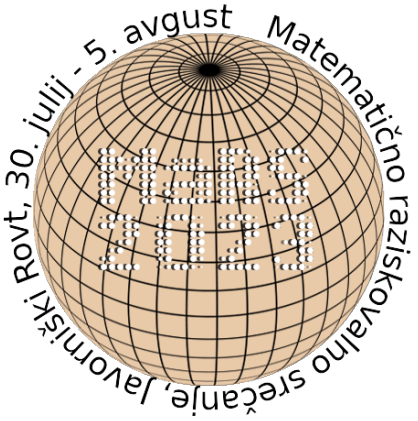
# Končni avtomati

Matic Bratina, Aleksander Kalacun, Vasja Žorž
Mentor: Nejc Zajc




**Povzetek**

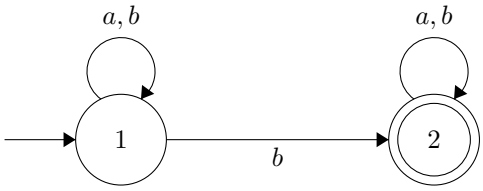
V članku so predstavljeni koncepti formalnega jezika in končnih avtomatov. Opisano je delovanje končnih avtomatov in njihove lastnosti. Povezava med njimi in formalnimi jeziki je pojasnjena na intuitiven in dostopen način.

## 1 Uvod

Končni avtomati so končne množice stanj in usmerjenih povezav. So ključni računalniški modeli za prepoznavanje vzorcev in reševanje problemov s formalnimi jeziki. Obstajata dve vrsti; deterministični in nedeterministični končni avtomati. Njihovo razumevanje je temelj za številne aplikacije, kot so prevajalniki in obdelava jezika. Avtomate lahko predstavimo z nazornim vizualnim gradivom.

## 2 Jeziki

Naj bo $A$ končna abeceda črk. Beseda $u = a_1 a_2 \cdots a_n$ je končen niz črk $a_i \in A$ za vse $i \in \{1, 2, \ldots, n\}$. Z $|u| = n$ označimo dolžino besede, ki je enaka številu črk v besedi. Besedi, ki ne vsebuje nobene črke, pravimo prazna beseda in jo označimo z $\varepsilon$. Množico vseh besed nad abecedo $A$ označimo z $A^*$. Podmnožici $L \subseteq A^*$ rečemo jezik.




Slika 1: Primer diagrama končnega avtomata.

Stik besed, ki ga pišemo kot množenje, je operacija združitve dveh besed. Za besedi $u = a_1a_2 \cdots a_n$ in $v = b_1b_2 \cdots b_m$ je njun stik beseda

$$u \cdot v = a_1 \cdots a_n b_1 \cdots b_m \in A^*.$$

## 2.1 Operacije na jezikih

Oglejmo si, kako lahko z jeziki računamo. Izvajamo lahko operacije, značilne za množice, kot sta unija in presek. Standardno ju označimo z $\cup$ in $\cap$. Za lažjo berljivost bomo operacijo unije zapisovali kot $+$. Nevtralen element unije je prazna množica $\emptyset$, ki jo označimo z 0.

Množenje jezikov $L_1$ in $L_2$ definiramo kot operacijo, ki nam da nov jezik, v katerem so vse besede, za katere je prvi del beseda iz $L_1$ in drugi del beseda iz $L_2$. Torej je $L_1 \cdot L_2 = \{uv \mid u \in L_1, v \in L_2\}$. Nevtralen element množenja je jezik s prazno besedo $\{\varepsilon\}$, ki ga označimo z 1.

Z definicijo množenja lahko vpeljemo tudi operacijo potenciranja. Definiramo jo kot

$$L^0 = 1, \qquad L^1 = L, \qquad L^n = L \cdot L^{n-1} \quad \text{za vse } n \geq 1.$$

Definiramo še operaciji

$$L^* = 1 + L + L^2 + L^3 + \cdots \quad \text{in} \quad L^+ = L + L^2 + L^3 + \cdots .$$

Za omenjene operacije na jezikih veljajo naslednje lastnosti. Unija, presek in množenje so asociativne operacije nad jeziki. Velja tudi distributivnost množenja nad unijo. Če preverimo še komutativnost, ugotovimo, da je poleg unije tudi presek komutativen. Pri množenju vidimo, da je pri obliki besede pomembno, katera sestavlja začetek produkta besed in katera konec, tako da množenje ni komutativna operacija.

Oglejmo si še operacijo, nasprotno množenju. Kvocient je operacija, ki jo za besedo $u$ in jezik $L$ zapišemo kot $u^{-1}L$. Kvocient je jezik, v katerem so končnice vseh besed jezika $L$, ki se začnejo z $u$, torej

$$u^{-1}L = \{v \in A^* \mid uv \in L\}.$$

Primer operacije kvocient je

$$(aba)^{-1}\{abaa, aabbb, abab\} = \{a, b\}.$$

Množica vseh jezikov je potenčna množica $P(A^*)$. Oglejmo si množico jezikov, ki jih lahko zgradimo s končnim številom osnovnih operacij iz jezikov, ki vsebujejo le eno črko.

**Definicija 1.** *Množica $F \subseteq P(A^*)$ **racionalnih jezikov** je najmanjša množica jezikov, za katero velja:*

- *$F$ vsebuje 0 in $\{a\}$ za vsak $a \in A$,*
- *$F$ je zaprta za končne unije, produkte in $^*$, torej*

$$\forall L_1, L_2 \in F: \quad L_1 + L_2 \in F, \quad L_1 \cdot L_2 \in F \quad \text{in} \quad L_1^* \in F.$$

Poljuben končen jezik je racionalen, saj velja

$$\{u_1, u_2, \ldots, u_n\} = \{u_1\} \cup \{u_2\} \cup \{u_3\} \cup \ldots \cup \{u_n\}.$$

Recimo, da je $L_1$ jezik besed sode dolžine abecede $A = \{a, b\}$. Naj bo jezik $X$ takšen, da so v njem vse različne besede dolžine 2 iz $A$. V našem primeru je $X = \{aa, ab, ba, bb\}$. Velja, da je dolžina zmnožka poljubnih dveh besed iz $X$ soda. Če dobljen produkt še naprej množimo z besedami iz $X$, bo dobljen produkt vedno sode dolžine. Tako lahko množico vseh možnih besed sode dolžine zapišemo kot vsoto potenc

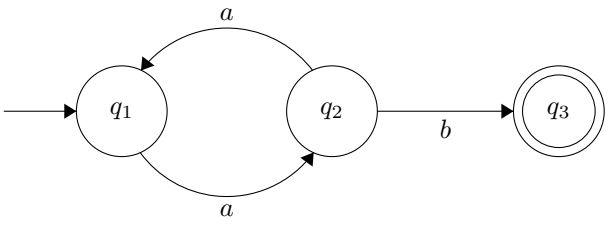
$$L_1 = 1 + X + X^2 + X^3 + \cdots ,$$

kjer so v $X^2$ vse besede dolžine 4, v $X^3$ so besede dolžine 6 in podobno. Spomnimo se, da je to enako $X^*$. Označimo z $L_2$ še jezik besed lihe dolžine. Zapišemo ga lahko kot produkt jezika $L_1$ z abecedo $A$. Tako je dolžina vseh besed za eno večja oziroma liha. Velja

$$L_2 = L_1 \cdot A = (1 + X + X^2 + X^3 + \cdots) \cdot A.$$

# 3 Končni avtomati

Končni avtomat je zgrajen iz stanj in usmerjenih prehodov med stanji. Formalno je avtomat $\mathcal{A}$ urejena peterica $\mathcal{A} = (A, Q, E, I, F)$. Stanja avtomata tvorijo množico $Q = \{q_1, q_2, \ldots, q_n\}$. Množica $E$ je množica prehodov med stanji avtomata $E \subseteq Q \times A \times Q$. Avtomat lahko predstavimo z diagramom, kot je prikazano na sliki 2. Na diagramih stanja označimo s krogi, prehode pa s puščico med dvema stanjema in črko, za katero se prehod izvede. Označimo še množici začetnih in končnih stanj $I, F \subseteq Q$.




Slika 2: Primer končnega avtomata z $I = \{q_1\}$ in $F = \{q_3\}$.

**Definicija 2.** ***Pot*** *v $\mathcal{A}$ je zaporedje stanj $q_1, q_2, \ldots, q_n$ s prehodi med njimi $(q_1, a_1, q_2), \ldots, (q_{n-1}, a_{n-1}, q_n)$ za neke črke $a_1, \ldots, a_{n-1}$.*

*Beseda $a_1 \cdots a_{n-1}$ je* ***oznaka poti****. Beseda je* ***sprejeta*** *s strani avtomata $\mathcal{A}$, če je oznaka poti iz začetnega stanja $q_i \in I$ v končno stanje $q_f \in F$. Takšna pot je v avtomatu $\mathcal{A}$ uspešna.*

Jezik, ki je prepoznan s strani avtomata $\mathcal{A}$, označimo z

$$\mathcal{L}(\mathcal{A}) = \{v \in A^* \mid \mathcal{A} \text{ sprejme besedo } v\}.$$

Za jezik $L$ rečemo, da je prepoznaven, če obstaja končni avtomat $\mathcal{A}$, da velja $L = \mathcal{L}(\mathcal{A})$.

**Lema 1** (Iteracijska lema)**.** *Naj bo $\mathcal{A}$ poljuben končni avtomat in $L = \mathcal{L}(\mathcal{A})$. Potem obstaja takšno naravno število $n \in \mathbb{N}$, da lahko vsak $u \in L$, $|u| \geq n$, zapišemo v obliki $u = xyz$, kjer so $x,\ y,\ z \in A^*$, $|xy| \leq n$ in $y \neq \varepsilon$. Beseda $xy^k z$ je element $L$ za vsak $k \in \mathbb{N}$.*

*Dokaz.* Naj bo $L = \mathcal{L}(\mathcal{A})$, kjer je $\mathcal{A} = (A, Q, E, I, F)$ končni avtomat. Naj bo $n = |Q|$ in beseda $u = a_1 a_2 \cdots a_m \in L$, tako da je $m \geq n$. Ker je $u \in L$, obstaja pot od $q_- \in I$ do $q_f \in F$, ki ima oznako $u$.


Slika 3: Pot iz dokaza leme 1.

Ker je stanj na poti $m + 1 > n$, se po Dirichletovem principu gotovo vsaj eno izmed stanj avtomata $\mathcal{A}$ na tej poti ponovi. Naj bo $i$ najmanjši indeks, da velja $q_i = q_j$ za $j > i$. Za $x$, $y$ in $z$ vpeljemo

$$\begin{aligned} x &= a_1 \cdots a_{i-1}, \\ y &= a_i \cdots a_{j-1}, \\ z &= a_j \cdots a_m. \end{aligned}$$

Velja $|xy| = j - 1 \leq n$, saj je $i$ najmanjši indeks ponovitve, $y$ pa ni prazna beseda, ker je $j > i$. Ponovitev stanja pomeni, da imamo na poti cikel. Ponavljanje cikla ne vpliva na to, da se pot začne v začetnem in konča v končnem stanju. Besedo $y$ smo definirali tako, da je oznaka tega cikla. Torej za poljuben $k \in \mathbb{N}_0$ velja $xy^k z \in L$. □

Oglejmo si primer uporabe leme, ki pokaže, da jezik $L = \{a^n b^n \mid n \geq 1\}$ ni prepoznaven. To storimo s protislovjem. Denimo, da je $L$ prepoznaven. Po lemi 1 obstaja tak $n_0 \in \mathbb{N}$, da lahko besedo $u = a^{n_0} b^{n_0} \in L$ zapišemo kot stik besed $x$, $y$, $z$, tako da je $u = xyz$. Velja $|a^{n_0} b^{n_0}| = 2n_0 \geq n_0$, ter $|xy| \leq n_0$. Torej tako $x$ kot $y$ vsebujeta samo črko $a$, torej

$$\begin{aligned} x &= a^t, \\ y &= a^s, \\ z &= a^{n_0-(t+s)} b^{n_0}, \end{aligned}$$

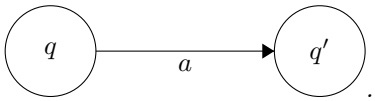
saj je $s + t \leq n_0$. Prav tako je $s \neq 0$, ker velja $y \neq \varepsilon$. Ker se pri potenciranju besede $y$ spreminja le število $a$-jev, ima beseda $xy^k z$ različno število $a$-jev in $b$-jev za $k \neq 1$. Ta beseda tako ni v jeziku $L$, kar je v protislovju z lemo in jezik ni prepoznaven.

# 4 Determinizacija končnega avtomata

Oglejmo si nekaj vrst končnih avtomatov. Smiselno si je želeti, da lahko v avtomatu od poljubnega stanja z dano besedo pridemo le do enega stanja. Avtomat, ki izpolnjuje to lastnost, je deterministično.

**Definicija 3.** *Avtomat $\mathcal{A} = (A, Q, E, I, F)$ je **deterministično**, če velja*

- $|I| = 1$,
- *za vsak $q \in Q$ in za vsak $a \in A$ obstaja največ en prehod oblike*




*V tem primeru zapišemo $q' = q \cdot a$.*

**Definicija 4.** *Avtomata $\mathcal{A}$ in $\mathcal{A}'$ sta **ekvivalentna**, če prepoznata isti jezik, torej velja*

$$\mathcal{L}(\mathcal{A}) = \mathcal{L}(\mathcal{A}').$$

**Trditev 1.** *Vsak avtomat je ekvivalenten nekemu determinističnemu avtomatu.*

Dokaz trditve uporabi potenčno konstrukcijo, ki zgradi deterministično avtomat, ki je ekvivalenten začetnemu. Naj bo $\mathcal{A} = (A, Q, E, I, F)$ poljuben avtomat. Zgradimo deterministično avtomat

$$\mathcal{A}' = (A, P(Q), E', I, \mathcal{F}),$$

kjer je $P(Q)$ potenčna množica množice $Q$. Stanja avtomata $\mathcal{A}'$ so torej množice stanj avtomata $\mathcal{A}$. Za novo stanje $P \in P(Q)$ in črko $a \in A$ vpeljemo $P \cdot a$ kot množico prvotnih stanj $q \in Q$, v katera lahko pridemo iz poljubnega stanja $p \in P$ preko prehoda $(p, a, q) \in E$, torej

$$P \cdot a = \{q \in Q \mid \exists\, p \in P :\ (p, a, q) \in E\}.$$

Množica prehodov potenčne konstrukcije je izbrana tako, da omogoča vse prehode, ki so bili na voljo v začetnem avtomatu $\mathcal{A}$

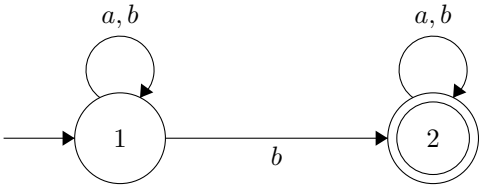
$$E' = \{(P, a, P \cdot a) \mid P \in P(Q), a \in A\}.$$

Končna stanja potenčne konstrukcije so tiste množice, ki vsebujejo katerega od končnih stanj avtomata $\mathcal{A}$,

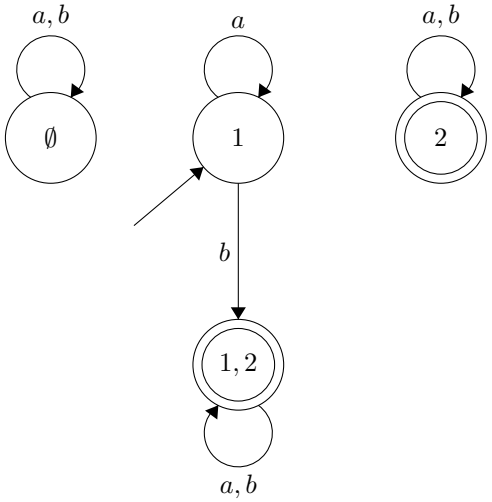
$$\mathcal{F} = \{P \subseteq Q \mid P \cap F \neq \emptyset\}.$$

Za potenčno konstrukcijo se izkaže, da prepozna isti jezik kot začetni avtomat, torej je $\mathcal{L}(\mathcal{A}) = \mathcal{L}(\mathcal{A}')$.

Na primeru si oglejmo, kako deluje potenčna konstrukcija. Slika 4 prikazuje avtomat $\mathcal{A}$. S potenčno konstrukcijo lahko zgradimo deterministično avtomat, prikazan na sliki 5, ki je ekvivalenten avtomatu $\mathcal{A}$.



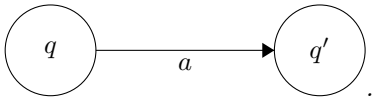

Slika 4: Avtomat $\mathcal{A}$.




Slika 5: Deterministični avtomat, ekvivalenten avtomatu $\mathcal{A}$.

**Definicija 5.** *Avtomat ima lahko naslednje lastnosti.*

- *Avtomat $\mathcal{A}$ je **poln**, če za vsak $q \in Q$ in za vsak $a \in A$ obstaja vsaj en prehod oblike*




.

- *Avtomat $\mathcal{A}$ je **dostopen**, če za vsak $q \in Q$ obstaja pot iz nekega začetnega stanja $q_i \in I$ v $q$.*
- *Avtomat $\mathcal{A}$ je **standarden**, če se noben prehod ne konča v začetnem stanju.*

Opazimo, da je potenčna konstrukcija poln in deterministično avtomat. Z novo konstrukcijo lahko ohranimo prepoznani jezik tudi pri prehodu na standarden avtomat.

**Trditev 2.** *Vsak deterministično avtomat je ekvivalenten nekemu standardnemu determinističnemu avtomat.*

Tudi tokrat le podamo konstrukcijo, ki jo uporabi dokaz, a ekvivalence med začetnim in novim avtomatom ne dokažemo. Naj bo $\mathcal{A} = (A, E, Q, q_-, F)$ deterministično avtomat. Če avtomat $\mathcal{A}$ ni standarden, dodamo novo stanje $p \notin Q$. Definiramo $\mathcal{A}' = (A, Q \cup \{p\}, E', p, F')$, kjer sta

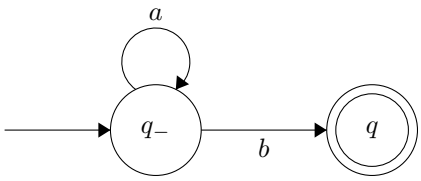
$$E' = E \cup \{(p, a, q) \mid (q_-, a, q) \in E\},$$
$$F' = \begin{cases} F; & q_- \notin F \\ F \cup \{p\}; & q_- \in F \end{cases}.$$

Avtomata $\mathcal{A}$ in $\mathcal{A}'$ sta ekvivalentna.

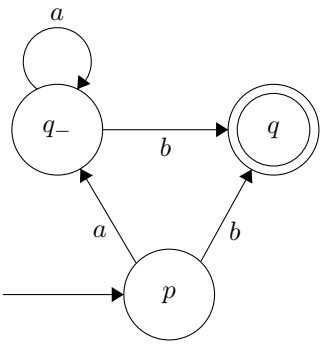
Na primeru si oglejmo še, kako deluje gradnja standardnega avtomata. Na sliki 6 vidimo deterministično avtomat $\mathcal{A}$, na sliki 7 pa standardni deterministični avtomat, ki je ekvivalenten avtomatu $\mathcal{A}$.

# 5 Konstrukcije avtomatov

Pokazali bomo, da so racionalni jeziki natanko prepoznavni jeziki. Najprej pokažemo, da so vsi racionalni jeziki tudi prepoznavni. V ta namen vse operacije, ki gradijo racionalne jezike, ponazorimo na avtomatih.




Slika 6: Deterministično avtomat $\mathcal{A}$.




Slika 7: Standardni deterministični avtomat ekvivalenten avtomatu $\mathcal{A}$.

## 5.1 Unija

Naj avtomat $\mathcal{A}_1 = (A_1, Q_1, E_1, I_1, F_1)$ prepozna jezik $L_1$ in naj avtomat $\mathcal{A}_2 = (A_2, Q_2, E_2, I_2, F_2)$ prepozna jezik $L_2$. Da dobimo avtomat $\mathcal{A}_U$, ki prepozna unijo jezikov $L_1$ in $L_2$, avtomata $\mathcal{A}_1$ in $\mathcal{A}_2$ preprosto združimo v enega. Velja torej, da avtomat

$$\mathcal{A}_U = (A_1 + A_2, Q_1 + Q_2, E_1 + E_2, I_1 + I_2, F_1 + F_2)$$

prepozna jezik $L_1 \cup L_2$.

## 5.2 Komplement

Naj deterministično in poln avtomat $\mathcal{A} = (A, Q, E, I, F)$ prepozna jezik $L$. Komplement jezika $L$ je jezik $L^C$, v katerem so vse besede, ki jih ni v jeziku $L$. Avtomat, ki prepozna jezik $L^C$, ne sme sprejeti nobene besede, ki jo sprejme $\mathcal{A}$, mora pa sprejeti vse besede, ki jih avtomat $\mathcal{A}$ ne. To lahko dosežemo s tem, da zamenjamo končna stanja. To pomeni, da tista stanja, ki so v $\mathcal{A}$ končna, v $\mathcal{A}_C$ niso in obratno. Velja torej, da avtomat

$$\mathcal{A}_C = (A, Q, E, I, Q \setminus F)$$

prepozna jezik $L^C$.

## 5.3 Zvezdica

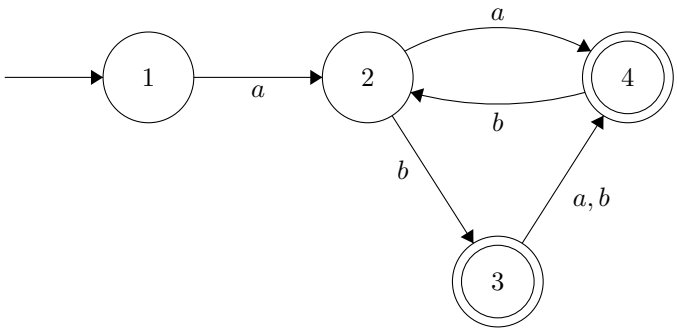
Naj deterministično in standarden avtomat $\mathcal{A} = (A, Q, E, q_-, F)$ prepozna jezik $L$. Če želimo zgraditi avtomat $\mathcal{A}_Z$, ki bo prepoznal jezik $L^*$, mora ta imeti možnost, da sprejme zaporedje besed iz jezika $L$. To lahko dosežemo tako, da dodamo prehode iz stanj, ki vodijo v končna stanja, nazaj do začetnega stanja. Primer te konstrukcije za končni avtomat s slike 8 vidimo na sliki 9. Natančneje, dodati moramo prehode oblike $(q, a, q_-)$ za $q \in Q$ in $a \in A$, če in samo če obstaja tak prehod $(q, a, q')$, da je $q' \in F$. Tako avtomat

$$\mathcal{A}_Z = (A, Q, E', q_-, F \cup q_-),$$

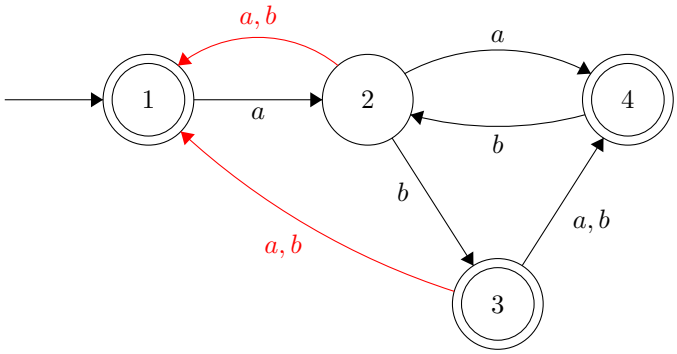
kjer velja $E' = E + \{(q, a, q_-) \mid \exists (q, a, q') \text{ za } q' \in F\}$, prepozna jezik $L^*$.

## 5.4 Produkt

Naj avtomat $\mathcal{A}_1 = (\mathcal{A}_1, Q_1, E_1, I_1, F_1)$ prepozna jezik $L_1$ in deterministično standarden avtomat $\mathcal{A}_2 = (\mathcal{A}_2, Q_2, E_2, q_-, F_2)$ prepozna jezik $L_2$. V produktu jezikov $L_1$ in $L_2$ so vse kombinacije besed, v katerih je prvi del besede iz jezika $L_1$ in drugi del besede iz jezika $L_2$. Zgraditi želimo avtomat, ki sprejme besedo, sestavljeno iz besede, ki jo sprejme avtomat $\mathcal{A}_1$, in besede, ki jo sprejme avtomat $\mathcal{A}_2$. Končna stanja avtomata $\mathcal{A}_1$ moramo združiti z avtomatom $\mathcal{A}_2$ tako, da nimamo nobenega dodatnega prehoda, saj bi nam ta med deli besed dodal črko. Zato iz avtomata $\mathcal{A}_2$ odstranimo začetno stanje $q_-$ in prehode, ki so izhajali iz začetnega stanja, povežemo tako, da se začnejo že v končnih stanjih $\mathcal{A}_1$. Primer te




Slika 8: Avtomat $\mathcal{A}$, ki prepozna jezik $L$.




Slika 9: Avtomat $\mathcal{A}_Z$, ki prepozna jezik $L^*$.

konstrukcije za končna avtomata s slike 10 vidimo na sliki 11. Avtomat $\mathcal{A}_P$, ki prepozna zmožek jezikov $L_1 \cdot L_2$, je tako

$$\mathcal{A}_P = (A_1 \cup A_2, Q', E', I', F'),$$

kjer so

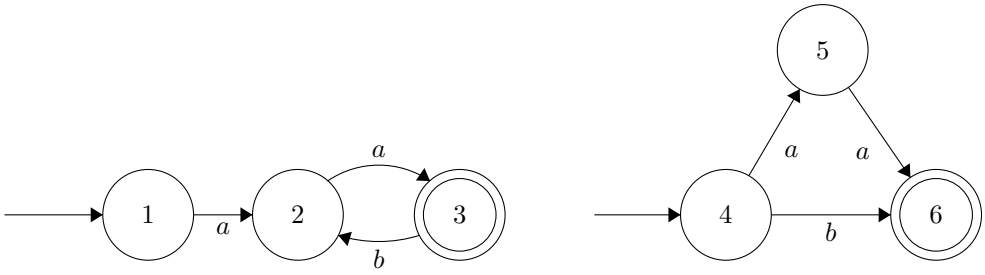
- $Q' = Q_1 + Q_2 \setminus \{q_-\}$,
- $E' = E_1 + \{(q, a, q') \in E_2 \mid q \neq q_-\}$
  $+ \{(q, a, q') \mid q \in F_1 : \exists (q_-, a, q') \in E_2\}$,
- $I' = I_1$,
- $F' = \begin{cases} F_2 & ; q_- \notin F_2 \\ F_1 + F_2 \setminus \{q_-\} & ; q_- \in F_2 \end{cases}$.

## 5.5 Kvocient

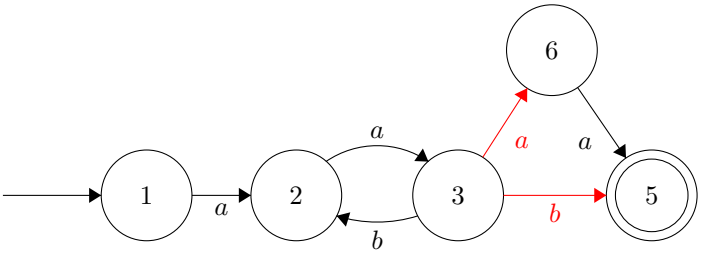
Naj deterministično standarden avtomat $\mathcal{A} = (A, Q, E, q_-, F)$ prepozna jezik $L$ in naj bo $u \in A^*$ poljubna beseda. Zgraditi želimo avtomat, ki prepozna jezik $u^{-1}L$. V avtomatu $\mathcal{A}$ moramo obiti del, ki prepozna besedo $u$, ko se pojavi na začetku. To naredimo tako, da začetno stanje avtomata $\mathcal{A}$ prestavimo v $q_- \cdot u$. Velja torej, da avtomat

$$\mathcal{A}_K = (A, Q, E, q_- \cdot u, F)$$

prepozna jezik $u^{-1}L$.




Slika 10: Avtomata, ki prepoznata jezika $L_1$ in $L_2$.




Slika 11: Avtomat $\mathcal{A}_P$, ki prepozna jezik $L_1 \cdot L_2$.

# 6 Sistem enačb jezikov

Pri delu na projektu smo pokazali tudi, da je jezik, ki ga prepozna poljuben avtomat, racionalen jezik. To smo storili tako, da smo definirali nekaj jezikov, ki so odvisni le od avtomata. Z njimi smo sestavili sistem enačb, katerega rešitev je enolična. Ta rešitev nam da jezik, ki ga avtomat prepozna. Ker je rešitev sestavljena iz končnih operacij množenja, unije in zvezdice, je ta jezik racionalen.

# 7 Zaključek

Ogledali smo si, kaj so formalni jeziki in definirali racionalne jezike. Spoznali smo končne avtomate in njihove različne lastnosti. Osrednji del projekta smo namenili premisleku, da so racionalni jeziki natanko jeziki, ki jih prepozna nek končni avtomat.

# Literatura

[1] J.-E. Pin, *Mathematical foundations of automata theory*, version of February 18, 2022

[2] Zapiski predavanj predmeta izbrana poglavja iz diskretne matematike: končni avtomati, profesorice dr. Ganne Kudryavtseve (Univerza v Ljubljani, Fakulteta za matematiko in fiziko, študijsko leto 2022/2023)