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.

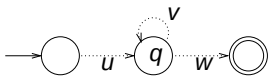
Esimerkki: sulkulausekekieli

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

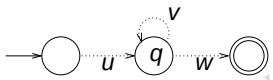
Formalisointi: "pumppauslemma".

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



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:

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

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 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**, ...).

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:

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

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

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

3.3 KIELIOPPIEN JÄSENNYSONGELMA

Ratkaistava tehtävä:

"Annettu yhteydetön kielioppi G ja merkkijono x. Onko $x \in L(G)$?"

Ratkaisumenetelmä = *jäsennysalgoritmi*.

Useita vaihtoehtoisia menetelmiä, erityisesti kun G on jotain rajoitettua (käytännössä esiintyvää) muotoa.

Johdot ja jäsennyspuut

Olkoon $\gamma \in V^*$ kieliopin $G = (V, \Sigma, P, S)$ lausejohdos.

Lähtösymbolista S merkkijonoon γ johtavaa suorien johtojen jonoa

$$S = \gamma_0 \Rightarrow \gamma_1 \Rightarrow \cdots \Rightarrow \gamma_n = \gamma$$

sanotaan γ:n *johdoksi* G:ssä.

Johdon *pituus* on siihen kuuluvien suorien johtojen määrä (edellä n).

Esimerkki: lauseen $a + a$ johtoja kieliopissa G_{expr}:

$$\begin{array}{lllllllll}
\text{(i)} & E & \Rightarrow & E+T & \Rightarrow & T+T & \Rightarrow & F+T \\
 & & \Rightarrow & a+T & \Rightarrow & a+F & \Rightarrow & a+a \\
\text{(ii)} & E & \Rightarrow & E+T & \Rightarrow & E+F & \Rightarrow & T+F \\
 & & \Rightarrow & F+F & \Rightarrow & F+a & \Rightarrow & a+a \\
\text{(iii)} & E & \Rightarrow & E+T & \Rightarrow & E+F & \Rightarrow & E+a \\
 & & \Rightarrow & T+a & \Rightarrow & F+a & \Rightarrow & a+a.
\end{array}$$

Johto $\gamma \Rightarrow^* \gamma'$ on *vasen johto*, merkitään

$$\gamma \underset{\text{lm}}{\Rightarrow}{}^* \gamma',$$

jos kussakin johtoaskelessa on produktiota sovellettu merkkijonon vasemmanpuoleisimpaan välikkeeseen (edellä johto (i)).

Vastaavasti määritellään *oikea johto* (edellä (iii)), jota merkitään

$$\gamma \underset{\text{rm}}{\Rightarrow}{}^* \gamma'$$

Suoria vasempia ja oikeita johtoaskelia merkitään $\gamma \underset{\text{lm}}{\Rightarrow} \gamma'$ ja $\gamma \underset{\text{rm}}{\Rightarrow} \gamma'$.

Olkoon $G = (V, \Sigma, P, S)$ yhteydetön kielioppi.

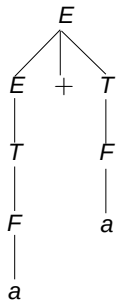
Kieliopin G mukainen *jäsennyspuu* on järjestetty puu, jolla on seuraavat ominaisuudet:

(i) puun solmut on nimetty joukon $V \cup \{\varepsilon\}$ alkioilla siten, että sisäsolmujen nimet ovat välikkeitä (so. joukosta $N = V - \Sigma$) ja juurisolmun nimenä on lähtösymboli S;

(ii) jos A on puun jonkin sisäsolmun nimi, ja $X_1, \ldots, X_k$ ovat sen jälkeläisten nimet järjestyksessä, niin $A \to X_1 \ldots X_k$ on G:n produktio.

Jäsennyspuun τ *tuotos* on merkkijono, joka saadaan liittämällä yhteen sen lehtisolmujen nimet esijärjestyksessä ("vasemmalta oikealle").

Esimerkki. Lauseen $a+a$ jäsennyspuu kieliopissa G_{expr}:



Lauseen johto:

$$\begin{aligned} E &\Rightarrow E+T \Rightarrow T+T \Rightarrow F+T \\ &\Rightarrow a+T \Rightarrow a+F \Rightarrow a+a \end{aligned}$$

Johtoa

$$S = \gamma_0 \Rightarrow \gamma_1 \Rightarrow \cdots \Rightarrow \gamma_n = \gamma$$

vastaavan jäsennyspuun muodostaminen:

(i) puun juuren nimeksi tulee S; jos $n = 0$, niin puussa ei ole muita solmuja; muuten

(ii) jos ensimmäisessä johtoaskelessa on sovellettu produktiota $S \to X_1 X_2 \ldots X_k$, niin juurelle tulee k jälkeläissolmua, joiden nimet vasemmalta oikealle ovat $X_1, X_2, \ldots, X_k$;

(iii) jos seuraavassa askelessa on sovellettu produktiota $X_i \to Y_1 Y_2 \ldots Y_l$, niin juuren i:nnelle jälkeläissolmulle tulee l jälkeläistä, joiden nimet vasemmalta oikealle ovat $Y_1, Y_2, \ldots, Y_l$; ja niin edelleen.

Konstruktiosta huomataan, että jos τ on jotakin johtoa $S \Rightarrow^* \gamma$ vastaava jäsennyspuu, niin τ:n tuotos on γ.

Olkoon τ kieliopin G mukainen jäsennyspuu, jonka tuotos on päätemerkkijono x.

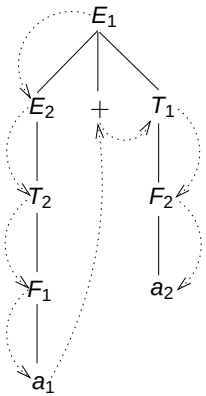
Tällöin τ:sta saadaan vasen johto x:lle käymällä puun solmut läpi esijärjestyksessä ("ylhäältä alas, vasemmalta oikealle") ja laventamalla vastaan tulevat välikkeet järjestyksessä puun osoittamalla tavalla.

Oikea johto saadaan käymällä puu läpi käänteisessä esijärjestyksessä ("ylhäältä alas, oikealta vasemmalle").

Muodostamalla annetusta vasemmasta johdosta $S \underset{\text{lm}}{\Rightarrow}^* x$ ensin jäsennyspuu edellä esitetyllä tavalla, ja sitten jäsennyspuusta vasen johto, saadaan takaisin alkuperäinen johto; vastaava tulos pätee myös oikeille johdoille.

Esimerkki. Lauseen $a+a$ vasemman johdon muodostaminen jäsennyspuusta.

Jäsennyspuu:



Solmut esijärjestyksessä:

$E_1 E_2 T_2 F_1 a_1 + T_1 F_2 a_2$

Vasen johto:

$$\begin{aligned} E &\underset{\text{lm}}{\Rightarrow} E+T \underset{\text{lm}}{\Rightarrow} T+T \underset{\text{lm}}{\Rightarrow} F+T \\ &\underset{\text{lm}}{\Rightarrow} a+T \underset{\text{lm}}{\Rightarrow} a+F \underset{\text{lm}}{\Rightarrow} a+a \end{aligned}$$

Lause 3.3 Olkoon $G = (V, \Sigma, P, S)$ yhteydetön kielioppi. Tällöin:

(i) jokaisella G:n lausejohdoksella γ on G:n mukainen jäsennyspuu τ, jonka tuotos on γ;

(ii) jokaista G:n mukaista jäsennyspuuta τ, jonka tuotos on päätemerkkijono x, vastaavat yksikäsitteiset vasen ja oikea johto $S \underset{\text{lm}}{\Rightarrow}^* x$ ja $S \underset{\text{rm}}{\Rightarrow}^* x$.

Seuraus 3.4 Jokaisella G:n lauseella on vasen ja oikea johto.

Siis: yhteydettömän kieliopin tuottamien lauseiden jäsennyspuut, vasemmat ja oikeat johdot vastaavat yksikäsitteisesti toisiaan.

Jäsennysongelman ratkaisuun katsotaan usein kuuluvan pelkän päätösongelman “Onko $x \in L(G)$?” ratkaisemisen lisäksi jonkin näistä jäsennysesityksistä tuottaminen.

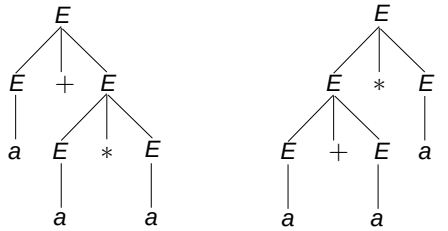
Kieliopin moniselitteisyys

Lauseella voi olla kieliopissa useita jäsennyksiä.

Esimerkki. Tarkastellaan yksinkertaisten aritmeettisten lausekkeiden kielioppia:

$$G'_{\text{expr}} = \{E \to E + E,\ E \to E * E,\ E \to a,\ E \to (E)\}.$$

Lauseella $a + a * a$ on tässä kieliopissa kaksi jäsennystä:



Yhteydetön kielioppi G on *moniselitteinen*, jos jollakin G:n lauseella x on kaksi erilaista G:n mukaista jäsennyspuuta. Muuten kielioppi on *yksiselitteinen*.

Moniselitteisyys on tietojenkäsittelysovelluksissa yleensä ei-toivottu ominaisuus, koska se merkitsee että annetulla lauseella on kaksi vaihtoehtoista “tulkintaa.”

Yhteydetön kieli, jonka tuottavat kieliopit ovat kaikki moniselitteisiä, on *luonnostaan moniselitteinen*.

Esimerkiksi kielioppi G'_{expr} on moniselitteinen, kieliopit G_{expr} ja G_{match} yksiselitteisiä. Kieli $L_{\text{expr}} = L(G'_{\text{expr}})$ ei ole luonnostaan moniselitteinen, koska sillä on myös yksiselitteinen kielioppi G_{expr}. Luonnostaan moniselitteinen on esimerkiksi kieli

$$\{a^i b^j c^k \mid i = j \text{ tai } j = k\}.$$

(Todistus sivuutetaan.)