Süntaksanalüüs

Süntaksanalüüs

- *Süntaksanalüüs* kontrollib programmi struktuuri vastavust keele grammatikale:
  - saab sisendina, skanneri poolt genereeritud, lekseemide jada;
  - väljastab programmi esitava (abstraktse) süntaksipuu;
  - süntaktiliste vigade korral, teeb kindlaks nende asukoha;
  - . . . teavitab võimalikest vea põhjustest;
  - . . . püüab veast toibuda ja jätkata analüüsi (et järgnevaid vigu avastada).
- Süntaksanalüüsi kutsutakse *parsimiseks* (parsing) ning vastavat analüsaatorit nimetatakse *parseriks* (parser).

Grammatikad

- Keelte süntaksi kirjeldatakse reeglina kontekstivaba grammatika abil.
- *Grammatika* on nelik $G = \langle N, T, P, S \rangle$, kus
  - N on lõplik *mitteterminaalide* tähestik;
  - T on lõplik *terminaalsümbolite* tähestik;
  - $N \cap T = \emptyset$ ja $V = N \cup T$;
  - $P \subset \{ \alpha \to \beta \mid \alpha \in V^+, \beta \in V^\star \}$ on lõplik *produktsioonireeglite* hulk;
  - $S \in N$ on *algsümbol*.
- Grammatika on *kontekstivaba* (context-free), kui produktsioonireeglid on kujul $A \to \alpha$, kus $A \in N$ ja $\alpha \in V^\star$.

Grammatikad

- Jada $w \in V^\star$ nimetatakse *lausevormiks* (sentential form).
- Lausevorm $u \in V^\star$ on *otsetuletatav* (directly derivable) lausevormist $v \in V^\star$ (tähistus $u \Longrightarrow v$), kui leiduvad $w_1, w_2, \alpha, \beta \in V^\star$ sellised, et $u = w_1 \alpha w_2$, $v = w_1 \beta w_2$ ja $\alpha \to \beta \in P$.
- Relatsiooni $\Longrightarrow$ refleksiivset transitiivset sulundit (tähistus $\Longrightarrow^\star$) nimetatakse *derivatsiooniks* (derivation) ehk *tuletuseks*.
- Grammatika $G = \langle N, T, P, S \rangle$ genereerib *keele*

$$L(G) = \{ w \in T^\star \mid S \Longrightarrow^\star w \}$$

- Grammatikad G_1 ja G_2 on *ekvivalentsed*, kui $L(G_1) = L(G_2)$.

Grammatikad

- Chomsky hierarhia:

	Produktsioonid	**Keelte tüüp**	**Automaat**
L_0	$\alpha \to \beta$	Semi-Thue süsteemid	Turingi masin
L_1	$\alpha A \beta \to \alpha \gamma \beta$	Kontekstist sõltuvad keeled	Tõkestatud Turingi masin
L_2	$A \to \alpha$	Kontekstivabad keeled	Magasinmäluga automaat
L_3	$A \to w, A \to wB$	Regulaarsed keeled	Lõplik automaat
(L_4)	$A \to w$	Lõplikud keeled	Tsükliteta lõplik automaat

kus $A, B \in N$, $\alpha, \beta, \gamma \in V^\star$ ja $w \in T^\star$.

- Lemma: Chomsky hierarhia on range; so.:

$$(L_4) \subset L_3 \subset L_2 \subset L_1 \subset L_0$$

Kontekstivabad grammatikad

- Edaspidi käsitleme ainult kontekstivabu grammatikaid.
- Kontekstivabade grammatikate produktsioonireegleid esitatakse tavaliselt *Backus-Naur'i kujul* (BNF).
- Näide: olgu $N = \{Exp\}$ ja $T = \{+, *, (,), \mathrm{id}\}$, siis

$$\begin{array}{lcl} Exp & ::= & Exp + Exp \\ & | & Exp * Exp \\ & | & (\; Exp \;) \\ & | & \mathrm{id} \end{array}$$

esitab produktsioonireeglite hulka

$$\begin{array}{lcll} P & = & \{ & Exp \to Exp + Exp, \quad Exp \to (\; Exp \;), \\ & & & Exp \to Exp * Exp, \quad Exp \to \mathrm{id} \quad \}. \end{array}$$

Kontekstivabad grammatikad

- Mitteterminaal A on *produktiivne* (productive), kui leidub $w \in T^\star$ selline, et $A \Longrightarrow^\star w$.
- Mitteterminaal A on *saavutatav* (reachable), kui leiduvad lausevormid $u, v \in V^\star$ sellised, et $S \Longrightarrow^\star uAv$.
- KV-grammatika $G = \langle N, T, P, S \rangle$ on *taandatud* (reduced), kui tema iga mitteterminaal on produktiivne ja saavutatav.
- Lemma: Iga KV-grammatika saab teisendada temaga ekvivalentseks taandatud KV-grammatikaks.

Kontekstivabad grammatikad

- Reeglina saab ühte ja sama lauset tuletada paljudel eri viisidel.
- Kanoonilised derivatsioonid:
  - *vasakderivatsioon* — derivatsiooni igal sammul asendatakse vasakpoolseim mitteterminaal;
  - *paremderivatsioon* — derivatsiooni igal sammul asendatakse parempoolseim mitteterminaal.
- Näide:

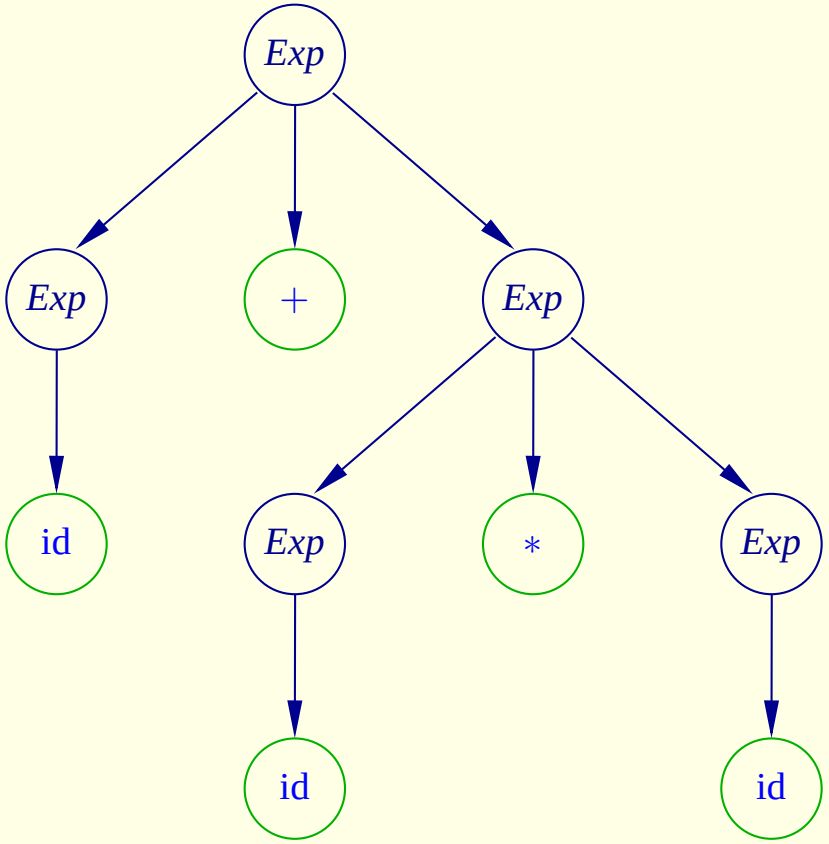
$$
\begin{aligned}
Exp &\Longrightarrow_{lm} Exp + Exp \\
&\Longrightarrow_{lm} \mathrm{id} + Exp \\
&\Longrightarrow_{lm} \mathrm{id} + Exp * Exp \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} * Exp \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} * \mathrm{id}
\end{aligned}
\qquad
\begin{aligned}
Exp &\Longrightarrow_{rm} Exp + Exp \\
&\Longrightarrow_{rm} Exp + Exp * Exp \\
&\Longrightarrow_{rm} Exp + Exp * \mathrm{id} \\
&\Longrightarrow_{rm} Exp + \mathrm{id} * \mathrm{id} \\
&\Longrightarrow_{rm} \mathrm{id} + \mathrm{id} * \mathrm{id}
\end{aligned}
$$

Kontekstivabad grammatikad

- Iga derivatsioon määrab üheselt *süntaksipuu* (syntax-tree, parse-tree), mis on järjestatud tippudega puu, kus:
  - puu juur on märgendatud algsümboliga S;
  - vahetipud on märgendatud mitteterminaalidega;
  - lehed on märgendatud terminaalsümboliga või tühisümboliga ε;
  - kui vahetipp on märgendatud mitteterminaaliga A ja tema alampuude (vasakult paremale) $t_1, \ldots, t_n$ juured on märgendatud vastavalt $A_1, \ldots, A_n$, siis $A \to A_1, \ldots, A_n \in P$.
- Tuletatud lause saadakse lugedes puu lehtede märgendid vasakult paremale.
- Süntaksipuu määrab üheselt kasutatud produktsioonireeglid, kuid mitte nende rakendamise järjekorda.

Kontekstivabad grammatikad

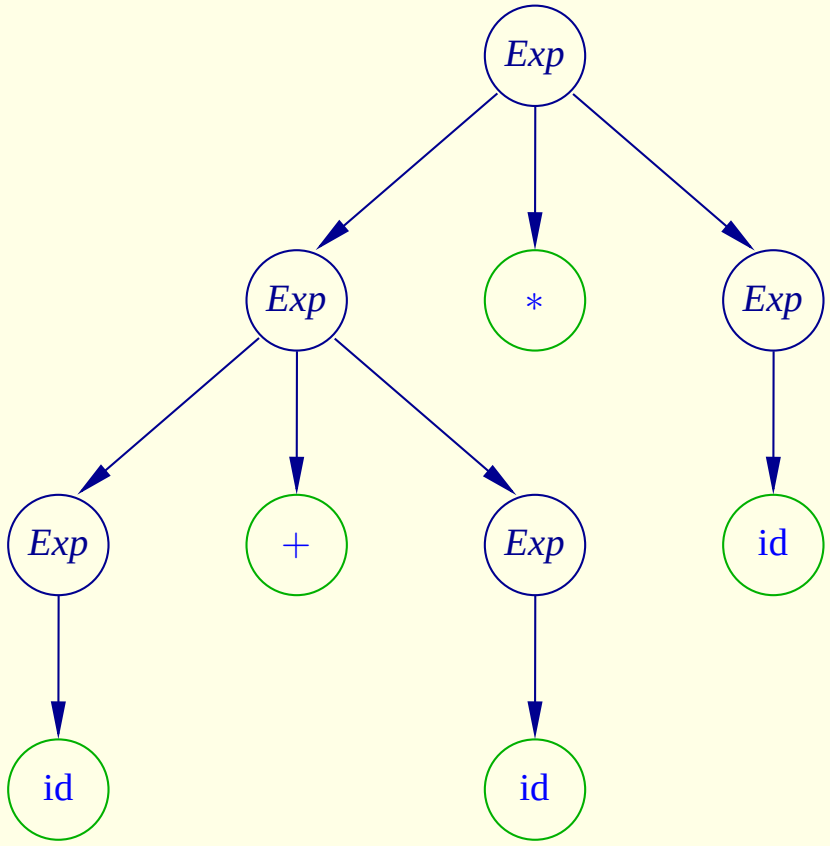
- Näide: eelnevalt toodud vasak- ja paremderivatsioonile vastab mõlemal juhul sama süntaksipuu



Kontekstivabad grammatikad

- NB! Ühel lausel $w \in L(G)$ võib olla mitu erinevat süntaksipuud!
- Näide:

$$
\begin{aligned}
Exp &\Longrightarrow_{lm} Exp * Exp \\
&\Longrightarrow_{lm} Exp + Exp * Exp \\
&\Longrightarrow_{lm} \mathrm{id} + Exp * Exp \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} * Exp \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} * \mathrm{id}
\end{aligned}
$$



Kontekstivabad grammatikad

- KV-grammatika on *mitmene* (ambiguous), kui ühe lause jaoks leidub mitu süntaksipuud.
- Iga süntaksipuu jaoks leidub täpselt üks vasak- ja üks paremderivatsioon; seega:
  - ühesel lausel on täpselt üks vasak- ja üks paremderivatsioon;
  - mitmesel lausel on vähemalt kaks vasak- ja paremderivatsiooni.
- Lause erinevad süntaksipuud vastavad reeglina lause semantiliselt erinevatele tõlgendusvõimalustele.
- Mitmese grammatika saab teatud juhtudel (aga mitte alati) teisendada ekvivalentseks üheseks grammatikaks.

Kontekstivabad grammatikad

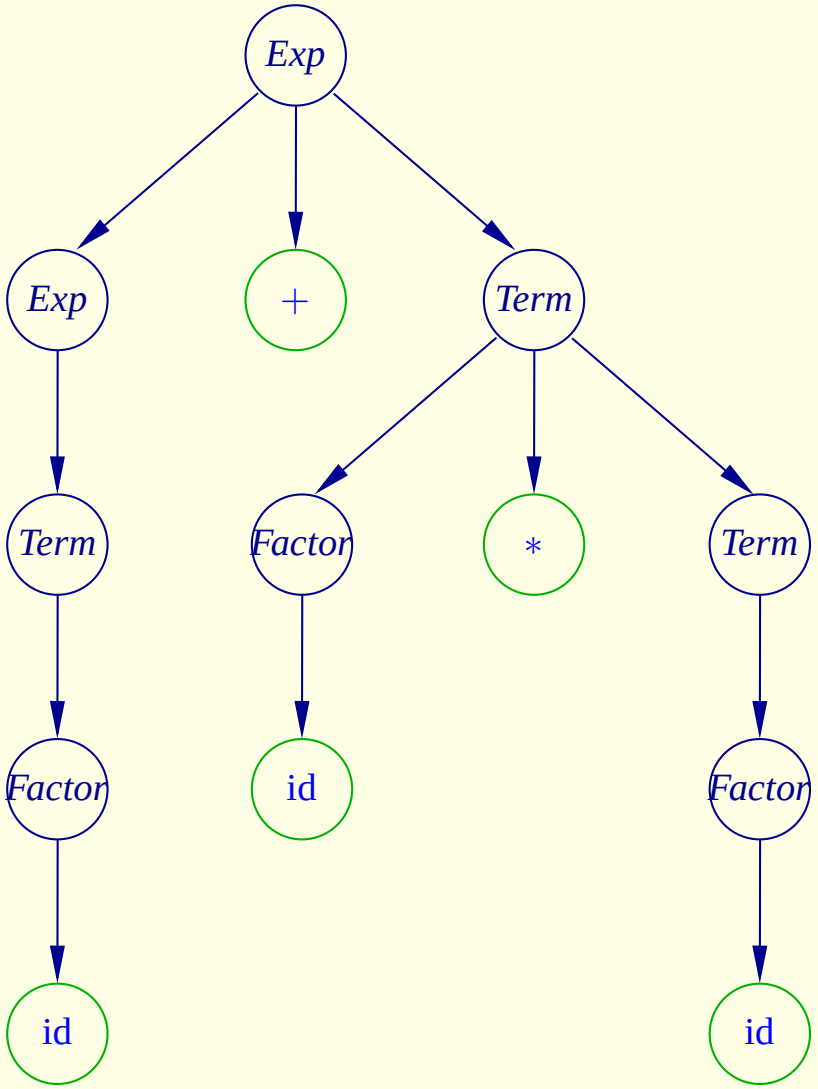
- Mitmesuse eemaldamine — binaarsed operaatorid:
  - iga prioriteeditaseme jaoks toome sisse uue mitteterminaali;
  - vasakassotsiatiivse operaatori korral kasutame vasakrekursiivset; paremassotsiatiivse korral paremrekursiivset produktsioonireeglit;
  - tugevama prioriteediga operaatoritele vastavad reeglid paigutame "sügavamale".
- Näide:

$$
\begin{array}{lcl}
\mathit{Exp} & ::= & \mathit{Exp} + \mathit{Term} \\
 & | & \mathit{Term} \\
\mathit{Term} & ::= & \mathit{Factor} * \mathit{Term} \\
 & | & \mathit{Factor} \\
\mathit{Factor} & ::= & (\ \mathit{Exp}\ ) \\
 & | & \mathrm{id}
\end{array}
$$

Kontekstivabad grammatikad

Näide:

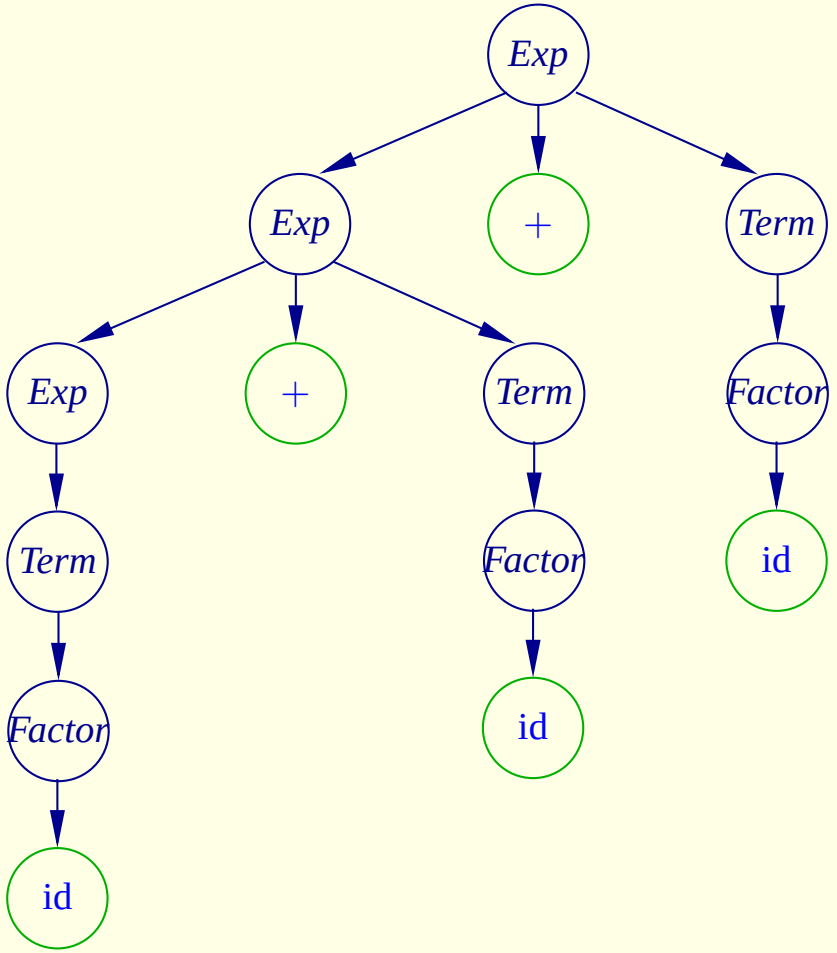
$$
\begin{aligned}
Exp &\Longrightarrow_{lm} Exp + Term \\
&\Longrightarrow_{lm} Term + Term \\
&\Longrightarrow_{lm} Factor + Term \\
&\Longrightarrow_{lm} \mathrm{id} + Term \\
&\Longrightarrow_{lm} \mathrm{id} + Factor * Term \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} * Term \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} * Factor \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} * \mathrm{id}
\end{aligned}
$$



Kontekstivabad grammatikad

Näide:

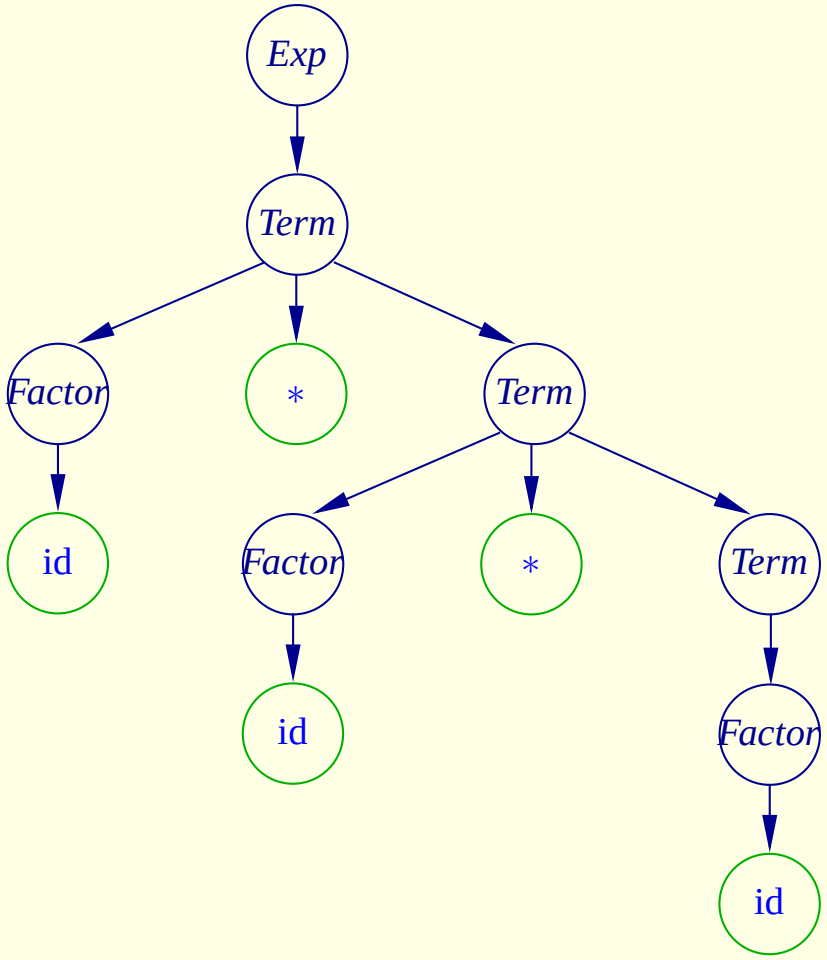
$$
\begin{aligned}
Exp &\Longrightarrow_{lm} Exp + Term \\
&\Longrightarrow_{lm} Exp + Term + Term \\
&\Longrightarrow_{lm} Term + Term + Term \\
&\Longrightarrow_{lm} Factor + Term + Term \\
&\Longrightarrow_{lm} \mathrm{id} + Term + Term \\
&\Longrightarrow_{lm} \mathrm{id} + Factor + Term \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} + Term \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} + Factor \\
&\Longrightarrow_{lm} \mathrm{id} + \mathrm{id} + \mathrm{id}
\end{aligned}
$$



Kontekstivabad grammatikad

Näide:

$$
\begin{aligned}
\mathit{Exp} &\Longrightarrow_{lm} \mathit{Term} \\
&\Longrightarrow_{lm} \mathit{Factor} * \mathit{Term} \\
&\Longrightarrow_{lm} \mathrm{id} * \mathit{Term} \\
&\Longrightarrow_{lm} \mathrm{id} * \mathit{Factor} * \mathit{Term} \\
&\Longrightarrow_{lm} \mathrm{id} * \mathrm{id} * \mathit{Term} \\
&\Longrightarrow_{lm} \mathrm{id} * \mathrm{id} * \mathit{Factor} \\
&\Longrightarrow_{lm} \mathrm{id} * \mathrm{id} * \mathrm{id}
\end{aligned}
$$



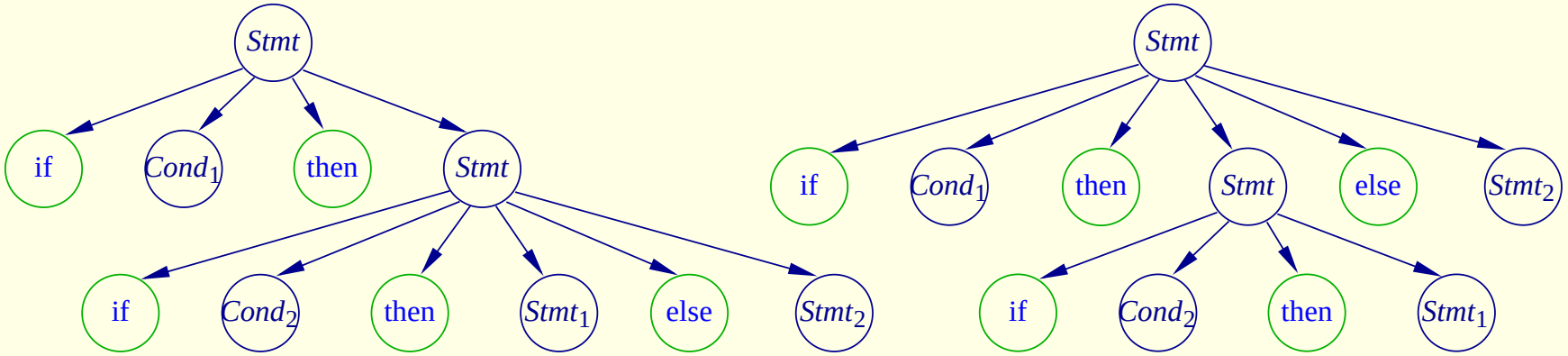
Kontekstivabad grammatikad

- Mitmesuse eemaldamine — tingimuslaused:

$$
\begin{array}{lcl}
\mathit{Stmt} & ::= & \text{if } \mathit{Cond} \text{ then } \mathit{Stmt} \\
& | & \text{if } \mathit{Cond} \text{ then } \mathit{Stmt} \text{ else } \mathit{Stmt} \\
& | & \mathit{Other}
\end{array}
$$

- Järgnev lausevorm omab kahte süntaksipuud:

$$\text{if } \mathit{Cond}_1 \text{ then if } \mathit{Cond}_2 \text{ then } \mathit{Stmt}_1 \text{ else } \mathit{Stmt}_2$$



Kontekstivabad grammatikad

- Tavaliselt loetakse korrektseks esimest süntaksipuud; so. else kuulub alati kõige sisemise võimaliku tingimuslause juurde:

$$
\begin{array}{lcl}
\mathit{Stmt} & ::= & \mathit{WithElse} \\
 & | & \mathit{NoElse} \\
\mathit{WithElse} & ::= & \text{if } \mathit{Cond} \text{ then } \mathit{WithElse} \text{ else } \mathit{WithElse} \\
 & | & \mathit{Other} \\
\mathit{NoElse} & ::= & \text{if } \mathit{Cond} \text{ then } \mathit{Stmt} \\
 & | & \text{if } \mathit{Cond} \text{ then } \mathit{WithElse} \text{ else } \mathit{NoElse}
\end{array}
$$

Kontekstivabad grammatikad