

2.8 Säännöllisten kielten rajoituksista

Kardinaliteettisyydestä on oltava olemassa (paljon) ei-säännöllisiä kieliä: kieliä on ylinumeroituva määrä, säännöllisiä lausekkeita vain numeroituvasti.

Voidaanko löytää konkreettinen, *mielenkiintoinen* esimerkki kielestä, joka ei olisi säännöllinen? Helposti.

Säännöllisten kielten perusrajoitus: äärellisillä automaateilla on vain rajallinen “muisti”. Siten ne eivät pysty ratkaisemaan ongelmia, joissa vaaditaan mielivaltaisen suurten lukujen tarkkaa muistamista.

Esimerkki: sulkulausekekieli

$$L_{\text{match}} = \{(^k)^k \mid k \geq 0\}.$$

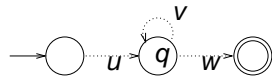
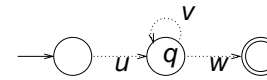
Formalisointi: “pumppauslemma”.



Lemma 2.6 (Pumppauslemma) Olkoon A säännöllinen kieli. Tällöin on olemassa sellainen $n \geq 1$, että mikä tahansa $x \in A$, $|x| \geq n$, voidaan jakaa osiin $x = uvw$ siten, että $|uv| \leq n$, $|v| \geq 1$, ja $uv^i w \in A$ kaikilla $i = 0, 1, 2, \dots$

Todistus. Olkoon M jokin A :n tunnistava deterministinen äärellinen automaatti, ja olkoon n M :n tilojen määrä. Tarkastellaan M :n läpikäymiä tiloja syötteellä $x \in A$, $|x| \geq n$. Koska M jokaisella x :n merkillä siirtyy tilasta toiseen, sen täytyy kulkea jonkin tilan kautta (ainakin) kaksi kertaa — itse asiassa jo x :n n :n ensimmäisen merkin aikana. Olkoon q ensimmäinen toistettu tila.

Olkoon u M :n käsittelemä x :n alkuosa sen tullessa ensimmäisen kerran tilaan q , v se osa x :stä jonka M käsittelee ennen ensimmäistä paluutaan q :hun, ja w loput x :stä. Tällöin on $|uv| \leq n$, $|v| \geq 1$, ja $uv^i w \in A$ kaikilla $i = 0, 1, 2, \dots$



Esimerkki. Tarkastellaan em. sulkulausekekieltä (merk. ‘(’ = a , ‘)’ = b):

$$L = L_{\text{match}} = \{a^k b^k \mid k \geq 0\}.$$

Oletetaan, että L olisi säännöllinen. Tällöin pitäisi pumppauslemman mukaan olla jokin $n \geq 1$, jota pitempiä L :n merkkijonoja voidaan pumpata. Valitaan $x = a^n b^n$, jolloin $|x| = 2n > n$. Lemman mukaan x voidaan jakaa pumpattavaksi osiin $x = uvw$, $|uv| \leq n$, $|v| \geq 1$; siis on oltava

$$u = a^i, v = a^j, w = a^{n-(i+j)} b^n, \quad i+j \leq n, j \geq 1.$$

Mutta esimerkiksi “0-kertaisesti” pumpattaessa:

$$uv^0 w = a^i a^{n-(i+j)} b^n = a^{n-j} b^n \notin L.$$

Siten L ei voi olla säännöllinen.



3. KIELIOPIT JA MERKKIJONOJEN TUOTTAMINEN

Kielioppi = muunnossysteemi merkkijonojen (kielen “sanojen”) tuottamiseen tietystä lähtöjonosta alkaen, osajonoja toistuvasti annettujen sääntöjen mukaan uudelleenkirjoittamalla.

Kielioppi on *yhteydetön*, jos kussakin uudelleenkirjoitusaskeleessa korvataan yksi erityinen muuttujat. *välikesymboli* jollakin siihen liitettyä korvausjonolla, ja korvaus voidaan aina tehdä symbolia ympäröivän merkkijonon rakenteesta riippumatta.

Sovelluksia: rakenteisten tekstien kuvaaminen (esim. ohjelmointikielten BNF-syntaksikuvaukset, XML:n DTD/Schema-määrittelyt), yleisemmin rakenteisten “olioiden” kuvaaminen (esim. syntaktinen hahmontunnistus).



Yhteydettömillä keliopilla voidaan kuvata (tuottaa) myös ei-säännöllisiä kieliä.

Esimerkki: yhteydetön kelioppi kielelle L_{match} (lähtösymboli S):

(i) $S \rightarrow \varepsilon$,

(ii) $S \rightarrow (S)$.

Esimerkiksi merkkijonon $((()))$ tuottaminen:

$$S \Rightarrow (S) \Rightarrow ((S)) \Rightarrow (((S))) \Rightarrow (((\varepsilon))) = ((())).$$



Toinen esimerkki: kelioppi C-tyyppisen ohjelmointikielen aritmeettisille lausekkeille (yksinkertaistettu).

$$\begin{array}{l} E \rightarrow T \mid E + T \\ T \rightarrow F \mid T * F \\ F \rightarrow a \mid (E). \end{array}$$

Esimerkiksi lausekkeen $(a + a) * a$ tuottaminen:

$$\begin{array}{l} \underline{E} \Rightarrow \underline{T} \Rightarrow \underline{T} * F \Rightarrow \underline{E} * F \\ \Rightarrow (\underline{E}) * F \Rightarrow (\underline{E} + T) * F \Rightarrow (\underline{T} + T) * F \\ \Rightarrow (\underline{F} + T) * F \Rightarrow (a + \underline{T}) * F \Rightarrow (a + \underline{F}) * F \\ \Rightarrow (a + a) * \underline{F} \Rightarrow (a + a) * a. \end{array}$$



Määritelmä 3.1 Yhteydetön kelioppi on nelikko

$$G = (V, \Sigma, P, S),$$

missä

- ▶ V on keliopin aakkosto;
- ▶ $\Sigma \subseteq V$ on keliopin *päätemerkkien* joukko; sen komplementti $N = V - \Sigma$ on keliopin *välikemerkkien t. -symbolien* joukko;
- ▶ $P \subseteq N \times V^*$ on keliopin *sääntöjen t. produktioiden* joukko;
- ▶ $S \in N$ on keliopin *lähtösymboli*.

Produktiota $(A, \omega) \in P$ merkitään tavallisesti $A \rightarrow \omega$.



Merkkijono $\gamma \in V^*$ *tuottaa t. johtaa suoraan* merkkijonon $\gamma' \in V^*$ keliopissa G , merkitään

$$\gamma \xrightarrow[G]{\Rightarrow} \gamma'$$

jos voidaan kirjoittaa $\gamma = \alpha A \beta$, $\gamma' = \alpha \omega \beta$ ($\alpha, \beta, \omega \in V^*$, $A \in N$), ja keliopissa G on produktio $A \rightarrow \omega$.

Jos kelioppi G on yhteydestä selvä, voidaan merkitä $\gamma \Rightarrow \gamma'$.

Merkkijono $\gamma \in V^*$ *tuottaa t. johtaa* merkkijonon $\gamma' \in V^*$ keliopissa G , merkitään

$$\gamma \xrightarrow[G]{\Rightarrow^*} \gamma'$$

jos on olemassa jono V :n merkkijonoja $\gamma_0, \gamma_1, \dots, \gamma_n$ ($n \geq 0$), siten että

$$\gamma = \gamma_0 \xrightarrow[G]{\Rightarrow} \gamma_1 \xrightarrow[G]{\Rightarrow} \dots \xrightarrow[G]{\Rightarrow} \gamma_n = \gamma'.$$

Erikoistapauksena $n = 0$ saadaan $\gamma \xrightarrow[G]{\Rightarrow^*} \gamma$ millä tahansa $\gamma \in V^*$.

Jälleen, jos G on yhteydestä selvä, voidaan merkitä $\gamma \Rightarrow^* \gamma'$.



Merkkijono $\gamma \in V^*$ on kieliopin G lausejohdos, jos on $S \xRightarrow{G}^* \gamma$.

Pelkästään päätemerkeistä koostuva G :n lausejohdos $x \in \Sigma^*$ on G :n lause.

Kieliopin G tuottama t. kuvaama kieli koostuu G :n lauseista:

$$L(G) = \{x \in \Sigma^* \mid S \xRightarrow{G}^* x\}.$$

Formaali kieli $L \subseteq \Sigma^*$ on yhteydetön, jos se voidaan tuottaa jollakin yhteydettömällä kieliopilla.



Esimerkiksi tasapainoisten sulkujonojen muodostaman kielen $L_{\text{match}} = \{(^k)^k \mid k \geq 0\}$ tuottaa kielioppi

$$G_{\text{match}} = (\{S, (,)\}, \{(,)\}, \{S \rightarrow \varepsilon, S \rightarrow (S)\}, S).$$

Yksinkertaisten aritmeettisten lausekkeiden muodostaman kielen L_{expr} tuottaa kielioppi

$$G_{\text{expr}} = (V, \Sigma, P, E),$$

missä

$$V = \{E, T, F, a, +, *, (,)\},$$

$$\Sigma = \{a, +, *, (,)\},$$

$$P = \{E \rightarrow T, E \rightarrow E + T, T \rightarrow F, T \rightarrow T * F, F \rightarrow a, F \rightarrow (E)\}.$$



Toinen kielioppi kielen L_{expr} tuottamiseen on

$$G'_{\text{expr}} = (V, \Sigma, P, E),$$

missä

$$V = \{E, a, +, *, (,)\},$$

$$\Sigma = \{a, +, *, (,)\},$$

$$P = \{E \rightarrow E + E, E \rightarrow E * E, E \rightarrow a, E \rightarrow (E)\}.$$

Huom: Vaikka kielioppi G'_{expr} näyttää yksinkertaisemmalta kuin kielioppi G_{expr} , sen ongelmana on ns. rakenteellinen moniselitteisyys, mikä on monesti ei-toivottu ominaisuus.



Vakiintuneita merkintätapoja

Väliksymboleita: A, B, C, \dots, S, T .

Päätemerkkejä: kirjaimet a, b, c, \dots, s, t ;

numerot $0, 1, \dots, 9$;

erikoismerkit; lihavoidut tai alleviivatut varatut sanat (**if**, **for**, **end**, ...).

Mielivaltaisia merkkejä (kun välikkeitä ja päätteitä ei erotella): X, Y, Z .

Päätemerkkijonoja: u, v, w, x, y, z .

Sekamerkkijonoja: $\alpha, \beta, \gamma, \dots, \omega$.



Produktiot, joilla on yhteinen vasen puoli A , voidaan kirjoittaa yhteen: joukon

$$A \rightarrow \omega_1, A \rightarrow \omega_2, \dots, A \rightarrow \omega_k$$

sijaan kirjoitetaan

$$A \rightarrow \omega_1 \mid \omega_2 \mid \dots \mid \omega_k.$$

Kielioppi esitetään usein pelkkänä sääntöjoukkona:

$$\begin{array}{l} A_1 \rightarrow \omega_{11} \mid \dots \mid \omega_{1k_1} \\ A_2 \rightarrow \omega_{21} \mid \dots \mid \omega_{2k_2} \\ \vdots \\ A_m \rightarrow \omega_{m1} \mid \dots \mid \omega_{mk_m}. \end{array}$$

Tällöin päätellään välikesymbolit edellisten merkintäsopimusten mukaan tai siitä, että ne esiintyvät sääntöjen vasempina puolina; muut esiintyvät merkit ovat päätemerkkejä.

Lähtösymboli on tällöin *ensimmäisen säännön vasempana puolena* esiintyvä väliske; tässä siis A_1 .



Eräitä konstruktioita

Olkoon $L(T)$ väliskeestä T johdettavissa olevien päätejonojen joukko. Olkoon meillä produktiokokoelma P , jossa ei esiinny välikettä A , ja jolla B :stä voidaan johtaa $L(B)$ (ja vastaavasti C :stä $L(C)$).

Lisäämällä P :hen jokin seuraavista produktioista saadaan uusia kieliä:

produktio	kieli
$A \rightarrow B \mid C$	yhdiste $L(A) = L(B) \cup L(C)$
$A \rightarrow BC$	katenaatio $L(A) = L(B)L(C)$, ja
$A \rightarrow AB \mid \epsilon$ (vasen rekursio) tai $A \rightarrow BA \mid \epsilon$ (oikea rekursio)	Kleenen sulkeuma $L(A) = L(B)^*$



Välikkeiden keskeisupotus on yhteydettömille kielioppeille ominainen konstruktio, joka tekee usein (muttei aina) kielestä epäsäännöllisen: lisäämällä produktio

$A \rightarrow BAC \mid \epsilon$ saadaan

$$L(A) = \bigcup_{i=0}^{\infty} L(B)^i L(C)^i.$$



3.2 Säännölliset kielet ja yhteydettömät kieliopit

Yhteydettömällä kielioppeilla voidaan siis kuvata joitakin ei-säännöllisiä kieliä (esimerkiksi kielet L_{match} ja L_{expr}). Osoitetaan, että myös kaikki säännölliset kielet voidaan kuvata yhteydettömällä kielioppeilla. Yhteydettömät kielet ovat siten säännöllisten kielten aito ylituokka.

Yhteydetön kielioppi on *oikealle lineaarinen*, jos sen kaikki produktiot ovat muotoa $A \rightarrow aB$ tai $A \rightarrow \epsilon$, ja *vasemmalle lineaarinen*, jos sen kaikki produktiot ovat muotoa $A \rightarrow Ba$ tai $A \rightarrow \epsilon$.

Osoittautuu, että sekä vasemmalle että oikealle lineaarisilla kielioppeilla voidaan tuottaa täsmälleen säännölliset kielet, minkä takia näitä kielioppeja nimitetään myös yhteisesti *säännöllisiksi*. Todistetaan tässä väite vain oikealle lineaarisille kielioppeille.



Lause 3.1 Jokainen säännöllinen kieli voidaan tuottaa oikealle lineaarisella kieliopilla.

Todistus. Olkoon L aakkoston Σ säännöllinen kieli, ja olkoon $M = (Q, \Sigma, \delta, q_0, F)$ sen tunnistava (deterministinen tai epädeterministinen) äärellinen automaatti. Muodostetaan kielioppi G_M , jolla on $L(G_M) = L(M) = L$.

Kieliopin G_M pääteaakkosto on sama kuin M :n syöteaakkosto Σ , ja sen välikeaakkostoon otetaan yksi välike A_q kutakin M :n tilaa q kohden. Kieliopin lähtösymboli on A_{q_0} , ja sen produktiot vastaavat M :n siirtymiä:

(i) kutakin M :n lopputilaa $q \in F$ kohden kielioppiin otetaan produktio $A_q \rightarrow \varepsilon$;

(ii) kutakin M :n siirtymää $q \xrightarrow{a} q'$ (so. $q' \in \delta(q, a)$) kohden kielioppiin otetaan produktio $A_q \rightarrow aA_{q'}$.



Konstruktion oikeellisuuden tarkastamiseksi merkitään välikkeestä A_q tuotettavien päätejonojen joukkoa

$$L(A_q) = \{x \in \Sigma^* \mid A_q \xRightarrow{G_M}^* x\}.$$

Induktiolla merkkijonon x pituuden suhteen voidaan osoittaa, että kaikilla q on

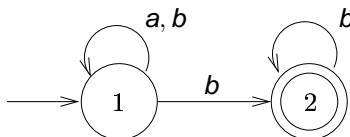
$$x \in L(A_q) \text{ joss } (q, x) \vdash_M^*(q_f, \varepsilon) \text{ jollakin } q_f \in F.$$

Erityisesti on siis

$$\begin{aligned} L(G_M) = L(A_{q_0}) &= \{x \in \Sigma^* \mid (q_0, x) \vdash_M^*(q_f, \varepsilon) \\ &\text{jollakin } q_f \in F\} \\ &= L(M) = L. \quad \square \end{aligned}$$



Esimerkki. Automaatti:



Vastaava kielioppi:

$$A_1 \rightarrow aA_1 \mid bA_1 \mid bA_2$$

$$A_2 \rightarrow \varepsilon \mid bA_2.$$



Lause 3.2 Jokainen oikealle lineaarisella kieliopilla tuotettava kieli on säännöllinen.

Todistus. Olkoon $G = (V, \Sigma, P, S)$ oikealle lineaarinen kielioppi. Muodostetaan kielen $L(G)$ tunnistava epädeterministinen äärellinen automaatti $M_G = (Q, \Sigma, \delta, q_S, F)$ seuraavasti: M_G :n tilat vastaavat G :n välikkeitä:

$$Q = \{q_A \mid A \in V - \Sigma\}.$$

M_G :n alkutila on lähtösymbolia S vastaava tila q_S .

M_G :n syöteaakkosto on G :n pääteaakkosto Σ .

M_G :n siirtymäfunktio δ jäljittelee G :n produktioita siten, että kutakin produktiota $A \rightarrow aB$ kohden automaatissa on siirtymä $q_A \xrightarrow{a} q_B$ (so. $q_B \in \delta(q_A, a)$).



M_G :n lopputiloja ovat ne tilat, joita vastaaviin välikkeisiin liittyy G :ssä ε -produktio:

$$F = \{q_A \in Q \mid A \rightarrow \varepsilon \in P\}.$$

Konstruktion oikeellisuus voidaan jälleen tarkastaa induktiolla G :n tuottamien ja M_G :n hyväksymien merkkijonojen pituuden suhteen. \square