

# Muodolliset kieliopit

Luonnollisen kielen lauseenmuodostuksessa esiintyy luonnollisia säännönmukaisuuksia. Esimerkiksi, on jokseenkin mielekäästä väittää, että luonnollisen kielen lauseet koostuvat nk. subjektiosasta (esim. "vanha kissa") ja nk. predikaattiosasta ("nukkuu sikeästi"), jotka liitetään peräkkäin. Toisin sanoen, sääntö

LAUSE := SUBJEKTIOSA PREDIKAATTIOSA

on voimassa. Muodolliset kieliopit abstrahoivat tällaiseen säännönmukaisuuteen liittyviä ilmiöitä. Muodollinen kielioppi on oleellisesti kokoelma abstrakteja lauseenmuodostussääntöjä. Muodollisilla kieliopilla on erittäin huomattavia sovelluksia esim. tietojenkäsittelytieteessä.

# Muodolliset kieliopit

Muodolliset kieliopit ovat systeemejä, joita käytettäessä muodostetaan askel kerrallaan piteneviä symbolijonoja. Muodostettua symbolijonoa (esim.  $ADx$ ) pidennetään korvaamalla symboleja (esim.  $D$ ) symbolijonoilla (esim.  $yT$ ). Näin saadaan uusi, kieliopin sallima symbolijono ( $AyTx$ ). Tämä edellyttää, että kyseiseen kielioppiin kuuluu sääntö

$$D := yT.$$

Myös symbolijonoja voi korvata symbolijonoilla. Käsittelemme kuitenkin lähes yksinomaan kielioppeja, joissa yksittäisiä symboleja korvataan symbolijonoilla.

# Muodolliset kieliopit

Olkoon käytössä seuraavat säännöt.

1.  $A := A_i$
2.  $A := F_s$
3.  $F := F_u$
4.  $F := u$

Aloitetaan nk. *aloitussymbolista*, joka tässä on  $A$ . Alussa meillä on siis symbolijono  $A$ . Korvataan sitten symboli  $A$  sääntöä  $A := A_i$  käyttäen. Näin saamme symbolijonon  $A_i$ . Korvataan tässä uudessa symbolijonossa symboli  $A$  symbolijonolla  $F_s$ , käyttäen sääntöä  $A := F_s$ . Saamme symbolijonon  $F_s i$ . Korvataan sitten symboli  $F$  käyttäen sääntöä  $F := F_u$ . Saamme symbolijonon  $F_u s i$ . Lopuksi, käytetään sääntöä  $F := u$ . Saamme symbolijonon  $u u s i$ .

# Muodolliset kieliopit

Käytimme seuraavia sääntöjä.

1.  $A := A_i$
2.  $A := F_s$
3.  $F := F_u$
4.  $F := u$

Aloitimme aloitussymbolista  $A$  ja päädyimme symbolijonoon *uusi*. Muodollinen kielioppi sisältää täsmälleen yhden aloitussymbolin. Lisäksi, säännöissä esiintyvät symbolit voidaan jakaa kahteen luokkaan, *terminaalisymboleihin* ja *nonterminaalisymboleihin*. Terminaalisymbolit ovat pieniä kirjaimia (tässä  $i, s, u$ ) ja nonterminaalisymbolit isoja kirjaimia (tässä  $A, F$ ). Vain nonterminaalisymboleja voi korvata symbolijonoilla.

Käytimme seuraavia sääntöjä ja päädyimme symbolijonoon *uusi*.

1.  $A := A_i$
2.  $A := F_s$
3.  $F := F_u$
4.  $F := u$

Jonon (tai sanan) *uusi* voi siis *johtaa* käyttämällä yllä olevia sääntöjä. Päättynyttä korvausprosessia eli *johtamista* kutsutaan *terminoiduksi* jos ja vain jos johtamisen tuloksena syntynyt symbolijono ei sisällä lainkaan nonterminaalisymboleja. Johdimme sanan *uusi*, joka sisältää ainoastaan terminaalisympoleja. Kyseessä oli siis terminoitunut johtaminen.

Sääntöjä merkittäessä käytetään usein symbolin  $:=$  sijasta symbolia  $\rightarrow$ . Esim. edellä esiintynyt sääntö  $F := u$  voidaan kirjoittaa myös käyttämällä merkintää  $F \rightarrow u$ . Tästä eteenpäin käytämme symbolia  $\rightarrow$  hyödyntävää merkintätapaa. Tämä merkintätapa on käytössä myös kirjassa.<sup>1</sup>

---

<sup>1</sup>Merikoski et al: Johdatus diskreettiin matematiikkaan

# Muodolliset kieliopit

Käytä alla olevaa kielioppia ja johda alkusymbolista LAUSE aloittaen jono vanha kissa nukkuu .

LAUSE → SUBJ.OSA VERBI

SUBJ.OSA → ADJ. SUBJ.OSA

SUBJ.OSA → SUBJ.

SUBJ. → kissa

SUBJ. → koira

VERBI → nukkuu

ADJ. → vanha

# Säännölliset kielet vs. muodolliset kieliopit

Olemme aiemmin todenneet, että muodollinen kieli  $L$  on säännöllinen jos ja vain jos jokin deterministinen automaatti tunnistaa kielen  $L$ . Säännöllisiin kieliin kuuluu varsin monimutkaisia kieliä. On kuitenkin olemassa muodollisia kieliä, jotka eivät ole säännöllisiä. Tämä on yksi syy, miksi muodolliset kieliopit ovat hyödyllisiä.



# Säännölliset kielet vs. muodolliset kielioipit

**Lause.** Olkoon  $\Sigma$  aakkosto. On olemassa kieli  $L \subseteq \Sigma^*$ , joka ei ole säännöllinen.

**Todistus** (Hahmotelma). Joukko  $\mathcal{P}(\Sigma^*)$  on kaikkien muodollisten kielten  $M \subseteq \Sigma^*$  joukko. Koska  $\Sigma$  on äärellinen ja epätyhjä joukko, on helppo osoittaa, että  $\Sigma^*$  ja  $\mathbb{N}$  ovat yhtämahtavat. Täten muodollisia kieliä  $L' \subseteq \Sigma^*$  on yhtä monta kuin joukossa  $\mathcal{P}(\mathbb{N})$  alkioita, eli ylinumeroituvasti ääretön määrä. On lisäksi helppo osoittaa, että säännöllisiä lausekkeita on numeroituvasti ääretön määrä, joten säännöllisiä kieliä on numeroituvasti ääretön määrä. Täten muodollisia kieliä ja säännöllisiä lausekkeita ei voida asettaa yksi-yhteen suhteeseen.  $\square$

# Säännölliset kielet vs. muodolliset kieliopit

Kieli  $L = \{a^n b^n \mid n \in \mathbb{N}\}$  ei ole säännöllinen.

**Todistus.** Oletetaan että  $L$  on säännöllinen. Tällöin on olemassa deterministinen automaatti  $M$  joka hyväksyy kielen  $L$ . Automaatin tilajoukko on äärellinen. Täten on oltava sellaiset  $i \in \mathbb{N}$  ja  $k \in \mathbb{N}$ ,  $i \neq k$ , että  $M$  päättyy samaan tilaan  $q$  sekä luettuaan prefiksin  $a^i$  että luettuaan prefiksin  $a^k$ . Täten, koska  $M$  hyväksyy sanat  $a^i b^i$  ja  $a^k b^k$ , hyväksyy  $M$  myös sanan  $a^i b^k$ . Koska  $i \neq k$ , tämä on ristiriita.  $\square$

# Muodollisen kieliopin määritelmä

Määrittelemme nyt muodollisen kieliopin formaalisti.

Muodollinen kielioppi on yhdelmä  $(\Sigma, V, R, S)$ .

1.  $\Sigma$  on äärellinen terminaalisyömbolien joukko.
2.  $V$  on äärellinen nonterminaalisyömbolien joukko.  
 $V \cap \Sigma = \emptyset$ .
3.  $R$  on äärellinen joukko sääntöjä. Formaalisti,  
 $R \subseteq (W^* \cdot V \cdot W^*) \times (W^* \cup \{\varepsilon\})$ , missä  $W = \Sigma \cup V$  ja  
 $\varepsilon \notin W$  on nk. tyhjä merkki.
4.  $S \in V$  on alkusyömboli.

$R$  on siis äärellinen joukko sääntöjä  $(\alpha, \beta) \in R$ . Merkitsemme sääntöä  $(\alpha, \beta)$  käyttäen merkintää  $\alpha \rightarrow \beta$ , ja yleensä listaamme tätä merkintää käyttäen joukon  $R$  säännöt kirjoittaen ne allekain.

# Muodollisen kieliopin määrittelemä kieli

Muodollisen kieliopin  $M = (\Sigma, V, R, S)$  mukainen johtaminen aloitetaan alkusymbolista  $S$ . Johtamisaskeleessa korvataan nonterminaalisympoli (tai jopa kokonainen symbolijono) symbolijonolla, noudattaen sääntökokoelman  $R$  sääntöjä. Esim. jos  $\alpha A \beta \rightarrow \alpha A A$  on sääntö jossa  $A$  on terminaalisympoli ja  $\alpha$  ja  $\beta$  symbolijonoja, ja olemme johtaneet jonon  $\alpha A \beta$ , voimme yhdellä johtamisaskeleella johtaa jonon  $\alpha A A$ .

Samaan tapaan, jos  $\beta A \rightarrow \beta$  on sääntö, ja olemme johtaneet jonon  $\alpha \beta A$ , saamme nyt johdettua lyhyemmän jonon  $\alpha \beta$  korvaamalla jonossa  $\alpha \beta A$  jono  $\beta A$  jonolla  $\beta$ .

# Muodollisen kieliopin määrittelemä kieli

Kieliopin  $M$  määrittelemä kieli koostuu täsmälleen niistä joukon  $\Sigma^*$  sanoista  $\alpha$ , joille voidaan esittää johtaminen joka alkaa alkusymbolista  $S$ , noudattaa joukon  $R$  sääntöjä ja lopulta päättyy sanaan  $\alpha$ . Huomaa, että  $\alpha$  sisältää pelkästään terminalisymboleja. (Myös tyhjän symbolin  $\varepsilon$  esiintymät ovat mahdollisia).

# Muodolliset kieliopit

Olkoon  $\Sigma = \{a, b\}$ . Määritellään kielioppi, jossa alkusymboli on  $S$  ja käytössä on seuraavat säännöt.

1.  $S \rightarrow aSb$
2.  $S \rightarrow \varepsilon$ .

Voimme nyt johtaa sanan  $aabb$  seuraavasti.

1.  $S$  (Alkusymboli)
2.  $aSb$  (Sääntö 1.)
3.  $aaSbb$  (Sääntö 1.)
4.  $aa\varepsilon bb$  (Sääntö 2.)

Näin olemme johtaneet sanan  $a \cdot a \cdot \varepsilon \cdot b \cdot b = aabb$ . Huomaa, että tämä kielioppi tuottaa kielen  $\{a^i b^i \mid i \in \mathbb{N}\}$ . Osoitimme aiemmin, että tätä kieltä ei vastaa mikään automaatti.

# Chomskyn kielityypit

Määrittelemme nyt neljä eri luokkaa kielioppeja. Näitä luokkia kutsutaan Chomskyn kielityypeiksi. Merkitsemme merkkijonoja kreikkalaisilla kirjaimilla.

*Tyyppin 0* kieliopin säännöille  $\alpha \rightarrow \beta$  ei aseteta mitään rajoituksia.

*Tyyppin 1* kieliopin säännöt ovat muotoa  $\alpha A \beta \rightarrow \alpha \gamma \beta$  ( $\gamma \neq \varepsilon$ ) tai  $A \rightarrow \varepsilon$ . Näitä kielioppeja kutsutaan *kontekstiriippuviksi kielioppeiksi*. Huomaa että  $A$ :n "ympäristö" ( $\alpha$  ja  $\beta$ ) tai "konteksti" säilyy, kun  $A$  muuttuu  $\gamma$ :ksi käytettäessä sääntöä  $\alpha A \beta \rightarrow \alpha \gamma \beta$ .

jatkuu...

*Tyyppin 2* kieliopin säännöt ovat muotoa  $A \rightarrow \gamma$ . Tällaisia kielioppeja kutsutaan *kontekstiriippumattomiksi* kielioppeiksi.

*Tyyppin 3* kieliopin säännöt ovat muotoa  $A \rightarrow aB$ ,  $A \rightarrow a$ ,  $A \rightarrow \varepsilon$ . On mahdollista todistaa, että kieli on säännöllinen jos ja vain jos sen tuottaa jokin tyyppin 3 kielioppi.  
(Sivuutamme todistuksen)



Määrittelemämme kielioppityypit muodostavat nk. *Chomskyn kielioppihierarkian*. On mahdollista todistaa, että kielioppityypit muodostavat aidon hierarkian järjestyksessä  $3 \subset 2 \subset 1 \subset 0$ , millä tässä tarkoitetaan sitä, että seuraavat kaksi ehtoa toteutuvat.

1. Jokainen kieli joka voidaan tunnistaa tyypin  $i$  kieliopilla, voidaan tuottaa myös tyypin  $i - 1$  kieliopilla.
2. On olemassa kieli, joka voidaan tunnistaa tyypin  $i - i$  kieliopilla, mutta sitä ei voida tunnistaa tyypin  $i$  kieliopilla.

**Harjoitustehtävä.** Muodosta kielioppi, joka tuottaa saman kielen kuin säännöllinen lauseke  $a^* \cdot b^*$ .

**Vastaus.** Kielioppi sisältää seuraavat säännöt.

1.  $A \rightarrow aA$ ,
2.  $A \rightarrow Ab$
3.  $A \rightarrow \varepsilon$ .

Aloitussymbolina on  $A$ .