2.8 Säännöllisten kielten rajoituksista

Kardinaliteettisyistä 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.

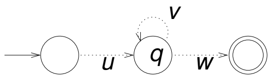
Esimerkki: sulkulausekekieli

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

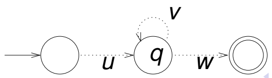
Formalisointi: “pumppauslemma”.

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 $uv^iw \in A$ kaikilla $i = 0, 1, 2, \ldots$

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 $uv^iw \in A$ kaikilla $i = 0, 1, 2, \ldots$ $\square$



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

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

Mutta esimerkiksi “0-kertaisesti” pumpattaessa:

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

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 uudelleenkirjoitusaskelessa korvataan yksi erityinen muuttuja- t. *välikesymboli* jollakin siihen liitetyllä 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 kieliopeilla voidaan kuvata (tuottaa) myös ei-säännöllisiä kieliä.

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

(i) $S \to \varepsilon$,
(ii) $S \to (S)$.

Esimerkiksi merkkijonon ((())) tuottaminen:

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

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

$$\begin{array}{lcl|l} E & \to & T & E+T \\ T & \to & F & T*F \\ F & \to & a & (E). \end{array}$$

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

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

Määritelmä 3.1 *Yhteydetön kielioppi* on nelikko

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

missä

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

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

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

$$\gamma \underset{G}{\Rightarrow} \gamma'$$

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

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

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

$$\gamma \underset{G}{\Rightarrow}^* \gamma'$$

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

$$\gamma = \gamma_0 \underset{G}{\Rightarrow} \gamma_1 \underset{G}{\Rightarrow} \ldots \underset{G}{\Rightarrow} \gamma_n = \gamma'.$$

Erikoistapauksena $n = 0$ saadaan $\gamma \underset{G}{\Rightarrow}^* \gamma$ 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 $S \underset{G}{\Rightarrow}^* \gamma$.

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

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

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

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

$$\begin{aligned} V &= \{E, T, F, a, +, *, (,)\}, \\ \Sigma &= \{a, +, *, (,)\}, \\ P &= \{E \rightarrow T,\ E \rightarrow E + T,\ T \rightarrow F,\ T \rightarrow T * F, \\ &\quad\ F \rightarrow a,\ F \rightarrow (E)\}. \end{aligned}$$

Toinen kielioppi kielen L_{expr} tuottamiseen on

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

missä

$$\begin{aligned} V &= \{E, a, +, *, (,)\}, \\ \Sigma &= \{a, +, *, (,)\}, \\ P &= \{E \rightarrow E + E,\ E \rightarrow E * E,\ E \rightarrow a,\ E \rightarrow (E)\}. \end{aligned}$$

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

Välikesymboleita: $A, B, C, \ldots, S, T$.

Päätemerkkejä: kirjaimet $a, b, c, \ldots, s, t$;
numerot $0, 1, \ldots, 9$;
erikoismerkit; lihavoidut tai alleviivatut varatut sanat (**if**, **for**, **end**, $\ldots$).

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

Päätemerkkijonoja: u, v, w, x, y, z.

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

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

$$A \to \omega_1,\ A \to \omega_2,\ \ldots A \to \omega_k$$

sijaan kirjoitetaan

$$A \to \omega_1 \mid \omega_2 \mid \ldots \mid \omega_k.$$

Kielioppi esitetään usein pelkkänä sääntöjoukkona:

$$\begin{array}{lclclcl} A_1 & \to & \omega_{11} & \mid & \ldots & \mid & \omega_{1k_1} \\ A_2 & \to & \omega_{21} & \mid & \ldots & \mid & \omega_{2k_2} \\ \vdots & & & & & & \\ A_m & \to & \omega_{m1} & \mid & \ldots & \mid & \omega_{mk_m}. \end{array}$$

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 välike; tässä siis A_1.

Eräitä konstruktioita

Olkoon $L(T)$ välikkeestä T johdettavissa olevien päätejonojen joukko. Olkoon annettu 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
$A \to B \mid C$	yhdiste $L(A) = L(B) \cup L(C)$
$A \to BC$	katenaatio $L(A) = L(B)L(C)$, ja
$A \to AB \mid \varepsilon$ (vasen rekursio) tai $A \to BA \mid \varepsilon$ (oikea rekursio)	Kleenen sulkeuma $L(A) = L(B)^*$

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

$A \to BAC \mid \varepsilon$ saadaan

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

3.2 Säännölliset kielet ja yhteydettömät kieliopit

Yhteydettömillä kieliopeilla voidaan siis kuvata joitakin ei-säännöllisiä kieliä (esimerkiksi kielet L_{match} ja L_{expr}). Osoitetaan, että myös kaikki säännölliset kielet voidaan kuvata yhteydettömillä kieliopeilla. Yhteydettömät kielet ovat siten säännollisten kielten aito yliluokka.

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

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

Lause 3.1 Jokainen säännöllinen kieli voidaan tuottaa oikealle lineaarisella kieliopilla.

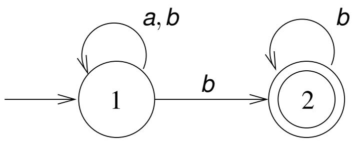
Todistus. Olkoon L aakkoston Σ säännöllinen kieli, ja olkoon $M = (Q, \Sigma, \delta, q_0, F)$ sen tunnistava (deterministinen tai epädeterministinen) äärellinen automaatti. Muodostetaan kielioppi G_M, jolla on $L(G_M) = L(M) = L$.

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 $A_q \to \varepsilon$;

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

Esimerkki. Automaatti:



Vastaava kielioppi:

$$\begin{aligned} A_1 &\to aA_1 \mid bA_1 \mid bA_2 \\ A_2 &\to \varepsilon \mid bA_2. \end{aligned}$$

Konstruktion oikeellisuuden tarkastamiseksi merkitään välikkeestä A_q tuotettavien päätejonojen joukkoa

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

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

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

Erityisesti on siis

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

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 $A \to aB$ kohden automaatissa on siirtymä $q_A \xrightarrow{a} q_B$ (so. $q_B \in \delta(q_A, a)$).

T–79.1001/1002 Tietojenkäsittelyteorian perusteet

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

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

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

Pekka Orponen kevät 2006