

4. **Tehtävä:** Osoita, että yhteydettömien kielten luokka on suljettu yhdiste-, katenaatio- ja sulkeumaoperaatioiden suhteen, so. jos kielet $L_1, L_2 \subseteq \Sigma^*$ ovat yhteydettömiä, niin samoin ovat myös kielet $L_1 \cup L_2$, L_1L_2 ja L_1^* .

Vastaus: Olkoon L_1 ja L_2 yhteydettömiä kieliä. Tällöin on olemassa kieliopit $G_1 = (V_1, \Sigma_1, R_1, S_1)$ ja $G_2 = (V_2, \Sigma_2, R_2, S_2)$, siten, että $L(G_1) = L_1$ ja $L(G_2) = L_2$. Vaaditaan lisäksi, että $(V_1 - \Sigma_1) \cap (V_2 - \Sigma_2) = \emptyset$, eli kieliopeissa ei esiinny samoja väliskeitä. Koska kieliopin väliskeet voidaan tarvittaessa nimetä uudelleen, ei tämä aseta oleellista rajoitusta.

Unioni: Olkoon S uusi väliske ja $G = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, R_1 \cup R_2 \cup \{S \rightarrow S_1 \mid S_2\}, S)$. Nyt $L(G) = L(G_1) \cup L(G_2) = L_1 \cup L_2$. Näin on, koska S :stä voidaan johtaa vain S_1 tai S_2 , joista voidaan edelleen johtaa vain sanoja jotka kuuluvat jompaan kumpaan aiemmista kielistä (sääntöjen sekaannukselta vältytään, koska väliskejoukot ovat pistevieraita).

Katenaatio: Tällä kertaa uusi kielioppi $G = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, R_1 \cup R_2 \cup \{S \rightarrow S_1S_2\}, S)$. Nyt $L(G) = L_1L_2$.

Kleenen tähti: Tällä kertaa uusi kielioppi $G = (V_1 \cup \{S\}, \Sigma_1, R_1 \cup \{S \rightarrow \epsilon \mid SS_1\}, S)$. Nyt $L(G) = L_1^*$

5. **Tehtävä:**

- (a) Osoita, että seuraava yhteydetön kielioppi on moniselitteinen:

$$\begin{aligned} S &\rightarrow \text{if } b \text{ then } S \\ S &\rightarrow \text{if } b \text{ then } S \text{ else } S \\ S &\rightarrow s. \end{aligned}$$

- (b) Muodosta (a)-kohdan kieliopin kanssa ekvivalentti, so. saman kielen tuottava yksiselitteinen kielioppi.

Vastaus: Yhteydetön kielioppi G on moniselitteinen, mikäli on olemassa sana $w \in L(G)$ siten, että w :llä on ainakin kaksi erilaista jäsennyyspuuta. Tehtävän kieliopille yksinkertaisin tällainen sana on:

$$\text{if } b \text{ then if } b \text{ then } s \text{ else } s,$$

joka voidaan jäsentää kahdella tapaa:

$$\begin{array}{ccccccc} & & & & S & & \\ & & & & & & \\ \text{if} & b & \text{then} & & S & & \\ & & & & & & \\ & \text{if} & b & \text{then} & S & \text{else} & S \\ & & & & & & \\ & & & & s & & s \end{array}$$

$$S$$

$$\text{if } b \text{ then } S \text{ else } S$$

$$\text{if } b \text{ then } S \text{ } s$$

$$s$$

Yleensä ohjelmointikielissä halutaan **else**-lause liittää lähimpään mahdolliseen **if**-lauseeseen. Ylläolevista puista ensimmäinen vastaa tätä käytäntöä.

Määritellään kielioppi seuraavasti:

$$G = (V, \Sigma, P, S)$$

$$V = \{S, B, U, s, b, \text{if}, \text{then}, \text{else}\}$$

$$\Sigma = \{s, b, \text{if}, \text{then}, \text{else}\}$$

$$P = \{S \rightarrow B \mid U$$

$$B \rightarrow \text{if } b \text{ then } B \text{ else } B \mid s$$

$$U \rightarrow \text{if } b \text{ then } S \mid \text{if } b \text{ then } B \text{ else } U\}$$

Tässä välikkeellä B saadaan johdettua vain ohjelmia, joissa kaikilla **if**-lauseilla on sekä **then**- että **else**-haarat. Välikkeellä U johdetaan sitten **if**-lauseet, joista puuttuu **else**-haara.

6. **Tehtävä:** Laadi rekursiivisesti etenevä jäsentäjä edellisten harjoitusten tehtävän 6 kieliopille.

Vastaus: Alla oleva C-ohjelma toteuttaa rekursiivisen jäsentäjän kieliopille:

$$C \rightarrow S \mid S; C$$

$$S \rightarrow a \mid \text{begin } C \text{ end} \mid \text{for } n \text{ times do } S$$

Tässä on yksinkertaistettu hieman edellisen laskuharjoituskerran 6. tehtävän kielioppia korvaamalla erilliset numerot terminaalilla n , joka tarkoittaa mitä tahansa numeroa.

Tärkeimmät ohjelmassa esiintyvät funktiot ovat:

- `C()`, `S()` — toteuttavat kieliopin varsinaiset säännöt
- `lex()` — lukee syötteestä seuraavan lekseemin ja tallettaa sen globaaliin muuttujaan `current_tok`.
- `expect(int token)` — yrittää lukea syötteestä lekseemin `token`. Mikäli lukeminen epäonnistuu annetaan virheilmoitus.
- `consume_token()` — merkitään tämänhetkinen lekseemi käytetyksi. Tämä (tai jokin muu vastaava funktio) tarvitaan siksi, että joissain tapauksissa täytyy syötettä lukea yksi lekseemi eteenpäin ennen kuin tiedetään, mitä sääntöä täytyy käyttää.

Käytännössä ohjelmointikielten jäsentäjät toteutetaan yleensä käyttäen *lex*- ja *yacc*-työkaluja¹. Näistä *lex* muodostaa tilakonepohjaisen selaaajan, joka tunnistaa säännöllisillä lausekkeilla määritellyt lekseemit, ja *yacc* tekee pinoautomaattipohjaisen jäsentimen annetulle yhteydettömälle kieliopille.

¹Tai niiden johdannaisia.

```

#include <stdio.h>
#include <stdlib.h>
#include <ctype.h>

/* Define the alphabet */
enum TOKEN { DO, FOR, END, BEGIN, TIMES, OP, SC, NUMBER, ERROR };
const char* tokens[] = { "do", "for", "end", "begin", "times", "a",
                        ";", "NUMBER", NULL };

/* A global variable holding the current token */
int current_tok = ERROR;

/* Maximum length of a token */
#define TOKEN_LEN 128

/* declare functions corresponding to nonterminals */
void S(void);
void C(void);

int lex(void);
void consume_token(void);
void error(char *st);
void expect(int token);

void C(void)
{
    S();
    lex();
    if (current_tok == SC) {
        consume_token();
        C();
        printf("C => S ; C\n");
    } else {
        printf("C => S\n");
    }
}

void S(void)
{
    lex();
    switch (current_tok) {
    case OP:
        consume_token();
        printf("S => a\n");
        break;
    case BEGIN:
        consume_token();
        C();
        expect(END);
        printf("S => begin C end\n");
        break;
    case FOR:
        consume_token();
        expect(NUMBER);

```

```

    expect(TIMES);
    expect(DO);
    S();
    printf("S => for N times do S\n");
    break;
default:
    error("Parse error");
}
}

/* int lex(void) returns the next token of the input. */
int lex(void)
{
    static char token_text[TOKEN_LEN];
    int pos = 0, c, i, next_token = ERROR;

    /* Is there an existing token already? */
    if (current_tok != ERROR)
        return current_tok;

    /* skip whitespace */
    do {
        c = getchar();
    } while (c != EOF && isspace(c));
    if (c != EOF) ungetc(c, stdin);

    /* read token */
    c = getchar();
    while (c != EOF && c != ';' && !isspace(c) && pos < TOKEN_LEN) {
        token_text[pos++] = c;
        c = getchar();
    }
    if (c == ';') {
        if (pos == 0) /* semicolon as token */
            next_token = SC;
        else /* trailing semicolon, leave it for future */
            ungetc(';', stdin);
    }
}
token_text[pos] = '\0'; /* trailing zero */

/* identify token */
if (isdigit(token_text[0])) { /* number? */
    next_token = NUMBER;
} else { /* not a number */
    for (i = DO; i < NUMBER; i++) {
        if (!strcmp(tokens[i], token_text)) {
            next_token = i;
            break;
        }
    }
}
current_tok = next_token;
return next_token;

```

```

}

void consume_token(void)
{
    current_tok = ERROR;
}

void error(char *st)
{
    printf(st);
    exit(1);
}

/* try to read a 'token' from input */
void expect(int token)
{
    int next_tok = lex();
    if (next_tok == token) {
        consume_token();
        return;
    } else
        error("Parse error");
}

int main(void)
{
    int i;
    C();
    return 0;
}

```

Liite: oikealle lineaariset kieliopit

Yhteydetön kielioppi G on oikealle lineaarinen, mikäli sen kaikki säännöt ovat muotoa:

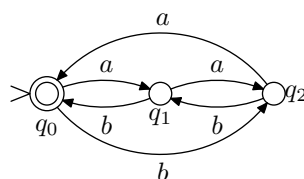
$$A \rightarrow \alpha B,$$

missä $\alpha \in \Sigma^*$ ja $B \in V \cup \{\varepsilon\}$. Toisin sanoen, säännön oikealla puolella saa esiintyä korkeintaan yksi välike, ja sen täytyy olla säännön lopussa. Esimerkiksi säännöt $A \rightarrow abC$ ja $A \rightarrow \varepsilon$ ovat oikealle lineaarisia, mutta $A \rightarrow Ba$ ja $A \rightarrow abCD$ eivät ole. Vastaavasti vasemmalle lineaarisen kieliopin kaikki säännöt ovat muotoa $A \rightarrow B\alpha$.

Lineaariset kieliopit ovat ilmaisuvoimaltaan samalla tasolla kuin äärelliset automaattit, eli niillä voidaan ilmaista kaikki säännölliset kielet.

Säännöllisen kielen tuottava oikealle lineaarinen kielioppi voidaan lukea suoraan kielen tunnistavasta automaatista. Kieliopin välikkeiksi otetaan automaatin tilat, ja kustakin automaatin siirtymästä $q_i \xrightarrow{a} q_j$ saadaan sääntö $Q_i \rightarrow aQ_j$. Lisäksi kielioppiin lisätään säännöt $Q_f \rightarrow \varepsilon$ kaikille hyväksyville lopputiloille $q_f \in F$.

Esimerkiksi automaattia:



vastaava kielioppi on:

$$Q_0 \rightarrow aQ_1 \mid bQ_2 \mid \varepsilon$$

$$Q_1 \rightarrow aQ_2 \mid bQ_0$$

$$Q_2 \rightarrow aQ_0 \mid bQ_1 .$$