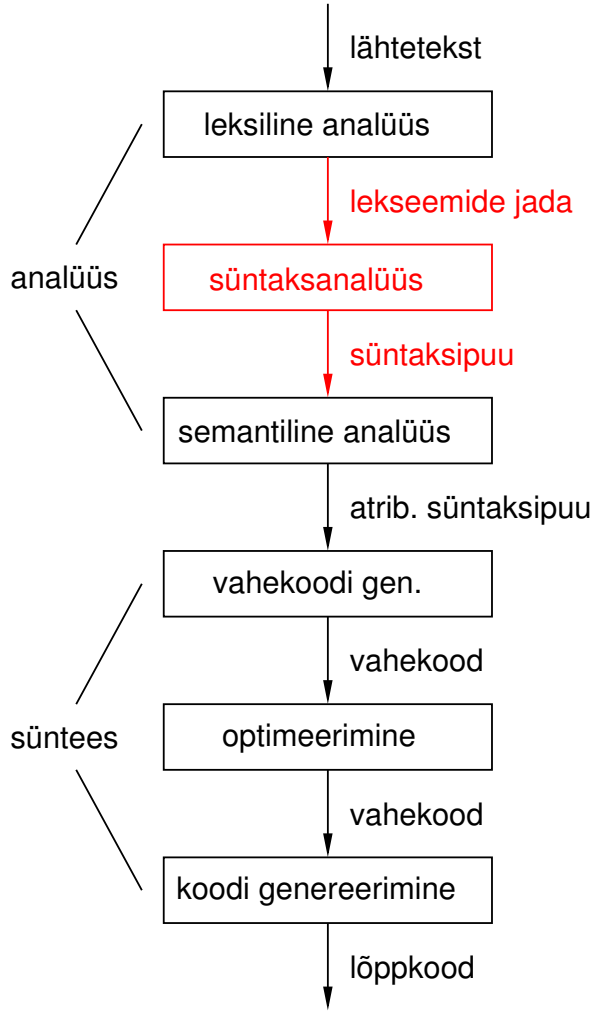
Süntaksanalüüs

Sissejuhatus

Kontekstivabad grammatikad

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 $v \in V^\star$ on otsetuletatav (directly derivable) lausevormist $u \in V^\star$ (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	Tõkestatud Turingi masin
L_2	$A \to \alpha$	Kontekstivabad	Magasinmäluga automaat
L_3	$A \to w,\ A \to wB$	Regulaarsed	Lõplik automaat
(L_4)	$A \to w$	Lõplikud	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 $N = \{\mathrm{Exp}\}$ ja $T = \{+, *, (,), id\}$, siis

$$\begin{array}{rcl} \mathrm{Exp} & \to & \mathrm{Exp} + \mathrm{Exp} \\ & | & \mathrm{Exp} * \mathrm{Exp} \\ & | & (\ \mathrm{Exp}\ ) \\ & | & id \end{array}$$

esitab produktsioonireeglite hulka

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

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:

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

Kontekstivabad grammatikad

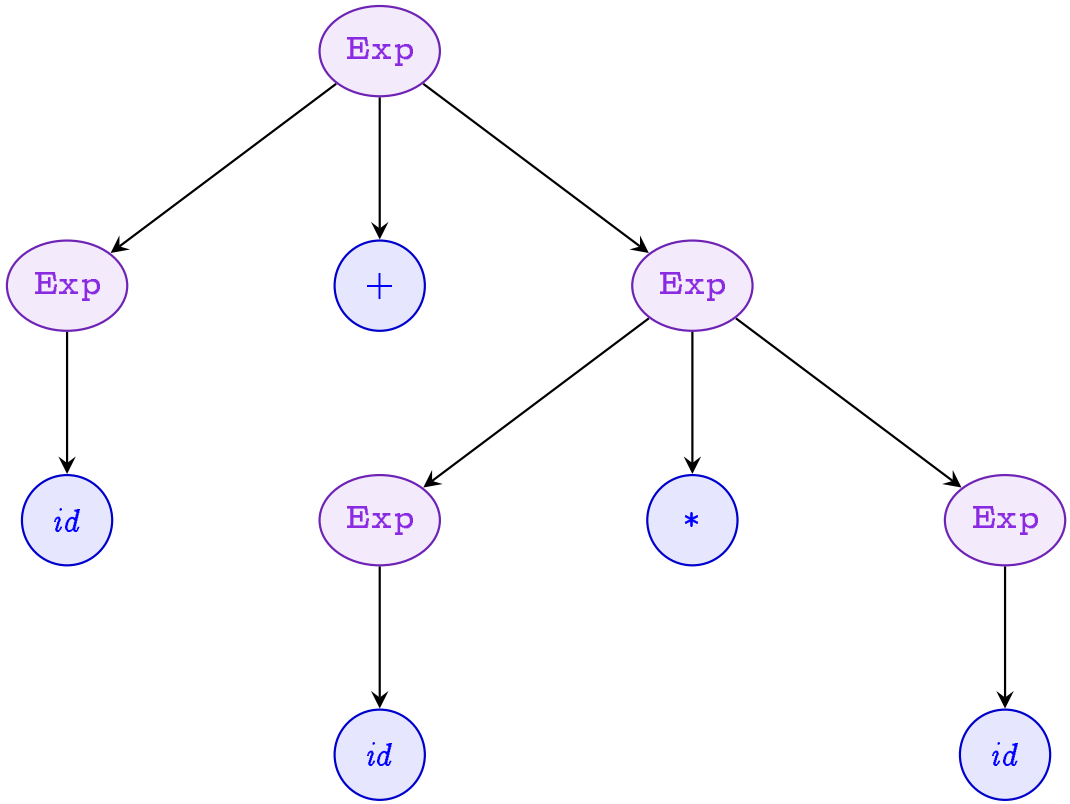
- Iga derivatsioon määrab üheselt süntaksipuu (syntax-tree, parse-tree), mis on järjestatud tippudega puu, kus:
  - puu juur on märgendatud algsümboliga S;
  - vahetipud on märgendatud mitteterminaalidega;
  - lehed on märgendatud terminaalsümboliga või tühisümboliga ε;
  - kui vahetipp on märgendatud 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 produktsiooni-reeglid, kuid mitte nende rakendamise järjekorda.

Kontekstivabad grammatikad

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

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

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



Kontekstivabad grammatikad

NB! Ühel lausel võib olla mitu erinevat süntaksipuud!

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