Tik-79.148 **Kevät 2000**
Tietojenkäsittelyteorian perusteet
Laskuharjoitus 3
Ratkaisut

1. (1.8.2 / 1.9.2)

   Säännölliset lausekkeet määritellään induktiivisesti:

   - $\emptyset$ ja kaikki $a \in \Sigma$ ovat säännöllisiä lausekkeita.
   - Mikäli α ja β ovat säännöllisiä lausekkeita, niin myös $(\alpha\beta), (\alpha \cup \beta)$ ja α^* ovat säännöllisiä lausekkeita.
   - Mitkään muut lausekkeet eivät ole säännöllisiä.

   Säännöllisestä lausekkeesta $\emptyset^*$ käytetään lyhennysmerkintää e. Yleensä jätetään ylimääräiset sulut pois: $((a(bc)) \cup (cd^*)) = abc \cup cd^*$.

   Säännöllinen lauseke α määrittelee kielen $L(\alpha)$, joka on joukon Σ^* osajoukko. Esimerkiksi:

   $$L(ab \cup ac^*d) = \{ab, ad, acd, accd, accd, acccd, accccd, \dots\} \ .$$

   Mikäli sekaantumisen vaaraa ei ole, jätetään $L()$ melkein aina pois säännöllisen lausekkeen ympäriltä ja käytetään lauseketta itseään tarkoittamaan niiden sanojen joukkoa, jotka kuuluvat sen kuvaamaan kieleen.

   Säännölliset lausekkeet on ehkä helpointa ajatella muottina. Sana kuuluu kieleen, jos sen saa sovitettua lausekkeen antamaan muottiin:

   - $ab \Rightarrow$ otetaan ensin a, sitten b
   - $a \cup b \Rightarrow$ otetaan joko a tai b
   - $a^* \Rightarrow$ otetaan kuinka monta a:ta tahansa (myös 0).

   Oppikirja esittää hieman epäselvästi lausekkeiden $\emptyset$ ja $\emptyset^*$ välisen eron ja siirtyy käyttämään jälkimmäisestä merkintää e määrittelemättä sitä sen tarkemmin. Määritelmän mukaan $L(\emptyset) = \emptyset$, eli tyhjä kieli, johon ei kuulu yhtään sanaa. Vastaavasti $L(\emptyset^*) = L(\emptyset)^*$, eli kieli, joka muodostetaan seuraavasti:

   $$L(\emptyset)^* = \{w \mid w = w_1 \circ w_2 \circ \cdots \circ w_3, \text{ jollekin } k \geq 0 \text{ ja } w_1, \dots, w_k \in \emptyset\}$$

   Koska tyhjään joukkoon ei kuulu yhtään sanaa, on ainoa mahdollinen k:n arvo yllä on 0. Koska määritelmän perusteella nollan merkkijonon katenaatio on tyhjä merkkijono e, saadaan kieleksi

   $$L(\emptyset^*) = \{e\}$$

   Toisin sanoen, säännöllinen lauseke $\emptyset$ tarkoittaa kieltä, johon ei kuulu yhtään sanaa, ja lauseke $\emptyset^*$ tarkoittaa kieltä, johon kuuluu ainoastaan tyhjä sana e.

   Säännöllisten lausekkeiden unionia voidaan merkitä monilla eri tavoilla. Yleisimmät ovat $\cup$, $+$ sekä $|$. Useimmissa käytännön toteutuksissa määritellään lisäksi ylimääräisiä operaattoreita yksinkertaistamaan lausekkeita. Lausekkeiden ilmaisuvoima pysyy kuitenkin samana.

Säännöllisiä lausekkeita käytetään mm. ohjelmointikielten kääntäjien syntaktiseen analyysiin sekä määrämuotoisen tekstin etsimiseen tekstitiedostoista (grep).

a) Lausekkeen $\emptyset \cup a^* \cup b^* \cup (a \cup b)^*$ kuvaamaan kieleen kuuluu tyhjä sana e sekä kaikki sanat, joissa on joko pelkästään a-kirjaimia, pelkästään b-kirjaimia tai molempia sekaisin. Koska $\emptyset \subset (a \cup b)^*, a^* \subset (a \cup b)^*$ ja $b^* \subset (a \cup b)^*$, lauseke supistuu muotoon $(a \cup b)^*$.

b) Lauseke $((a^*b^*)^*(b^*a^*)^*)^*$ kuvaa myös kielen $(a \cup b)^*$. Tämän voi nähdä jo ensimmäisestä sulkulausekkeesta: ensin otetaan 0 tai useampia a-kirjaimia, sitten 0 tai useampia b-kirjaimia ja toistetaan tätä niin monta kertaa kuin on tarpeen minkä tahansa sanan generoimiseksi.

c) Lausekkeen $(a^*b)^* \cup (b^*a)^*$ ensimmäinen osa kuvaa kaikki sanat, jotka päättyvät b-kirjaimeen, ja jälkimmäinen kaikki a-kirjaimeen päättyvät sanat. Näiden kahden unioni on taas kerran kaikkien sanojen joukko $(a \cup b)^*$.

d) Lauseke $(a \cup b)^*a(a \cup b)^*$ kuvaa kaikki sanat, joissa on ainakin yksi a. Keskellä olevaa a-kirjainta ei voi siirtää Kleenen tähden sisään, sillä silloin hyväksyttäisiin myös sanat, joissa ei ole yhtään a:ta. Yksi muutos lausekkeelle voidaan kuitenkin tehdä. Nykyisessä muodossa voidaan mikä tahansa sanassa olevista a-kirjaimista valita "pakolliseksi" merkiksi. Voimme kuitenkin valita aina viimeisen a-kirjaimen "pakolliseksi", jolloin lauseke supistuu muotoon: $(a \cup b)^*ab^*$ (tai vastaavasti $b^*a(a \cup b)^*$, kun ensimmäinen a on kiinnitetty).

2. (1.8.3 / 1.9.3)

Kaikissa tämän tehtävän kohdissa käytetään aakkostona $\Sigma = \{a, b\}$.

(a) Kieltä $L = \{w \mid w \in \Sigma^*, w\text{:ssä on korkeintaan 3 } a\text{:ta.}\}$ vastaava säännölinen lauseke on: $b^*(\emptyset^* \cup a)b^*(\emptyset^* \cup a)b^*(\emptyset^* \cup a)b^*$

Perusajatuksena on luetella kaikki mahdolliset a-kirjamen esiintymät, ja ympäröidä ne b-kirjaimilla. Jokaisen a:n kohdalla on myös mahdollisuus jättää se pois valitsemalla $\emptyset^*$.

(b) Kieltä $L = \{w \mid w : \text{ssä olevien } a : \text{den määrä on jaollinen kolmella}\}$ vastaa lauseke $b^*(ab^*ab^*ab^*)^*$

(c) Kieltä $L = \{w \mid w\text{:ssä on tasan yksi alimerkkinono } aaa\}$ vastaa lauseke: $(b \cup ab \cup aab)^*aaa(b \cup ba \cup baa)^*$. Alimerkkijonoiksi lasketaan myös päällekäiset esiintymät, eli aaa sisältyy kaksi kertaa merkkijonoon $aaaa$.

3. (1.8.5 / 1.9.5)

a) Väite $baa \in a^*b^*a^*b^*$ on tosi, sillä valitsemalla yksi ensimmäisistä b-kirjaimista ja kaksi toisista a-kirjaimista saadaan sana baa.

b) Väite $b^*a^* \cap a^*b^* = a^* \cup b^*$ on tosi, sillä

$$\begin{gathered} b^*a^* \cap a^*b^* = \\ \{e, b^+, a^+, b^+a^+\} \cap \{e, a^+, b^+, a^+b^+\} = \\ \{e, a^+, b^+\} = a^* \cup b^* \end{gathered}$$

Merkintä $a^+ = aa^*$

c) Väite $a^*b^* \cap c^*d^* = \emptyset$ on väärä, sillä

$$\begin{gathered} a^*b^* \cup c^*d^* = \\ \{e, b^+, a^+, b^+a^+\} \cap \{e, c^+, d^+, c^+d^+\} = \\ \{e\} \neq \emptyset \end{gathered}$$

Tässä kannattaa huomata, että yhtälön vasemmalla puolella oleva merkintä **ei** ole säännöllinen lauseke, sillä säännöllisessä lausekkeessa ei voi esiintyä leikkausta, vaan kyseessä on kahden säännöllisen kielen leikkaus.

d) Väite $abcd \in (a(cd)^*b)^*$ on myös väärä, sillä kaikki säännöllisen lausekkeen hyväksymät sanat (tyhjää sanaa lukuunottamatta) päättyvät b-kirjaimeen.

4. Väite: Jos kieli L on säännöllinen, niin myös kieli

$$L' = \{w \mid uw \in L \text{ jollekin merkkijonolle } u\}$$

on säännöllinen.

Kieleen L' kuuluvat kaikki ne merkkijonot, jotka esiintyvät kieleen L kuuluvien sanojen suffikseina. Esimerkiksi: jos

$$\begin{gathered} L = \{a, aba, abb\}, \text{ niin} \\ L' = \{e, a, b, ba, bb, aba, abb\} \end{gathered}$$

Tässä konstruoidaan systemaattinen tapa muodostaa L:n määrittelemästä säännöllisestä lausekkeesta R (eli $L(R) = L$) uusi lauseke R', jonka määrittelemä kieli on L' (eli $L(R') = L'$).

Formaalimmin, merkitään kaikkien säännöllisten lausekkeiden joukkoa symbolilla S. Nyt halutaan löytää jokin funktio $f : S \to S$, siten, että $f(R) = R'$.

Määritellään funktio f rekursiivisesti:

i) (perustapaus) $f(\emptyset) = \emptyset$ ja $f(a) = a \cup e, a \in \Sigma \cup \{e\}$. Toisin sanoen, mikäli säännöllinen lauseke R on tyhjä ei sillä ole yhtään suffiksia, ja jos se on pelkästään yksi aakkoston merkki tai e, on lauseke R' saman merkin ja e:n unioni.

ii) $f(\alpha \cup \beta) = f(\alpha) \cup f(\beta), \alpha, \beta \in S$. Toisin sanoen, mikäli R on kahden lausekkeen α ja β unioni, niin sen määrittelemän kielen sanojen lopussa voi olla mitä tahansa, mitä voi olla jomman kumman alilausekkeen lopussa.

iii) $f(\alpha^*) = f(\alpha)\alpha^*$, eli sanan lopussa voi olla tähden vaikutusalueella oleva alilauseke kokonaisena mielivaltaisen monta kertaa, ja sen edellä jotain, mihin alilauseke voi päättyä

iv) $f(\alpha\beta) = f(\alpha)\beta \cup f(\beta)$, eli katenaatiolla muodostetun lausekkeen kuvaaman kielen sanojen lopussa voi joko olla jotain, mitä voi olla jälkimmäisen osan lopussa, tai sitten viimeinen osa on kokonaisuudessaan mukana, ja sen edellä on jotain, mihin ensimmäinen osa voi loppua.

Tämän jälkeen puuttuu enää formaali todistus ylläolevien sääntöjen oikeellisuudesta. Todistus sivuutetaan tässä, mutta sen voi rakentaa osoittamalla erikseen jokaisen säännön oikeellisuuden induktiolla.

Esimerkki: Olkoon $R = ba(aa \cup bab)^*$. Nyt

$$\begin{aligned}
f(R) &= f(ba(aa \cup bab)^*) \\
&= f(ba)(aa \cup bab)^* \cup f((aa \cup bab)^*) \\
&= (f(b)a \cup f(a))(aa \cup bab)^* \cup f(aa \cup bab)(aa \cup bab)^* \\
&= ((e \cup b)a \cup (e \cup a))(aa \cup bab)^* \cup (f(aa) \cup f(bab))(aa \cup bab)^* \\
&= (e \cup a \cup ba)(aa \cup bab)^* \cup ((f(a)a \cup f(a)) \cup (f(b)ab \cup f(a)b \cup f(b)))(aa \cup bab)^* \\
&= (e \cup a \cup ba)(aa \cup bab)^* \cup (e \cup a \cup aa \cup b \cup ab \cup bab)(aa \cup bab)^* \\
&= (e \cup a \cup aa \cup b \cup ab \cup bab \cup ba)(aa \cup bab)^* \\
&= (e \cup a \cup b \cup ab \cup ba)(aa \cup bab)^*
\end{aligned}$$

5. Perl-kielessä on sisäänrakennettuna laaja tuki säännöllisille lausekkeille. Normaalien operaatioiden (katenaatio, unioni ja Kleenen tähti) lisäksi Perlissä on koko joukko lyhennysmerkintöjä, joista tärkeimmät ovat:

   - '.', millä tarkoitetaan mitä merkkiä tahansa.
   - '[]', millä kohtaa saa olla mikä tahansa hakasulkujen sisältä löytyvä merkki. Esimerkiksi merkintä `[a-cX]` on lyhennysmerkintä lausekkeelle $(a \cup b \cup c \cup X)$. Laittamalla hakasulkujen sisään ensimmäiseksi merkiksi '`^`' voidaan määritellä haluttujen kirjainten komplementti. Lauseke `[^a-z]` tunnistaa kaikki merkit, jotka eivät ole pieniä kirjaimia.
   - '`^`' ja '`$`', jotka esittävät rivin alku- ja loppukohtia.
   - '?', jota käytetään merkityksessä ”korkeintaan yksi”. Esimerkiksi lauseketta `a?` vastaa lauseke $(a \cup e)$.
   - '+', joka tarkoittaa “vähintään yksi”. Esimerkiksi `a+`, jota vastaa aa^*.
   - Lausekkeiden unionia $\cup$ esitetään merkillä '`|`'.

   Tyhjät salasanat löytävä säännöllinen lauseke on seuraavanlainen:

   `/^[^:]*::/`

   Lauseke rakentuu seuraavista osista:

   - '/'-merkit toimivat lausekkeen alku- ja loppumerkkeinä.

- '^'-merkki lausekkeen alussa tunnistaa syöteestä rivien alkukohdat.
- Rakenne '`[^:]*`' tunnistaa kaikki merkkijonot, joissa ei ole kaksoispistettä. Yhdistettynä '^'-merkkiin, lauseke tunnistaa syöterivin ensimmäisen kentän sisällön.
- '`::`': tunnistaa tyhjät kentät.

Varsinainen ohjelmakoodi on yksinkertaisesti:

```
#!/usr/bin/perl
while (<>) {
  if ( /^[^:]*::/ ){
    print $_;
  }
}
```