$$M_k = M_{k+1} = M_{k+2} = \dots,$$

deci M_k este deci mulțimea neterminalelor productive. Putem elimina acum producțiile care conțin neterminale neobservabile, adică neterminalele din mulțimea $N \setminus M_k$.

Exemplu 1.1. Se consideră gramatica independentă de context $G = (N, T, S, P)$, unde $N = \{S, A, B, C, D\}$, $T = \{a, b\}$ și mulțimea producțiilor P este următoarea:

$$S \rightarrow AB|aBC \quad (1)$$

$$A \rightarrow BA|a \quad (2)$$

$$B \rightarrow b|AC \quad (3)$$

$$C \rightarrow AC|CB \quad (4)$$

$$D \rightarrow AD|a \quad (5)$$

$$M_0 = \{X \in N \mid \exists X \rightarrow \alpha \in P \text{ cu } \alpha \in T^*\} = \{A, B, D\}$$

$$\begin{aligned} M_1 &= M_0 \cup \{X \in N \mid \exists X \rightarrow \alpha \in P \text{ cu } \alpha \in (T \cup M_0)^*\} = \\ &= \{A, B, D\} \cup \{X \in N \mid \exists X \rightarrow \alpha \in P \text{ cu } \alpha \in \{A, B, D, a, b\}^*\} = \\ &= \{A, B, D\} \cup \{S, A, B, D\} = \{S, A, B, D\} \end{aligned}$$

$$\begin{aligned} M_2 &= M_1 \cup \{X \in N \mid \exists X \rightarrow \alpha \in P \text{ cu } \alpha \in (T \cup M_1)^*\} = \\ &= \{S, A, B, D\} \cup \{X \in N \mid \exists X \rightarrow \alpha \in P \text{ cu } \alpha \in \{S, A, B, D, a, b\}^*\} = \\ &= \{S, A, B, D\} \cup \{S, A, B, D\} = \{S, A, B, D\} \end{aligned}$$

Se observă că în acest moment șirul de mulțimi s-a stabilizat deoarece $M_2 = M_1$, deci neterminalele observabile sunt cele din $M_2 = \{S, A, B, D\}$. Rezultă că singurul neterminal neobservabil este C . Putem simplifica gramatica G , îl eliminăm pe C din N și eliminăm din P producțiile în care acesta apare, obținând astfel o gramatică independentă de context

$G' = (N', T, S, P')$ echivalentă cu G și care nu conține neterminale neobservabile, unde

$N' = \{S, A, B, D\}$, iar mulțimea producțiilor P' este următoarea:

$$S \rightarrow AB \quad (1)$$

$$A \rightarrow BA|a \quad (2)$$

$$B \rightarrow b|AC \quad (3)$$

$$D \rightarrow AD|a \quad (4)$$

Teorema 1.1. Fie $G = (N, T, S, P)$ o gramatică independentă de context. Există algoritm pentru a verifica dacă $L(G)$ este o mulțime vidă sau nu.

Ca urmare, este suficient să se verifice dacă simbolul de start S este observabil adică:

$$L(G) \neq \emptyset \iff S \text{ este observabil.}$$

- **Simboluri inaccesibile**

Definiția 1.2. Fie $G = (N, T, S, P)$ o gramatică independentă de context și $X \in N \cup T$. Spunem că $X \in N \cup T$ este **simbol accesibil** dacă există cel puțin o derivare de forma

$$S \xRightarrow{*} \alpha X \beta, \text{ unde } \alpha, \beta \in (N \cup T)^*.$$

În caz contrar spunem că X este **simbol inaccesibil**.

Este evident faptul că numai simbolurile accesibile sunt utile în generarea cuvintelor din $L(G)$, deci producțiile care conțin în membrul stâng sau în cel drept simboluri inaccesibile pot fi eliminate din P .

Simbolurile accesibile ale unei gramatici independente de context $G = (N, T, S, P)$ se pot obține prin metoda șirului crescător de mulțimi, astfel:

$$\begin{cases} M_0 = \{S\} \\ M_{n+1} = M_n \cup \{X \in N \cup T \mid \exists Y \in M_n \cap N \text{ astfel încât } Y \rightarrow \alpha X \beta \in P \text{ cu } \alpha, \beta \in (N \cup T)^*\} \end{cases}$$

Se observă că $M_0 \subset M_1 \subset \dots \subset M_n \subset M_{n+1} \subset N$. Deoarece mulțimea N este finită, rezultă că există un indice $k \geq 0$ pentru care șirul de mulțimi se stabilizează, respectiv $M_k = M_{k+1} = M_{k+2} = \dots$, deci M_k este mulțimea simbolurilor accesibile. Putem elimina acum producțiile care conțin simboluri inaccesibile, adică simbolurile din mulțimea $(N \cup T) \setminus M_k$.

Exemplul 1.2. Se consideră gramatica independentă de context $G = (N, T, S, P)$, unde $N = \{S, A, B, D\}$, $T = \{a, b, c\}$ și mulțimea producțiilor P este următoarea:

$$S \rightarrow AB \quad (1)$$

$$A \rightarrow BA \mid a \quad (2)$$

$$B \rightarrow b \quad (3)$$

$$D \rightarrow AD \mid c \quad (4) .$$

Simbolurile accesibile ale gramaticii independente de context se pot obține prin metoda șirului crescător de mulțimi, astfel:

$$M_0 = \{S\}$$

$$M_1 = M_0 \cup \{A, B\} = \{S, A, B\}$$

$$M_2 = M_1 \cup \{a, b\} = \{S, A, B, a, b\}$$

$$M_3 = M_2.$$

Se observă că în acest moment șirul de mulțimi s-a stabilizat deoarece $M_3 = M_2$, deci simbolurile accesibile sunt cele din $M_3 = \{S, A, B, a, b\}$. Rezultă că simbolurile inaccesibile sunt D și c . Eliminăm pe D din N și pe c din T . Eliminăm din P producțiile în care acestea apar, obținând astfel o gramatică independentă de context $G' = (N', T, S, P')$ echivalentă cu G și care nu conține neterminale inaccesibile, unde $N' = \{S, A, B\}$, $T' = \{a, b\}$ iar și mulțimea producțiilor P' este următoarea:

$$S \rightarrow AB \quad (1)$$

$$A \rightarrow BA \mid a \quad (2)$$

$$B \rightarrow b \mid AC \quad (3) .$$

2. Gramatici regulate. Automate finite deterministe

2.1. Gramatici regulate

Definiția 2.1.1. O gramatică $G = (N, T, S, P)$ este **gramatică regulată** dacă orice producție a sa este fie de forma $A \rightarrow aB$, fie de forma $A \rightarrow a$ cu $A, B \in N$ și $a \in T$.

Exemplul 2.1.1. Fie gramatica regulată $G = (N, T, S, P)$, unde $N = \{S\}$, $T = \{a\}$, iar mulțimea producțiilor este $P = \{S \rightarrow a|aS\}$. Se observă foarte ușor că limbajul generat de gramatica G este $L(G) = \{a^n \mid n \geq 1\}$.

Exemplul 2.1.2. Fie gramatica regulată $G = (N, T, S, P)$, unde $N = \{S, A\}$, $T = \{a, b, c\}$, iar mulțimea producțiilor P este următoarea:

$$S \rightarrow aS \quad (1)$$

$$S \rightarrow bA \quad (2)$$

$$A \rightarrow cA \quad (3)$$

$$A \rightarrow c \quad (4)$$

Cuvântul $w_1 = abc \in L(G)$ deoarece:

$$\begin{array}{ccccccc} & * & & * & & * & \\ S & \Rightarrow & aS & \Rightarrow & abA & \Rightarrow & abc = w_1 \in L(G). \\ (1) & & (2) & & (4) & & \end{array}$$

Cuvântul $w_2 = aaabccccc \in L(G)$ deoarece:

$$\begin{array}{ccccccccccc} & * & & * & & * & & * & & * & & * & & * \\ S & \Rightarrow & aS & \Rightarrow & aaS & \Rightarrow & aaaS & \Rightarrow & aaabA & \Rightarrow & aaabcA & \Rightarrow & aaabccA & \Rightarrow & aaabccccA & \Rightarrow \\ (1) & & (1) & & (1) & & (2) & & (3) & & (3) & & (3) & & (4) \\ & & & & & & * & & & & & & & & \\ & & & & & & \Rightarrow & aaabccccc = w_2 \in L(G). \\ & & & & & & (4) & & & & & & & & \end{array}$$

Deoarece după aplicarea producției (2) se poate aplica producția (3) cel puțin o dată, iar producția (4) doar o singură dată, rezultă că limbajul generat de gramatica G este:

$$L(G) = \{a^n bc^m \mid n, m \geq 1\}.$$

Exemplul 2.1.3. Fie gramatica regulată $G = (N, T, S, P)$, unde $N = \{S, A\}$, $T = \{a, b, c\}$, iar mulțimea producțiilor P este următoarea:

$$S \rightarrow aA \quad (1)$$

$$A \rightarrow aA|aB \quad (2.1|2.2)$$

$$B \rightarrow bC \quad (3)$$

$$C \rightarrow cB|c \quad (4.1|4.2)$$

Se poate afirma că limbajul generat de gramatica G conține cuvântul $w = aabc$, într-adevăr are loc următoarea derivare:

$$\begin{array}{ccccccc} * & & * & & * & & * \\ S \Rightarrow aA & \Rightarrow & aaB & \Rightarrow & aabC & \Rightarrow & aabc = w \in L(G). \\ (1) & (2.2) & (3) & (4.2) & & & \end{array}$$

Deoarece producția (2.1) introduce cel puțin două simboluri a , iar producția (2.2) urmată de producția (3), la rândul ei urmată de (4.1) de oricâte ori și la final urmată de (4.2), ce introduc perechea bc de oricâte ori, rezultă că limbajul generat de gramatica G este:

$$L(G) = \{a^n(bc)^m \mid n \geq 2, m \geq 1\}.$$

2.2. Automate finite deterministe

Definiția 2.2.1. Se numește **automat finit determinist** un cvintuplu $A = (\Sigma, Q, \delta, q_0, F)$, unde:

1. Σ se numește **alfabetul de intrare**;
2. Q se numește **mulțimea stărilor** și este o mulțime finită nevidă;
3. $\delta : Q \times \Sigma \rightarrow Q$ se numește **funcția de tranziție**;
4. $q_0 \in Q$ reprezintă **starea inițială**;
5. $F \subseteq Q$ se numește **mulțimea stărilor finale**.

Observația 2.2.1. Funcția de tranziție δ se poate prelungi pe mulțimea $Q \times \Sigma^*$ obținându-se funcția $\bar{\delta} : Q \times \Sigma^* \rightarrow Q$, care poate fi aplicată și unui cuvânt, nu numai unui simbol, definită astfel:

$$\begin{cases} \bar{\delta}(q, \varepsilon) = q, \forall q \in Q \\ \bar{\delta}(q, wa) = \bar{\delta}(\bar{\delta}(q, w), a), \forall q \in Q, w \in \Sigma^*, a \in \Sigma \end{cases}$$

Observația 2.2.2. Pentru a se simplifica scrierea, în continuare se va nota $\bar{\delta}$ tot prin δ .

Observația 2.2.3. (S. Marcus) Se poate verifica că, pentru $x, y \in \Sigma^*$ și oricare $q \in Q$ avem:

$$\delta(q, xy) = \delta(\delta(q, x), y),$$

și în particular, pentru $x = a \in \Sigma$ avem:

$$\delta(q, ay) = \delta(\delta(q, a), y). \quad (**)$$

Definiția 2.2.2. Cuvântul $w \in \Sigma^*$ este **admis/ acceptat de automatul A** dacă

$$\delta(q_0, w) \in F.$$

Mulțimea tuturor cuvintelor acceptate de automatul A se notează prin $\mathcal{T}(A)$ și se numește **limbajul acceptat de automatul A** , adică:

$$\mathcal{T}(A) = \{w \in \Sigma^* \mid \delta(q_0, w) \in F\}.$$

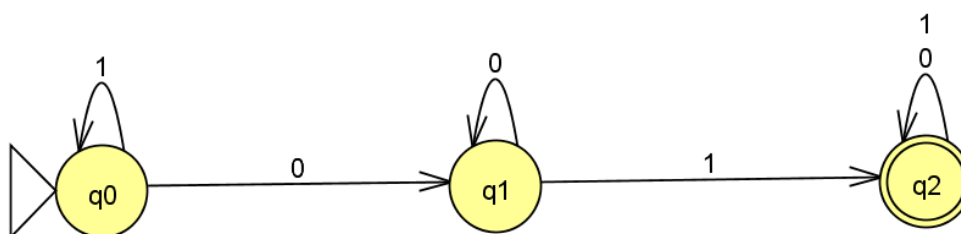
La un automat ne interesează limbajul **acceptat** de acesta, în contrast,
la o gramatică ne interesa limbajul **generat** de ea!

Un automat finit determinist poate fi reprezentat prin două metode:

1. **analitic**, prin precizarea mulțimilor Σ și Q , funcția de tranziție δ fiind dată cu ajutorul unui tabel, format din linii și coloane, în care capetele liniilor conțin stările automatului, capetele de coloane conțin simbolurile din alfabetul de intrare, astfel încât starea inițială este marcată cu \rightarrow (săgeată), iar stările finale sunt marcate cu $*$.

2. **grafic**, folosind un graf orientat în care fiecare nod reprezintă o stare a automatului, un arc între două noduri ce reprezintă stările s_i și s_j este etichetat cu simbolul a dacă are loc tranziția $\delta(s_i, a) = s_j$. Starea inițială este marcată cu o săgeată, iar o stare finală este reprezentată de un nod dublu încercuit.

Exemplul 2.2.1. Graful finit orientat de mai jos este reprezentarea grafică a automatului finit determinist $A = (\Sigma, Q, \delta, q_0, F)$ care acceptă toate șirurile binare care conțin subșirul 01.



Se pot verifica următoarele afirmații: $w_1 = 110110 \in \mathcal{T}(A)$, iar $w_2 = 1100 \notin \mathcal{T}(A)$.

Mai întâi, se construiește reprezentarea analitică a automatului A , astfel:

- $\Sigma = \{0,1\}$;
- $Q = \{q_0, q_1, q_2\}$;
- $F = \{q_2\}$;
- funcția de tranziție δ este definită în următorul tabel:

	δ	0	1
\rightarrow	q_0	q_1	q_0
	q_1	q_1	q_2
*	q_2	q_2	q_2

Folosind relația (**) de mai sus se verifică relația $w_1 = 110110 \in \mathcal{T}(A)$:

$$\begin{aligned}\delta(q_0, 110110) &= \delta(\delta(q_0, 1), 10110) = \delta(q_0, 10110) = \delta(\delta(q_0, 1), 0110) = \\ &= \delta(q_0, 0110) = \delta(\delta(q_0, 0), 110) = \delta(q_1, 110) = \delta(\delta(q_1, 1), 10) = \delta(q_2, 10) = \\ &= \delta(\delta(q_1, 1), 0) = \delta(q_2, 0) = q_2 \in F.\end{aligned}$$

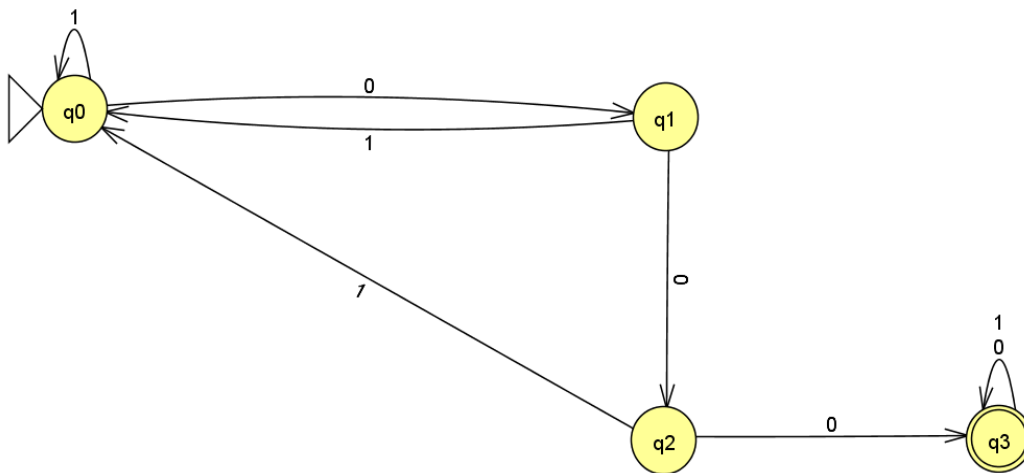
Deci $w_1 = 110110 \in \mathcal{T}(A)$.

Cuvântul w_2 nu este acceptat de automatul A . Într-adevăr, are loc:

$$\begin{aligned}\delta(q_0, 1100) &= \delta(\delta(q_0, 1), 100) = \delta(q_0, 100) = \delta(\delta(q_0, 1), 00) = \\ &= \delta(q_0, 00) = \delta(\delta(q_0, 0), 0) = \delta(q_1, 0) = q_1 \notin F.\end{aligned}$$

Deci $w_2 \notin \mathcal{T}(A)$, adică w_2 nu este acceptat de automatul A .

Exemplul 2.2.2. Graful finit orientat de mai jos este reprezentarea grafică a automatului finit determinist $A = (\Sigma, Q, \delta, q_0, F)$ care acceptă toate șirurile binare care conțin subșirul 000.



Se pot verifica următoarele afirmații: $w_1 = 11000 \in \mathcal{T}(A)$, iar $w_2 = 1001001 \notin \mathcal{T}(A)$.

Mai întâi, se construiește reprezentarea analitică a automatului A , astfel:

- $\Sigma = \{0, 1\}$;
- $Q = \{q_0, q_1, q_2, q_3\}$;
- $F = \{q_3\}$;
- funcția de tranziție δ este definită în următorul tabel:

	δ	0	1
\rightarrow	q_0	q_1	q_0
	q_1	q_2	q_0
	q_2	q_3	q_0
*	q_3	q_3	q_3

Folosind relația (***) de mai sus se verifică relația $w_1 = 11000 \in \mathcal{T}(A)$:

$$\begin{aligned} \delta(q_0, 11000) &= \delta(\delta(q_0, 1), 1000) = \delta(q_0, 1000) = \delta(\delta(q_0, 1), 000) = \\ &= \delta(q_0, 000) = \delta(\delta(q_0, 0), 00) = \delta(q_1, 00) = \delta(\delta(q_1, 0), 0) = \delta(q_2, 0) = q_3 \in F. \end{aligned}$$

Deci $w_1 = 11000 \in \mathcal{T}(A)$.

Cuvântul w_2 nu este acceptat de automatul **A**. Într-adevăr, are loc:

$$\begin{aligned} \delta(q_0, 1001001) &= \delta(\delta(q_0, 1), 001001) = \delta(q_0, 001001) = \delta(\delta(q_0, 0), 01001) = \\ &= \delta(q_1, 01001) = \delta(\delta(q_1, 0), 1001) = \delta(q_2, 1001) = \delta(\delta(q_1, 1), 001) = \\ &= \delta(q_0, 001) = \delta(\delta(q_0, 0), 01) = \delta(q_1, 01) = \delta(\delta(q_1, 0), 1) = \delta(q_2, 1) = q_0 \notin F. \end{aligned}$$

Deci $w_2 \notin \mathcal{T}(A)$, adică w_2 nu este acceptat de automatul **A**.