

Sissejuhatus teoreetilisse informaatikasse

Kevad 2022

7. Regulaaravaldised. Mitteregulaarsed keeled

Regulaaravaldise mõiste

Eespool, 5. loengus defineerisime tehted keeltega: ühend \cup , konkatenatsioon \circ ja tärnoperatsioon $*$. Nagu algebraliste tehete abil moodustatakse algebralisi avaldisi, saab nende tehete abil moodustada keelavaldisi, näiteks $(L_1 \cup L_2) \circ L_1^*$.

Olgu keelteks $L_1 = \{0\}$ ja $L_2 = \{1\}$. Tähistame keelt L_1 lihtsalt sümbooliga 0 ja keelt L_2 sümbooliga 1. Avaldise $(L_1 \cup L_2) \circ L_1^*$ võime siis kirja panna kujul $(0 \cup 1)0^*$. Selliseid avaldisi nimetatakse regulaaravaldisteks.

Regulaaravaldise üldine definitsioon on järgmine. Olgu Σ lõplik tähestik.

Definitsioon. *Avaldist R nimetame regulaaravaldiseks, kui R on*

1. *a mingi $a \in \Sigma$ korral;*
2. *tühi sõne ε ;*
3. *tühi hulk \emptyset ;*
4. *$(R_1 \cup R_2)$, kus R_1 ja R_2 on regulaaravaldised;*
5. *$(R_1 \circ R_2)$, kus R_1 ja R_2 on regulaaravaldised;*
6. *R_1^* , kus R_1 on regulaaravaldis.*

See on *induktiivne definitsioon*: punktid 1–3 defineerivad kõige lihtsamad regulaaravaldised ning punktid 4–6 annavad reeglid, mille järgi saab juba olemasolevatest regulaaravaldistest järk-järgult konstrueerida uusi.

Meil läheb veel sagedasti vaja veel tehet $+$, mille defineerime järgmiselt: iga regulaaravaldise R puhul $R^+ = R \circ R^*$. Avaldistes sulgude vähendamiseks on kokku lepitud tehete prioriteedijärjekord: kõige kõrgema prioriteediga on $*$ ja $+$, järgmise prioriteediga \circ ja kõige madalama prioriteediga \cup . Kui ei teki segadust, siis võime lühiduse mõttes jätta avaldistes tehemärgi \circ kirjutamata.

Iga regulaaravaldis tähistab ehk *kirjeldab* mingit keelt (sõnede hulka). Avaldis a kirjeldab keelt $\{a\}$, avaldis ε keelt $\{\varepsilon\}$ ja avaldis \emptyset keelt \emptyset . Tehted \cup , \circ , $*$ kanduvad regulaaravaldistelt loomulikul viisil üle vastavateks teheteks keeltega. Regulaaravaldise R poolt kirjeldatavat keelt märgitakse tähisega $L(R)$. Kahte regulaaravaldist loeme võrdseks, kui nad kirjeldavad sama keelt. Näide 1. Vaatleme mitmesuguseid tähestikus $\Sigma = \{0, 1\}$ määratud regulaaravaldisi.

- Regulaaravaldis $(0\cup 1)^*$ kirjeldab kõigi kahendsõnede hulka (kaasa arvatud tühisõne):

$$L((0\cup 1)^*) = (\{0\} \cup \{1\})^* = \{0, 1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \dots\}.$$

Seda regulaaravaldist tähistatakse tihti ka tähisega Σ^* .

- Regulaaravaldis $(01^*)\cup(10^*)$ kirjeldab kõigi selliste kahendsõnede hulka, kus esimene sümbol erineb kõigist järgnevatest.
- Regulaaravaldis $(01^+)^*$ kirjeldab kõigi selliste kahendsõnede hulka, milles igale 0-le järgneb vähemalt üks 1.
- Regulaaravaldis $(0\Sigma^*0)\cup(1\Sigma^*1)\cup 0\cup 1$ kirjeldab kõigi selliste kahendsõnede hulka, mis algavad ja lõpevad sama sümboliga.
- Regulaaravaldis $(0\cup\varepsilon)(1\cup\varepsilon)$ kirjeldab keelt $\{\varepsilon, 0, 1, 01\}$.

Näide 2. Järgnevas on esitatud mõned lihtsad regulaaravaldiste omadused, mis kehtivad igasuguse regulaaravaldise R jaoks.

- $R\cup\emptyset = R$, sest iga keele ühend tühja keelega on keel ise.
- $R\circ\emptyset = \emptyset$, sest iga keele konkatenatsioon tühja keelega on tühi keel.
- $R\circ\varepsilon = R$, sest keele sõnedele tühja sõne lõppu kirjutamine sõnesid ei muuda.
- $\emptyset^* = \varepsilon$.
- $R^* = R^+\cup\varepsilon$, sest vasak pool R^* on regulaaravaldise R konkatenatsioon iseendaga null või rohkem korda, millest konkatenatsioon null korda on ε ja konkatenatsioon vähemalt üks kord R^+ .

Regulaaravaldiste seos regulaarsete keeltega

Järgnevas tõestame, et regulaaravaldisega kirjeldatavate keelte klass langeb parajasti kokku regulaarsete keelte klassiga. Tõestuse esitame kahe lemmana: et iga regulaaravaldisega kirjeldatav keel on regulaarne ja et iga regulaarne keel on kirjeldatav mingi regulaaravaldisega.

Lemma. *Kui keelt saab kirjeldada regulaaravaldisega, siis see keel on regulaarne.*

Tõestus. Olgu antud keel ja seda kirjeldav regulaaravaldis R . Teisendame regulaaravaldise R mittedetermineeritud lõplikuks automaadiks N , mis antud keelt ära tunneb.

1. Kui $R = a$, siis $L(R) = \{a\}$. Automaat N on



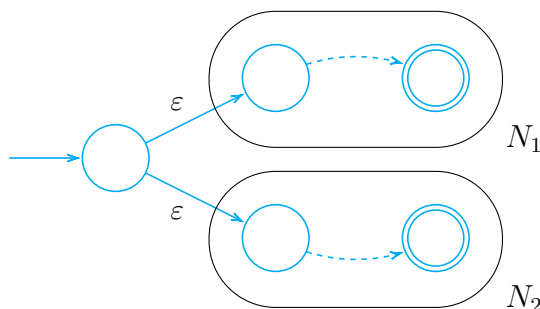
2. Kui $R = \varepsilon$, siis $L(R) = \{\varepsilon\}$. Automaat N on



3. Kui $R = \emptyset$, siis $L(R) = \emptyset$. Automaat N on



4. Kui $R = R_1 \cup R_2$, siis olgu N_1 ja N_2 mittedetermineeritud lõplikud automaadid, mis vastavalt tunnevad ära keeli $L(R_1)$ ja $L(R_2)$. Automaat N on



5. Kui $R = R_1 \circ R_2$, siis olgu N_1 ja N_2 vastavalt keeli $L(R_1)$ ja $L(R_2)$ äratundvad automaadid. Automaat N on selline nagu eelmise nädala praktikumiülesandes 2.
6. Kui $R = R_1^*$, siis olgu N_1 keelt $L(R_1)$ äratundev automaat. Automaat N on selline nagu eelmise nädala praktikumiülesandes 3.

Kõigil juhtudel aktsepteerib automaat N parajasti keelt $L(R)$. Korrates neid samme, saame järk-järgult üles ehitada sobiva mittedetermineeritud lõpliku automaadi. \square

Lemma. *Kui keel on regulaarne, siis seda keelt saab kirjeldada regulaaravaldisega.*

Selle lemma tõestus on antud lisas.

Keele mitteregulaarsus

Tuletame meelde, et keel on regulaarne parajasti siis, kui leidub lõplik automaat, mis seda keelt ära tunneb. Vaatleme keelt

$$B = \{0^n 1^n \mid n \geq 0\}.$$

Kas see keel on regulaarne ehk kas leidub lõplik automaat, mis tunneb ära keelt B ?

Vastus on ei. Intuitiivselt selgitades peaks selline automaat sõne lugemise käigus „meeles pidama“, mitut nulli on ta juba näinud. Sedalaadi teavet saab automaat säilitada ainult olekute abil: minnes iga järgmist nulli kohates uude olekusse, kus ta varem pole viibinud. Kuid et sõned võivad olla kui tahes pikad, vajaks automaat keele kõigi sõnede äratundmiseks lõpmata palju olekuid ega oleks siis enam lõplik automaat.

See muidugi ei ole range tõestus, miks niisugust automaati ei leidu, kuid heidab mõningat valgust probleemi olemusele. Teine näide keelest, mida äratundvat automaati samuti ei leidu, on

$$C = \{w \mid \text{sõnes } w \text{ on võrdne arv sümboleid } 0 \text{ ja } 1\}.$$

Pumpamislemma

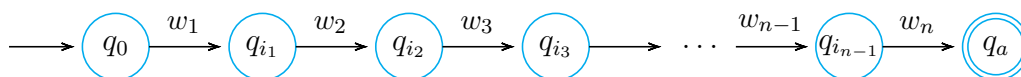
Igal regulaarsel keelel on üks teatav omadus: iga piisavalt pikk keele sõne sisaldab alamsõnet, mida võib selles sõnes korrata järjest ükskõik mitu korda nii, et saadav sõne jääb ikka keelde kuuluvaks. Seega kui mõnel keelel see omadus puudub, siis see keel regulaarne ei ole. Täpsemalt väljendab seda omadust järgmine lemma.

Pumpamislemma. Kui L on regulaarne keel, siis leidub selline positiivne täisarv p , et iga sõne $w \in L$, mille puhul $|w| \geq p$, esitub kujul $w = xyz$, kus täidetud on järgmised kolm tingimust:

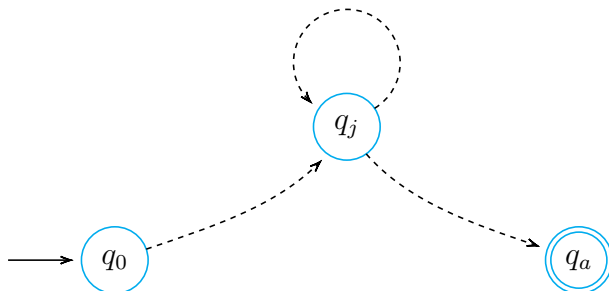
1. iga $i \geq 0$ korral $xy^iz \in L$;
2. $|y| > 0$;
3. $|xy| \leq p$.

Kirjutis $|w|$ tähistab siin sõne w pikkust. Esituses $w = xyz$ võib x või z (või mõlemad) olla tühisõne ε , kuid y on kindlasti mittetühi.

Tõestuse idee. Olgu p keelt äratundva automaadi olekute arv. Vaatleme selle automaadi töötamist sisendsõnel w :



Kui $n \geq p$, siis läbib automaat vähemalt $p + 1$ olekut. Esimese $p + 1$ oleku hulgas peab mingi olek q_j esinema vähemalt kaks korda. Selle oleku kahe esinemise vahel asuvat alamsõnet võib korrata ükskõik milline arv kordi.



Tõestus. Olgu $M = (Q, \Sigma, \delta, q_0, F)$ determineeritud lõplik automaat, mis tunneb ära keelt L , ja p automaadi M olekute arv. Olgu $w = w_1w_2 \dots w_n$ suvaline sõne keeles L , kusjuures $n \geq p$. Olgu $q_0 = q_{i_0} \rightarrow q_{i_1} \rightarrow q_{i_2} \rightarrow \dots \rightarrow q_{i_{n-1}} \rightarrow q_{i_n} = q_a$ olekute järjend, mille automaat M sõnet w lugedes läbib, st milles iga $j = 0, 1, \dots, n - 1$ puhul $\delta(q_{i_j}, w_j) = q_{i_{j+1}}$.

Selle olekute järjendi pikkus on $n + 1 \geq p + 1$. Et automaadil M on ainult p olekut, siis leidub esimese $p + 1$ oleku $q_0, q_{i_1}, \dots, q_{i_p}$ hulgas olek q_j , mis esineb seal vähemalt kaks korda:

$$q_0 = q_{i_0} \xrightarrow{w_1} \dots \xrightarrow{w_l} q_{i_l} = q_j \xrightarrow{w_{l+1}} \dots \xrightarrow{w_r} q_{i_r} = q_j \xrightarrow{w_{r+1}} \dots \xrightarrow{w_n} q_{i_n} = q_a.$$

Oleku q_j esimene esinemine olgu q_{i_l} ja teine esinemine q_{i_r} . Tähistame

$$\begin{aligned} x &= w_1w_2 \dots w_l, \\ y &= w_{l+1}w_{l+2} \dots w_r, \\ z &= w_{r+1}w_{r+2} \dots w_n. \end{aligned}$$

Alamsõne x viib automaadi M olekust q_0 olekusse q_j , alamsõne y (ja järelikult ka iga sõne y^i) viib automaadi M olekust q_j olekusse q_j ning alamsõne z viib automaadi M olekust q_j olekusse q_a . Seega aktsepteerib automaat M igasugust sõnet kujul xy^iz , kus $i \geq 0$. See tähendab, tingimus 1 on täidetud. Et $q_{i_l} \neq q_{i_r}$, siis $|y| > 0$ ehk tingimus 2 on samuti täidetud. Lõpuks, $r \leq p$, seega $|xy| \leq p$, mis tähendab, et ka tingimus 3 on täidetud. \square

Näide 3. Tõestame, et keel

$$B = \{0^n 1^n \mid n \geq 0\}$$

ei ole regulaarne.

Oletame väitevastaselt, et B on regulaarne. Olgu p pumpamislemmast saadav pikkus. Vaatleme sõnet $w = 0^p 1^p \in B$. Et $|w| \geq p$, siis võime kirjutada $w = xyz$, kusjuures iga $i \geq 0$ puhul $xy^iz \in B$.

- Kui y koosneb ainult nullidest, siis $xyyz \notin B$, sest selles sõnes on nulle rohkem kui ühtesid.
- Kui y koosneb ainult ühtedest, siis $xyyz \notin B$, sest selles sõnes on ühtesid rohkem kui nulle.
- Kui y koosneb nii nullidest kui ka ühtedest, siis $xyyz \notin B$, sest selles sõnes on nullid ja ühed segamini: esimene y lõpeb kindlasti ühega ja teine y algab kindlasti nulliga.

Kõigil kolmel juhul $xyyz \notin B$, mis on vastuolus pumpamislemmaga. Seega oletus, et B on regulaarne, ei kehti.

Praktikumiülesanded

1. Mis keelt kirjeldavad järgmised regulaaravaldised?

- (a) $(0 \cup \varepsilon)^*(1 \cup \varepsilon)$
- (b) $0\Sigma^*1$
- (c) $0\emptyset 10^*$
- (d) $0\varepsilon 10^*$

2. Kas lause on tõene või väär?

- (a) $R \cup \varepsilon = R$

(b) $R \circ \varepsilon = R$

(c) $R \cup \emptyset = R$

(d) $R \circ \emptyset = R$

3. Leida mittedetermineeritud lõplik automaat, mis tunneb ära keelt, mida kirjeldab regulaaravaldis $(0 \cup 1)^*010$.
4. Tõestada, et keel

$$L = \{w \mid \text{sõnes } w \text{ on võrdne arv sümboleid } 0 \text{ ja } 1\}$$

ei ole regulaarne.

5. Tõestada, et keel $L = \{ss \mid s \in \{0, 1\}^*\}$ ei ole regulaarne.

Lahendused

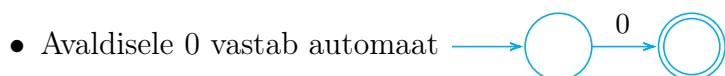
1. Lahendus.

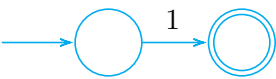
- (a) Kõik sõned, mille alguses on teatud arv (võib olla ka null) sümboleid 0 ja nende järel teatud arv (võib olla ka null) sümboleid 1.
- (b) Kõik sõned, mis algavad 0-ga ja lõpevad 1-ga. Esimese ja viimase sümboli vahel võib olla suvaline sõne (sealhulgas ka tühisõne).
- (c) Tühi hulk, sest tühja hulga konkatenatsioon ükskõik millega on tühi hulk.
- (d) Kõik sõned, mille alguses olevale 01-le järgneb teatud arv nulle.

2. Lahendus.

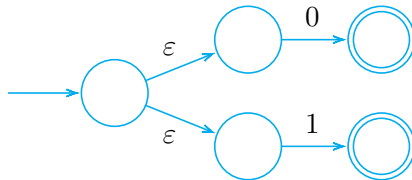
- (a) Väär. Näiteks kui $R = 0$, siis $L(R \cup \varepsilon) = \{0, \varepsilon\}$, aga $L(R) = \{0\}$.
- (b) Tõene. Sõnede lõppu sümboli ε lisamine sõnesid ei muuda.
- (c) Tõene. Keelele tühja hulga lisamine keelt ei muuda.
- (d) Väär. Näiteks kui $R = 0$, siis $L(R \circ \emptyset) = \emptyset$, aga $L(R) = \{0\}$.

3. Lahendus. Konstrueerime selle automaadi sammukaupa loengu esimese lehma tõestuse kohaselt.

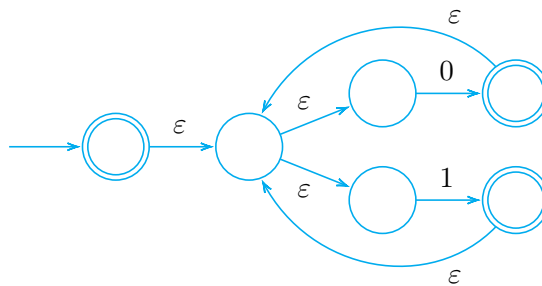


- Avaldisele 1 vastab automaat 

- Avaldisele $0 \cup 1$ vastab automaat



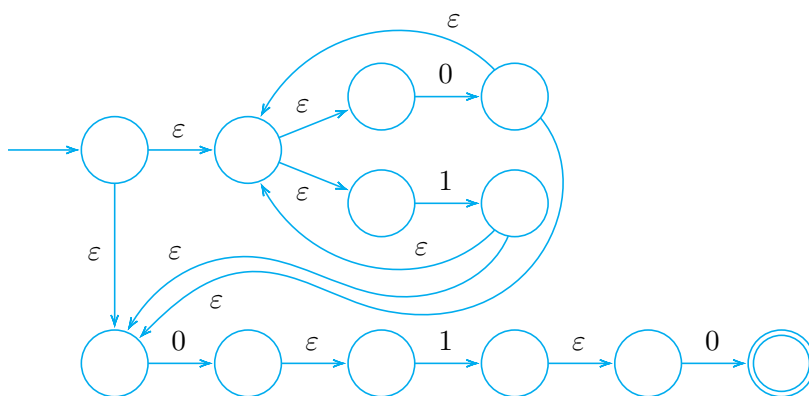
- Avaldisele $(0 \cup 1)^*$ vastab automaat



- Avaldisele 010 vastab automaat



- Avaldisele $(0 \cup 1)^*010$ vastab automaat



4. *Lahendus.* Oletame väitevastaselt, et L on regulaarne. Olgu p pumpamislemmast saadav „pumpamis pikkus“. Vaatleme sõnet $w = 0^p 1^p \in L$, mille puhul $|w| \geq p$. Pumpamislemma põhjal saab selle sõne esitada kujul $w = xyz$ nii, et lemma tingimused 1–3 on täidetud.

Et tingimuse 3 põhjal $|xy| \leq p$, siis koosneb y ainult nullidest. Nüüd ühelt poolt tingimuse 1 põhjal $xyyz \in L$, teiselt poolt aga sisaldab $xyyz$ nulle rohkem kui ühtesid, mistõttu $xyyz \notin L$. Vastuolu. Järelikult oletus, et keel L on regulaarne, ei pea paika.

5. *Lahendus.* Oletame väitevastaselt, et L on regulaarne. Olgu p „pumpamis pikkus“. Võtame $w = 0^p 10^p 1 \in L$. Ilmselt $|w| \geq p$. Seega leiduvad sõned x, y, z nii, et $w = xyz$ ja kehtivad pumpamislemma tingimused. Kuna $|xy| \leq p$, siis koosneb y ainult nullidest. Järelikult sõne $xyyz$ ei ole kujul ss , sest üks nullide plokk on pikem kui teine. Seega $xyyz \notin L$. Vastuolu lemma tingimusega 1. Järelikult oletus, et L on regulaarne, oli väär.

Lisa (iseõppimiseks)

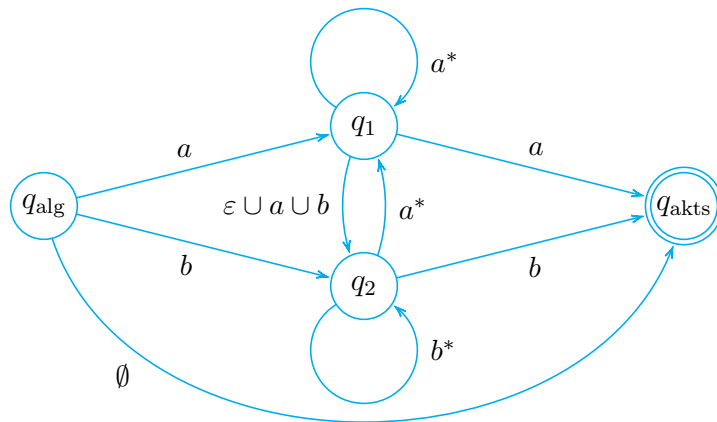
Lemma. *Kui keel on regulaarne, siis seda keelt saab kirjeldada regulaaravaldisega.*

Allolevas tõestuses kasutame uut automaatide tüüpi, mille nimi on üldistatud mittedetermineeritud lõplik automaat (GNFA). See on mittedetermineeritud lõplik automaat, mille kaarte märgenditeks võivad olla üldised regulaaravaldised. Selline automaat loeb sisendist sümboleid plokikaupa, mitte enam ühekaupa, ja siirdub ühest olekust teise, kui sisendplokk on sõne, mida kirjeldab vastaval kaarel asuv regulaaravaldis.

Üldistatud mittedetermineeritud lõplikul automaadil on järgmine kuju.

- Algolekusse ei sisene ühtegi kaart ning sealt väljuvad kaared kõikidesse teistesse olekutesse.
- Aktsepteerivast olekust ei välju ühtegi kaart ning sinna sisenevad kaared kõigist teistest olekutest.
- Ülejäänud olekute puhul peale algoleku ja aktsepteeriva oleku on olemas kaared igast olekust igasse olekusse, kaasa arvatud kaared igast olekust iseendasse.

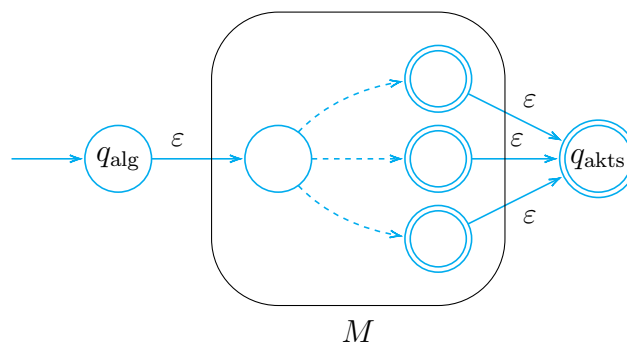
Näiteks võib üldistatud mittedetermineeritud lõplik automaat olla selline (lihtsuse mõttes jätame joonistamata kaared, mille märgend on \emptyset):



Tõestus. Kui keel on regulaarne, siis leidub determineeritud lõplik automaat M , mis seda keelt ära tunneb.

1) Teisendame kõigepealt automaadi M üldistatud mittedetermineeritud lõplikuks automaadiks.

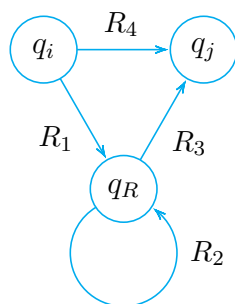
- Lisame uue algoleku ja uue aktsepteeriva oleku.
- Tõmbame uuest algolekust ε -kaare vanasse algolekusse.
- Tõmbame igast vanast aktsepteerivast olekust ε -kaare uude aktsepteerivasse olekusse.
- Kui leidub paralleelseid kaari, st kaari, millel on mitu märgendit, siis asendame iga sellise ühe kaarega, mille märgendiks on seniste märgendite ühend (regulaaravaldis).
- Kui mingist olekust mingisse olekusse kaart ei vii, siis lisame sinna kaare märgendiga \emptyset .



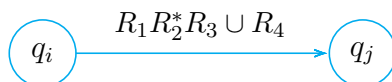
Niimoodi saadud üldistatud mittedetermineeritud automaat tunneb ära sedasama keelt.

2) Edasi teisendame selle üldistatud mittedetermineeritud lõpliku automaadi regulaaravaldiseks.

- Kui automaadil on ainult algolek ja aktsepteeriv olek, siis on tal üksainus kaar, mille märgend ongi vajalik regulaaravaldis.
- Kui automaadil leidub lisaks neile veel mõni olek q_R , siis konstrueerime selle automaadiga ekvivalentse automaadi, millel seda olekut pole. Selleks kustutame automaadist oleku q_R ja muudame kõigi järelejäävate kaarte märgendeid. Vaatleme suvalisi allesjäävaid olekuid q_i ja q_j . Eeldame, et kaarte märgendid on järgmised: $q_i \xrightarrow{R_1} q_R$, $q_R \xrightarrow{R_2} q_R$, $q_R \xrightarrow{R_3} q_j$, $q_i \xrightarrow{R_4} q_j$:



Siis uues automaadis viib olekust q_i olekusse q_j kaar $q_i \rightarrow q_j$ märgendiga $R_1 R_2^* R_3 \cup R_4$:



Selline märgendite asendamine ei mõjuta aktsepteeriva arvutustee olemasolu ja seega jätab aktsepteeritavate sõnede hulga samaks. \square

Definitsioon. Üldistatud mittedetermineeritud lõplik automaat on viisik $(Q, \Sigma, \delta, q_{alg}, q_{akts})$, kus

- Q on lõplik olekute hulk;
- Σ on lõplik tähestik;
- $\delta: (Q \setminus \{q_{akts}\}) \times (Q \setminus \{q_{alg}\}) \rightarrow \mathcal{R}$ on üleminekufunktsioon, kus \mathcal{R} on kõigi regulaaravaldiste hulk;
- $q_{alg} \in Q$ on algolek;
- $q_{akts} \in Q$ on aktsepteeriv olek.

Definitsioon. Olgu $M = (Q, \Sigma, \delta, q_{alg}, q_{akts})$ üldistatud mittedetermineeritud lõplik automaat ja w sõne tähestikus Σ . Ütleme, et automaat M aktsepteerib sõnet w , kui sõne saab esitada kujul $w = w_1 w_2 \dots w_k$, kus $w_i \in \Sigma^*$, ja leidub olekute järjend q_0, q_1, \dots, q_k , et

1. $q_0 = q_{alg}$;
2. iga $i = 1, 2, \dots, k$ puhul $w_i \in L(R_i)$, kus $R_i = \delta(q_{i-1}, q_i)$;
3. $q_k = q_{akts}$.

Algoritm üldistatud mittedetermineeritud lõpliku automaadi teisendamiseks regulaaravaldiseks

1. Olgu k etteantud automaadi G olekute arv.
2. Kui $k = 2$, siis G koosneb algolekust, aktsepteerivast olekust ja neid ühendavast kaarest, mille märgend on regulaaravaldis R . Tagasta R .
3. Kui $k > 2$, siis vali suvaline olek $q_R \in Q$, $q_R \neq q_{alg}$, $q_R \neq q_{akts}$. Konstrueeri üldistatud lõplik automaat $G' = (Q', \Sigma, \delta', q_{alg}, q_{akts})$, kus

(a) $Q' = Q \setminus \{q_R\}$;

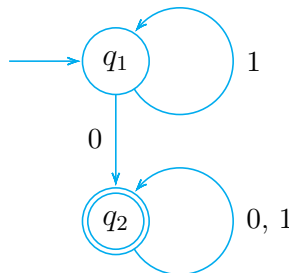
(b) iga $q_i \in Q' \setminus \{q_{akts}\}$, $q_j \in Q' \setminus \{q_{alg}\}$ korral

$$\delta'(q_i, q_j) = R_1 R_2^* R_3 \cup R_4,$$

kus $R_1 = \delta(q_i, q_R)$, $R_2 = \delta(q_R, q_R)$, $R_3 = \delta(q_R, q_j)$, $R_4 = \delta(q_i, q_j)$.

4. Rakenda algoritmi rekursiivselt automaadile G' ja tagasta tulemus.

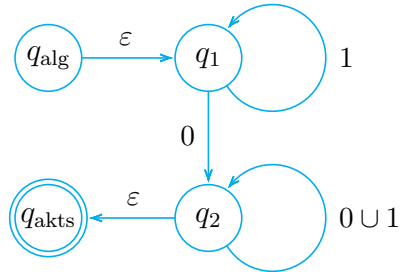
6. Antud on järgmine determineeritud lõplik automaat:



Milline on sellega samaväärne regulaaravaldis?

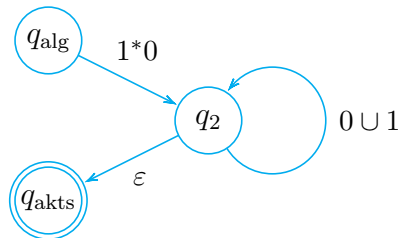
Lahendus. Leiame selle regulaaravaldise vastavalt loengu teise lemma tõestusele ja eetoodud algoritmile.

1. Lisame uue algoleku ja uue aktsepteeriva oleku:

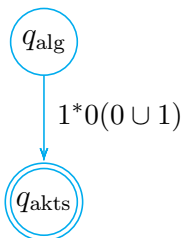


Märkus: selguse mõttes jätame kaared märgendiga \emptyset joonistamata.

2. Eemaldame oleku q_1 :

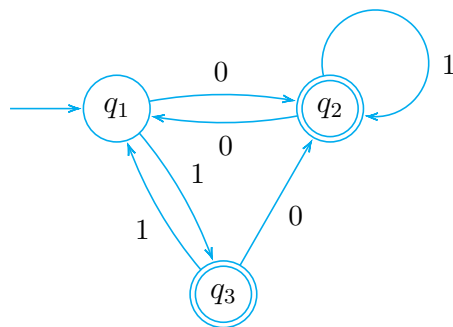


3. Eemaldame oleku q_2 :



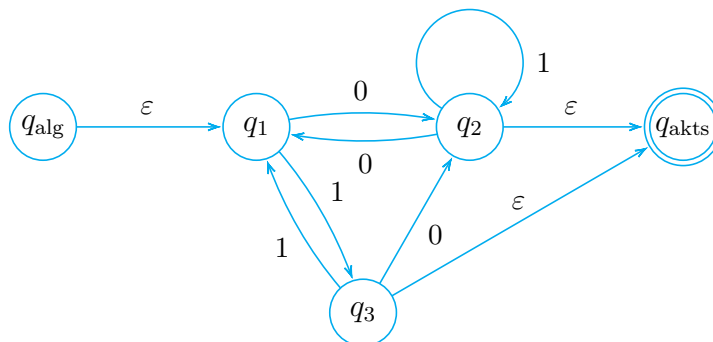
4. Tulemuseks saame regulaaravaldise $1^*0(0 \cup 1)^*$.

7. Leida järgmise determineeritud lõpliku automaadiga samaväärne regulaaravaldis:

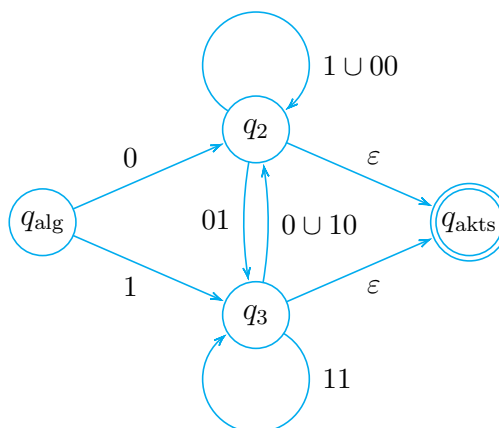


Lahendus.

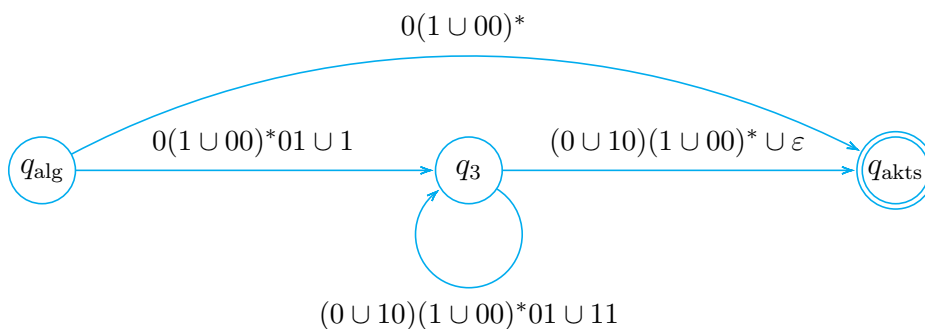
1. Lisame algoleku ja aktsepteeriva oleku:



2. Eemaldame oleku q_1 :



3. Eemaldame oleku q_2 :



4. Eemaldame oleku q_3 :

