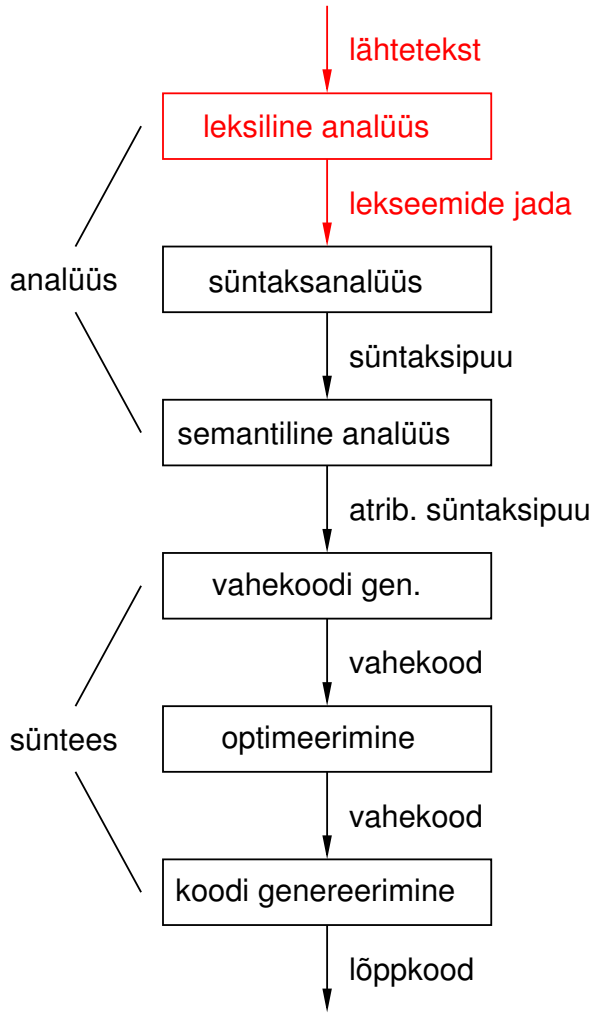
Leksiline analüüs

Sissejuhatus
Regulaaravaldised

Leksiline analüüs

- Leksiline analüüs kontrollib programmi sõnade (literaalsümbolite) vastavust leksilistele reeglitele ning teisendab programmi sümbolite (tokens) jadaks:
  - eemaldab tühisümbolid ja kommentaarid;
  - identifitseerib võtmesõnad, identifikaatorid ja konstandid;
  - konstrueerib sümbolite tabeli;
  - leiab sümbolite rea- ja veerunumbrid;
  - teavitab vajadusel leksilistest vigadest.
- Leksilist analüüsi kutsutakse skaneerimiseks (scanning) ning vastavat analüsaatorit nimetatakse skanneriks (scanner).



Käsitsi kodeeritud skanner

Näide: võtmesõna `if` äratundmine:

```
c = readchar();
if (c != 'i')
    error();
else {
    c = readchar();
    if (c != 'f')
        error();
    else
        return IF_TOKEN;
}
```

Käsitsi kodeeritud skanner

Probleemid:

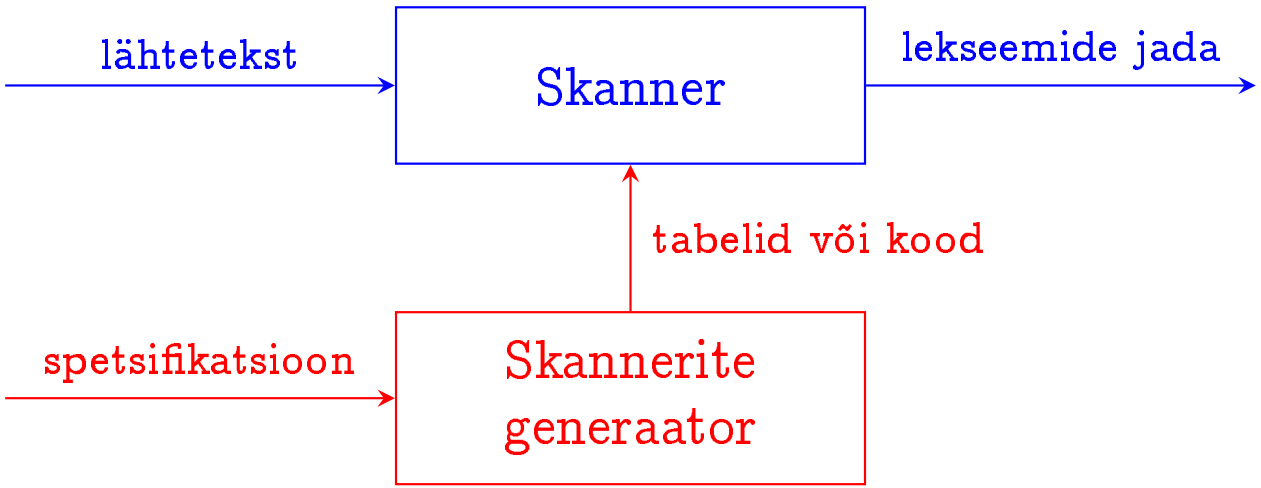
- Vaja kombineerida erinavat liiki lekseemide skannereid
  - Fikseeritud sõnad ja märgijadad (võtmesõnad, operaatorid)
  - Reeglite abil defineeritud lekseemid (identifikaatorid, arvud)
- Lekseemid võivad kattuda
  - = vs. ==
  - `if` vs. `if0`

Käsitsi kodeeritud skanner

Näide: `if` on võtmesõna, kuid `if0` identifikaator:

```
c = readchar();
if (c != 'i') { /* teised lekseemid... */ }
else {
    c = readchar();
    if (c != 'f') { /* teised lekseemid... */ }
    else {
        c = readchar();
        if (c not alpha-numeric) {
            putback(c);
            return IF_TOKEN;
        }
        while (c alpha-numeric) {
            /* ehitada identifikaator */
        }
    }
}
```

Skannerite generaator



- Kuna käsitsi kodeerimine on tülikas ja vigade aldis, siis tavaliselt genereeritakse automaatselt skannerite generaatori abil.
- Programmi sõnade leksilised reeglid esitatakse tavaliselt regulaaravaldiste abil.

Regulaaravaldised

Regulaaravaldised

- Regulaaravaldised üle (lõpliku) tähestiku Σ

$$E ::= \varepsilon \mid \mathsf{a} \mid (E\ E) \mid (E \mid E) \mid E^\star$$

  kus $\mathsf{a} \in \Sigma$.

- Regulaaravaldis E defineerib keele $L(E) \subseteq \Sigma^\star$

$$\begin{array}{lcl} L(\varepsilon) & = & \{""\} \\ L(\mathsf{a}) & = & \{"\mathsf{a}"\} \\ L(E_1\ E_2) & = & \{vw \mid v \in L(E_1),\ w \in L(E_2)\} \\ L(E_1 \mid E_2) & = & L(E_1) \cup L(E_2) \\ L(E^\star) & = & \{""\} \cup \{vw \mid v \in L(E), w \in L(E^\star)\} \end{array}$$

Regulaaravaldised

Näiteid:

Regulaaravaldis	Defineeritav keel
$a \mid b$	{"a", "b"}
abba	{"abba"}
$ab^\star a$	{"aa", "aba", "abba", "abbba", ...}
$(ab)^\star$	{"", "ab", "abab", "ababab", ...}
$(a \mid b)^\star$	{"", "a", "b", "aa", "ab", "ba", "aaa", "aab", "aba", "baa", "abb", "bab", "bba", "bbb", ...}

Regulaaravaldised

- Regulaaravaldistes esinevate sulgude vähendamiseks on operaatoritele määratud prioriteedid:
  - sulundioperaator $(\cdot)^\star$ seob kõige tugevamalt;
  - valikuoperaator $(\cdot \mid \cdot)$ seob kõige nõrgemalt.
- Lühendavaid tähistusi regulaaravaldiste esitamiseks:
  - *mittetühi sulund*: $E^+ = EE^\star$;
  - *optsioon*: $E? = \varepsilon \mid E$;
  - *märgihulgad*: $[\mathsf{abc}] = \mathsf{a} \mid \mathsf{b} \mid \mathsf{c}$;
  - *märgivahemikud*: $[\mathsf{a} - \mathsf{z}] = \mathsf{a} \mid \ldots \mid \mathsf{z}$.

Regulaaravaldised

Regulaarne kirjeldus tähestikus Σ on reeglite hulk

$$\begin{aligned} d_1 &\to E_1 \\ d_2 &\to E_2 \\ &\dots \\ d_n &\to E_n \end{aligned}$$

kus d_i on (unikaalne) nimi ja E_i on regulaaravaldis tähestikus $\Sigma \cup \{d_1, \dots, d_{i-1}\}$.

Regulaaravaldised

Näiteid regulaarsetest kirjeldustest:

Identifikaatorid:

$$\begin{aligned} \text{Letter} &\rightarrow [\texttt{a} - \texttt{z}\ \texttt{A} - \texttt{Z}] \\ \text{Digit} &\rightarrow [\texttt{0} - \texttt{9}] \\ \text{Identifier} &\rightarrow \text{Letter}\,(\text{Letter} \mid \text{Digit})^{\star} \end{aligned}$$

Arvkonstandid:

$$\begin{aligned} \text{Sign} &\rightarrow (\texttt{+} \mid \texttt{-})? \\ \text{Integer} &\rightarrow \texttt{0} \mid \text{Sign}\,[\texttt{1} - \texttt{9}]\,\text{Digit}^{\star} \\ \text{Decimal} &\rightarrow \text{Integer}\,\texttt{.}\,\text{Digit}^{+} \\ \text{Real} &\rightarrow (\text{Integer} \mid \text{Decimal})\,\texttt{E}\,\text{Integer} \end{aligned}$$