Luku 4

Kontekstittomat kielet ja pinoautomaatit

- Kuten edellä nähtiin, esim. tasapainoisten sulkulausekkeiden muodostama kieli

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

  if-else-parien muodostama kieli

  $L_{if-else} = \{\text{if}^k\text{else}^l | l \leq k\}$

  eivät ole säännöllisiä (Pumppauslemma)

- Mutta näitä kieliä voi kuvata *kontekstittomilla kieliopeilla*
- Kuten edellä säännöllistiä kieliä tunnistava kone oli äärellinen automaatti, kontekstitonta kieltä tunnistava kone on *pinoautomaatti*

4.1 Kontekstittomat kieliopit ja kielet

- Esim. sulkulausekekielellä ysinkertainen rekursiivinen kuvaus: Merk. $S =$ "mielivaltainen tasapainoinen sulkumerkkijono". Tällöin S on tasapainoinen sulkumerkkijono, jos

  (i) $S = \epsilon$ tai
  (ii) S on muotoa (S'), missä S' on tasapainoinen sulkumerkkijono

- Toinen kuvaustapa: seuraavat *muunnossäännöt tuottavat* täsmälleen kielen L_{match} merkkijonot symbolista S:

  (i) $S \to \epsilon$,
  (ii) $S \to (S)$

- Esim. merkkijonon $((()))$ tuottaminen:

$$S \Rightarrow (S) \Rightarrow ((S)) \Rightarrow (((S))) \Rightarrow (((\epsilon))) = ((()))$$

- Kontekstiton kielioppi on muunnossysteemi, jossa kuvattavat merkkijonot tuotetaan korvaamalla erityisiä muuttuja- t. *välikesymboleita* annettujen sääntöjen mukaan yksi kerrallaan, symbolia ympäröivän merkkijonon rakenteesta riippumatta

- Merkinnällä

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

  lyhennetään

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

- Esim. Yksinkertainen kielioppi tietyille aritmeettisille lausekkeille:

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

Esim. 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: Kontekstiton kielioppi (engl. context-free grammar) on nelikko

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

missä

- V on kieliopin aakkosto;
- $\Sigma \subseteq V$ on kieliopin *päätemerkkien* joukko; sen komplementti $N = V \setminus \Sigma$ 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 $A \to \omega$

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

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

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

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

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

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

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

- Erikoistapaus: $n = 0$, $\gamma \underset{G}{\Rightarrow}^* \gamma$ millä tahansa $\gamma \in V^*$

- Merkkijono $\gamma \in V^*$ on kieliopin G *lausejohdos*, jos on $S \underset{G}{\Rightarrow}^* \gamma$
- G:n *lause* on pelkästään päätemerkeistä koostuva G:n lausejohdos $x \in \Sigma^*$
- Kieliopin G *tuottama* t. *kuvaama kieli* koostuu G:n lauseista:

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

Määritelmä: Formaali kieli $L \subseteq \Sigma^*$ on *kontekstiton*, jos se voidaan tuottaa jollakin kontekstittomalla kieliopilla

- Esim. Tasapainoisten sulkujonojen muodostaman kielen $L_{\text{match}} = \{(^k)^k \mid k \geq 0\}$ tuottaa kielioppi

$$G_{\text{match}} = (\{S, (,)\}, \{(,)\}, \{S \to \epsilon, S \to (S)\}, S)$$

- Esim. Aiemmin tarkisteltujen yksinkertaisten aritmeettisten lausekkeiden muodostaman kielen L_{expr} tuottaa kielioppi

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

missä

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

Toinen kielioppi kielen L_{expr} tuottamiseen on

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

missä

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

- Esim. ($\sim$ Orponen teht.44) Tarkastellaan suomen kielen virkettä, joka koostuu yksinkertaisesta päälauseesta + 0 tai useammasta sisäkkäisestä relatiivilauseesta:

$$L_{rel} = \{subj(\text{ joka } pred\ attr\ obj)^*\ pred\ attr\ obj\}$$

Tällaisia virkkeitä voidaan tuottaa esim. seuraavilla kontekstittoman kieliopin G_{rel} säännöillä:

(Esitetään yksinkertaisuuden vuoksi säännöt merkkijonoilla)

VIRKE $\rightarrow$ SUBJ SL PRED ATTR OBJ
SL $\rightarrow$ joka PRED ATTR OBJ SL | ϵ
SUBJ $\rightarrow$ poika | tyttö | jänis | susi | peikko
PRED $\rightarrow$ pelkäsi | metsästi
ATTR $\rightarrow$ suurta | pientä | vihaista | hirmuista | arkaa
OBJ $\rightarrow$ poikaa | tyttöä | jänistä | sutta | peikkoa

Kieleen kuuluvat mm. seuraavat virkkeet:

  - poika joka metsästi sutta joka pelkäsi peikkoa joka pelkäsi suurta tyttöä pelkäsi hirmuista jänistä
  - tyttö joka metsästi arkaa poikaa pelkäsi vihaista jänistä

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

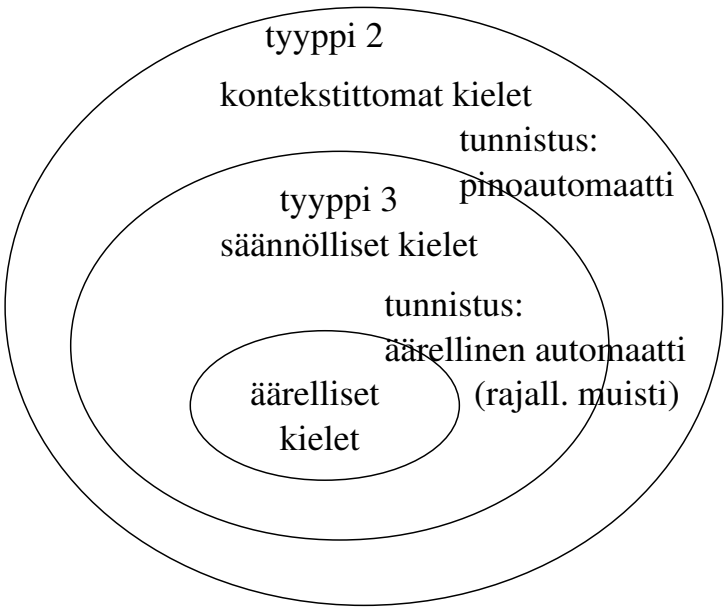
- Kielioppi esitetään usein pelkkänä sääntöjoukkona:

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

- 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

4.1.1 Säännölliset kielet ja kontekstittomat kieliopit

- Kontekstittomilla kieliopeilla voidaan kuvata joitakin ei-säännöllisiä kieliä (esimerkiksi kielet L_{match} ja L_{expr})
- Osoitetaan, että myös kaikki säännölliset kielet voidaan kuvata kontekstittomilla kieliopeilla
- Kontekstittomat kielet ovat siten säännollisten kielten aito yliluokka



4.1.2 Oikealle ja vasemmalle lineaariset kieliopit

- *Määritelmä:* Kontekstiton kielioppi on *oikealle lineaarinen*, jos sen kaikki produktiot ovat muotoa $A \to \epsilon$ tai $A \to aB$, ja *vasemmalle lineaarinen*, jos sen kaikki produktiot ovat muotoa $A \to \epsilon$ tai $A \to Ba$

- Osoittautuu, että sekä vasemmalle että oikealle lineaarisilla kieliopeilla voidaan tuottaa täsmälleen säännölliset kielet

- $\Rightarrow$ lineaarisia kielioppeja nimitetään myös yhteisesti *säännöllisiksi* kieliopeiksi

- Todistetaan tässä väite vain oikealle lineaarisille kieliopeille:

Lause: 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 $L(G_M) = L(M) = L$.

- G_M:n pääteaakkosto = M:n syöteaakkosto Σ

- G_M:n välikeaakkostoon otetaan yksi välike A_q kutakin M:n tilaa q kohden.

- G_M:n lähtösymboli on A_{q_0}

- G_M:n produktiot vastaavat M:n siirtymiä:

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

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

- Tarkastetaan konstruktion oikeellisuus:

  - Merk. A_q:sta tuotettavien päätejonojen joukkoa

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

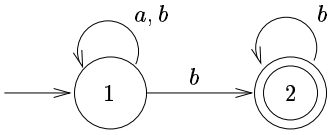
  - Induktiolla merkkijonon x pituuden suhteen voidaan osoittaa, että kaikilla q on

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

– Erityisesti on siis

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

Esim.



Kuva 4.1: Yksinkertainen äärellinen automaatti.

Automaattia vastaava kielioppi on:

$$\begin{aligned} A_1 &\rightarrow aA_1 \mid bA_1 \mid bA_2 \\ A_2 &\rightarrow \epsilon \mid bA_2 \end{aligned}$$

Lause: 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 $q_A \overset{a}{\rightarrow} q_B$ (so. $q_B \in \delta(q_A, a)$)

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

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

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

Esimerkkejä

Esimerkkejä säännöllisistä lausekkeista ja vastaavista kontekstittomista kieliopeista:

1. Lauseke: a^*
   Kielioppi:$S \rightarrow aS|\epsilon$

2. Lauseke: $a^+ = aa^*$
   Kielioppi: $S \rightarrow aS|a$

3. Lauseke: $(aa)^*$
   Kielioppi: $S \rightarrow aSa|\epsilon$
   (merkkijono koostuu vain parillisesta määrästä a:ta)

4. Lauseke: $(b^*ab^*ab^*)^*$
   Kielioppi:
   $S \rightarrow BaBaBS|\epsilon$
   $B \rightarrow bB|\epsilon$
   (merkkijono sisältää parillisen näärän a:ta, lisäksi saa olla b:tä missä tahansa)

5. Lauseke: $(0 \bigcup 1 \bigcup ... \bigcup 9)(0 \bigcup 1 \bigcup ... \bigcup 9)^*$
   Kielioppi:
   $S \rightarrow DN$
   $N \rightarrow DN|\epsilon$
   $D \rightarrow 0|1|...|9$
   (vähintään yhdestä digitistä koostuva numero)

4.2 *Ekskursio: Kasvikieliopit (Lindenmayer Systems)

Esimerkki 1 Olkoon symbolit $\Sigma = \{SIEMEN, SIRKKALEHDET, LEHDET, VARSI, NUPPU, PUNKUKKA, SINKUKKA\}$. Kukkaketo voidaan nyt esittää Σ^*:n merkkijonona. Kukin kasvi kehittyy seuraavien sääntöjen mukaan:

SIEMEN $\rightarrow$ SIRKKALEHDET
SRKKALEHDET $\rightarrow$ LEHDET | VARSI
LEHDET $\rightarrow$ VARSI
VARSI $\rightarrow$ NUPPU | LEHDET
NUPPU $\rightarrow$ PUNKUKKA | SINKUKKA
PUNKUKKA $\rightarrow$ SIEMEN | SIEMEN SIEMEN $|\epsilon$
SINKUKKA $\rightarrow$ SIEMEN | SIEMEN SIEMEN $|\epsilon$

missä ϵ kuvaa kasvin kuolemaa (se katoaa kylvämättä edes siemeniä).

Kyseessä on *kontekstiton kielioppi*, sillä säännön seuraus määräytyy ainoastaan edeltävän symbolin perusteella (jos useita vaihtoehoisia seurauksia, niin valitaan jokin niistä).

Esimerkki 2 Laajennetaan symbolijoukkoa: $\Sigma_2 = \Sigma \bigcup \{PAIVA, YO\}$ ja määritellään uudet säännöt:

SIEMEN $\rightarrow$ SIRKKALEHDET
SRKKALEHDET PAIVA$\rightarrow$ LEHDET | VARSI
LEHDET PAIVA$\rightarrow$ VARSI
VARSI PAIVA$\rightarrow$ NUPPU | LEHDET
NUPPU YO$\rightarrow$ PUNKUKKA | SINKUKKA
PUNKUKKA $\rightarrow$ SIEMEN | SIEMEN SIEMEN $|\epsilon$
SINKUKKA $\rightarrow$ SIEMEN | SIEMEN SIEMEN $|\epsilon$
PAIVA $\rightarrow$ YO
YO $\rightarrow$ PAIVA

Nyt kukat kasvavat vain päivisin, mutta kukat aukeavat vain öisin. Tämä kielioppi on *kontekstillinen*, sillä myös symbolin konteksti (ympäristötekijät) vaikuttavat säännön seuraukseen.

- L-järjestelmät $\sim$ yksinkertaistettu abstrakti kielioppi, jonka tuottama kieli kuvaa kasvien (tms. elollisten organismien) kasvua ja kehitystä
- alkujaan botanisti Aristid Lindenmayerin kehittämä matemaattinen malli, jolla voi ennustaa kasvien kasvua
- sittemmin käytetty keinotekoisten organismien luomiseen ja kaikkeen mahdoliseen!
- L-järjestelmä tuottaa vain kieliopin mukaisen merkkijonon $\rightarrow$ voidaan tulkita graafisena kuvana
- muistuttavat fraktaaleja, mutta eivät välttämättä fraktaaleja
- Yleinen idea: jokaisella aika-askelella merkkijonon symboli lavennetaan naapurisymbolien ja ehtoihin sopivan säännön perusteella (voi pysyä ennallaankin) $\leftrightarrow$ *s*oluautomaatti, jossa kukin hilan (vektorin, 2- tai useampiulotteisen taulukon) solu on äärellinen automaatti
- Leikkaavat Chomskyn hierarkian kaikkia luokkia:
  - voivat noudattaa säännöllisen kielen sääntöjä, ts. ovat muotoa $A \rightarrow aB$ tai $A \rightarrow Ba$ (muunnoksen tuloksena yksi päätesymboli ja yksi välikesymboli) tai $A \rightarrow \epsilon$ (”solun kuolema”, symboli katoaa) tai
  - kontekstitonta kielioppia, ts. säännöt muotoa $A \rightarrow V^*$ (ts. muunnetaan vain yhtä symbolia kerrallaan, riippumatta sen naapureista, muunnoksen tuloksena voi tulla mitä tahansa välike- ja päätesymboleja, myös ϵ) tai
  - kontekstillista kielioppia, ts. säännöt muotoa $\alpha A \beta \rightarrow \alpha \omega \beta$ (ts. muunnos vain tietyssä ”kontekstissa”, kun naapureina α ja β)
  - rajoittamatonta kielioppia, ts. säännöt muotoa $\omega \rightarrow \omega'$ (ts. mikä tahansa merkkijono voi muuntua miksi tahansa merkkijonoksi, merkkijonot voivat sisältää niin pääte- kuin välikesymboleja)

- esim. erään levämuodon kasvun mallinnus:
  - alkuvaiheessa voi olla kahdenlaisia soluja, merk. A ja B (”aksioomat”)
  - säännöt:
    $A \to AB$
    $B \to A$
  - esim. $AB \Rightarrow ABA \Rightarrow ABAAB \Rightarrow ABAABABA$ (3 aika-askelta)
- yksinkertaisimmat L-järjestelmät ns. DOL-järjestelmiä (kontekstiton ja deterministinen L-järjestelmä)
- esim. 2: säännöt:
  $A \to CB$
  $B \to A$
  $C \to DA$
  $D \to C$
  - lähtökohta: mikä tahansa aakkoston sallittu ”siemen” (lähtösymboli)
  - esim. $A \Rightarrow CB \Rightarrow DAA \Rightarrow CCBCB \Rightarrow DADAADAA$
- esim. 3: Puurakenteen muodostaminen:
  $A \to C[B]D$
  $B \to A$
  $C \to C$
  $D \to C(E)A$
  $E \to D$
  - tulkitaan hakasulut haarana vasemmalle ja tavalliset sulut haarana oikealle
  - esim. $A \Rightarrow C[B]D \Rightarrow C[A]C(E)A$ (2 aika-askelta)
  - puuna:

- Sovellus (WH): muodostetaan kukkaketoa grafiikkaohjelmalla, joka noudattaa seuraavia sääntöjä:
  $SIEMEN \rightarrow SIRKKALEHDET|VVARSI|OVARSI|KVARSI$
  $VVARSI \rightarrow VLEHTI|VNUPPU$ (satunnainen väri)
  $OVARSI \rightarrow OLEHTI|ONUPPU$
  $VNUPPU \rightarrow KUKKA$
  $ONUPPU \rightarrow KUKKA$
  $KNUPPU \rightarrow KUKKA$
  $KUKKA \rightarrow SIRKKALEHDET$
  $KVARSI \rightarrow VVARSI|OVARSI|KNUPPU|SIRKKALEHDET$
  $|VKIELONVARSI|OKIELONVARSI|HAVU$
  $SIRKKALEHTI \rightarrow VVARSI|OVARSI|KVARSI|LEHDET$
  $VKIELOVARSI \rightarrow VKIELO$
  $OKIELOVARSI \rightarrow OKIELO$
  $VKIELO \rightarrow VMARJAT$
  $OKIELO \rightarrow OMARJAT$

- Kuvaruutu on jaettu soluihin, joita käydään läpi jossain järjestyksessä ja piirretään ko. paikkaan (tai naapurisoluihin) symbolin mahdollinen seuraaja

Kirjallisuutta

- Prusinkiewicz & Lindenmayer: The Algorithmic Beaty of Plants
- Prusinkiewicz & Hanan: Lindenmayer Systems, Fractals, and Plants
- Rozenberg & Salomaa (toim.): Lindenmayer Systems. Impacts on Theoretical Computer Science, Computer Graphics, and Developmental Biology