Luonnolliset vs. muodolliset kielet

Luonnollisia kieliä ovat esim.

1. englanti,
2. suomi,
3. ranska.

Muodollisia kieliä ovat esim.

1. lauselogiikan kieli (ilmaisut $\neg p$, $p \wedge q$ jne.),
2. C++, FORTRAN,
3. bittijonokokoelma
   $\{0, 1, 10, 11, 100, 101, 110, 111, 1000, ...\}$.

Aakkosto Σ

Olkoon Σ epätyhjä ja äärellinen joukko. Joukon Σ alkioita kutsumme *aakkosiksi, kirjaimiksi* tai *merkeiksi*. Joukko Σ^* on kaikkien niiden äärellisten kirjainjonojen joukko, joka saadaan kirjoittamalla joukon Σ aakkosia peräkkäin. Myös *tyhjä jono* ε on äärellinen kirjainjono.

Esimerkki. Olkoon $\Sigma = \{a\}$. Nyt
$\Sigma^* = \{\varepsilon, a, aa, aaa, aaaa, ...\}$.
Huomaa, että $\Sigma^* \neq \{a, aa, aaa, aaaa, ...\}$.

Esimerkki. Olkoon $\Sigma = \{a, b\}$. Nyt
$\Sigma^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, aab...\}$.

Muodollisen kielen määritelmä

Olkoon Σ aakkosto. Kutsumme joukon Σ^* osajoukkoja muodollisiksi kieliksi.

Esimerkki. Olkoon $\Sigma = \{a\}$. Nyt esim. $L = \{a, aaaaa, aaaaaaaaaaaaa\}$ on muodollinen kieli, sillä $L \subseteq \Sigma^*$. Kieli L sisältää kolme alkiota. Kielen alkioita kutsumme *sanoiksi*. Kieli L sisältää sanat

1. a,
2. $aaaaa$,
3. $aaaaaaaaaaaaa$.

Myös *tyhjä kieli* $\emptyset$ on, muodollinen kieli, kuten myös joukko $\{\varepsilon\}$. Huomaa, että $\emptyset$ ja $\{\varepsilon\}$ ovat eri kielet.

Muodollisen kielen määritelmä

Esimerkki. Olkoon $\Sigma = \{a, b\}$. Nyt esim. $L = \{a, abbb, bbbb, aaaaa\}$ on muodollinen kieli. Myös $K = \{a, ab, abb, abbb, abbbb, abbbbb, abbbbbb, abbbbbbbb, ...\}$ on muodollinen kieli. Kieli L sisältää neljä sanaa. Kieli K sisältää ääretön sanaa. Se on siis ääretön kieli.

Esimerkki. Olkoon $\Sigma = \{a, b, c\}$. Nyt esim. $M = \{\varepsilon, abc, abcc\}$ on muodollinen kieli, kuten myös $N = \{abc, abcc\}$. Nämä ovat eri kielet. Kieli L sisältää kolme sanaa ja kieli N kaksi.

Kuinka monta sanaa sisältää kieli $\{aaa, aaa, aaa, aaa\}$? Miksi?

Sanojen konkatenaatio

Olkoon Σ aakkosto. Olkoon α ja β kaksi Σ-sanaa, eli aakkoston Σ kirjaimista muodostettua sanaa. Sanojen α ja β *ketju* eli *konkatenaatio* $\alpha \cdot \beta$ saadaan kirjoittamalla α ja β peräkkäin.

Esimerkki. Olkoon $\alpha = aa$ ja $\beta = bb$. Tällöin $\alpha \cdot \beta = aabb$ ja $\beta \cdot \alpha = bbaa$.

Esimerkki. Olkoon $\alpha = \varepsilon$ ja $\beta = ab$. Tällöin $\alpha \cdot \beta = ab = \beta \cdot \alpha$. Tällöin lisäksi $\alpha \cdot \alpha = \varepsilon \cdot \varepsilon = \varepsilon$.

Selvästi konkatenaatiolle on voimassa liitäntälaki: $\alpha \cdot (\beta \cdot \gamma) = (\alpha \cdot \beta) \cdot \gamma$. Voimmekin jättää sulut kirjoittamatta ja merkitä $\alpha \cdot (\beta \cdot \gamma) = (\alpha \cdot \beta) \cdot \gamma = \alpha \cdot \beta \cdot \gamma$.

Kielten konkatenaatio

Olkoon Σ aakkosto. Olkoon L ja K kaksi Σ-kieltä, eli $L \subseteq \Sigma^*$ ja $K \subseteq \Sigma^*$. Määrittelemme

$$L \cdot K = \{\, \alpha \cdot \beta \mid \alpha \in L,\ \beta \in K \,\}.$$

$L \cdot K$ on kielten L ja K konkatenaatio. $L \cdot K$ on Σ-kieli. On tärkeää ymmärtää, että kielten konkatenaatio ja sanojen konkatenaatio ovat eri käsitteitä.

Esimerkki. Olkoot $L = \{a, ab\}$ ja $K = \{aaa, bb\}$. Nyt $L \cdot K = \{aaaa, abb, abaaa, abbb\}$.

Esimerkki. Olkoot $L = \{\varepsilon, a\}$ ja $K = \{aaa, bb\}$. Nyt $L \cdot K = \{aaa, bb, aaaa, abb\}$.

Esimerkki. Olkoot $L = \emptyset$ ja $K = \{aaa, bb\}$. Nyt $L \cdot K = \emptyset$.

Konkatenaatio

Esimerkki. Olkoot $L = \{\varepsilon, a\}$ ja $K = \{\varepsilon, bb\}$. Nyt $L \cdot K = \{\varepsilon, bb, a, abb\}$.

Miten kielet L ja K olisi määriteltävä, että olisi voimassa $L \cdot K = \{\varepsilon\}$?

Olkoon $\{a\}$ aakkosto. Määrittele jokin kieli L siten että $a \in L$ ja $L \cdot L = L$.

Myös kielien konkatenaatiolle on voimassa liitäntälaki: $L \cdot (K \cdot H) = (L \cdot K) \cdot H$. Voimmekin merkitä $L \cdot (K \cdot H) = (L \cdot K) \cdot H = L \cdot K \cdot H$.

Olkoot α ja β sanoja sekä L ja K kieliä. Merkinnän $\alpha \cdot \beta$ sijaan voimme käytää lyhyempää merkintää $\alpha\beta$. Samaan tapaan kielien L ja K konkatenaatioon voidaan viitata merkinnällä LK.

Sanojen potenssi

Olkoon α sana. Määrittelemme sanan α *potenssin* seuraavalla tavalla.

1. $\alpha^0 = \varepsilon$.
2. $\alpha^1 = \alpha$.
3. Kaikille kokonaisluvuille $n \geq 2$ pätee

$$\alpha^n = \alpha \cdot \ldots \cdot \alpha\,,$$

   eli α^n on sanan α n-kertainen konkatenaatio itsensä kanssa.

Esimerkki. Olkoon $\alpha = ab$. Tällöin $\alpha^3 = \alpha \cdot \alpha \cdot \alpha = ababab$.

Esimerkki. Olkoon $\beta = \varepsilon$. Tällöin $\beta^{19} = \varepsilon$.

Kielten potenssi

Olkoon L kieli. Määrittelemme kielen L *potenssin* seuraavalla tavalla.

1. $L^0 = \{\varepsilon\}$.
2. $L^1 = L$.
3. Kaikille kokonaisluvuille $n \geq 2$ pätee

$$L^n = L \cdot \ldots \cdot L\,,$$

eli L^n on kielen L n-kertainen konkatenaatio itsensä kanssa.

Esimerkki. Olkoon $L = \{ab\}$. Tällöin $L^3 = \{ababab\}$.

Esimerkki. Olkoon $L = \{a, aa, aaa, aaaa, \ldots\}$. Tällöin $L^3 = \{aaa, aaaa, aaaaa, aaaaaa, \ldots\}$.

Kleenen sulkeuma

Olkoon L kieli. Kielen L *Kleenen sulkeuma* L^* on L:n kaikkien potenssien yhdiste, eli

$$L^* = L^0 \cup L^1 \cup L^2 \cup L^3 \cup \dots \ .$$

Esimerkki. Olkoon $L = \{aa\}$. Tällöin $L^* = \{\varepsilon, aa, aaaa, aaaaaa, aaaaaaaa, \dots\}$.

Esimerkki. $\{a, b\}^* = \{\varepsilon, a, b, a^2, ab, ba, b^2, a^3, a^2b, aba, ab^2, ba^2, bab, b^2a, b^3, a^4, \dots\}$.

Esimerkki. $\emptyset^* = \{\varepsilon\}$.

Esimerkki. $\{\varepsilon\}^* = \{\varepsilon\}$.

Olkoot L kieli. Päteekö välttämättä, että $(L^*)^* = L^*$?

Säännölliset lausekkeet

Säännölliset lausekkeet muodostavat täsmällisen tavan määritellä muodollisia kieliä. Olkoon Σ aakkosto.

1. $\emptyset$ on säännöllinen lauseke.
2. ε on säännöllinen lauseke.
3. Jos $a \in \Sigma$, niin a on säännöllinen lauseke.
4. Jos ξ ja η ovat säännöllisiä lausekkeita, niin $(\xi \cup \eta)$ on säännöllinen lauseke.
5. Jos ξ ja η ovat säännöllisiä lausekkeita, niin $(\xi \cdot \eta)$ on säännöllinen lauseke.
6. Jos ξ on säännöllinen lauseke, niin ξ^* on säännöllinen lauseke.
7. Muita säännöllisiä lausekkeita ei ole.

Säännölliset lausekkeetkin ovat siis muodollisia sanoja. Tässä säännöllisiin lausekkeisiin liittyvä aakkosto on joukko $\Sigma \cup \{\emptyset, \varepsilon, (,), \cup, \cdot, ^*\}$.

Säännölliset lausekkeet

Määrittelimme säännöllisten lausekkeiden syntaksin. Määrittelemme nyt tulkinnan eli semantiikan säännöllisille lausekkeille. Tarkoitamme lyhennyksellä s.l. säännöllistä lauseketta.

1. S.l. $\emptyset$ tarkoittaa joukkoa $\emptyset$
2. S.l. ε tarkoittaa joukkoa $\{\varepsilon\}$.
3. S.l. a, missä $a \in \Sigma$, tarkoittaa joukkoa $\{a\}$.
4. Jos s.l. ξ tarkoittaa joukkoa X ja s.l. η tarkoittaa joukkoa Y, niin s.l. $(\xi \cup \eta)$ tarkoittaa joukkoa $X \cup Y$.
5. Jos s.l. ξ tarkoittaa joukkoa X ja s.l. η tarkoittaa joukkoa Y, niin s.l. $(\xi \cdot \eta)$ tarkoittaa joukkoa $X \cdot Y$.
6. Jos s.l. ξ tarkoittaa joukkoa X, niin ξ^* tarkoittaa joukkoa X^*.

Jokainen säännöllinen lauseke siis tarkoittaa jotakin muodollista kieltä.

Säännölliset lausekkeet

Esimerkki. S.l. $(a \cup (b \cup c))$ tarkoittaa joukkoa $\{a\} \cup (\{b\} \cup \{c\}) = \{a, b, c\}$.

Esimerkki. S.l. $(a \cdot c^*)$ tarkoittaa joukkoa $\{a\} \cdot (\{c\}^*) = \{a\} \cdot \{\varepsilon, c, cc, ccc, cccc, ...\} = \{a, ac, acc, accc, acccc, ...\}$.

Esimerkki. S.l. $(a \cup \emptyset)$ tarkoittaa joukkoa $\{a\}$, mutta s.l. $(\varepsilon \cup a)$ tarkoittaa joukkoa $\{\varepsilon, a\}$.

Säännölliset lausekkeet

Esimerkki. S.l. $(a^* \cdot b^*)$ tarkoittaa joukkoa
$\{a\}^* \cdot \{b\}^* = \{\varepsilon, a, aa, aaa, ...\} \cdot \{\varepsilon, b, bb, bbb, ...\} =$
$\{\varepsilon, a, b, aa, ab, bb, aaa, aab, abb, bbb, aaaa, aaab, aabb, ...\}$.

Esimerkki. S.l. $(\emptyset^* \cdot (a \cdot a))$ tarkoittaa joukkoa
$\{\varepsilon\} \cdot (\{a\} \cdot \{a\}) = \{\varepsilon\} \cdot \{aa\} = \{aa\}$.

Esimerkki. S.l. $(a \cup b)^*$ tarkoittaa joukkoa
$(\{a\} \cup \{b\})^* = \{a, b\}^* = \{\varepsilon, a, b, aa, ab, ba, bb, ...\}$.

Säännölliset lausekkeet

Liitäntälakien nojalla voimme viitata säännölliseen lausekkeeseen $(\xi \cdot (\eta \cdot \zeta))$ ilmaisulla $\xi \cdot \eta \cdot \zeta$. Tässä myös uloimmat sulut on jätetty kirjoittamatta. Vastaavia lyhennyksiä voidaan käyttää myös operaation $\cup$ tapauksessa.

Esimerkki. Ilmaisu $(a \cup b \cup c) \cdot d^* \cdot e \cdot f$ viittaa niiden sanojen joukkoon, joilla on ensimmäisenä kirjaimena a, b tai c, tämän jälkeen on jono (mahdollisesti tyhjä) d-kirjaimia, jonka jälkeen sana päättyy kahden kirjaimen muodostamaan jonoon ef.

Säännölliset kielet

Olkoon Σ aakkosto. *Säännölliset kielet* ovat muodollisia kieliä $L \subseteq \Sigma^*$. Aakkoston ollessa Σ, määrittelemme säännölliset kielet seuraavasti.

1. $\emptyset$ on säännöllinen kieli.
2. $\{\varepsilon\}$ on säännöllinen kieli.
3. Jos $a \in \Sigma$, niin $\{a\}$ on säännöllinen kieli.
4. Jos X ja Y ovat säännöllisiä kieliä, niin $X \cup Y$ on säännöllinen kieli.
5. Jos X ja Y ovat säännöllisiä kieliä, niin $X \cdot Y$ on säännöllinen kieli.
6. Jos X on säännöllinen kieli, niin X^* on säännöllinen kieli.
7. Muita säännöllisiä kieliä ei ole.

Säännölliset kielet

Lause. Kieli L on säännöllinen kieli jos ja vain jos on olemassa säännöllinen lauseke, joka määrittelee kielen L.

Sivuutamme todistuksen.

Esimerkki. Kieli $L = \{\varepsilon, a, aa, aaa, aaaa, ...\}$ on yllä olevan lauseen nojalla säännöllinen kieli, sillä L voidaan määritellä säännöllisellä lausekkeella a^*.

Harjoitustehtäviä

Harjoitustehtävä. Olkoon $\Sigma = \{0, 1\}$ aakkosto. Muodosta säännöllinen ilmaus joka määrittelee kielen L, johon kuuluvat täsmälleen kaksi kirjainta pitkät sanat.

Vastaus. $00 \cup 01 \cup 10 \cup 11$.

Harjoitustehtävä. Olkoon $\Sigma = \{0, 1\}$ aakkosto. Muodosta säännöllinen ilmaus joka määrittelee kielen L, johon kuuluvat parillista pituutta olevat sanat. Tyhjä sana ε on pituudeltaan parillinen.

Vastaus. $(00 \cup 01 \cup 10 \cup 11)^*$.

Harjoitustehtäviä

Harjoitustehtävä. Onko säännöllisille lausekkeille yleisesti voimassa $(\xi \cup \eta)^* \equiv (\xi^* \cup \eta^*)$? Tässä symboli $\equiv$ tarkoittaa, että lausekkeet määrittelevät täsmälleen saman kielen. Perustele vastauksesi lyhyesti.

Vastaus. Ehto ei ole voimassa, sillä $(a \cup b)^* \not\equiv a^* \cup b^*$. Esimerkiksi *ba* kuuluu ilmauksen $(a \cup b)^*$ määrittelemään kieleen, mutta on helppo nähdä että *ba* ei kuulu ilmauksen $a^* \cup b^*$ määrittelemään kieleen.

Harjoitustehtävä Vastaa perustelematta, onko säännöllisille lausekkeille voimassa ehto $\xi^{**} \equiv \xi^*$.

Vastaus. On.