2.6 SÄÄNNÖLLISET LAUSEKKEET

Automaattimalleista poikkeava tapa kuvata yksinkertaisia kieliä.

Olkoot A ja B aakkoston Σ kieliä. Perusoperaatioita:

- *Yhdiste:* $A \cup B = \{x \in \Sigma^* \mid x \in A \text{ tai } x \in B\}$;
- *Katenaatio:* $AB = \{xy \in \Sigma^* \mid x \in A,\ y \in B\}$;
- *Potenssit:*

$$\begin{cases} A^0 = \{\varepsilon\}, \\ A^k = AA^{k-1} = \{x_1 \dots x_k \mid x_i \in A \quad \forall i = 1, \dots, k\} \qquad (k \geq 1); \end{cases}$$

- *Sulkeuma t. “Kleenen tähti”:*

$$\begin{aligned} A^* &= \bigcup_{k \geq 0} A^k \\ &= \{x_1 \dots x_k \mid k \geq 0,\ x_i \in A \quad \forall i = 1, \dots, k\}. \end{aligned}$$

Määritelmä 2.3 Aakkoston Σ *säännölliset lausekkeet* määritellään induktiivisesti säännöillä:

1. $\emptyset$ ja ε ovat Σ:n säännöllisiä lausekkeita;
2. a on Σ:n säännöllinen lauseke kaikilla $a \in \Sigma$;
3. jos r ja s ovat Σ:n säännöllisiä lausekkeita, niin $(r \cup s)$, (rs) ja r^* ovat Σ:n säännöllisiä lausekkeita;
4. muita Σ:n säännöllisiä lausekkeita ei ole.

Kukin Σ:n säännöllinen lauseke r *kuvaa* kielen $L(r)$, joka määritellään:

- $L(\emptyset) = \emptyset$;
- $L(\varepsilon) = \{\varepsilon\}$;
- $L(a) = \{a\}$ kaikilla $a \in \Sigma$;
- $L((r \cup s)) = L(r) \cup L(s)$;
- $L((rs)) = L(r)L(s)$;
- $L(r^*) = (L(r))^*$.

Aakkoston $\{a, b\}$ säännöllisiä lausekkeita:

$$r_1 = ((ab)b), \quad r_2 = (ab)^*,$$

$$r_3 = (ab^*), \quad r_4 = (a(b \cup (bb)))^*.$$

Lausekkeiden kuvaamat kielet:

$$\begin{aligned} L(r_1) &= (\{a\}\{b\})\{b\} = \{ab\}\{b\} = \{abb\}; \\ L(r_2) &= \{ab\}^* = \{\varepsilon, ab, abab, ababab, \dots\} \\ &= \{(ab)^i \mid i \geq 0\}; \\ L(r_3) &= \{a\}(\{b\})^* = \{a, ab, abb, abbb, \dots\} \\ &= \{ab^i \mid i \geq 0\}; \\ L(r_4) &= (\{a\}\{b, bb\})^* = \{ab, abb\}^* \\ &= \{\varepsilon, ab, abb, abab, ababb, \dots\} \\ &= \{x \in \{a,b\}^* \mid \text{ kutakin } a\text{-kirjainta } x\text{:ssä} \\ &\qquad \text{seuraa 1 tai 2 } b\text{-kirjainta } \}. \end{aligned}$$

Sulkumerkkien vähentämissääntöjä:

Operaattoreiden prioriteetti:

$$* \quad \succ \quad \cdot \quad \succ \quad \cup$$

Yhdiste- ja tulo-operaatioiden assosiatiivisuus:

$$\begin{aligned} L(((r \cup s) \cup t)) &= L((r \cup (s \cup t))) \\ L(((rs)t)) &= L((r(st))) \end{aligned}$$

$\Rightarrow$ peräkkäisiä yhdisteitä ja tuloja ei tarvitse suluttaa.

Käytetään tavallisia kirjasimia, mikäli sekaannuksen vaaraa merkkijonoihin ei ole.

Yksinkertaisemmin siis:

$$r_1 = abb, \quad r_2 = (ab)^*, \quad r_3 = ab^*, \quad r_4 = (a(b \cup bb))^*.$$

Esimerkki: C-kielen etumerkittömät reaaliluvut

$$number = (dd^*.d^* \cup .dd^*)(e(+ \cup - \cup \varepsilon)dd^* \cup \varepsilon) \cup (dd^*e(+ \cup - \cup \varepsilon)dd^*),$$

missä d on lyhennemerkintä lausekkeelle

$$d = (\mathtt{0} \cup \mathtt{1} \cup \mathtt{2} \cup \mathtt{3} \cup \mathtt{4} \cup \mathtt{5} \cup \mathtt{6} \cup \mathtt{7} \cup \mathtt{8} \cup \mathtt{9})$$

ja e on lyhennemerkintä lausekkeelle

$$e = (\mathtt{E} \cup \mathtt{e}).$$

Usein merkitään myös lyhyesti $rr^* \equiv r^+$. Esim.:

$$(d^+.d^* \cup .d^+)(e(+ \cup - \cup \varepsilon)d^+ \cup \varepsilon) \cup (d^+e(+ \cup - \cup \varepsilon)d^+).$$

Määritelmä 2.4 Kieli on *säännöllinen*, jos se voidaan kuvata säännöllisellä lausekkeella.

Säännöllisten lausekkeiden sieventäminen

Säännöllisillä kielillä on yleensä useita vaihtoehtoisia kuvauksia, esim.:

$$\begin{aligned} \Sigma^* &= L((a \cup b)^*) \\ &= L((a^*b^*)^*) \\ &= L(a^*b^* \cup (a \cup b)^*ba(a \cup b)^*). \end{aligned}$$

Määritelmä. Säännölliset lausekkeet r ja s ovat *ekvivalentit*, merk. $r = s$, jos $L(r) = L(s)$.

Lausekkeen sieventäminen = “yksinkertaisimman” ekvivalentin lausekkeen määrittäminen.

Säännöllisten lausekkeiden ekvivalenssitestaus on epätriviaali, mutta periaatteessa mekaanisesti ratkeava ongelma.

Sievennyssääntöjä:

$$
\begin{aligned}
r \cup (s \cup t) &= (r \cup s) \cup t \\
r(st) &= (rs)t \\
r \cup s &= s \cup r \\
r(s \cup t) &= rs \cup rt \\
(r \cup s)t &= rt \cup st \\
r \cup r &= r
\end{aligned}
\qquad
\begin{aligned}
r \cup \emptyset &= r \\
\varepsilon r &= r \\
\emptyset r &= \emptyset \\
r^* &= \varepsilon \cup r^* r \\
r^* &= (\varepsilon \cup r)^*
\end{aligned}
$$

Mikä tahansa säännöllisten lausekkeiden tosi ekvivalenssi voidaan johtaa näistä laskulaeista, kun lisätään päättelysääntö: jos $r = rs \cup t$, niin $r = ts^*$, edellyttäen että $\varepsilon \notin L(s)$.

Kahden lausekkeen ekvivalenssin toteamiseksi kannattaa usein päätellä erikseen kummankin kuvaaman kielen sisältyminen toiseen.

Merkitään lyhyesti: $r \subseteq s$, jos $L(r) \subseteq L(s)$.
Tällöin siis $r = s$ joss $r \subseteq s$ ja $s \subseteq r$.

Esimerkki: todetaan, että $(a^*b^*)^* = (a \cup b)^*$.

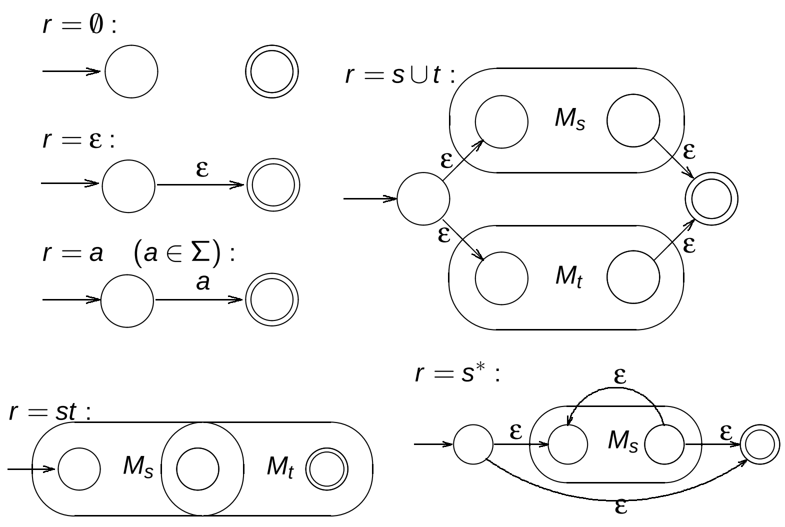
1. Selvästi $(a^*b^*)^* \subseteq (a \cup b)^*$, koska $(a \cup b)^*$ kuvaa *kaikkia* aakkoston $\{a, b\}$ merkkijonoja.
2. Koska selvästi $(a \cup b) \subseteq a^*b^*$, niin myös $(a \cup b)^* \subseteq (a^*b^*)^*$.

2.7 ÄÄRELLISET AUTOMAATIT JA SÄÄNNÖLLISET KIELET

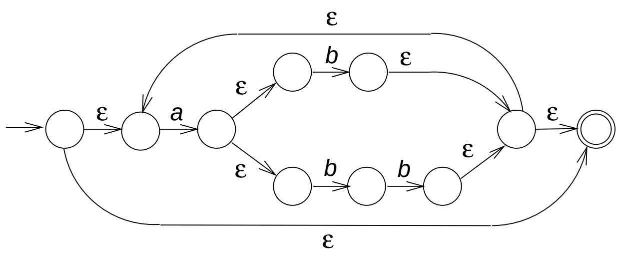
Lause 2.3 Jokainen säännöllinen kieli voidaan tunnistaa äärellisellä automaatilla.

Todistus. Seuraavan kalvon induktiivisen konstruktion avulla voidaan mielivaltaisen säännöllisen lausekkeen r rakennetta seuraten muodostaa ε-automaatti M_r, jolla $L(M_r) = L(r)$. Tästä automaatista voidaan poistaa ε-siirtymät Lemman 2.4 mukaisesti, ja tarvittaessa voidaan syntyvä epädeterministinen automaatti determinisoida Lauseen 2.2 konstruktiolla.

Esitettävästä konstruktiosta on syytä huomata, että muodostettaviin ε-automaatteihin tulee aina yksikäsitteiset alku- ja lopputila, eikä minkään osa-automaatin lopputilasta lähde eikä alkutilaan tule yhtään ko. osa-automaatin sisäistä siirtymää. □

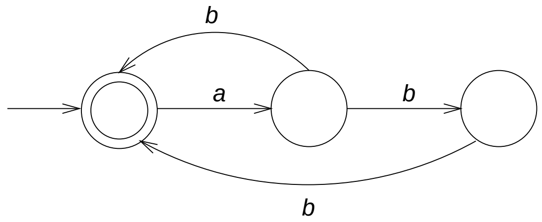


Esimerkiksi lausekkeesta $r = (a(b \cup bb))^*$ saadaan näiden sääntöjen mukaan seuraava ε-automaatti:



Automaatti on selvästi hyvin redundantti. Käsin automaatteja suunniteltaessa ne kannattaakin usein muodostaa suoraan.

Esim. lausekkeen $r = (a(b \cup bb))^*$ perusteella on helppo muodostaa seuraava yksinkertainen epädeterministinen tunnistaja-automaatti:



Lause 2.4 Jokainen äärellisellä automaatilla tunnistettava kieli on säännöllinen.

Todistus. Tarvitaan vielä yksi äärellisten automaattien laajennus: *lausekeautomaatissa* voidaan siirtymien ehtoina käyttää mielivaltaisia säännöllisiä lausekkeita.

Formalisointi: Merk. RE_Σ = aakkoston Σ säännöllisten lausekkeiden joukko. *Lausekeautomaatti* on viisikko

$$M = (Q, \Sigma, \delta, q_0, F),$$

missä siirtymäfunktio δ on äärellinen kuvaus

$$\delta : Q \times \mathrm{RE}_\Sigma \to \mathcal{P}(Q)$$

(so. $\delta(q,r) \neq \emptyset$ vain äärellisen monella parilla $(q,r) \in Q \times \mathrm{RE}_\Sigma$).

Yhden askelen tilannejohto määritellään:

$$(q,w) \underset{M}{\vdash} (q',w')$$

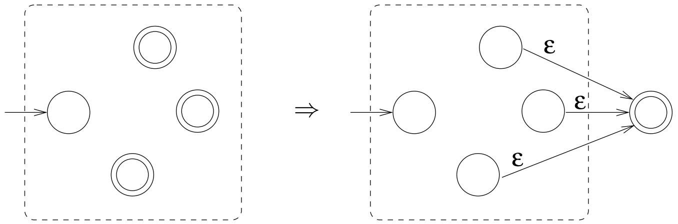
jos on $q' \in \delta(q,r)$ jollakin sellaisella $r \in \mathrm{RE}_\Sigma$, että $w = zw'$, $z \in L(r)$. Muut määritelmät samat kuin aiemmin.

Todistetaan: jokainen lausekeautomaatilla tunnistettava kieli on säännöllinen.

Olkoon M jokin lausekeautomaatti. Säännöllinen lauseke, joka kuvaa M:n tunnistaman kielen, muodostetaan kahdessa vaiheessa:

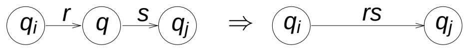
1. Tiivistetään M ekvivalentiksi enintään 2-tilaiseksi lausekeautomaatiksi seuraavilla muunnoksilla:

(i) Jos M:llä on useita lopputiloja, yhdistetään ne seuraavasti.

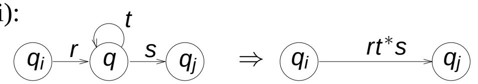


(ii) Poistetaan M:n muut kuin alku- ja lopputila yksi kerrallaan seuraavasti. Olk. q jokin M:n tila, joka ei ole alku- eikä lopputila; tarkastellaan kaikkia “reittejä”, jotka M:ssä kulkevat q:n kautta. Olk. q_i ja q_j q:n välittömät edeltäjä- ja seuraajatila jollakin tällaisella reitillä. Poistetaan q reitiltä $q_i \to q_j$ oheisen kuvan (i) muunnoksella, jos tilasta q ei ole siirtymää itseensä, ja kuvan (ii) muunnoksella, jos tilasta q on siirtymä itseensä:

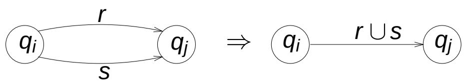
(i):



(ii):

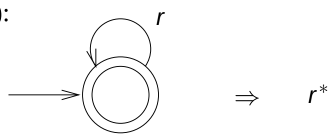


Samalla yhdistetään rinnakkaiset siirtymät seuraavasti:



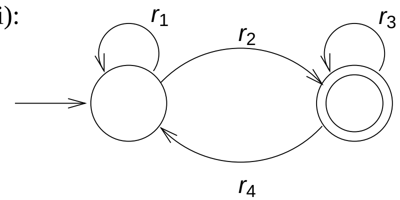
2. Tiivistyksen päättyessä jäljellä olevaa 2-tilaista automaattia vastaava säännöllinen lauseke muodostetaan seuraavan kuvan esittämällä tavalla:

(i):



$\Rightarrow \quad r^*$

(ii):



$\Rightarrow \quad r_1^* r_2 (r_3 \cup r_4 r_1^* r_2)^*$

Esimerkki: