

2.5 Säännöllisten kielten rajoituksista

- Minkä tahansa aakkoston formaaleja kieliä (= päätös-ongelmia, tunnistusongelmia) on ylinumeroituva määrä
- kun taas säännöllisiä lausekkeita (= merkkijonoja) on numeroituva määrä
- Näin ollen kaikki kielet eivät voi olla säännöllisiä
- Onko olemassa intuitiivista esimerkkiä kielestä, joka ei ole säännöllinen?
- Tasapainoisten sulkujonojen kieli

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

Lause 1.70 (Pumppauslemma) *Olk. A säännöllinen kieli. Tällöin on olemassa $p \geq 1$ s.e. mikä tahansa $w \in A$, $|w| \geq p$, voidaan jakaa osiin $w = x^i y^i z$ s.e.*

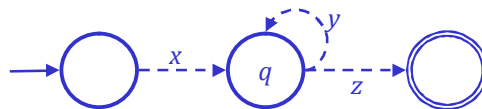
- $|x^i y^i| \leq p$,
- $|y| \geq 1$ ja
- $x^i y^i z \in A \quad \forall i = 0, 1, 2, \dots$

Todistus. Olk. $M = (Q, \Sigma, \delta, q_0, F)$ jokin A :n tunnistava deterministinen äärellinen automaatti s.e. $|Q| = p$. Automaatin tunnistaessa merkkijonoa $w \in A$, $|w| \geq p$, sen täytyy kulkea w :n p :tä ensimmäistä merkkiä käsitellessään jonkin tilan kautta vähintään kaksi kertaa. Olkoon q ensimmäinen sellainen tila.

Valitaan

- x on M :n käsittelemä w :n alkuosa sen tullessa q :hun ensimmäisen kerran,
- y on se osa w :n loppuosasta, jonka M käsittelee ennen seuraavaa paluutaan tilaan q ja
- z on loput merkkijonosta w .

Selvästi $|x^i y^j z| \leq p$, $|y| \geq 1$ ja $x^i y^j z \in A$ kaikilla $i = 0, 1, 2, \dots$



□

Huomio: Pumpauslemma ei sano, että voisimme valita x :n ja y :n haluamallamme tavalla.

Esimerkki

Oletetaan, että L_{sulut} olisi säännöllinen.

Pumppauslemman mukaan on tällöin olemassa jokin p , jota pidempiä merkkijonoja voidaan pumpata. Valitaan $w = (p)^p$, jolloin $|w| = 2p > p$.

Lemman perusteella w voidaan jakaa pumpattavaksi osiin $w = xyz$ s.e. $|xy| \leq p$ ja $|y| \geq 1$. On siis oltava

- $x = (i \quad i \leq p-1,$
- $y = (j \quad j \geq 1$ ja
- $z = (p-(i+j))^p.$

Oletuksen perusteella $xy^kz \in L_{\text{sulut}}$ kaikilla $k = 0, 1, 2, \dots$, mutta esim.

$$xy^0z = xz = (i (p-(i+j))^p = (p-j)^p \notin L_{\text{sulut}}$$

sillä $p-j \neq p$ koska $j \geq 1$. Siten L_{sulut} ei voi olla säännöllinen

3. Kontekstittomat kielet

- Tasapainoisten sulkumerkkijonojen kieli ei ole säännöllinen
- Toisaalta se on kuvattavissa seuraavin *muunnossäännöin*
 1. $S \rightarrow \varepsilon$ ja
 2. $S \rightarrow (S)$
- Nämä *produktiosäännöt* tuottavat kielen L_{sulut} merkkijonot symbolista S

$$S \xrightarrow{2} (S) \xrightarrow{2} ((S)) \xrightarrow{2} (((S))) \xrightarrow{1} (((\varepsilon))) = ((()))$$

- Kuvattava merkkijono tuotetaan korvaamalla *välikesymboleita* yksi kerrallaan annettujen sääntöjen mukaan
- Symbolia ympäröivän merkkijonon rakenne ei määrää käytettyä muunnossääntöä \Leftrightarrow *kontekstiton kielioppi*
- Usein käytetään lyhennysmerkintää

$$A \rightarrow w_1 \mid \dots \mid w_k$$

kuvaamaan välikesymboliin A liittyviä vaihtoehtoisia sääntöjä

$$A \rightarrow w_1, \dots, A \rightarrow w_k$$

- $S \rightarrow \varepsilon \mid (S)$

Yksinkertaiset aritmeettiset lausekkeet
(E = expression, T = term ja F = factor)

$$E \rightarrow E + T \mid T$$

$$T \rightarrow T \times F \mid F$$

$$F \rightarrow (E) \mid a$$

Lausekkeen $(a + (a)) \times a$ tuottaminen

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

Määritelmä Kontekstiton kielioppi on nelikko $G = (V, \Sigma, R, S)$, missä

- V on kieliopin *muuttujien* eli *välikemerkkien* joukko,
- Σ on kieliopin *päätemerkkien* joukko, se on pistevieras V :n kanssa,
- $V \cup \Sigma$ on G :n *aakkosto*,
- $R \subseteq V \times (V \cup \Sigma)^*$ on kieliopin *sääntöjen* joukko ja
- $S \in V$ on kieliopin *lähtösymboli*

$(A, w) \in R$ merk. $A \rightarrow w$

- Olk. $G = (V, \Sigma, R, S)$ ja merkkijonot $u, v, w \in (V \cup \Sigma)^*$ ja $A \rightarrow w$ produktio R :ssä
- uAv tuottaa suoraan merkkijonon uvw kieliopissa G ,
 $uAv \succ_G uvw$
- Merkkijono u tuottaa merkkijonon v kieliopissa G ,
 $u \ggg_G v$,
jos on olemassa jono $u_1, u_2, \dots, u_k \in (V \cup \Sigma)^*$ ($k \geq 0$) s.e.
 $u \succ_G u_1 \succ_G u_2 \succ_G \dots \succ_G u_k \succ_G v$
- $k = 0$: $u \ggg_G u$ millä tahansa $u \in (V \cup \Sigma)^*$

- $u \in (V \cup \Sigma)^*$ on G :n *lausejohdos*, jos
 $S \ggg_G u$
- Pelkistä päätemerkeistä koostuva lausejohdos $w \in \Sigma^*$ on G :n *lause*
- G :n *tuottama kieli* koostuu lauseista
 $L(G) = \{ w \in \Sigma^* \mid S \ggg_G w \}$
- Formaali kieli $L \subseteq \Sigma^*$ on *kontekstiton*, jos se voidaan tuottaa kontekstittomalla kieliopilla

Kontekstiton kielioppi on *oikealle lineaarinen*, jos kaikki sen produktiot ovat muotoa $A \rightarrow \varepsilon$ tai $A \rightarrow aB$

Lause Jokainen säännöllinen kieli voidaan tuottaa oikealle lineaarisella kontekstittomalla kieliopilla

Lause Jokainen oikealle lineaarinen kontekstiton kieli on säännöllinen

- Näin ollen siis oikealle lineaarisilla kieliopilla voidaan tuottaa täsmälleen säännölliset kielet
- Kuitenkin on olemassa kontekstittomia kieliä, jotka eivät ole säännöllisiä; esim. tasapainoisten sulkumerkkijonojen muodostama kieli
- Näin ollen *kontekstittomat kielet ovat säännöllisten kielten aito ylliluokka*

Moniselitteisyys

- Lähtösymbolista S merkkijonoon w johtavaa suorien johtojen jonoa

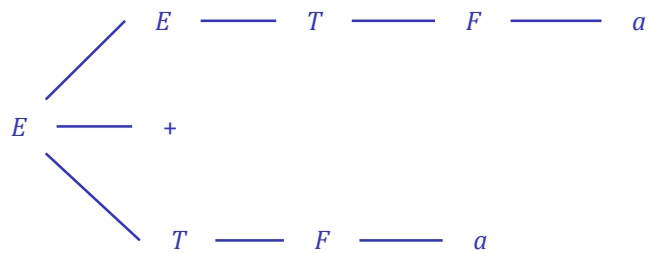
$$S \triangleright w_1 \triangleright \dots \triangleright w_k \triangleright w$$

sanotaan w :n *johdoksi*

Aritmeettisten lausekkeiden kieliopissa lause $a+a$ voidaan johtaa monin tavoin:

1. $E \triangleright E+T \triangleright T+T \triangleright F+T \triangleright a+T \triangleright a+F \triangleright a+a$
2. $E \triangleright E+T \triangleright E+F \triangleright T+F \triangleright F+F \triangleright F+a \triangleright a+a$
3. $E \triangleright E+T \triangleright E+F \triangleright E+a \triangleright T+a \triangleright F+a \triangleright a+a$

- Välikkeiden laventamisjärjestyksen aiheuttamat erot abstrahoituvat pois tarkasteltaessa *jäsennyspuuta*



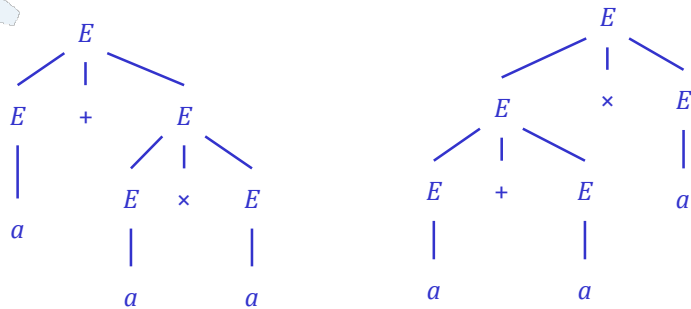
- Kontekstiton kielioppi G on *moniselitteinen*, jos jollakin G :n lauseella on kaksi erilaista jäsenyyspuuta
- Muuten kielioppi on *yksiselitteinen*
- Kieli, jonka tuottavat kieliopit ovat kaikki moniselitteisiä, on *luonnostaan* moniselitteinen
- Esim. kieli $\{a^i b^j c^k \mid i=j \vee j=k\}$ on luonnostaan moniselitteinen

Vaihtoehtoinen kielioppi yksinkertaisille aritmeettisille lausekkeille

$$E \rightarrow E + E \mid E \times E \mid (E) \mid a$$



$$a + a \times a$$



Chomskyn normaalimuoto



Kontekstiton kielioppi on Chomskyn normaalimuodossa (CNF), jos

- väliskeistä enintään S on tyhjentyvä,
- produktiot ovat muotoa $A \rightarrow BC$ tai $A \rightarrow a$ (paitsi mahd. $S \rightarrow \epsilon$) ja
- lähtösymboli S ei esiinny minkään produktion oikealla puolella

Lause 2.9 Mistä tahansa kontekstittomasta kieliopista voidaan muodostaa ekvivalentti CNF-kielioppi.

Todistus Konstruktio etenee vaiheittain. Ensin kieleen lisätään uusi lähtösymboli, sen jälkeen poistetaan ϵ -produktiot ja yksikköproduktiot.

ϵ -produktioiden poistaminen

Lemma *Mistä tahansa kontekstittomasta kieliopista voidaan muodostaa ekvivalentti kielioppi, jossa enintään lähtösymboli on tyhjentyvä.*

Todistus

Olk. $G = (V, \Sigma, R, S)$.

G :n tyhjentyvät välikkeet:

- $NULL = \{A \in V \mid A \rightarrow \epsilon \in R\}$
- Toista kunnes joukko $NULL$ ei enää kasva

$$NULL += \{A \in V \mid A \rightarrow B_1 \dots B_k \in R, B_i \in NULL \forall i = 1, \dots, k\}$$

Korvataan kukin G :n produktio $A \rightarrow X_1 \dots X_k$ kaikkien sellaisten produktioiden joukolla, jotka ovat muotoa $A \rightarrow \alpha_1 \dots \alpha_k$, missä

$$\alpha_i = \begin{cases} X_i & \text{jos } X_i \notin NULL \\ X_i | \epsilon & \text{jos } X_i \in NULL \end{cases}$$

Lopuksi poistetaan kaikki muotoa $A \rightarrow \epsilon$ olevat produktiot.

Jos myös $S \rightarrow \epsilon$ on poistettavien joukossa, niin otetaan kieliopille uusi lähtösymboli S' ja sille produktiot $S' \rightarrow S \mid \epsilon$.

□

$$\begin{aligned} S &\rightarrow A \mid B \\ A &\rightarrow aBa \mid \varepsilon \\ B &\rightarrow bAb \mid \varepsilon \end{aligned}$$

$$\text{NULL} = \{A, B, S\}$$

$$\begin{aligned} S &\rightarrow A \mid B \mid \varepsilon \\ A &\rightarrow aBa \mid aa \mid \varepsilon \\ B &\rightarrow bAb \mid bb \mid \varepsilon \end{aligned}$$

$$\begin{aligned} S' &\rightarrow S \mid \varepsilon \\ S &\rightarrow A \mid B \\ A &\rightarrow aBa \mid aa \\ B &\rightarrow bAb \mid bb \end{aligned}$$

Yksikköproduktioiden poistaminen

Yksikköproduktio on muotoa $A \rightarrow B$, missä A ja B ovat välitteitä.

Lemma *Mistä tahansa kontekstittomasta kieliopista voidaan muodostaa ekvivalentti kielioppi, jossa ei ole yksikköproduktioita.*

Todistus Olk. $G = (V, \Sigma, R, S)$.


G :n kunkin välitteen yksikköseuraajat:

1. $F(A) = \{B \in V \mid A \rightarrow B \in R\}$

2. Kunnes F -joukot eivät enää kasva


$$F(A) += \{F(B) \mid A \rightarrow B \in R\}$$

Lopuksi poistetaan kaikki G :n yksikköproduktiot ja lisätään niiden sijaan kaikki mahdolliset muotoa $A \rightarrow \Omega$ olevat produktiot, missä $B \in F(A)$ ja $B \rightarrow \Omega$. \square



$$\begin{aligned} S' &\rightarrow S \mid \varepsilon \\ S &\rightarrow A \mid B \\ A &\rightarrow aBa \mid aa \\ B &\rightarrow bAb \mid bb \end{aligned}$$

$$\begin{aligned} F(S') &= \{ S, A, B \} \\ F(S) &= \{ A, B \} \\ F(A) &= \emptyset \\ F(B) &= \emptyset \end{aligned}$$

$$\begin{aligned} S' &\rightarrow aBa \mid aa \mid bAb \mid bb \mid \varepsilon \\ S &\rightarrow aBa \mid aa \mid bAb \mid bb \\ A &\rightarrow aBa \mid aa \\ B &\rightarrow bAb \mid bb \end{aligned}$$


Kun ε - ja yksikköproduktiot on poistettu, niin kaikki produktiot ovat muotoa $A \rightarrow a$, $A \rightarrow X_1 \dots X_k$, $k \geq 2$, tai $S \rightarrow \varepsilon$.

Lisätään kielioppiin kutakin $a \in \Sigma$ vastaten välike C_a ja produktio $C_a \rightarrow a$.

Produktiot $A \rightarrow X_1 \dots X_k$, $k \geq 2$, korvataan produktiojoukoilla

$$\begin{aligned} A &\rightarrow X'_1 A_1 \\ A_1 &\rightarrow X'_2 A_2 \\ &\dots \\ A_{k-2} &\rightarrow X'_{k-2} A_{k-1} \\ A_{k-1} &\rightarrow X'_{k-1} X'_{k'} \end{aligned}$$

missä

$$X'_i = \begin{cases} X_i & \text{jos } X_i \in V \\ C_a & \text{jos } X_i = a \in \Sigma \end{cases}$$

$$S' \rightarrow aBa \mid aa \mid bAb \mid bb \mid \varepsilon$$

$$S \rightarrow aBa \mid aa \mid bAb \mid bb$$

$$A \rightarrow aBa \mid aa$$

$$B \rightarrow bAb \mid bb$$

$$S' \rightarrow C_a S'_1$$

$$S'_1 \rightarrow BC_a$$

$$S' \rightarrow C_a C_a$$

$$S' \rightarrow C_b S'_2$$

$$S'_2 \rightarrow AC_b$$

$$S' \rightarrow C_b C_b$$

$$S' \rightarrow \varepsilon$$

$$S \rightarrow C_a S_1$$

$$S_1 \rightarrow BC_a$$

$$S \rightarrow C_a C_a$$

$$S \rightarrow C_b S_2$$

$$S_2 \rightarrow AC_b$$

$$S \rightarrow C_b C_b$$

$$A \rightarrow C_a A_1$$

$$A_1 \rightarrow BC_a$$

$$A \rightarrow C_a C_a$$

$$B \rightarrow C_b B_1$$

$$B_1 \rightarrow AC_b$$

$$B \rightarrow C_b C_b$$

$$C_a \rightarrow a$$

$$C_b \rightarrow b$$

CYK-algoritmi

- CNF-muotoon muunnetun kontekstittoman kieliopin merkkijonot voidaan jäsentää $\theta(n^3)$ ajassa Cocke-Younger-Kasami –algoritmeilla
- Kontekstittomat kielet voidaan siis tunnistaa tehokkaasti
- CYK-algoritmin toimintaperiaate on dynaaminen ohjelmointi
- Osajonoille taulukoidaan ne välikemerkit, jotka voivat tuottaa ko. merkkijonon
- Jos lopulta kieliopin lähtösymboli kuuluu koko merkkijonon tuottavien välikkeiden joukkoon, niin tarkasteltava merkkijono kuuluu kieleen

3.1 Pinoautomaatit

- Pinoautomaatti on äärellinen automaatti, jolla on lisäksi yksi (ääretön) pinona käsiteltävä työnauha
- Pinon päällimmäiseksi alkioksi voidaan lisätä uusi alkio `push`-operaatiolla ja pinon päällimmäinen alkio voidaan poistaa `pop`-operaatiolla
- Pinoautomaatissa siirtymiin liittyy aina myös pinon käsittely
- Työnauha antaa automaatille "muistin", jonka avulla voidaan välttää äärellisten automaattien rajoituksia

Formaalisti pinoautomaatti on kuusikko $M=(Q, \Sigma, \Gamma, \delta, q_0, F)$, missä

- Q on tilojen äärellinen joukko,
- Σ on syöteaakkosto,
- Γ on pinoaakkosto,
- $q_0 \in Q$ on alkutila,
- $F \subseteq Q$ on lopputilojen joukko ja
- δ on joukkoarvoinen siirtymäfunktio:

$$\delta: Q \times \Sigma_\epsilon \times \Gamma_\epsilon \rightarrow \mathcal{P}(Q \times \Gamma_\epsilon)$$

- Pinoautomaatit ovat yleisessä tapauksessa epädeterministisiä:

$$\delta(r, x, a) = \{ (r_1, b_1), \dots, (r_k, b_k) \}$$

- lukemalla syötemerkin x ja pinomerkin a
 - automaatti voi siirtyä tilasta r johonkin tiloista r_1, \dots, r_k ja
 - samalla korvata pinon päällimmäisen merkin jollakin merkeistä b_1, \dots, b_k .
- Jos $x = \varepsilon$, niin automaatti tekee siirtymän syötemerkkiä lukematta;
 - jos $a = \varepsilon$, niin automaatti ei lue pinomerkkiä, vaan kirjoittaa uuden merkin pinon päällimmäiseksi alkioksi jättäen vanhan päällimmäisen merkin ennalleen (push);
 - jos $a \neq \varepsilon$ ja $b_i = \varepsilon$, niin pinon päällimmäinen alkio luetaan ja poistetaan, mutta uutta merkkiä ei kirjoiteta pinoon sen tilalle (pop)

Pinoautomaatti $M = (Q, \Sigma, \Gamma, \delta, q_0, F)$ hyväksyy merkkijonon $w \in \Sigma^*$ jos

- se voidaan kirjoittaa muotoon $w = w_1 w_2 \dots w_m$, missä kukin $w_i \in \Sigma_\varepsilon$, ja lisäksi
- on olemassa jono tiloja $r_0, r_1, \dots, r_m \in Q$ ja
- merkkijonoja $s_0, s_1, \dots, s_m \in \Gamma^*$

s.e. seuraavat kolme ehtoa pätevät.

- Lähtötilanteessa automaatin tila on alkutila ja pino on tyhjä:
 $r_0 = q_0$ ja $s_0 = \varepsilon$;
- $(r_{i+1}, b) \in \delta(r_i, w_{i+1}, a)$ kaikilla $i \in \{0, \dots, m-1\}$,
missä $s_i = at$ ja $s_{i+1} = bt$ joillain $a, b \in \Gamma_\varepsilon$ ja $t \in \Gamma^*$;
- $r_m \in F$.

Lause 2.20 *Kieli on kontekstiton jos ja vain jos se voidaan tunnistaa pinoautomaatilla.*

- Pinoautomaatti M on deterministinen, jos jokaisella tilanteella (r, x, a) on enintään yksi mahdollinen seuraaja (r', x', a') , jolla $(r, x, a) \succ_M (r', x', a')$
- Epädeterministiset pinoautomaatit ovat aidosti vahvempia kuin deterministiset. Esimerkiksi kieltä $\{ ww^R \mid w \in \{a, b\}^* \}$ ei voida tunnistaa deterministisellä pinoautomaatilla
- Deterministiset kontekstittomat kielet voidaan jäsentää tehokkaammin kuin yleiset kontekstittomat kielet