

Lemma (Säännöllisten kielten pumppauslemma). Olkoon A säännöllinen kieli. Tällöin on olemassa $n \geq 1$ siten, että kaikki A :n merkkijonot x , joiden pituus $|x| \geq n$ ovat ilmaistavissa muodossa $x = uvw$, missä $|uv| \leq n$, $|v| \geq 1$ ja merkkijonot muotoa $uv^i w$ kuuluvat kieleen A kaikilla $i \geq 0$.

Kompaktimmin, painottaen pumppauslemman asettamia vaatimuksia, voimme kirjoittaa seuraavasti:

\forall säännöllisille kielille A

$\exists n \geq 1$ s.e.

$\forall x \in A : |x| \geq n$

\exists osinjako $x = uvw$, missä $|uv| \leq n$ ja $|v| \geq 1$

$\forall i \geq 0 \ uv^i w \in A$.

Pumppauslemmaa voidaan käyttää hyväksi, kun halutaan osoittaa kieli L *ei-säännölliseksi*. Tehdään ensin vasta oletus, eli oletetaan L säännölliseksi kieleksi. Tavoitteena on päästä ristiriitaan tämän oletuksen kanssa seuraten pumppauslemman asettamia vaatimuksia säännöllisille kielille.

Pumppauslemmaa käytettäessä täytyy aina muistaa, että sillä voi osoittaa vain kielen epäsäännöllisyyden, ja sitä *ei* voi käyttää toiseen suuntaan. Esimerkiksi kieli

$$I = \{c^i a^n b^n \mid i > 0 \wedge n \geq 0\} \cup L(a^* b^*)$$

ei ole säännöllinen, mutta kaikki siihen kuuluvat sanat (tyhjää sanaa lukuunottamatta) voidaan osoittaa pumppauslemman ehtojen mukaisesti. Näin ollen kieltä I ei voida suoraan todistaa epäsäännölliseksi, vaan todistuksessa täytyy käyttää apuna säännöllisten kielten sulkeumaominaisuuksia.

Jos halutaan osoittaa kieli säännölliseksi, voidaan muodostaa sen hyväksyvä äärellinen automaatti, sillä pätee: Kieli L on säännöllinen \Leftrightarrow on olemassa äärellinen automaatti M , joka hyväksyy kielen L (merkitään $L(M) = L$).

4. **Tehtävä:** *Hahmolausekkeet* ovat esimerkiksi UN^*X -järjestelmien tekstityökaluissa käytetty säännöllisten lausekkeiden yleistys, jossa sallitaan merkkijonoarvoisten muuttujien käyttö lausekkeissa. Sovitettaessa merkkijonoa annettuun lausekkeeseen vaaditaan, että tietynnimisen muuttujan arvoksi tulee eri kohdissa sama osamerkkijono. Siten esimerkiksi aXb^*Xa ja $aX(a \cup b)^*YX(a \cup b)^*Ya$ ovat aakkoston $\{a, b\}$ hahmolausekkeitä, joista ensimmäinen kuvaa kielen $\{awb^n wa \mid w \in \{a, b\}^*, n \geq 0\}$. Osoita, että hahmolausekkeet ovat säännöllisten lausekkeiden aito yleistys, so. niillä voidaan kuvata myös joitakin ei-säännöllisiä kieliä.

Vastaus: Osoitaksemme, että tehtävän hahmolausekkeet on säännöllisten lausekkeiden aito yleistys, tulee meidän löytää hahmolauseke, jonka määrittämä kieli ei ole säännöllinen.

Tarkastellaan hahmolausekettä XX vastaavaa kieltä $L = \{zz \mid z = \{a, b\}^*\}$. Oletetaan, että L on säännöllinen. Valitaan $x = a^n b a^n \in L$, missä n on pumppauslemmassa esiintyvä kielestä L riippuva kokonaisluku. Nyt $|x| = 2n + 2 > n$. Pumppauslemman mukaan voidaan kirjoittaa $x = uvw$, missä $|uv| \leq n$ ja $|v| \geq 1$. Siis $u = a^{n-|v|-k}$, $v = a^{|v|}$ ja $w = a^k b a^n b$, missä $0 \leq k < n$. Nyt pumppauslemman mukaan kaikille $i \geq 0$ tulisi päteä $uv^i w \in L$. Kuitenkin $uv^0 w = uw = a^{n-|v|} b a^n b \notin L$, sillä se ei ole muotoa zz , koska vaadittiin $|v| \geq 1$. Päädyttiin siis ristiriitaan oletuksen kanssa. L ei näin ollen voi olla säännöllinen.

Löydettiin siis ei-säännöllinen kieli, joka voidaan kuvata hahmolausekkeella. Näin ollen hahmolausekkeet ovat säännöllisten lausekkeiden aito yleistys. \square

5. **Tehtävä:** Osoita, että kieli $\{w \in \{a, b\}^* \mid w\text{:ssä on yhtä monta } a\text{:ta ja } b\text{:tä}\}$ ei ole säännöllinen, ja laadi yhteydetön kielioppi sen kuvaamiseen.

Vastaus: Kielen $L = \{w \in \{a, b\}^* \mid w\text{:ssä on yhtä monta } a\text{:ta ja } b\text{:tä}\}$ voi todistaa ei-säännölliseksi suoraan pumppauslemmalla. Tässä esitetään kuitenkin hieman monimutkaisempi ratkaisu esimerkkinä siitä, miten ”hankalia” kieliä voidaan käsitellä.

Määritellään kieli $L' = L \cap L(a^*b^*)$. Oletetaan, että L on säännöllinen. Koska $L(a^*b^*)$ on säännöllinen ja säännöllisten kielten joukko on suljettu leikkauksen suhteen, täytyy myös L' :n olla säännöllinen. (Toisinpäin ehto ei päde: L' voi olla säännöllinen vaikka L ei olisi, sillä esim. $A \cap \emptyset = \emptyset$ kaikille kielille A).

Huomataan, että $L' = \{a^k b^k \mid k \geq 0\}$. Tarkastellaan sanaa $w = a^n b^n$, missä n on pumppauslemmassa esiintyvä parametri. Yritetään osoittaa w lemman ehtojen mukaisesti. Koska $|xy| \leq n$, osituksen täytyy olla muotoa;

$$\begin{aligned}x &= a^{n-i-k} \\y &= a^i \\z &= a^k b^n,\end{aligned}$$

missä $0 < i \leq n$ ja $i + k \leq n$. Nyt $xz = a^{n-i} b^n$, joten $xz \notin L'$. Näin ollen sanaa w ei voida pumpata, eikä L' ole säännöllinen, joten myöskään L ei ole säännöllinen.

Alla kielen L kuvaava yhteydetön kielioppi G :

$$S \rightarrow aSbS \mid bSaS \mid \varepsilon$$

6. **Tehtävä:** Laadi algoritmi, joka testaa onko annetun yhteydetön kieliopin $G = (V, \Sigma, P, S)$ tuottama kieli epätyhjä, so. voidaanko kieliopin lähtösymbolista S johtaa yhtään päätteenä $x \in \Sigma^*$.

Vastaus: Allaoleva proseduri `?GENERATESNONEMPTYLANGUAGE(G)` ottaa syötteenä yhteydetön kieliopin G , ja palauttaa arvon `true`, jos G :n generoima kieli ei ole tyhjä.

`?GENERATESNONEMPTYLANGUAGE(G = (V, Σ, P, S): context-free grammar)`

```
T ← Σ
repeat |V - Σ| times
  for each A → X1...Xk ∈ P
    if A ∉ T ∧ X1...Xk ∈ Tk
      T ← T ∪ {A}
if S ∈ T
  return true
else
  return false
```

Algoritmin idea on lähteä terminaalisyönteiden joukosta Σ ja testata, onko näistä mahdollista perääntyä S :ään käyttäen joukon P produktioita käänteisesti. Perääntymistä simuloidaan iteroimalla $|V - \Sigma|$ kertaa saavutettavien symbolien joukkoa T .

Perusteluksi sille, että $|V - \Sigma|$ askelta riittää, tarkastellaan sanaa $z \in L(G)$, jolla on kielen sanoista kaikkein pienin jäsennyyspuu. Jos z :lla on muotoa

$$S \rightarrow^* uAy \rightarrow^* uvAxy \rightarrow^* uvwxy$$

oleva johto, missä $u, v, w, x, y \in \Sigma^*$, niin myös sana $z' = uvw$ voidaan johtaa kieliopin säännöllillä¹. Tällöin kuitenkin z' :n jäsennyyspuu on pienempi kuin z :n jäsennyyspuu, mikä on ristiriidassa sen oletuksen kanssa, että z :n puu on pienin. Tästä seuraa se, että missään z :n minimaalisen jäsennyyspuun haaroissa ei voi esiintyä sama välikahta kertaan,

¹Vertaa yhteydetön kielten pumppauslemmaa.

joten algoritmissa riittää käydä sääntöjoukko läpi yhtä monta kertaa kuin kieliopissa on välitteitä.

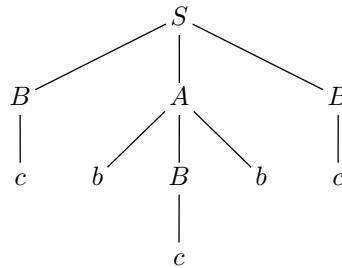
Tarkastellaan esimerkiksi kielioppia:

$$\begin{aligned} S &\rightarrow BAB \mid ABA \\ A &\rightarrow aAS \mid bBa \\ B &\rightarrow bBS \mid c \end{aligned}$$

Algoritmin laskenta etenee joukon T osalta seuraavasti:

$$\begin{aligned} T_0 &= \{a, b, c\} \\ T_1 &= \{a, b, c, B\} && (B \rightarrow c) \\ T_2 &= \{a, b, c, A, B\} && (A \rightarrow bBa) \\ T_3 &= \{a, b, c, A, B, C, S\} && (S \rightarrow BAB, S \rightarrow ABA) \end{aligned}$$

Koska $|V - \Sigma| = 3$, algoritmin suoritus päättyy ja $T = T_3$. Huomataan, että $S \in T$, joten kieli ei ole tyhjä. Pienin kieleen kuuluvan sanan jäsenyspuu on:



Liite: Chomskyn normaalimuoto

Muutetaan esimerkin vuoksi seuraava kielioppi Chomskyn normaalimuotoon:

$$\begin{aligned} P &= \{S \rightarrow aAS \mid bBS \mid \varepsilon \\ &A \rightarrow aAA \mid b, \\ &B \rightarrow bBB \mid a\} \end{aligned}$$

Kielioppi on Chomskyn normaalimuodossa, mikäli seuraavat ehdot toteutuvat:

- Ainoastaan alkuvälite S voi olla tyhjentyvä.
- Alkuvälite S ei esiinny säännöissä oikealla puolella.
- Mahdollisesti esiintyvää sääntöä $S \rightarrow \varepsilon$ lukuunottamatta kaikki säännöt ovat muotoa $A \rightarrow BC$ tai $A \rightarrow a$, missä A, B ja C ovat välitteitä ja a terminaalisympoli.

Kielioppi muutetaan normaalimuotoon vaiheittain:

- Poistetaan lähtösymboli sääntöjen oikealta puolelta.**

Koska kieliopissa on säännöt $S \rightarrow aAS$ ja $S \rightarrow bBS$, lisätään uusi lähtösymboli S' ja sääntö $S' \rightarrow S$. Saadaan tulokseksi sääntöjoukko:

$$\begin{aligned} S' &\rightarrow S, \\ S &\rightarrow aAS \mid bBS \mid \varepsilon \\ A &\rightarrow aAA \mid b, \\ B &\rightarrow bBB \mid a \end{aligned}$$

(b) **Poistetaan ε -produktiot.**

Koska Chomskyn normaalimuodossa ainoastaan lähtösymboli S' saa olla tyhjentyvä, täytyy muut ε -säännöt poistaa kieliopista. Lasketaan aluksi tyhjentävien välikkeiden joukko NULL:

$$\begin{aligned} \text{NULL}_0 &= \{S\} & (S \rightarrow \varepsilon) \\ \text{NULL}_1 &= \{S, S'\} & (S' \rightarrow S) \\ \text{NULL}_2 &= \{S, S'\} = \text{NULL} \end{aligned}$$

Tämän jälkeen korvataan säännöt $A \rightarrow X_1 \cdots X_n$ joukolla sääntöjä

$$A \rightarrow \alpha_1 \cdots \alpha_n, \quad \text{missä } \alpha_i = \begin{cases} X_i, X_i \notin \text{NULL} \\ X_i \text{ tai } \varepsilon, X_i \in \text{NULL} \end{cases}$$

Lopuksi poistetaan kaikki säännöt muotoa $A \rightarrow \varepsilon$ (lukuunottamatta sääntöä $S' \rightarrow \varepsilon$). Saadaan tulokseksi sääntöjoukko²:

$$\begin{aligned} S' &\rightarrow S \mid \varepsilon \\ S &\rightarrow aAS \mid aA \mid bBS \mid bB \\ A &\rightarrow aAA \mid b, \\ B &\rightarrow bBB \mid a \end{aligned}$$

(c) **Poistetaan yksikköproduktiot.**

Seuraavaksi poistetaan kieliopista kaikki muotoa $A \rightarrow B$ olevat säännöt, missä sekä A että B ovat välikkeitä.

Lasketaan ensin joukot $F(A)$ kaikilla $A \in V - \Sigma$:

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

Välike B kuuluu joukkoon $F(A)$ täsmälleen silloin, kun A :sta voidaan johtaa B käyttäen pelkkiä yksikköproduktioita.

Sääntö $A \rightarrow B$ korvataan sääntöjoukolla $\{A \rightarrow w \mid \exists C \in F(B) \cup \{B\} : C \rightarrow w \in P\}$. Tulokseksi saadaan sääntöjoukko:

$$\begin{aligned} S' &\rightarrow aAS \mid aA \mid bBS \mid bB \mid \varepsilon \\ S &\rightarrow aAS \mid aA \mid bBS \mid bB \\ A &\rightarrow aAA \mid b, \\ B &\rightarrow bBB \mid a \end{aligned}$$

(d) **Poistetaan liian pitkät produktiot.**

Viimeisessä vaiheessa lisätään kielioppiin uusi välike C_σ sekä sääntö $C_\sigma \rightarrow \sigma$ kaikille $\sigma \in \Sigma$ sekä jaetaan kaikki säännöt $A \rightarrow w$ ($|w| > 2$) ketjuksi sääntöjä, jotka kaikki johtavat tismalleen kaksi symbolia.

²Tarkkaan ottaen tässä vaiheessa pitäisi lisätä vielä uusi aloitusvälike S'' ja säännöt $S'' \rightarrow \varepsilon \mid S'$, mutta tässä tapauksessa ei synny ongelmia, vaikka käytetään S' :a lähtösymbolina.

Annetun kieliopin Chomskyn normaalimuodoksi saadaankin seuraava sääntöjoukko:

$$\begin{aligned}
 S' &\rightarrow C_a S'_1 \mid C_a A \mid C_b S'_2 \mid C_b B \mid \varepsilon \\
 S'_1 &\rightarrow AS \\
 S'_2 &\rightarrow BS \\
 S &\rightarrow C_a S_1 \mid C_a A \mid C_b S_2 \mid C_b B \\
 S_1 &\rightarrow AS \\
 S_2 &\rightarrow BS \\
 A &\rightarrow C_a A_1 \mid b \\
 A_1 &\rightarrow AA \\
 B &\rightarrow C_a B_1 \mid a \\
 B_1 &\rightarrow BB \\
 C_a &\rightarrow a \\
 C_b &\rightarrow b
 \end{aligned}$$

Liite: CYK-algoritmi

CYK-algoritmillä voidaan tutkia kuuluuko sana $x = x_1 \cdots x_n$ kieliopin G määrittelemään kieleen. Algoritmin kuluessa lasketaan välikeyoukko $N_{i,k}$. Joukko $N_{i,k}$ käsittää kaikki ne välitteet, joista voidaan johtaa osajono $x_i \cdots x_{i+k}$, eli sanan x kohdasta i alkava k :n merkin mittainen osajono. Joukkojen laskemisessa voidaan käyttää apuna dynaamista ohjelmointia seuraavaan tapaan:

$$\begin{aligned}
 N_{i,1} &= \{A \mid (A \rightarrow x_i) \in P\} \\
 N_{i,k} &= \bigcup_{j=1}^{k-1} \{A \in V - \Sigma \mid G\text{:ssä on sääntö } A \rightarrow BC, \\
 &\quad \text{missä } B \in N_{ij} \text{ ja } C \in N_{i+j,k-j}\}
 \end{aligned}$$

Tarkastellaan kielioppia G :

$$\begin{aligned}
 S &\rightarrow C_a D \mid C_a A \mid C_a E \mid BC_b \mid a \mid b \\
 A &\rightarrow C_a D \mid C_a A \mid a \\
 B &\rightarrow C_a E \mid BC_b \mid b \\
 D &\rightarrow AC_b \\
 E &\rightarrow BC_b \\
 C_a &\rightarrow a \\
 C_b &\rightarrow b
 \end{aligned}$$

Tarkistetaan, kuuluvatko sanat $w_1 = aabbb$ ja $w_2 = aabb$ kieleen $L(G)$. Koska w_2 on w_1 :n prefiksi, täytyy taulukko tehdä ainoastaan sanalle w_1 .

Lasketaan ensin joukot N_{i1} , $i \leq 5$:

		$i \rightarrow$				
		1 : a	2 : a	3 : b	4 : b	5 : b
$k \downarrow$	1	<u>a</u> abbb {S, A, C _a }	a <u>a</u> bbb {S, A, C _a }	aa <u>b</u> bb {S, B, C _b }	aab <u>b</u> b {S, B, C _b }	aabb <u>b</u> {S, B, C _b }

Kussakin taulukon ruudussa on alleviivattuna sitä vastaava sanan osajono.

Lasketaan seuraavaksi N_{12} , eli niiden välikkeiden joukko, jolla voidaan johtaa sanan alussa oleva aa . Koska normaalimuotoisessa kieliopissa ei ole yhtään sääntöä, jolla voisi suoraan johtaa useamman kuin yhden terminaalisympolin, voidaan aa saada ainoastaan siten, että johdetaan jostain välikkeestä kaksi a :n tuottavaa välikettä. Käytännössä tämä tapahtuu

siten, että etsitään kaikki sellaiset välitteet X , joille on olemassa sääntö $X \rightarrow YZ$, missä $Y \in N_{11}$ ja $X \in N_{21}$.

$$\begin{aligned} N_{11} = \{S, A, C_a\} \\ N_{21} = \{S, A, C_a\} \end{aligned} \Rightarrow N_{12} = \{S, A\} .$$

Tässä käytettiin sääntöjä $S \rightarrow C_a A$ ja $A \rightarrow C_a A$. Ruudulle N_{22} saadaan vastaavasti:

$$\begin{aligned} N_{21} = \{S, A, C_a\} \\ N_{31} = \{S, B, C_b\} \end{aligned} \Rightarrow N_{22} = \{D\} .$$

Ainoa ehdot täyttävä sääntö on $D \rightarrow AC_b$. Kokonaisuudessaan taulukon toiseksi riviksi muodostuu:

		$i \rightarrow$				
		$1 : a$	$2 : a$	$3 : b$	$4 : b$	$5 : b$
	1	<u>a</u> abb {S, A, C _a }	<u>a</u> abb {S, A, C _a }	<u>a</u> abb {S, B, C _b }	<u>a</u> abb {S, B, C _b }	<u>a</u> abb {S, B, C _b }
$k \downarrow$	2	<u>a</u> abb {S, A}	<u>a</u> abb {D}	<u>a</u> abb {S, B, E}	<u>a</u> abb {S, B, E}	

Ruudun N_{13} kohdalla huomataan, että sanan kolme ensimmäistä merkkiä (aab) voidaan johtaa kahdella eri tavalla:

- (a) johdetaan a välitteellä $X \in N_{11}$ ja ab välitteellä $Y \in N_{22}$; tai
- (b) johdetaan aa välitteellä $X \in N_{12}$ ja b välitteellä $Y \in N_{31}$.

Vastaavat joukot ovat:

$$\begin{aligned} j = 1 \Rightarrow N_{11} = \{S, A, C_a\} \\ N_{22} = \{D\} \end{aligned} \quad \begin{aligned} j = 2 \Rightarrow N_{12} = \{S, A\} \\ N_{31} = \{S, B, C_b\} \end{aligned}$$

Tapausta $j = 1$ vastaava välitejoukko on $\{S, A\}$ ($S \rightarrow C_a D, A \rightarrow C_a D$) ja tapausta $j = 2$ vastaava on $\{D\}$ ($D \rightarrow AC_b$), joten $N_{13} = \{S, A, D\}$. Samaan tapaan jatkamalla saadaan lopulta koko taulukoksi:

		$i \rightarrow$				
		$1 : a$	$2 : a$	$3 : b$	$4 : b$	$5 : b$
	1	<u>a</u> abb {S, A, C _a }	<u>a</u> abb {S, A, C _a }	<u>a</u> abb {S, B, C _b }	<u>a</u> abb {S, B, C _b }	<u>a</u> abb {S, B, C _b }
	2	<u>a</u> abb {S, A}	<u>a</u> abb {D}	<u>a</u> abb {S, B, E}	<u>a</u> abb {S, B, E}	
$k \downarrow$	3	<u>a</u> abb {A, S, D}	<u>a</u> abb {S, B}	<u>a</u> abb {S, B, E}		
	4	<u>a</u> abb \emptyset	<u>a</u> abb {S, B, E}			
	5	<u>a</u> abb {S, B}				

Viimeisessä vaiheessa täytyi käydä läpi neljä eri kombinaatiota: (N_{11}, N_{24}) , (N_{12}, N_{33}) , (N_{13}, N_{42}) sekä (N_{14}, N_{51}) .

Koska $S \in N_{15}$, niin $aabb \in L(G)$. Toisaalta, koska $S \notin N_{1,4}$, niin $aabb \notin L(G)$. Taulukosta voidaan konstruoida suoraan sanan w_1 jäsenyspuu käymällä se läpi takaperin: S saadaan ruutuun N_{15} ruuduista N_{11} ja N_{24} käyttäen välitteitä C_a ja E , joten ensimmäisenä käytetään sääntöä $S \rightarrow C_a E$. Ylläolevassa taulukossa on sanan w_1 johdossa käytetyt välitteet alleviivattu. Kokonaisuudessaan jäsenyspuu on seuraavanlainen:

