Syntaks og semantik

Lektion 6

1 marts 2007

Kontekstfrie grammatikker

En kontekstfri grammatik:

$$S \xrightarrow{1} ASB$$
$$S \xrightarrow{2} \varepsilon$$
$$A \xrightarrow{3} 0$$
$$B \xrightarrow{4} 1$$

- S, A, B: variable
- 0, 1: terminaler
- startvariablen: S

At *generere ord*:

- $S \stackrel{2}{\Longrightarrow} \varepsilon$ ✓
- $S \stackrel{1}{\Longrightarrow} ASB \stackrel{2}{\Longrightarrow} A\varepsilon B \stackrel{3}{\Longrightarrow} 0B \stackrel{4}{\Longrightarrow} 01$ ✓
- $S \stackrel{1}{\Longrightarrow} ASB \stackrel{1}{\Longrightarrow} AASBB \stackrel{2}{\Longrightarrow} AA\varepsilon BB \stackrel{3,3,4,4}{\Longrightarrow} 0011$ ✓
- $S \stackrel{1,\dots,1}{\Longrightarrow} A^n S B^n \stackrel{2}{\Longrightarrow} A^n \varepsilon B^n \stackrel{3,4}{\Longrightarrow} 0^n 1^n$
- grammatikken genererer sproget $\{0^n 1^n \mid n \in \mathbb{N}_0\}$

Definition 2.2: En kontekstfri grammatik (CFG) er en 4-tupel $G = (V, \Sigma, R, S)$, hvor delene er

1. V : en endelig mængde af variable
2. Σ : en endelig mængde af terminaler, med $V \cap \Sigma = \emptyset$
3. $R : V \to \mathcal{P}\big((V \cup \Sigma)^*\big)$: produktioner / regler
4. $S \in V$: startvariablen

– produktioner skrives $A \to w$ i stedet for $w \in R(A)$

- Hvis $u, v, w \in (V \cup \Sigma)^*$ er ord og $A \to w$ er en produktion, siges uAv at frembringe uwv: $uAv \Rightarrow uwv$.
- Hvis $u, v \in (V \cup \Sigma)^*$ er ord, siges u at derivere v: $u \overset{*}{\Rightarrow} v$, hvis $u = v$ eller der findes en følge $u_1, u_2, \dots, u_k$ af ord således at $u \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \dots \Rightarrow u_k \Rightarrow v$.
- Sproget som G genererer er $[\![G]\!] = \{w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w\}$.

– dvs. et ord $w \in \Sigma^*$ genereres af G hvis og kun hvis der findes en derivation $S \Rightarrow w_1 \Rightarrow w_2 \Rightarrow \dots \Rightarrow w_k \Rightarrow w$, hvor alle $w_i \in (V \cup \Sigma)^*$.

Eksempel 2.3: $G_3 = (\{S\}, \{a, b\}, R, S)$ med produktioner

$$S \to aSb \mid SS \mid \varepsilon$$

Et par derivationer:

- $S \Rightarrow \varepsilon$
- $S \Rightarrow aSb \Rightarrow ab$
- $S \Rightarrow aSb \Rightarrow aSSb \Rightarrow aaSbSb \Rightarrow aaSbaSbb \Rightarrow aababb$

Generelt er det nok at opskrive *produktionerne* for at specificere en kontekstfri grammatik:

- de variable er venstresiderne
- terminalerne er alle andre bogstaver
- startvariablen er venstresiden af den *øverste* produktion

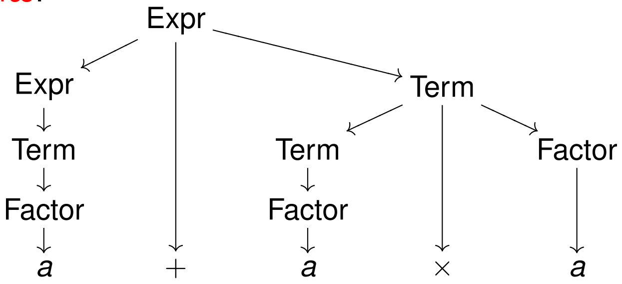
Eksempel 2.4: Aritmetiske udtryk

$$
\begin{aligned}
\text{Expr} &\longrightarrow \text{Expr} + \text{Term} \mid \text{Term} \\
\text{Term} &\longrightarrow \text{Term} \times \text{Factor} \mid \text{Factor} \\
\text{Factor} &\longrightarrow (\text{Expr}) \mid a
\end{aligned}
$$

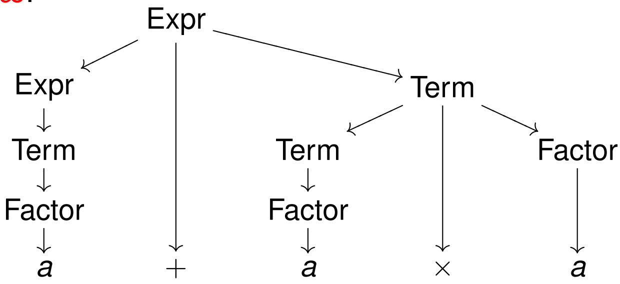
En derivation:

$$
\begin{aligned}
\text{Expr} \Rightarrow \text{Expr}+\text{Term} \Rightarrow \text{Term}+\text{Term} &\overset{*}{\Rightarrow} \text{Factor}+\text{Term}\times\text{Factor} \\
&\Rightarrow \text{Factor}+\text{Factor}\times\text{Factor} \overset{*}{\Rightarrow} a+a\times a
\end{aligned}
$$

Et parse-træ:



Et parsetræ:



- Parsetræer udtrykker betydningen af et ord
- At parse: programkode $\rightsquigarrow$ parsetræ $\rightsquigarrow$ ...
- En kontekstfri grammatik i hvilken der er et ord der har to *forskellige* parsetræer kaldes tvetydig.
- to forskellige parsetræer $\Rightarrow$ to forskellige *betydninger* $\Rightarrow$ BAD

Opsummering:

- CFG: et (endeligt) antal produktioner af formen $A \to s_1 \mid s_2 \mid \dots s_k$ for symboler A og strenge $s_1, s_2, \dots, s_k$.
- "|" kendetegner *alternativer* (nondeterminisme!)
- symboler på venstre side af produktionerne: variable (eller non-terminaler)
- alle andre symboler: terminaler
- venstre side af *første* produktion: startsymbolet
- at frembringe: $uAv \Rightarrow uwv$ hvis $A \to w$ er en produktion
- hvis w er en streng af *terminaler*: grammatikken genererer w hvis der findes en derivation $S \Rightarrow w_1 \Rightarrow \dots \Rightarrow w_k \Rightarrow w$, hvor alle w_i er strenge af terminaler og variable.
- vigtigt: parsetræer
- Definition: Et sprog siges at være kontekstfrit hvis det kan genereres af en CFG.

Eksempel: En CFG til sproget

$$\{w \in \{a,b\}^* \mid \text{antallet af } a \text{ i } w = \text{antallet af } b \text{ i } w\}$$

Idé: Variable som *tilstande*:

- S : Jeg har set lige mange a og b
- A : Jeg mangler et a
- B : Jeg mangler et b

$$\begin{aligned} S &\to aB \mid bA \mid \varepsilon \\ A &\to aS \mid bAA \\ B &\to bS \mid aBB \end{aligned}$$

(opgave 2.21!)

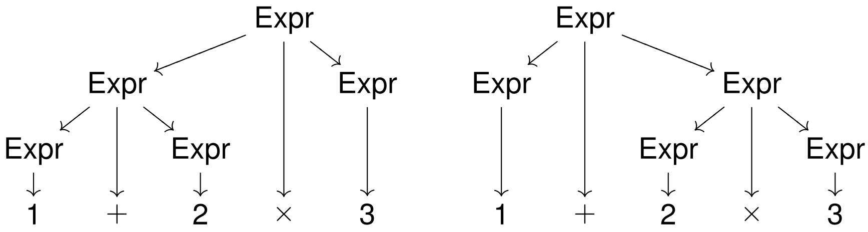
Eksempel: En (ufuldstændig og ikke helt rigtig) CFG for **Sok**

$$
\begin{aligned}
\text{ProGram} &\to \text{VarErkList}\,;\,\text{MetErkList}\\
\text{VarErkList} &\to \text{VarErk}\,;\,\text{VarErkList} \mid \varepsilon\\
\text{VarErk} &\to \textit{var}\ \text{VarNavn} = \text{HelTal}\\
\text{MetErkList} &\to \text{MetErk}\,;\,\text{MetErkList} \mid \varepsilon\\
\text{MetErk} &\to \textit{metode}\ \text{MetNavn}\ \text{StateMentList}\ \textit{end}\\
\text{StateMentList} &\to \text{StateMent}\,;\,\text{StateMentList} \mid \varepsilon\\
\text{StateMent} &\to \text{MetKald} \mid \text{TilSkriv}\\
\text{TilSkriv} &\to \text{VarNavn} := \text{ArUdtryk}\\
\text{MetKald} &\to \textit{kald}\ \text{MetNavn}
\end{aligned}
$$

Eksempel: Grammatikken G_5, ca.:

$$\text{Expr} \to \text{Expr} + \text{Expr} \mid \text{Expr} \times \text{Expr} \mid (\text{Expr}) \mid \text{Heltal}$$

To forskellige parsetræer for $1 + 2 \times 3$:



$= 9$ $= 7$

Definition: En derivation $S \Rightarrow w_1 \Rightarrow w_2 \Rightarrow \ldots \Rightarrow w_k$ i en grammatik kaldes en venstre-derivation hvis det i ethvert skridt er den variable *længst til venstre* der erstattes.

Eksempel:

- $S \Rightarrow AB \Rightarrow aB \Rightarrow ab$ er en venstre-derivation,
- $S \Rightarrow AB \Rightarrow Ab \Rightarrow ab$ er ikke.

Bemærk: Til ethvert parsetræ svarer en entydig venstre-derivation.

Definition 2.7:

- Et ord siges at være genereret tvetydigt hvis det har to forskellige venstre-derivationer.
- En grammatik er tvetydig hvis den genererer et ord på en tvetydig måde.
- Et kontekstfrit sprog er *inherently ambiguous* hvis enhver CFG der genererer det er tvetydig.

Mål: specielle former for kontekstfrie grammatikker som er nemme at håndtere

Definition 2.8: En CFG med startvariabel S er i Chomsky-normalform hvis hver produktion er af formen $A \to BC$ eller $A \to a$, hvor a er en terminal, A, B og C er variable og $B, C \neq S$. Desuden tillades produktionen $S \to \varepsilon$.

Sætning 2.9: Ethvert kontekstfrit sprog genereres af en CFG i Chomsky-normalform.

Bevis: Lad (V, Σ, R, S) være en CFG. Vi konverterer den til Chomsky-normalform:

1. S må ikke forekomme på højresider.
   Introducér en ny startvariabel S_0 og en produktion $S_0 \rightarrow S$.

Bevis: Lad (V, Σ, R, S) være en CFG. Vi konverterer den til Chomsky-normalform:

1. S må ikke forekomme på højresider.
2. Vi vil ikke have ε-produktioner $A \to \varepsilon$, medmindre $A = S$.
   - Tag en produktion $A \to \varepsilon$ og fjern den.
   - For alle produktioner $R \to uAv$: introducér en ny produktion $R \to uv$.
   - Men hvis der er en produktion $R \to A$, introduceres $R \to \varepsilon$ *kun hvis den ikke allerede før er blevet fjernet*.
   - Gentag indtil alle ε-produktioner er væk (undtaget måske $S \to \varepsilon$).

Bevis: Lad (V, Σ, R, S) være en CFG. Vi konverterer den til Chomsky-normalform:

1. S må ikke forekomme på højresider.
2. Vi vil ikke have ε-produktioner $A \to \varepsilon$, medmindre $A = S$.
3. Vi vil ikke have *unit rules*: produktioner af formen $A \to B$.
    - Tag en produktion $A \to B$ og fjern den.
    - For alle produktioner $B \to u$: introducér en ny produktion $A \to u$.
    - Men hvis der er en produktion $B \to C$, introduceres $A \to C$ *kun hvis den ikke allerede før er blevet fjernet*.
    - Gentag indtil alle *unit rules* er væk.

Bevis: Lad (V, Σ, R, S) være en CFG. Vi konverterer den til Chomsky-normalform:

1. S må ikke forekomme på højresider.
2. Vi vil ikke have ε-produktioner $A \to \varepsilon$, medmindre $A = S$.
3. Vi vil ikke have *unit rules*: produktioner af formen $A \to B$.
4. Vi vil ikke have produktioner af formen $A \to u_1 u_2 \dots u_k$ for $k \geq 3$.
   - Lad $A \to u_1 u_2 \dots u_k$ være en sådan produktion. (Her er $u_i$erne variable eller terminaler.)
   - Erstat den med produktioner $A \to u_1 A_1$, $A_1 \to u_2 A_2, \dots, A_{k-2} \to u_{k-1} u_k$, hvor $A_i$erne er nye variable.
   - Gentag.

Bevis: Lad (V, Σ, R, S) være en CFG. Vi konverterer den til Chomsky-normalform:

1. S må ikke forekomme på højresider.
2. Vi vil ikke have ε-produktioner $A \to \varepsilon$, medmindre $A = S$.
3. Vi vil ikke have *unit rules*: produktioner af formen $A \to B$.
4. Vi vil ikke have produktioner af formen $A \to u_1 u_2 \ldots u_k$ for $k \geq 3$.
5. Vi vil ikke have produktioner af formen $A \to bC$, $A \to Bc$ eller $A \to bc$.
   - Erstat $A \to bC$ med $A \to BC$ og $B \to b$, og gør lignende for de andre to. (Igen introduceres nye variable.)
6. Færdig!