

## Hahmon etsiminen syötteestä (johdatteleva esimerkki)

Unix-komennolla

```
grep hahmo [ tiedosto ]
```

voidaan etsiä hahmon esiintymiä tiedostosta (tai syötevirrasta):

```
$ grep Kisaveikot SM-tulokset.txt  
$ ps aux | grep mmeikala  
$ ps aux | grep '\(mmeikala\|tteikala\)'
```

Tässä siis haetaan

- SM-tuloksista rivit, joilla esiintyy seura Kisaveikot
- prosessilistasta ne rivit, joilla esiintyy käyttäjätunnus mmeikala, ja
- prosessilistasta ne rivit, joilla esiintyy käyttäjätunnus mmeikala tai tteikala.

Eräs idea `grep`-toiminnon toteuttamiseksi olisi seuraava:

1. Muodostetaan äärellinen automaatti, joka hyväksyy tasan sellaiset merkkijonot, joissa esiintyy *hahmo*.
2. **Selataan** syöte rivi kerrallaan käyttämällä tätä automaattia, ja tulostetaan hyväksytyt rivit.

Selausvaihe toimii nopeasti eikä edellytä syötteen esiprosessointia, mikä on tässä tärkeää.

**Kysymys:** Kuinka monimutkaisia hahmoja tällä periaatteella voidaan käsitellä?

Esim. edellä muodostettiin hahmoista **mmeikala** ja **tteikala** uusi hahmo tai-operaattorilla `" | "`. Kuinka voimakkaat operaattorit voidaan siis sallia?

## Operaatiot (säännöllisillä) kielillä [Sipser s. 44–47]

Kuten muistetaan, kielet ovat merkkijonojoukkoja.

Niillä voidaan siis suorittaa normaaleja joukko-opillisia operaatioita, kuten yhdiste, leikkaus ja komplementointi: jos  $A$  ja  $B$  ovat kieliä, niin

- kieli  $A \cup B$  koostuu niistä merkkijonoista, jotka kuuluvat ainakin toiseen kielistä  $A$  ja  $B$ ,
- kieli  $A \cap B$  koostuu niistä merkkijonoista, jotka kuuluvat sekä kieleen  $A$  että  $B$ , ja
- kieli  $\overline{A}$  koostuu niistä merkkijonoista, jotka eivät kuulu kieleen  $A$ .

(Komplementin  $\overline{A}$  määrittelyssä oletetaan, että puhutaan jonkin tietyn aakkoston  $\Sigma$  merkkijonoista, ja  $\Sigma$  selviää asiayhteydestä.)

Nimenomaan merkkijonojoukoille on kätevää määritellä myös konkatenaatio ja tähti:

- kieli  $A \circ B$  koostuu merkkijonoista  $w$ , jotka voidaan esittää muodossa  $w = xy$  joillakin  $x \in A$  ja  $y \in B$  ja
- kieli  $A^*$  koostuu merkkijonoista  $w_1 \dots w_k$ , missä  $k \geq 0$  ja  $w_i \in A$  kaikilla  $i$ .

Intuitiivisesti

$$A^* = A^0 \cup A^1 \cup A^2 \cup A^3 \cup \dots,$$

missä  $A^k$  on kielen  $A$  konkatenaatio itsensä kanssa  $k$  kertaa (ja  $A^0 = \{\varepsilon\}$ ).

**Esimerkki** Tarkastellaan aakkoston  $\{a, \dots, z, 0, \dots, 9\}$  kieliä  $A = \{aa, bb\}$  ja  $B = \{01, 02\}$ . Nyt

$$A \cup B = \{aa, bb, 01, 02\}$$

$$A \circ B = \{aa01, aa02, bb01, bb02\}$$

$$A^* = \{\epsilon, aa, bb, aaaa, aabb, bbaa, bbbb, \\ aaaaa, aaaabb, aabbaa, aabbbb, bbaaaa, \dots\}.$$



Seuraavana tavoitteenamme on osoittaa, että säännöllisten kielten luokka on **suljettu** operaatioiden  $\cup$ ,  $\circ$  ja  $*$  suhteen. Toisin sanoen jos  $A$  ja  $B$  ovat säännöllisiä, niin myös  $A \cup B$ ,  $A \circ B$  ja  $A^*$  ovat. Tapaus  $A \cup B$  on melko suoraviivainen, ja aloitamme siitä.

**Lause 1.1:** [Sipser Thm. 1.25] Jos kielet  $A$  ja  $B$  ovat säännöllisiä, niin myös  $A \cup B$  on.

**Todistus:** Oletetaan siis, että  $A = L(M_1)$  ja  $B = L(M_2)$  äärellisillä automaateilla  $M_1 = (Q_1, \Sigma, \delta_1, q_1, F_1)$  ja  $M_2 = (Q_2, \Sigma, \delta_2, q_2, F_2)$ . Muodostamme äärellisen automaatin  $M$ , jolla  $L(M) = A \cup B$ .

Automaatin  $M$  pitää siis hyväksyä  $w$ , jos ainakin toinen automaateista  $M_1$  ja  $M_2$  hyväksyy. Ongelmaa ei voi ratkaista simuloimalla automaatteja  $M_1$  ja  $M_2$  peräjälkeen, sillä äärellinen automaatti käsittelee kunkin syötemerkin vain kerran.

Ratkaisu on simuloida kumpaakin automaattia **samanaikaisesti**. Jos  $|Q_1| = m$  ja  $|Q_2| = n$ , niin automaatin  $M$  tilajoukoksi valitaan  $m \times n$  taulukko, jossa rivi esittää  $M_1$ :n ja sarake  $M_2$ :n tilaa. Vastaavasti  $\delta_1$  kertoo, mille riville mennään seuraavaksi, ja  $\delta_2$  mille sarakkeelle.

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

- $Q = Q_1 \times Q_2$ ,
- aakkosto  $\Sigma$  pysyy samana,
- kun  $r = (r_1, r_2) \in Q_1 \times Q_2$  ja  $a \in \Sigma$ , niin  $\delta(r, a) = (s_1, s_2)$  missä
$$s_1 = \delta_1(r_1, a) \quad \text{ja} \quad s_2 = \delta_2(r_2, a),$$
- $q_0 = (q_1, q_2)$  ja
- $F = \{ (r_1, r_2) \mid r_1 \in F_1 \text{ tai } r_2 \in F_2 \}$ .

Tarkastellaan mielivaltaista merkkijonoa  $w = w_1 \dots w_k$ . On olemassa automaatin  $M_1$  laskentaa esittävä (yksikäsitteinen) tilajono  $(r_0, \dots, r_k) \in Q_1^{k+1}$  ja automaatin  $M_2$  laskentaa esittävä (yksikäsitteinen) tilajono  $(s_0, \dots, s_k) \in Q_2^{k+1}$ , missä

1.  $r_0 = q_1$  ja  $s_0 = q_2$  ja
2.  $r_{i+1} = \delta_1(r_i, w_{i+1})$  ja  $s_{i+1} = \delta_2(s_i, w_{i+1})$  kun  $i = 0, \dots, n - 1$ .

Siis valitsemalla  $p_i = (r_i, s_i)$  saadaan automaatin  $M$  laskentaa kuvaava tilajono, jolla  $p_0 = q_0$  ja  $p_{i+1} = \delta(p_i, w_{i+1})$ . Nyt

$M$  hyväksyy merkkijonon  $w$

$$\Leftrightarrow (r_n, s_n) \in F$$

$$\Leftrightarrow r_n \in F_1 \text{ ja } s_n \in F_2$$

$$\Leftrightarrow M_1 \text{ hyväksyy merkkijonon } w \text{ ja } M_2 \text{ hyväksyy merkkijonon } w.$$

Siis  $L(M) = A \cup B$ .  $\square$

Jatkossa osoitetaan myös

**Lause 1.2:** [Sipser Thm. 1.26] Säännöllisten kielten joukko on suljettu konkatenation suhteen; ts. jos kielet  $A$  ja  $B$  ovat säännöllisiä, niin myös  $A \circ B$  on.

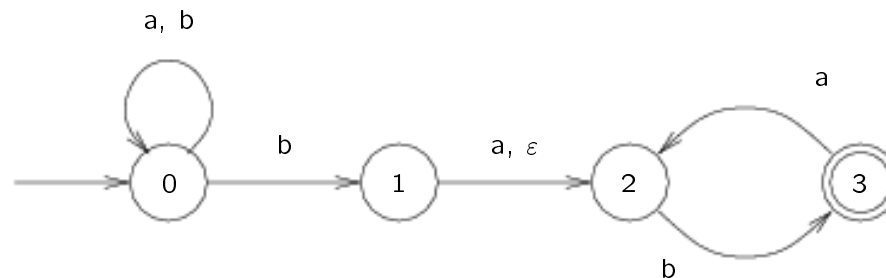
Todistaminen on kuitenkin edellistä lausetta hankalampaa. Oletaan  $A = L(M_1)$  ja  $B = L(M_2)$ , ja halutaan muodostaa  $M$ , jolla  $L(M) = A \circ B$ . Luonnollinen ajatus olisi pistää automaattit  $M_1$  ja  $M_2$  peräkkäin. Tässä tulee ongelmaksi, milloin pitäisi siirtyä automaattista  $M_1$  automaattiin  $M_2$ .

Eryteisesti jollain  $w \in A \circ B$  voi olla kaksi esitystä  $w = uv = u'v'$ , missä  $u \in A$  ja  $u' \in A$ , ja  $v \in B$  mutta  $v' \notin B$ . Jos tässä vielä  $u'$  on merkkijonon  $u$  alkuosa (eli  $u = u'x$  jollain  $x$ ), niin emme voi ilman muuta siirtyä automaattiin  $M_2$  heti, kun  $M_1$  pääsee hyväksyvään tilaan.

Jotta pääsemme eteenpäin, yleistämme äärellistä automaattia sallimalla epädeterminismin.

## Epädeterministiset äärelliset automaattit [Sipser s. 47–54]

Tarkastellaan esimerkkinä seuraavaa automaattia:



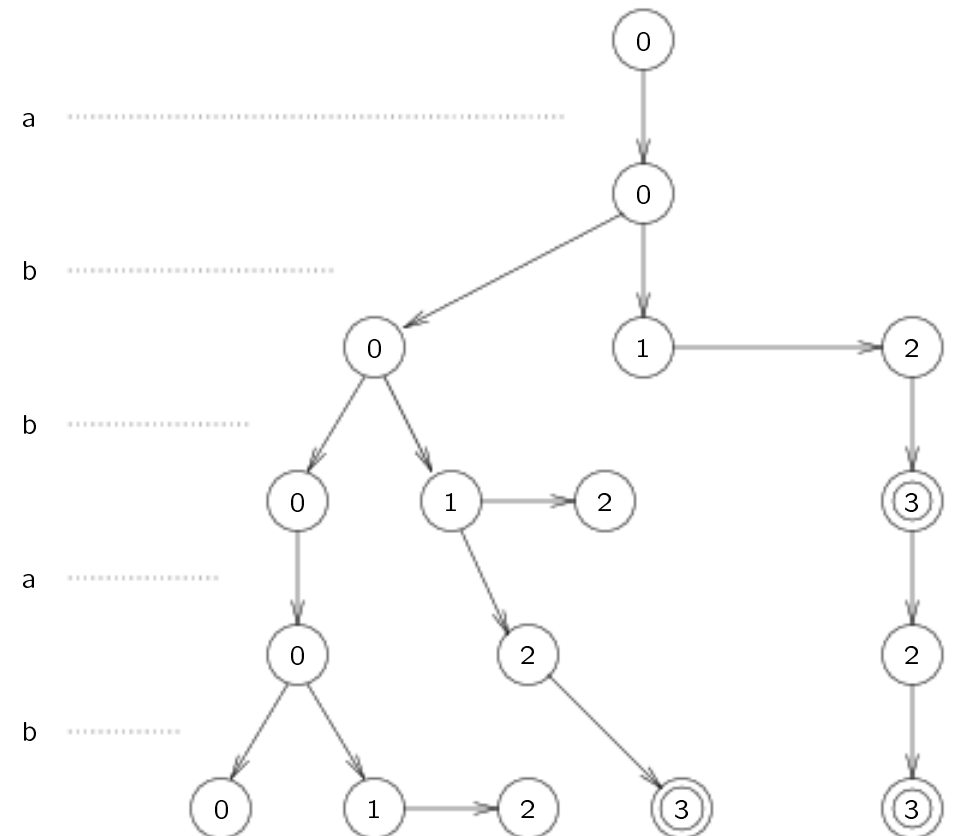
Tämä ei noudata edellä esitettyä formalismia:

- tilasta 0 pääsee merkillä b sekä tilaan 0 että tilaan 1,
- tilasta 1 on  $\epsilon$ -siirtymä, jolla pääsee tilaan 2 "käyttämättä" yhtään syötemerkkiä ja
- joitain siirtymiä ei ole määritelty.

Aiottu tulkinta on, että annetulla syötteellä "kokeillaan" **kaikkia mahdollisia** siirtymävaihtoehtoja. Automaatti hyväksyy, jos **yksikin** näistä vaihtoehtoista päättyy hyväksyvään tilaan.

Tällaisen automaatin laskentaa voidaan havainnollistaa ryhmittämällä vaihtoehdot puuksi:

- esimerkkinä syöte abbab
- vaakanuolet kuvaavat  $\epsilon$ -siirtymiä
- viimeisellä rivillä esiintyy kahteenkin kertaan hyväksyvä tila 3, joten automaatti hyväksyy merkkijonon abbab



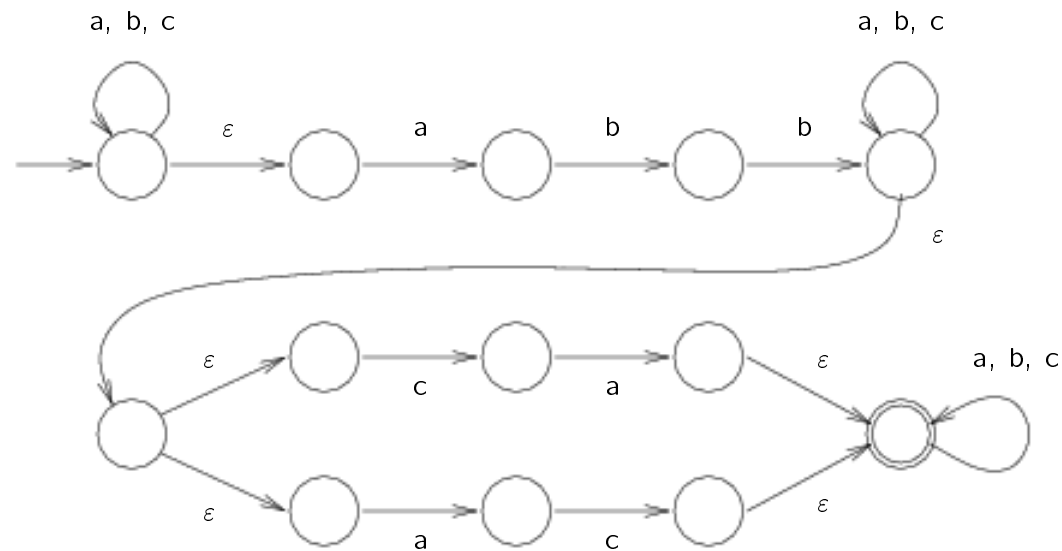
Kutsumme tällaista yleisempää automaattia epädeterministiseksi äärelliseksi automaatiksi (nondeterministic finite automaton, NFA). Aluksi esitetty perusversio on vastaavasti deterministinen äärellinen automaatti (deterministic finite automaton, DFA).

NFA:lle ei ole mitään samalla lailla ilmeistä fyysistä toteutusmallia kuin DFA:lle. Se kannattaa mieltää lähinnä kuvausformalismiksi.

Näemme jatkossa, että mille tahansa NFA:lle voidaan muodostaa DFA, joka tunnistaa saman kielen. Siis epädeterminismi ei anna tässä mielessä lisää ilmaisuvoimaa.

Epädeterminismiä käyttämällä esitystä voidaan kuitenkin usein selkeyttää ja yksinkertaistaa.

Seuraava NFA hyväksyy merkkijonot, joissa on osajonona ensin **abb** ja sen jälkeen **ca** tai **ac**. Huomaa konstruktion modulaarisuus. (Juuri tämän kielen voisi tunnistaa DFA:lla helpomminkin.)



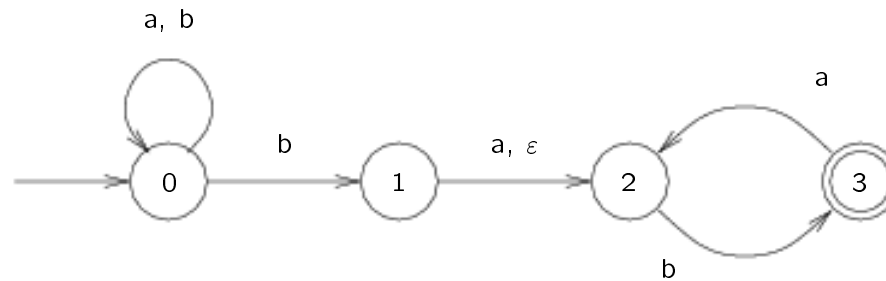
Mille tahansa aakkostolle  $\Sigma$  merkitsemme  $\Sigma_\epsilon = \Sigma \cup \{\epsilon\}$ . Muistetaan, että joukon  $A$  **potenssijoukkoa** eli kaikkien osajoukkojen joukkoa merkitään  $\mathcal{P}(A)$ .

Muodollisesti NFA on viisikko  $(Q, \Sigma, \delta, q_0, F)$ , missä

1.  $Q$  on äärellinen tilajoukko,
2.  $\Sigma$  on äärellinen aakkosto,
3.  $\delta: Q \times \Sigma_\epsilon \rightarrow \mathcal{P}(Q)$  on siirtymäfunktio,
4.  $q_0 \in Q$  on alkutila ja
5.  $F \subseteq Q$  on hyväksyvien tilojen joukko.

Erotukseksi DFA:sta siirtymäfunktio siis antaa yhden seuraajatilan sijaan **joukon** tiloja. Nämä tulkitaan ”mahdollisiksi” seuraajatiloiiksi.

## Esimerkkiautomaattimme



formaali esitys on siis  $(\{0, 1, 2, 3\}, \{a, b\}, \delta, 0, \{3\})$ , missä  $\delta$  saadaan taulukosta

$\delta$	a	b	$\varepsilon$
0	{0}	{0, 1}	$\emptyset$
1	{2}	$\emptyset$	{2}
2	$\emptyset$	{3}	$\emptyset$
3	{2}	$\emptyset$	$\emptyset$ .

Määrittelemme nyt, että NFA  $(Q, \Sigma, \delta, q_0, F)$  hyväksyy merkkijonon  $w$ , jos jollain  $k$  voidaan valita jonot  $(y_1, \dots, y_n) \in \Sigma_\varepsilon^n$  ja  $(r_0, \dots, r_n) \in Q^{n+1}$ , joilla

- $w = y_1 \dots y_n$ ,
- $r_0 = q_0$ ,
- $r_{i+1} \in \delta(r_i, y_{i+1})$  kun  $i = 0, \dots, n - 1$  ja
- $r_n \in F$ .

Edellisen sivun automaatilla ja merkkijonolla  $abbab$  voidaan valita esim.

$$\begin{aligned} n &= 6 \\ (y_1, y_2, y_3, y_4, y_5, y_6) &= (a, b, \varepsilon, b, a, b) \\ (r_0, r_1, r_2, r_3, r_4, r_5, r_6) &= (0, 0, 1, 2, 3, 2, 3), \end{aligned}$$

mistä nähdään, että automaatti hyväksyy merkkijonon.