2. Yhteydettömät kielet

Yhteydettömät eli kontekstittomat kielet (context-free language, CFL) ovat säännöllisiä kieliä laajempi luokka formaaleja kieliä. Ne voidaan esittää yhteydettömillä kieliopeilla (context-free grammar, CFG). Täsmälleen tämän kieliluokan tunnistamiseen kykeneväksi laskentalaitteeksi osoittautuu pinoautomaatti (push-down automaton, PDA).

Tämän luvun jälkeen opiskelija

- osaa määritellä kontekstittoman kieliopin ja pinoautomaatin sekä selittää niiden välisen suhteen,

- osaa muodostaa kieliopin yksinkertaiselle kontekstittomalle kielelle ja

- tuntee joitain perusesimerkkejä ei-yhteydettömistä kielistä ja osaa yksinkertaisissa tapauksissa osoittaa tämän ominaisuuden.

Yhteydettömät kieliopit [Sipser luku 2.1]

Johdantoesimerkkinä tarkastelemme kieltä $L = \{ \mathrm{a}^n\mathrm{b}^m\mathrm{a}^n \mid n > 0, m > 0 \}$, joka on yhteydetön (mutta ei säännöllinen). Vastaavan kieliopin ytimenä on säännöt eli produktiot

$$\begin{aligned} S &\to \mathrm{a}S\mathrm{a} \\ S &\to \mathrm{a}B\mathrm{a} \\ B &\to \mathrm{b}B \\ B &\to \mathrm{b}. \end{aligned}$$

Kieliopin muuttujat eli välikkeet (variable, nonterminal symbol) ovat S ja B. Näistä S on erityisasemassa lähtösymbolina. Päätemerkit eli -symbolit (terminal) ovat a ja b. Kieliopista saadaan esim. merkkijonolle aabbaa $\in L$ seuraava johto (derivation):

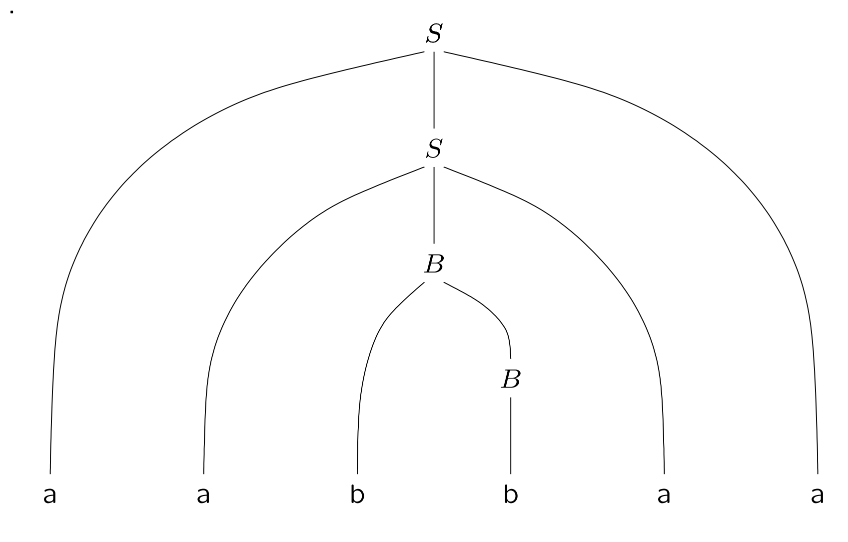
$$S \Rightarrow \mathrm{a}S\mathrm{a} \Rightarrow \mathrm{aa}B\mathrm{aa} \Rightarrow \mathrm{aab}B\mathrm{aa} \Rightarrow \mathrm{aabbaa}.$$

Johdossa siis aloitetaan lähtösymbolista ja korvataan muuttujat yksi kerrallaan vastaavan produktion oikealla puolella olevalla merkkijonolla.

Merkkijonon aabbaa johtoa

$$S \Rightarrow \mathsf{a}S\mathsf{a} \Rightarrow \mathsf{aa}B\mathsf{aa} \Rightarrow \mathsf{aab}B\mathsf{aa} \Rightarrow \mathsf{aabbaa}$$

vastaa seuraava jäsennyspuu (parse tree):



Lehtinä on johdetun merkkijonon merkit esijärjestyksessä ("vasemmalta oikealle"), juurena lähtösymboli ja sisäsolmuina välikkeitä. Kunkin välikkeen lapset vastaavat välikkeeseen liittyvän säännön oikeaa puolta.

Muodollisesti yhteydetön eli kontekstiton kielioppi (context-free grammar, CFG) on nelikko (V, Σ, R, S), missä

1. V on äärellinen muuttujien eli välikesymbolien joukko,
2. Σ on äärellinen päätesymbolien joukko, jolla $\Sigma \cap V = \emptyset$,
3. R on äärellinen joukko sääntöjä eli produktioita muotoa $A \to w$, missä $A \in V$ ja $w \in (V \cup \Sigma)^*$, ja
4. $S \in V$ on lähtösymboli.

Jos $x = uAv$ ja $y = uwv$, missä $(A \to w) \in R$, niin x johtaa suoraan eli tuottaa suoraan (yields, derives directly) merkkijonon y. Tätä merkitään $x \Rightarrow y$.

Jos on olemassa merkkijonot $w_0, \ldots, w_k \in (V \cup \Sigma)^*$, joilla

$$x = w_0 \Rightarrow w_1 \Rightarrow \ldots \Rightarrow w_k = y,$$

niin x johtaa eli tuottaa (derives) merkkijonon y. Tätä merkitään $x \overset{*}{\Rightarrow} y$. Jos tällöin $k > 0$, voidaan merkitä $x \overset{+}{\Rightarrow} y$.

Merkkijono $w \in (V \cup \Sigma)^*$ on lausejohdos (sentential form), jos $S \overset{*}{\Rightarrow} w$. Jos lisäksi w sisältää vain päätesymboleja, se on lause.

Kieliopin G tuottama eli kuvaama kieli on sen lauseiden joukko

$$L(G) = \left\{ w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w \right\}.$$

Kieli on yhteydetön, jos jokin yhteydetön kielioppi tuottaa sen.

Jos $w \in L(G)$ ja

$$S \Rightarrow w_1 \Rightarrow \ldots \Rightarrow w_k = w,$$

tätä ketjua sanotaan merkkijonon w johdoksi (derivation), ja sen pituus on n. Tyypillisesti lauseella on useita eri johtoja; palaamme tähän pian.

Johdosta voidaan muodostaa jäsennyspuu ilmeisellä tavalla: kirjoitetaan ensin juureksi lähtösymboli, ja sitten laajennetaan solmuja soveltamalla johdon mukaisia sääntöjä.

Merkintäkonventioita

Yhteydettömiä kielioppeja käytetään mm. luonnollisten ja ohjelmointikielten syntaksin kuvaamiseen, jolloin niistä voi tulla hyvin laajoja.

Merkintöjen yksinkertaistamiseksi kieliopista riittää listata pelkkä sääntöjoukko:

- muuttujia ovat ne, jotka ovat jonkin säännön vasempana puolena,
- lähtösymboli on listassa ensimmäisen säännön vasen puoli ja
- muut symbolit ovat päätesymboleita.

Samaan muuttujaan liittyvät säännöt

$$A \rightarrow w_1, \quad A \rightarrow w_2, \quad \ldots, \quad A \rightarrow w_n$$

voidaan tiivistää muotoon

$$A \rightarrow w_1 \mid w_2 \mid \ldots \mid w_n.$$

Esimerkki 2.1: Yhteydetön kielioppi

$$S \to SS \mid (S) \mid \varepsilon$$

(jonka päätesymbolit ovat siis "(" ja ")") tuotaa kaikki oikein muodostetut sulkulausekkeet. $\square$

Esimerkki 2.2: Akkoston $\{\,\mathsf{a}, \mathsf{b}\,\}$ palindromien joukko $\left\{\, w \in \{\,\mathsf{a}, \mathsf{b}\,\}^* \mid w = w^{\mathcal{R}} \,\right\}$ voidaan tuottaa yhteydettömällä kieliopilla

$$S \to \varepsilon \mid \mathsf{a} \mid \mathsf{b} \mid \mathsf{a}S\mathsf{a} \mid \mathsf{b}S\mathsf{b}.$$

Vastaavasti kieli $\left\{\, ww^{\mathcal{R}} \mid w \in \{\,\mathsf{a}, \mathsf{b}\,\}^* \,\right\}$ voidaan tuottaa yhteydettömällä kieliopilla

$$S \to \varepsilon \mid \mathsf{a}S\mathsf{a} \mid \mathsf{b}S\mathsf{b}.$$

Sen sijaan kieli $\{\, ww \mid w \in \{\,\mathsf{a}, \mathsf{b}\,\}^* \}$ ei ole yhteydetön, kuten jatkossa näemme. $\square$

Huomautus: "Yhteydetön" viittaa siihen, että sääntöä $A \to w$ voidaan soveltaa riippumatta siitä, missä "yhteydessä" A esiintyy. Yleisemmissä yhteysherkissä kieliopeissa sallitaan säännöt muotoa $uAv \to uwv$, jolloin A saadaan muuttaa w:ksi vain "kontekstissa" u_v.

Esimerkki 2.3: Luonnollinen kieli on ollut tärkeä motivaatio kielioppiformalismien kehittämiselle. Leikkikaluesimerkki:

$$
\begin{array}{rcl}
\langle \texttt{sentence} \rangle & \to & \langle \texttt{noun-phrase} \rangle \langle \texttt{verb-phrase} \rangle \\
& & \mid \langle \texttt{noun-phrase} \rangle \langle \texttt{verb} \rangle \langle \texttt{direct-object-phrase} \rangle \\
\langle \texttt{noun-phrase} \rangle & \to & \langle \texttt{proper-noun} \rangle \mid \langle \texttt{determiner} \rangle \langle \texttt{common-noun} \rangle \\
\langle \texttt{proper-noun} \rangle & \to & \mathbf{John} \mid \mathbf{Jill} \\
\langle \texttt{common-noun} \rangle & \to & \mathbf{car} \mid \mathbf{hamburger} \\
\langle \texttt{determiner} \rangle & \to & \mathbf{a} \mid \mathbf{the} \\
\langle \texttt{verb-phrase} \rangle & \to & \langle \texttt{verb} \rangle \langle \texttt{adverb} \rangle \mid \langle \texttt{verb} \rangle \\
\langle \texttt{verb} \rangle & \to & \mathbf{drives} \mid \mathbf{eats} \\
\langle \texttt{adverb} \rangle & \to & \mathbf{slowly} \mid \mathbf{frequently} \\
\langle \texttt{direct-object-phrase} \rangle & \to & \langle \texttt{adjective-list} \rangle \langle \texttt{proper-noun} \rangle \\
& & \mid \langle \texttt{determiner} \rangle \langle \texttt{adjective-list} \rangle \langle \texttt{common-noun} \rangle \\
\langle \texttt{adjective-list} \rangle & \to & \varepsilon \mid \langle \texttt{adjective} \rangle \langle \texttt{adjective-list} \rangle \\
\langle \texttt{adjective} \rangle & \to & \mathbf{big} \mid \mathbf{juicy} \mid \mathbf{brown}
\end{array}
$$

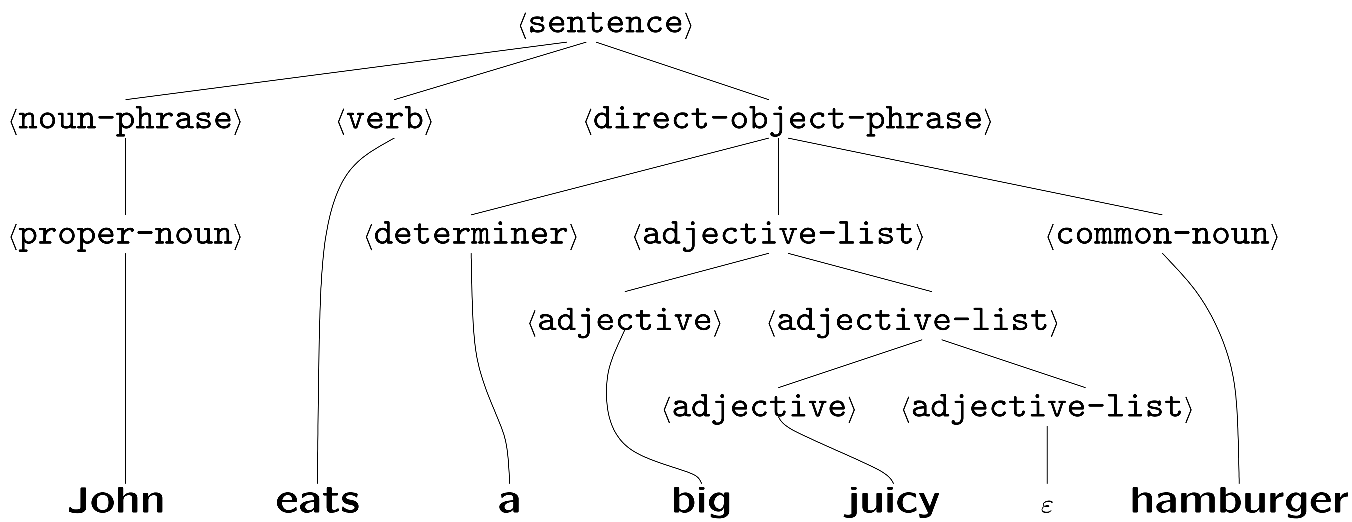
Tässä siis $\langle \texttt{sentence} \rangle$ jne. ovat muuttujasymboleja ja **John** jne. päätesymboleja.

Edellisestä kieliopista saadaan esim. seuraavanlaisia johtoja:

$$
\begin{aligned}
\langle\texttt{sentence}\rangle &\Rightarrow \langle\texttt{noun-phrase}\rangle\langle\texttt{verb}\rangle\langle\texttt{direct-object-phrase}\rangle\\
&\Rightarrow \langle\texttt{proper-noun}\rangle\langle\texttt{verb}\rangle\langle\texttt{direct-object-phrase}\rangle\\
&\Rightarrow \textbf{John } \langle\texttt{verb}\rangle\langle\texttt{direct-object-phrase}\rangle\\
&\Rightarrow \textbf{John eats } \langle\texttt{direct-object-phrase}\rangle\\
&\Rightarrow \textbf{John eats } \langle\texttt{determiner}\rangle\langle\texttt{adjective-list}\rangle\langle\texttt{common-noun}\rangle\\
&\Rightarrow \textbf{John eats a } \langle\texttt{adjective-list}\rangle\langle\texttt{common-noun}\rangle\\
&\Rightarrow \textbf{John eats a } \langle\texttt{adjective}\rangle\langle\texttt{adjective-list}\rangle\langle\texttt{common-noun}\rangle\\
&\Rightarrow \textbf{John eats a big } \langle\texttt{adjective-list}\rangle\langle\texttt{common-noun}\rangle\\
&\Rightarrow \textbf{John eats a big } \langle\texttt{adjective}\rangle\ \langle\texttt{adjective-list}\rangle\langle\texttt{common-noun}\rangle\\
&\Rightarrow \textbf{John eats a big juicy } \langle\texttt{adjective-list}\rangle\langle\texttt{common-noun}\rangle\\
&\Rightarrow \textbf{John eats a big juicy } \langle\texttt{common-noun}\rangle\\
&\Rightarrow \textbf{John eats a big juicy hamburger}
\end{aligned}
$$

Tämä on esimerkki vasemmasta johdosta (leftmost derivation), mihin palataan pian.

Edellistä johtoa vastaava jäsennyspuu:



Esimerkki 2.4: Toinen sekä historiallisesti että nykykäytännön kannalta tärkeä yhteydettömien kielioppien sovellus on ohjelmointikielten syntaksin kuvaaminen. Tällöin käytetään usein ns. Backus-Naur-muotoa (BNF), joka on suunnilleen sama kuin yhteydetön kielioppi.

Kielioppi toimii paitsi ohjelmoijan muistilistana myös kääntäjän (tarkemmin jäsentäjän) laatimisen lähtökohtana.

Tarkastellaan esimerkkinä pientä Pascal-kielen osaa:

$$
\begin{array}{rcl}
\langle \texttt{lause} \rangle & \rightarrow & \langle \texttt{ehtolause} \rangle \mid \langle \texttt{koottu-lause} \rangle \mid \langle \texttt{sijoitus} \rangle \mid \langle \texttt{kutsu} \rangle \\
\langle \texttt{ehtolause} \rangle & \rightarrow & \textbf{if} \langle \texttt{ehto} \rangle \textbf{then} \langle \texttt{lause} \rangle \textbf{else} \langle \texttt{lause} \rangle \\
\langle \texttt{ehto} \rangle & \rightarrow & \texttt{x=0} \\
\langle \texttt{koottu-lause} \rangle & \rightarrow & \textbf{begin} \langle \texttt{lausejono} \rangle \textbf{end} \\
\langle \texttt{lausejono} \rangle & \rightarrow & \langle \texttt{lause} \rangle \mid \langle \texttt{lause} \rangle \texttt{;} \langle \texttt{lausejono} \rangle \\
\langle \texttt{sijoitus} \rangle & \rightarrow & \texttt{x:= 0} \\
\langle \texttt{kutsu} \rangle & \rightarrow & \mathsf{a} \mid \mathsf{b} \mid \mathsf{c}
\end{array}
$$

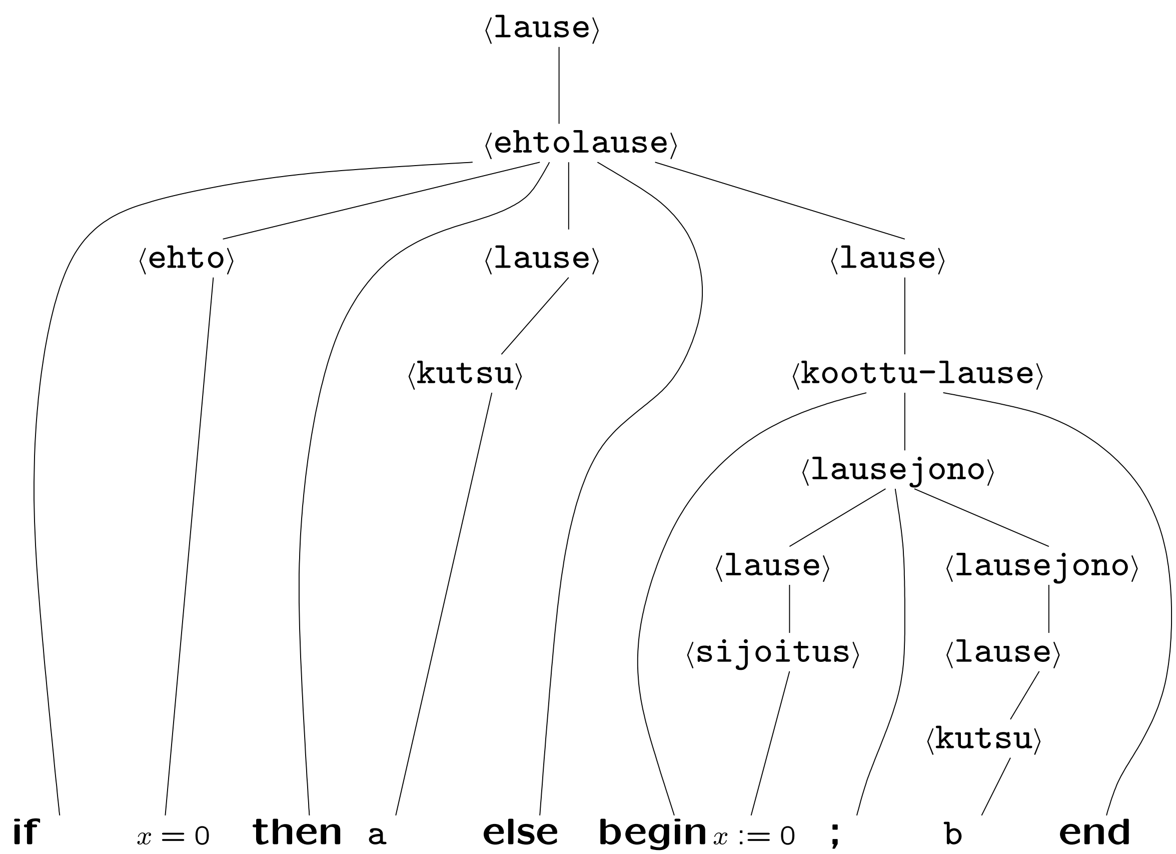
Esimerkkijohto:

$$
\begin{aligned}
\langle\texttt{lause}\rangle &\Rightarrow \langle\texttt{ehtolause}\rangle\\
&\Rightarrow \textbf{if } \langle\texttt{ehto}\rangle \textbf{ then } \langle\texttt{lause}\rangle \textbf{ else } \langle\texttt{lause}\rangle\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \langle\texttt{lause}\rangle \textbf{ else } \langle\texttt{lause}\rangle\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \langle\texttt{kutsu}\rangle \textbf{ else } \langle\texttt{lause}\rangle\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else } \langle\texttt{lause}\rangle\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else } \langle\texttt{koottu-lause}\rangle\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else begin } \langle\texttt{lausejono}\rangle \textbf{ end}\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else begin } \langle\texttt{lause}\rangle \textbf{ ; } \langle\texttt{lausejono}\rangle \textbf{ end}\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else begin } \langle\texttt{sijoitus}\rangle \textbf{ ; } \langle\texttt{lausejono}\rangle \textbf{ end}\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else begin } \texttt{x:=}\mathsf{0} \textbf{ ; } \langle\texttt{lausejono}\rangle \textbf{ end}\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else begin } \texttt{x:=}\mathsf{0} \textbf{ ; } \langle\texttt{lause}\rangle \textbf{ end}\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else begin } \texttt{x:=}\mathsf{0} \textbf{ ; } \langle\texttt{kutsu}\rangle \textbf{ end}\\
&\Rightarrow \textbf{if } \mathsf{x{=}0} \textbf{ then } \mathsf{a} \textbf{ else begin } \texttt{x:=}\mathsf{0} \textbf{ ; } \mathsf{b} \textbf{ end}
\end{aligned}
$$

Vastaava ohjelmanpätkä:

```
if x=0 then
      a
else begin
      x:= 0;
      b
end
```

Vielä jäsennyspuu samalle lauseelle:



Yhteydettömien kielten sulkeumaominaisuuksista

Yhteydettömien kielten luokka on suljettu yhdisteen, konkatenaation ja tähtioperaation suhteen:

Lause 2.5: Jos A ja B ovat yhteydettömiä kieliä, niin myös $A \cup B$, $A \circ B$ ja A^* ovat.

Todistus: suoraviivainen harjoitustehtävä. $\square$

Tästä seuraa erityisesti

Korollaari 2.6: Kaikki säännölliset kielet ovat yhteydettömiä. $\square$

Tulemme jatkossa todistaneeksi tämän korollaarin myös toista kautta osoittamalla, että äärellisiä automaatteja voimakkaammat pinoautomaatit tunnistavat kuitenkin vain yhteydettömiä kieliä.

Jatkossa osoittautuu myös, että yhteydettömien kielten luokka **ei** ole suljettu leikkauksen ja komplementin suhteen, ts. $A \cap B$ ja $\overline{A}$ eivät välttämättä ole yhteydettömiä.

Moniselitteisyys [Sipser s. 107–108]

Kontekstittoman kielen lauseella on tyypillisesti useita johtoja. Esim. kieliopissa

$$\begin{aligned} S &\rightarrow AB \\ A &\rightarrow \mathsf{a} \\ B &\rightarrow \mathsf{b} \end{aligned}$$

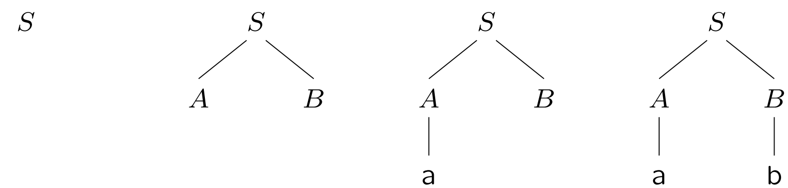
merkkijono ab voidaan johtaa kahdella eri tavalla: $S \Rightarrow AB \Rightarrow aB \Rightarrow ab$ ja $S \Rightarrow AB \Rightarrow Ab \Rightarrow ab$. Intuitiivisesti näissä johdoissa on vain käytetty samoja sääntöjä eri järjestyksessä. Johdot myös vastaavat samaa jäsennyspuuta.

Tällaisen triviaalin monijohtoisuuden poissulkemiseksi tarkastelemme vasempia johtoja (leftmost derivation). Vasemmassa johdossa sovelletaan aina produktiota merkkijonon vasemmanpuoleisimpaan muuttujaan. Siis em. kieliopissa merkkijonon ab ainoa vasen johto on

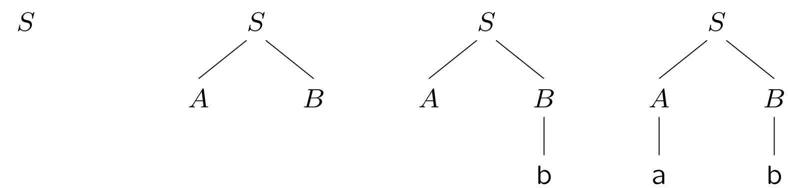
$$S \Rightarrow AB \Rightarrow aB \Rightarrow ab.$$

Kielioppi on moniselitteinen (ambiguous), jos siinä jollain lauseella on useampi kuin yksi vasemmanpuoleinen johto. Muuten kielioppi on yksiselitteinen. Tähänastiset esimerkkimme ovat kaikki olleet yksiselitteisiä.

Vasen johto $S \Rightarrow AB \Rightarrow aB \Rightarrow ab$ esittää jäsennyspuun kasvattamista vasemmalta alkaen:



Samaan jäsennyspuuhun päästään myös johdolla $S \Rightarrow AB \Rightarrow Ab \Rightarrow ab$:

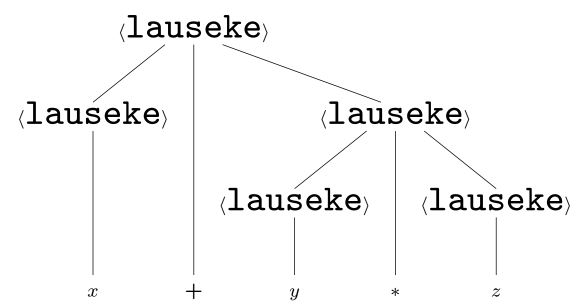


Esimerkki 2.7: Muuttujien x, y ja z aritmeettiset lausekkeet voidaan kuvata kieliopilla

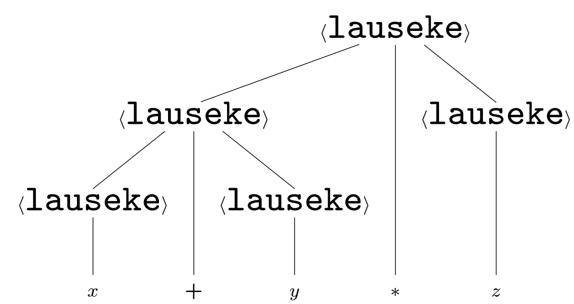
$$\langle\texttt{lauseke}\rangle \to \langle\texttt{lauseke}\rangle + \langle\texttt{lauseke}\rangle \mid \langle\texttt{lauseke}\rangle * \langle\texttt{lauseke}\rangle \mid (\ \langle\texttt{lauseke}\rangle\ ) \mid x \mid y \mid z.$$

Kielioppi on moniselitteinen, sillä lauseella $x + y * z$ on kaksi vasenta johtoa ja jäsennyspuuta:

$$\begin{aligned}
\langle\texttt{lauseke}\rangle &\Rightarrow \langle\texttt{lauseke}\rangle + \langle\texttt{lauseke}\rangle \\
&\Rightarrow x + \langle\texttt{lauseke}\rangle \\
&\Rightarrow x + \langle\texttt{lauseke}\rangle * \langle\texttt{lauseke}\rangle \\
&\Rightarrow x + y * \langle\texttt{lauseke}\rangle \\
&\Rightarrow x + y * z
\end{aligned}$$



$$\begin{aligned}
\langle\texttt{lauseke}\rangle &\Rightarrow \langle\texttt{lauseke}\rangle * \langle\texttt{lauseke}\rangle \\
&\Rightarrow \langle\texttt{lauseke}\rangle + \langle\texttt{lauseke}\rangle * \langle\texttt{lauseke}\rangle \\
&\Rightarrow x + \langle\texttt{lauseke}\rangle * \langle\texttt{lauseke}\rangle \\
&\Rightarrow x + y * \langle\texttt{lauseke}\rangle \\
&\Rightarrow x + y * z.
\end{aligned}$$

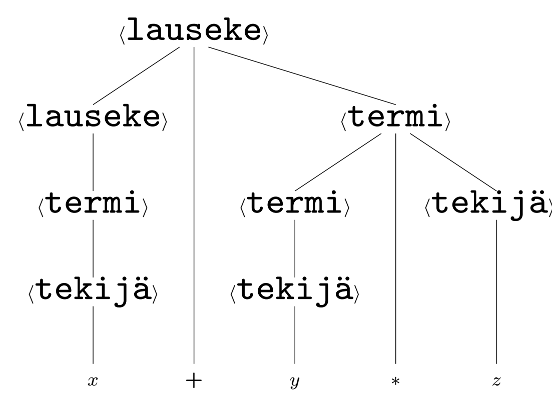


Sama kieli voidaan tuottaa seuraavalla yksiselitteisellä kieliopilla

$$\begin{aligned}
\langle\texttt{lauseke}\rangle &\rightarrow \langle\texttt{lauseke}\rangle + \langle\texttt{termi}\rangle \mid \langle\texttt{termi}\rangle \\
\langle\texttt{termi}\rangle &\rightarrow \langle\texttt{termi}\rangle * \langle\texttt{tekijä}\rangle \mid \langle\texttt{tekijä}\rangle \\
\langle\texttt{tekijä}\rangle &\rightarrow \texttt{(}\ \langle\texttt{termi}\rangle\ \texttt{)} \mid x \mid y \mid z.
\end{aligned}$$

Lauseen $x + y * z$ yksikäsitteiset vasen johto ja jäsennyspuu ovat

$$\begin{aligned}
\langle\texttt{lauseke}\rangle &\Rightarrow \langle\texttt{lauseke}\rangle + \langle\texttt{termi}\rangle \\
&\Rightarrow \langle\texttt{termi}\rangle + \langle\texttt{termi}\rangle \\
&\Rightarrow \langle\texttt{tekijä}\rangle + \langle\texttt{termi}\rangle \\
&\Rightarrow x + \langle\texttt{termi}\rangle \\
&\Rightarrow x + \langle\texttt{termi}\rangle * \langle\texttt{tekijä}\rangle \\
&\Rightarrow x + \langle\texttt{tekijä}\rangle * \langle\texttt{tekijä}\rangle \\
&\Rightarrow x + y * \langle\texttt{tekijä}\rangle \\
&\Rightarrow x + y * z
\end{aligned}$$



Aritmeettisen lausekkeen jäsennyspuun avulla voidaan helposti laskea lausekkeen arvo, kun muuttujien arvot tunnetaan. Yleisemmin kääntäjä voi jäsennyspuun avulla generoida koodia lausekkeen evaluoimiseksi. Tätä sovellusta silmällä pitäen edellisen sivun kielioppi noudattaa koulusta tuttua presedenssisääntöä, jonka mukaan kertolaskut lasketaan ennen yhteenlaskuja.

Jäsennyspuun hyödyntämiseksi pitää tietysti ensinnä osata muodostaa annetulle merkkijonolle jäsennyspuu (eli yhtäpitävästi johto) annetussa kieliopissa, tai todeta, että merkkijono ei kuulu kieleen. Palaamme jatkossa lyhyesti tähän yhteydettömän kieliopin jäsennysongelmaan. Todetaan tässä vaiheessa, että ongelma on ei-triviaali, ja tehokkaat jäsennysmenetelmät jäävät tämän kurssin ulkopuolelle.

Todetaan vielä, että joillekin yhteydettömille kielille **ei ole olemassa** yksiselitteistä kielioppia. Esimerkki tällaisesta luonnostaan moniselitteisestä (inherently ambiguous) kielestä on $\left\{ \mathrm{a}^i\mathrm{b}^j\mathrm{c}^k \mid i = j \text{ tai } j = k \right\}$ (todistus sivuutetaan).

Chomskyn normaalimuoto [Sipser s. 108–111]

Kielioppeja algoritmisesti käsiteltäessä on hyvä, jos ne ovat ”siistejä”.

Määritelmä 2.8: Yhteydetön kielioppi (V, Σ, R, S) on Chomskyn normaalimuodossa, jos joukon R jokainen sääntö on jotain seuraavista muodoista:

1. $A \rightarrow BC$, missä $A, B, C \in V$ ja $B \neq S$ ja $C \neq S$,
2. $A \rightarrow a$, missä $A \in V$ ja $a \in \Sigma$, tai
3. $S \rightarrow \varepsilon$.

Normaalimuodosta seuraa erityisesti, että jos $S \overset{*}{\Rightarrow} w$ ja $|w| = n > 0$, niin johdon pituus on tasan $2n - 1$.

Lause 2.9: [Sipser Thm. 2.9] Mikä tahansa yhteydetön kieli voidaan tuottaa Chomskyn normaalimuodossa olevalla yhteydettömällä kieliopilla.

Tehdään todistus kahdessa vaiheessa. Tarkastellaan ensin ε-sääntöjen poistamista.

Lemma 2.10: Mikä tahansa yhteydetön kieli voidaan tuottaa yhteydettömällä kieliopilla, jossa

1. lähtösymboli ei esiinny minkään säännön oikealla puolella ja

2. kieliopissa ei ole sääntöä $A \rightarrow \varepsilon$ millään A, joka ei ole lähtösymboli.

Todistus: Olkoon $G = (V, \Sigma, R, S)$ yhteydetön kielioppi. Muodostamme ehdot toteuttavan kieliopin $G' = (V', \Sigma, R', S_0)$, jolla $L(G') = L(G)$.

Otetaan käyttöön uusi lähtösymboli $S_0 \notin V$ ja asetetaan $V' = V \cup \{ S_0 \}$. Laitetaan joukkoon R' sääntö $S_0 \rightarrow S$. Jos $\varepsilon \in L(G)$, lisätään myös sääntö $S_0 \rightarrow \varepsilon$. Muuttujalle S_0 ei tule muita sääntöjä, eikä se tule minkään säännön oikealle puolelle.

Määritellään nyt nollautuvien muuttujien joukko

$$\mathsf{Null} = \left\{ A \in V \mid A \overset{*}{\Rightarrow} \varepsilon \right\}.$$

Joukko Null voidaan helposti laskea iteratiivisella algoritmilla.

Perusidea on, että jos $B \in$ Null, niin sääntö $A \to uBv$ korvataan säännöillä

$$A \to uv \mid uBv.$$

Vastaavasti jos $B, C \in$ Null, niin sääntö $A \to uBvCw$ korvataan säännöillä

$$A \to uvw \mid uvCw \mid uBvw \mid uBvCw,$$

jne. Tämän jälkeen ε-säännöt ovat jääneet turhiksi ja voidaan poistaa.

Käydään siis läpi kaikki joukon R säännöt $A \to w$. Kirjoitetaan w muotoon

$$w = u_1 A_1 u_2 A_2 \ldots u_k A_k u_{k+1},$$

missä kukin A_i on nollautuva muuttuja ja toisaalta mikään u_i ei sisällä nollautuvia muuttujia. Jos w ei sisällä nollautuvia muuttujia, niin $k = 0$. Lisätään nyt joukkoon R' kaikki 2^k sääntöä

$$A \to u_1 x_1 u_2 x_2 \ldots u_k x_k u_{k+1},$$

missä x_i on joko A_i tai ε, paitsi ei sääntöä $A \to \varepsilon$ mikäli se esiintyy.

On helppo nähdä, että saatu kielioppi toteuttaa vaaditut ehdot. $\square$

Esimerkki 2.11: Sovelletaan konstruktiota kielioppiin

$$\begin{array}{rcl} S & \to & BSB \mid A \\ A & \to & \mathsf{a}A \mid \mathsf{aa} \\ B & \to & \mathsf{b}B \mid \varepsilon. \end{array}$$

Nyt Null $= \{ B \}$. Siis säännöstä $S \to BSB$ saadaan

$$S \to S \mid SB \mid BS \mid BSB,$$

ja säännöstä $B \to \mathsf{b}B$ saadaan

$$B \to \mathsf{b} \mid \mathsf{b}B.$$

Lisätään vielä alkusymboli, ja otetaan mukaan kaikki vanhat säännöt lukuunottamatta ε-sääntöjä. Nyt $S \stackrel{*}{\not\Rightarrow} \varepsilon$, joten kieliopiksi tulee

$$\begin{array}{rcl} S_0 & \to & S \\ S & \to & S \mid SB \mid BS \mid BSB \mid A \\ A & \to & \mathsf{a}A \mid \mathsf{aa} \\ B & \to & \mathsf{b} \mid \mathsf{b}B. \end{array}$$

$\square$

Seuraava vaihe on poistaa yksikkösäännöt eli säännöt muotoa $A \to B$, missä $A, B \in V$.

Lemma 2.12: Mikä tahansä yhteydetön kieli voidaan tuottaa yhteydettömällä kieliopilla, jossa ei ole yksikkösääntöjä.

Todistus: Kaikilla $A \in V$ olkoon Unit(A) niiden muutujien joukko, jotka voidaan johtaa A:sta pelkillä yksikkösäännöillä, A itse mukaanlukien. Joukot Unit(A) on helppo laskea iteratiivisesti.

Kieliopin muuttujat ja päätesymbolit pidetään ennallaan, samoin lähtösymboli. Sääntöihin laitetaan $A \to w$ kaikilla $A \in V$ ja $w \in (V \cup \Sigma)^*$, joilla

1. $B \in$ Unit(A),
2. $B \to w$ on sääntö alkuperäisessä kieliopissa ja
3. w ei ole muuttujasymboli.

Ehtojen 1 ja 2 perusteella on selvää, että uusien sääntöjen joukkoon ei tule mitään, mitä siellä ei saisi olla. Jos $B \in \mathsf{Unit}(A)$ ja $B \Rightarrow w$, niin $A \overset{*}{\Rightarrow} w$, joten $A \rightarrow w$ voidaan lisätä.

Ainoa ongelma näyttäisi olevan, että ehdon 3 nojalla pudotetaan pois muotoa $B \rightarrow C$ olevat säännöt. Kuitenkin jos $B \in \mathsf{Unit}(A)$ ja $B \rightarrow C$ on sääntö, niin $C \in \mathsf{Unit}(A)$, joten vuorostaan muotoa $C \rightarrow w$ olevat säännöt tuovat mukaan uudet säännöt $A \rightarrow w$. Siis uudet säännöt tuottavat samat merkkijonot kuin vanhat. $\square$

Esimerkki 2.13: Jatketaan edellisen esimerkin lopputuloksesta

$$\begin{aligned} S_0 &\rightarrow S \\ S &\rightarrow S \mid SB \mid BS \mid BSB \mid A \\ A &\rightarrow \mathsf{a}A \mid \mathsf{aa} \\ B &\rightarrow \mathsf{b} \mid \mathsf{b}B. \end{aligned}$$

Nyt Unit(S_0) $= \{ S_0, S, A \}$, Unit(S) $= \{ S, A \}$, ja Unit(X) $= \{ X \}$ muuten. Siis muuttujalle S_0 tulee omien alkuperäisten lisäksi kaikki muuttujien S ja A ei-yksikkösäännöt:

$$S_0 \rightarrow SB \mid BS \mid BSB \mid \mathsf{a}A \mid \mathsf{aa}.$$

Samoin S:lle lisätään A:n säännöt:

$$S \rightarrow SB \mid BS \mid BSB \mid \mathsf{a}A \mid \mathsf{aa}.$$

Ottamalla vielä kaikki vanhat ei-yksikkösäännöt saadaan kielioppi

$$\begin{aligned} S_0 &\rightarrow SB \mid BS \mid BSB \mid \mathsf{a}A \mid \mathsf{aa} \\ S &\rightarrow SB \mid BS \mid BSB \mid \mathsf{a}A \mid \mathsf{aa} \\ A &\rightarrow \mathsf{a}A \mid \mathsf{aa} \\ B &\rightarrow \mathsf{b} \mid \mathsf{b}B. \end{aligned}$$

$\square$

Lauseen 2.9 todistus: Olkoon $G = (V, \Sigma, R, S)$ yhteydetön kielioppi. Muodostamme Chomskyn normaalimuodossa olevan kieliopin, joka tuottaa saman kielen. Lemmojen 2.10 ja 2.11 nojalla voimme olettaa kieliopista G, että

1. S ei esiinny minkään säännön oikealla puolella,
2. $A \rightarrow \varepsilon \notin R$ kaikilla $A \in V - \{ S \}$, ja
3. yksikkösääntöjä ei ole.

Siis sääntö $A \rightarrow u_1 \ldots u_k$, missä $u_i \in V \cup \Sigma$, voi rikkoa normaalimuotoa kahdella tavalla:

1. joko $k = 2$ ja jokin u_i on päätesymboli tai
2. $k \geq 3$.

Otetaan jokaista päätesymbolia $a \in \Sigma$ kohti käyttöön uusi muuttuja X_a ja lisätään sääntö $X_a \to a$.

Korvataan jokainen normaalimuotoa rikkova sääntö $A \to u_1 u_2$ säännöllä $A \to U_1 U_2$, missä $U_k = X_a$ jos $u_k = a \in \Sigma$ ja $U_k = u_k$ jos $u_k \in V$.

Sääntö $A \to u_1 \ldots u_k$, missä $k \geq 3$, korvataan $k - 1$ uudella säännöllä

$$\begin{array}{rcl} A & \to & U_1 A_2 \\ A_2 & \to & U_2 A_3 \\ & \ldots & \\ A_{k-2} & \to & U_{k-2} A_{k-1} \\ A_{k-1} & \to & U_{k-1} U_k \end{array}$$

missä $A_2, \ldots, A_{k-1}$ ovat uusia muuttujia ja U_k kuten edellä. $\square$

Esimerkki 2.14: Jatketaan edellistä esimerkkiä

$$\begin{aligned}
S_0 &\longrightarrow SB \mid BS \mid BSB \mid \mathsf{a}A \mid \mathsf{aa} \\
S &\longrightarrow SB \mid BS \mid BSB \mid \mathsf{a}A \mid \mathsf{aa} \\
A &\longrightarrow \mathsf{a}A \mid \mathsf{aa} \\
B &\longrightarrow \mathsf{b} \mid \mathsf{b}B.
\end{aligned}$$

Kielioppi tulee muotoon

$$\begin{aligned}
S_0 &\longrightarrow SB \mid BS \mid BA_1 \mid X_{\mathsf{a}}A \mid X_{\mathsf{a}}X_{\mathsf{a}} \\
A_1 &\longrightarrow SB \\
S &\longrightarrow SB \mid BS \mid BA_2 \mid X_{\mathsf{a}}A \mid X_{\mathsf{a}}X_{\mathsf{a}} \\
A_2 &\longrightarrow SB \\
A &\longrightarrow X_{\mathsf{a}}A \mid X_{\mathsf{a}}X_{\mathsf{a}} \\
B &\longrightarrow \mathsf{b} \mid X_{\mathsf{b}}B \\
X_{\mathsf{a}} &\longrightarrow \mathsf{a} \\
X_{\mathsf{b}} &\longrightarrow \mathsf{b}.
\end{aligned}$$

$\square$