Proprietăți speciale ale limbajelor regulate

Atunci când am discutat despre familiile de limbaje din Clasificarea Chomsky a limbajelor am văzut că aceste familii sunt închise la operațiile regulate: Reuniune, Produs, Iterație.

Reamintim "rețetele" folosite în demonstrația constructivă a acestor proprietăți.

Fie $L_1, L_2 \in \mathcal{L}_i$, $i=\overline{3,2}$ și $G_j = (V_{N_j}, V_{T_j}, x_{0_j}, P_j) \in \mathcal{G}_i$ ce generează $L_j = L(G_j)$, $j=\overline{1,2}$

Presupunem $\boxed{V_{N_1} \cap V_{N_2} = \phi.}$

Construim gramatici ce generează $L_1 \cup L_2$, $L_1 L_2$, L_1^*

$\boxed{\cup}$ $G_\cup = (V_{N_1} \cup V_{N_2} \cup \{S\}, V_{T_1} \cup V_{T_2}, S, \{S \to x_{0_1} | x_{0_2}\} \cup P_1 \cup P_2)$

$L(G_\cup) = L_1 \cup L_2$ (valabil la orice familie)

$\boxed{\cdot}$ Tipul 2:

$G_\cdot = (V_{N_1} \cup V_{N_2} \cup \{S\}, V_{T_1} \cup V_{T_2}, S, \{S \to x_{0_1} x_{0_2}\} \cup P_1 \cup P_2)$

Tipul 3:

$G_\cdot = (V_{N_1} \cup V_{N_2}, V_{T_1} \cup V_{T_2}, x_{0_1}, P_1' \cup P_2)$

unde P_1' se obține din P_1 prin înlocuirea regulilor ce termină o derivare

$C \to g \in P_1$ cu $C \to g x_{0_2} \in P_1'$

$\boxed{*}$ Tipul 2:

$G_* = (V_{N_1} \cup \{S\}, V_{T_1}, S, \{S \to \lambda | S x_{0_1}\} \cup P_1)$

Tipul 3:

$G_* = (V_{N_1} \cup \{S\}, V_{T_1}, S, \{S \to \lambda | x_{0_1}\} \cup P_1' \cup P_1)$

Deci limbajele regulate sunt închise la operațiile regulate.

Exemplu de aplicație imediată a teoremelor. Combinarea AF.

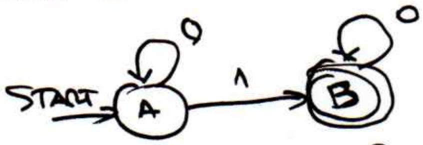
Problemă: Găsiți un AF ce recunoaște limbajul

$L=\{w\in\{0,1\}^* \mid w$ conține un 1 sau două simboluri 1$\}$

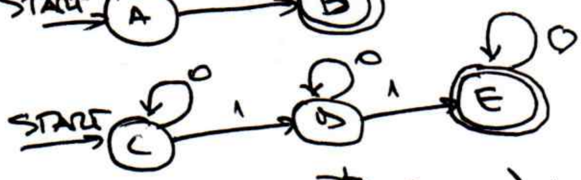
Șabloanele asociate cuvintelor sunt

L_1 $\boxed{0..0}\boxed{1}\boxed{0..0}$ și $\boxed{0..0}\boxed{1}\boxed{0...0}\boxed{1}\boxed{0..0}$ L_2

Avem 2 AF ce recunosc limbajele descrise L_1 și L_2.

AF_1: 

$L(AF_1)=L_1$

AF_2: 

$L(AF_2)=L_2$.

Extrag gramatici echivalente ce generează L_1 și L_2

$$G_1: \begin{cases} A \to 0A \mid 1B \mid 1 \\ B \to 0B \mid 0 \end{cases} \qquad G_2: \begin{cases} C \to 0C \mid 1D \\ D \to 0D \mid 1E \mid 1 \\ E \to 0E \mid 0 \end{cases}$$

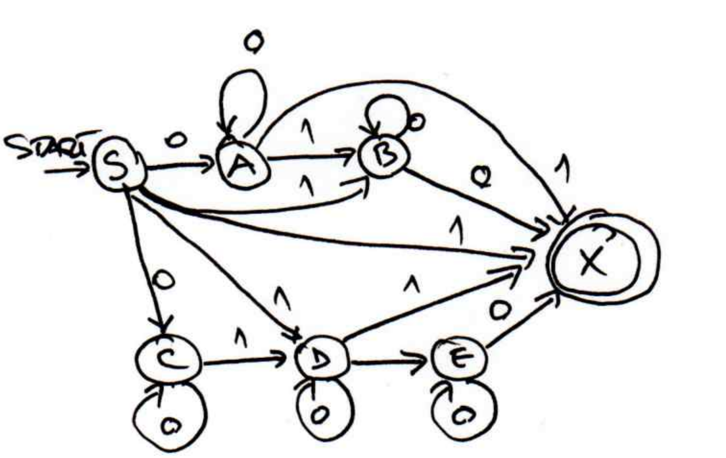
Construiesc gramatica ptr reuniunea $L_1 \cup L_2$

$$G_U = \begin{cases} S \to A \mid C \\ A \to 0A \mid 1B \mid 1 \\ B \to 0B \mid 0 \\ C \to 0C \mid 1D \\ D \to 0D \mid 1E \mid 1 \\ E \to 0E \mid 0 \end{cases}$$

Normalizez gramatica (elimin redenumirile lui S)

$$G'_U = \begin{cases} S \to 0A \mid 1B \mid 1 \mid 0C \mid 1D \\ A \to 0A \mid 1B \mid 1 \\ B \to 0B \mid 0 \\ C \to 0C \mid 1D \\ D \to 0D \mid 1E \mid 1 \\ E \to 0E \mid 0 \end{cases}$$

Desenez AF echivalent



Închiderea la operația de prefixare.

Dacă $w = i_1 i_2 .. i_n \in I^*$ atunci înțelegem prin prefix pt w orice secvență de litere de la începutul cuvântului w, adică

$$Pref(w) = \{ p \in I^* \mid \exists q \in I^* \text{ a.î. } w = pq \}$$

Obs: În definiție au fost incluse și prefixele "improprii" λ, w!

LEMA: Dacă $L \in R$ atunci $Pref(L) \stackrel{def}{=} \bigcup_{p \in L} Pref(p) \in R$

Demonstrație

Fie $L \in R$ și $AFD = (S, I, f, s_0, S_f)$ ce recunoaște L.

Construim un AF ce recunoaște prefixele cuvintelor din L

$$AF = (S, I, f, s_0, S_f')$$

unde $S_f' = \{ s \in S \mid s = f(s_0, q),\ q \in Pref(L) \}$

Este evident că $L(AF) = Pref(L)$. (Demonstrația detaliată o găsiți în cursul de pe pagina web!)

Închiderea la complementară

Dacă $L \in R$ este limbaj peste I atunci $C_L \stackrel{def}{=} I^* \setminus L$.

LEMA: Dacă $L \in R$ atunci $C_L \in R$

Demonstrație

Fie $L \in R$ și $AFD = (S, I, f, s_0, S_f)$ ai $L = L(AFD)$

Construim imediat un AFD_C ce recunoaște C_L.

$$AFD_L = (S, I, f, s_0, S_f') \text{ unde } S_f' = S \setminus S_f$$

Evident $L(AFD_L) = C_L$

Corolar: Închiderea la intersecție

Dacă $L_1, L_2 \in R$ atunci $L_1 \cap L_2 \in R$

Demonstrație

Imediată prin folosirea formulelor lui DeMORGAN ce transformă intersecția în Complementară și Reuniune

$$L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}} \quad \text{q.ed.}$$

LEMA DE POMPARE (PTR. LIMBAJE REGULATE)

Lema pune în evidență o proprietate caracteristică a limbajelor regulate și se folosește ptr. a arăta că un limbaj dat nu este regulat.

LEMA: $\forall L \in \mathcal{R}$ $\exists n \in \mathbb{N}$ a.î. $\forall p \in L$ cu $|p| \geq n$ $\exists$ partiționarea

$p = uvw$ cu proprietățile:

1) $v \neq \lambda$

2) Cuvântul nou format $p' = uv^{k'}w \in L$, $\forall k' \in \mathbb{N}$

(cuvântul format prin "pomparea" lui v de oricâte ori rămâne în limbaj)

#

Fie $L \in \mathcal{R}$ și $AFD = (S, I, f, s_0, S_f)$ a.î. $L(AFD) = L$.

Fie $n = |S|$ și un cuvânt arbitrar $p \in L$ cu lungime $\geq n$.

pres. $p = i_1 i_2 \dots i_k$, $k \geq n$ (k suficient de mare)

$p \in L \iff$ avem o traiectorie de la s_0 la o stare finală cu arcele etichetate cu literele lui p.

Detaliem traiectoria

$$s_0 \xrightarrow{i_1} s_1 \xrightarrow{i_2} s_2 \longrightarrow \dots \xrightarrow{i_k} s_k \in S_f.$$

Pe traiectorie apar $k+1 > n$ stări, deci sigur o etichetă se repetă cel puțin o dată pe traiectorie. Fie X eticheta ce se repetă și $0 \leq t < s \leq k$ pozițiile ei care apare. Rescriu traiectoria evidențiind apariția lui X.

$$s_0 \xrightarrow{i_1} s_1 \rightarrow \dots \xrightarrow{i_t} \underset{X}{\underset{\parallel}{s_t}} \xrightarrow{i_{t+1}} \dots \longrightarrow \underset{X}{\underset{\parallel}{s_s}} \xrightarrow{i_{s+1}} \dots \xrightarrow{i_k} s_k \in S_f.$$

Considerăm partiționarea cuvântului p determinată de apariția etichetei $X = s_t = s_s$, $0 \leq t < s \leq k$

$$\underbrace{s_0 \xrightarrow{i_1} s_1 \rightarrow \dots \xrightarrow{i_t} s_t}_{u} = X \underbrace{\xrightarrow{i_{t+1}} \dots \longrightarrow s_s}_{v} = X \underbrace{\xrightarrow{i_{s+1}} \dots \xrightarrow{i_k} s_k}_{w} \in S_f$$

Avem partiționarea $p = uvw$ cu

$$u = i_1 \dots i_t$$
$$v = i_{t+1} \dots i_s \quad , \; 0 \leq t < s \leq k$$
$$w = i_{s+1} \dots i_k$$

Arătăm că partiționarea are proprietățile din lemă

1) $v \neq \lambda$ evident din condiția impusă $t < s \Rightarrow s - t > 0 \Rightarrow$
$|v| = s - t \geq 1$

2) Cuvântul nou format prin repetarea lui v rămâne în limbaj.

Evident $f(X, v) = X$ (adică $f(s_t, v) = s_s$).

Deci $f(X, v^2) = f(f(X, v), v) = f(X, v) = X$

Și imediat $f(X, v^{k'}) = X$, $\forall k' \in \mathbb{N}$

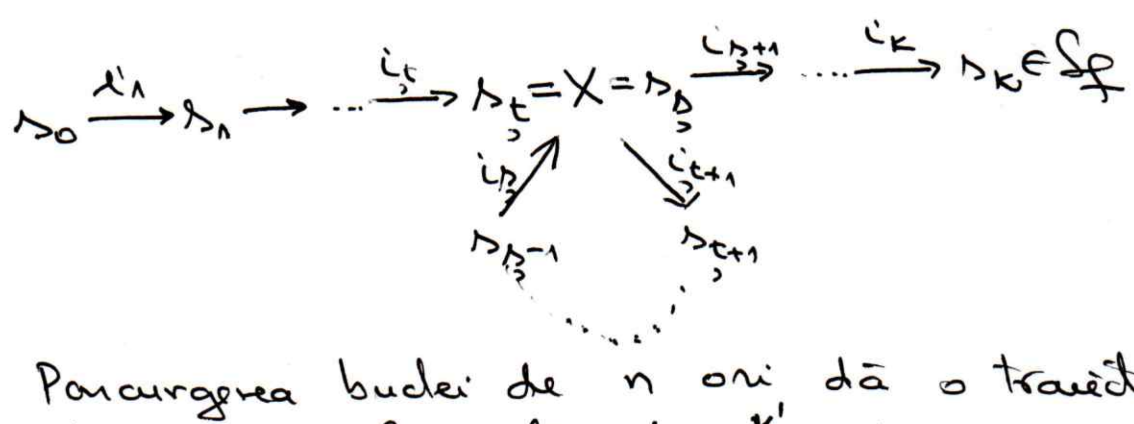
$f(s_0, u) = X$, $f(X, w) = s_k$

Consider cuvântul nou format $p' = u v^{k'} w$.

$$f(s_0, p') = f(s_0, u v^{k'} w) = f(f(s_0, u v^{k'}), w) =$$
$$= f(f(\underline{f(s_0, u)}, v^{k'}), w) =$$
$$= f(\underline{f(X, v^{k'})}, w) = f(X, w) = s_k \in S_f.$$

adică $p' \in L(AFD)$. q.e.d.

Grafic, putem redesena traiectoria de recunoaștere a lui p sub forma unei traiectorii ce conține un ciclu între $s_t = X = s_s$.



Parcurgerea buclei de n ori dă o traiectorie de recunoaștere pt $p' = u v^{k'} w \in L$

Corolar: $\mathcal{L}_3 \underset{\neq}{\subset} \mathcal{L}_2$ este strictă

Vom arăta că limbajul $L = \{a^n b^n \mid n \geq 1\} \in \mathcal{L}_2 - \mathcal{L}_3$.

Evident că gramatica G cu regulile

$$S \to aSb \mid ab \in \mathcal{G}_2$$

generează L, adică $L(G) = L$. Deci $L \in \mathcal{L}_2$.

Arătăm că $L \notin \mathcal{L}_3$ (folosim lema de pompare!)

Pres. prin reducere la absurd că $L \in \mathcal{R}$.

Cf. lemei $\exists\, n \in \mathbb{N}$ (fixat, depinde doar de limbaj!) a.î. pentru orice cuvânt suficient de lung ($|p| \geq n$) $\exists$ partiționarea $p = uvw$, cu $v \neq \lambda$ și orice cuvânt nou format $p' = uv^k w$ rămâne în limbaj.

Considerăm <u>o partiționare oarecare</u> a lui p (obținută prin extragerea unui subcuvânt nevid v)

$p = uvw$. Pres. $p = a\ldots ab\ldots b = a^m b^m$, $m \geq n$.

și pentru fiecare posibilitate de extragere a lui v găsim un nr. de repetări pt. care $p' = uv^{n_0}w \notin L$, adică avem o contradicție cu lema (presupusă valabilă!)

<u>Caz 1.</u> în cuvântul v intră un singur tip de simbol

$v = a^t$, $t \geq 1$. $p = \underbrace{a..a}_{u}\underbrace{a..a}_{v}\underbrace{a..ab\ldots b}_{w}$

Aleg $n_0 = 2$. Cuvântul nou format $p' = uv^2w \notin L$

$p' = uv^2w = \underbrace{a\ldots a}_{u}\underbrace{a\ldots a}_{v}\underbrace{a\ldots a}_{v}\underbrace{a\ldots ab\ldots b}_{w} = a^{m+t}b^m \notin L$

CONTRADICȚIE cu L.P.

Idem cazul $v = b^t$, $t \geq 1$.

<u>Caz 2</u> în v intră ambele litere $v = a^t b^s$ $t, s \geq 1$

$p = \underbrace{a\ldots a}_{u}\underbrace{a\ldots ab\ldots b}_{v}\underbrace{b\ldots b}_{w}$

Aleg $n_0 = 2$. Cuvântul nou format $p' = uv^2w \notin L$

Desenăm cuvântul

$p' = \underbrace{a\ldots a}_{u}\underbrace{a\ldots ab\ldots b}_{v}\underbrace{a\ldots ab\ldots b}_{v}\underbrace{b\ldots b}_{w} \notin L$ deoarece a urmează după b

CONTRADICȚIE cu L.P.

Obs: 1) Fie $L \in \mathcal{R}$ și $AFD = (S, I, f, s_0, S_f)$ ai $L = L(AFD)$.

Atunci $|L| = \infty \iff \exists\, p \in L$ a.î. $|p| \geq |S|$

#

Dacă L este infinit atunci evident am cuvinte de lungimi oricât de mari, adică $|p| \geq |S|$.

Dacă $\exists\, p \in L$ a.î. $|p| \geq |S|$ atunci orice cuvânt $p' = uv^n w \in L$, $\forall n \in \mathbb{N}$, adică $|L| = \infty$

Deci în cazul limbajelor finite, care sunt regulate, orice cuvânt din limbaj are mai puține litere "decât un AF" ce recunoaște limbajul, mai exact decât numărul stărilor oricărui automat pe care îl pot asocia recunoașterii limbajului

2) Sunt limbaje $L \in \mathcal{L}_2 \setminus \mathcal{L}_3$ pentru care încercarea de a arăta că $L \notin \mathcal{L}_3$ folosind procedeul din rezolvarea $\{a^n b^n \mid n \geq 1\} \notin \mathcal{L}_3$ nu funcționează direct.

Exemplu: Arătați că $L = \{a^n b^k a^n \mid n, k \geq 1\} \notin \mathcal{L}_3$.

Incercam. Presupunem prin reducere la absurd că $L \in \mathcal{L}_3$. Aplicăm lema de pompare, deci avem o constantă $m \in \mathbb{N}$ și pentru orice partiționare a unui cuvânt p suficient de lung ($|p| \geq m$) încerc să găsesc un nr. convenabil de repetări n_0 astfel încât cuvântul nou format $p' = uv^{n_0}w$ să nu rămână în limbaj (adică să contrazic lema!).

Ptr. cazul în care extrag un subcuvânt v de forma b^t, $t \geq 1$ nu pot găsi n_0 convenabil deoarece nr. de sb. b crește și cuvântul rămâne în limbaj.

???

Idee: Se consideră un AFD ce recunoaște L și un cuvânt cu lungime suficient de mare astfel încât n (nr. de apariții a lui a în prima parte a cuvântului) să depășească numărul de stări ptr AFD. Astfel, pot construi traiectorii ptr. cuvinte ce au mai multe simboluri a în stânga lui b, fără să afectez nr. simboluri din dreapta!

Teorema de caracterizare algebrică a lb. regulate

Dă o caracterizare a structurii cuvintelor unui limbaj regulat indiferent de mecanismul de recunoaștere sau generare.

TEOREMĂ: Fie $L \subset I^*$ un limbaj. Următoarele afirmații sunt echivalente.

(a) $L \in \mathcal{R}$

(b) L este o reuniune de clase de echivalență a unei congruențe de rang finit (congruența $\overset{def}{=}$ echivalență liniară la dreapta și la stânga în raport cu concatenarea!)

(c) Următoarea congruență

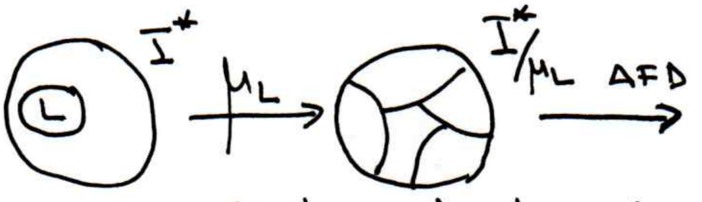
$$\mu_L = \{ (p,q) \mid \varphi_L(r_1 p r_2) = \varphi_L(r_1 q r_2), \forall r_1, r_2 \in I^* \}$$

unde φ_L este funcția caracteristică a lui L, este de rang finit

#

O găsiți detaliat în cursul de pe web.

De reținut că teorema pune în corespondență clasele de echivalență I^*/μ_L cu stările unui automat finit determinist ce recunoaște limbajul L. În plus, automatul astfel determinat are cel mai mic număr de stări posibile, deci este AF minimal asociat unui limbaj L (unic până la un izomorfism!)



$AFD = (I^*/\mu_L, I, f, C_\lambda, S_f)$

unde stările sunt clase de echivalență ($|I^*/\mu_L| < \infty$)

C_λ este clasa lui λ

$S_f = \{ C_p \mid p \in L \}$ – acele clase asociate cuvintelor limbajului (nr. finit)

$f(C_p, i) \overset{def}{=} C_{pi}$ este funcția de evoluție

Interesant! Eficient? Cam greu ptr că sunt mulțimi infinite de cuvinte!

Corolar: Familia $\mathcal{R}$ este închisă la oglindire

Dacă $p = i_1 i_2 \ldots i_n \in I^*$ atunci oglinditul este $\tilde{p} = i_n \ldots i_1 \in I^*$

$Mi(L) \overset{def}{=} \{\tilde{p} \mid p \in L\}$, $L \in I^*$

Fie $L \in \mathcal{R}$. Din teorema de caracterizare avem

$$\mu_L = \{(p,q) \mid \varphi_L(r_1 p r_2) = \varphi_L(r_1 q r_2), \forall r_1, r_2 \in I^*\}$$

are rang finit (adică un nr. finit de clase de echivalență).

Dar $\mu_{\tilde{L}} = \{(p,q) \mid \varphi_{\tilde{L}}(r_1 p r_2) = \varphi_{\tilde{L}}(r_1 q r_2) \ \forall r_1, r_2 \in I^*\}$

are rang finit.

deoarece $(p,q) \in \mu_L \Rightarrow \varphi_L(r_1 p r_2) = \varphi_L(r_1 q r_2) \ \forall r_1, r_2 \in L$

$$\widetilde{r_1 p r_2} = \tilde{r}_2 \tilde{p} \tilde{r}_1 \quad , \forall r_1, r_2 \in L \Rightarrow$$

$$\varphi_{\tilde{L}}(\widetilde{r_1 p r_2}) = \varphi_{\tilde{L}}(\widetilde{r_1 q r_2})$$

$$\Downarrow$$

$$\varphi_{\tilde{L}}(\tilde{r}_2 \tilde{p} \tilde{r}_1) = \varphi_{\tilde{L}}(\tilde{r}_2 \tilde{q} \tilde{r}_1) \qquad \forall \tilde{r}_1, \tilde{r}_2 \in I^*$$

$$\Downarrow$$

$$\varphi_{\tilde{L}}(r'_1 \tilde{p} r'_2) = \varphi_{\tilde{L}}(r'_1 \tilde{q} r'_2) \qquad \forall r'_1, r'_2 \in I^*$$

$$\Downarrow$$

$$(\tilde{p}, \tilde{q}) \in \mu_{\tilde{L}} \quad \text{q.e.d.}$$

$\Rightarrow \tilde{L}$ este regulat.

Obs: Dacă folosim definiția $\begin{cases} A \to iB \\ C \to j \end{cases}$ pt

limbaje regulate, atunci la fel de bună este

și $\begin{cases} A \to Bi \\ C \to j \end{cases}$ ce generează chiar oglinditul limbajului

Această observație justifică definirea echivalentă a gramaticilor regulate.

Problemă: Fie $L \in \mathcal{R}$. Găsiți un AF ce recunoaște $\tilde{L}$ pornind de la un AF ce recunoaște L! Idem cu gramatici!!