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.

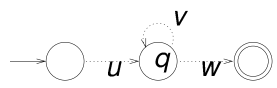
Esimerkki: sulkulausekekieli

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

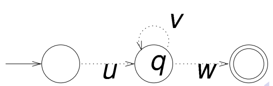
Formalisointi: "pumppauslemma".

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):

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

Oletetaan, että L olisi säännöllinen. Tällöin pitäisi pumppauslemman mukaan olla jokin $n \geq 1$, jota pitempiä L:n merkkijonoja voidaan pumpata. Valitaan $x = a^nb^n$, 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:

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

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).

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

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

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

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**, ...).

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

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

$A \longrightarrow BAC \mid \epsilon$ saadaan

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