Limbaje Formale, Automate şi Compilatoare

Curs 1

Limbaje Formale, Automate şi Compilatoare

Curs:

- O.Captarencu: otto@infoiasi.ro,
  `http://www.infoiasi.ro/~otto/lfac.html`
- Gh. Grigoraş: grigoras@info.uaic.ro

Laboratoare:

- O.Captarencu
- A. Moruz: mmoruz@info.uaic.ro
- C. Liţă: clita@bitdefender.com

Pagina cursului:

- `http://www.infoiasi.ro/~otto/lfac.html`

Evaluare

- 7 seminarii, 6 laboratoare;
- AS = activitatea la seminar (max 10 puncte);
- AL = activitatea la laborator (max 10 puncte);
- T1,T2 teste scrise în săptămânile 8, respectiv în sesiune;
- Punctajul final se obţine astfel:

$$P = 3 * AS + 3 * AL + 2 * T1 + 2 * T2$$

- Condiţii miminale de promovare: $AS \geq 5$, $AL \geq 5$;
- Punctaj minim pentru promovare: $P \geq 50$;
- Nota finală se va stabili conform criteriilor ECTS;

Evaluare

- AS = activitatea la seminar (max 10 puncte)
  - 8 puncte din 2 teste scrise
  - 2 puncte activitatea la seminar
- AL = activitatea la laborator (max 10 puncte)
  - 1 test scris, 2 teme laborator (note de la 0 la 10)
  - AL = media celor 3 note

Tematica cursului I

- Limbaje şi gramatici
- Limbaje regulate; gramatici, automate , expresii regulate
- Limbaje independente de context; gramatici, automate pushdown
- Maşini Turing

Tematica cursului II

- Limbaje de programare: proiectare şi implementare
- Analiza lexicală
- Analiza sintactică
- Traducere în cod intermediar

Tematica seminarului

- Exemple de limbaje şi gramatici
- Automate finite deterministe, nedeterministe, cu epsilon-tranziţii - Exemple
- Expresii regulate
- Gramatici independente de context, arbori de derivare, eliminarea simbolurilor inutile, eliminarea regulilor de ştergere, a redenumirilor
- Forma normală Chomsky, algoritmul CYK
- Automate pushdown - exemple

Tematica laboratorului

- Analiza lexicală folosind instrumente de tip LEX
- Analiza sintactică folosind instrumente de tip YACC
- Interpretor construit cu LEX si YACC

Bibliografie

- Grigoras, Gh. Constructia compilatoarelor - Algoritmi fundamentali, Ed. Universitatii Al. I. ”Cuza Iasi”, ISBN 973-703-084-2, 274 pg., 2005
- Jucan Toader - Limbaje formale si automate, Editura Matrix Rom, Bucuresti, 1999, 162 p.
- Jucan Toader, Stefan Andrei Limbaje formale si teoria automatelor. Teorie si practica, Editura Universitatii Al. I. Cuza, Iasi, 2002, 327p.
- Hopcroft, John E.; Motwani, Rajeev; Ullman, Jeffrey D. (2006).Introduction to Automata Theory, Languages, and Computation(3rd ed.). Addison-Wesley
- Stoughton Alley, Formal Language Theory, Kansas State University, Draft of Fall 2007.
- Manual LEX, Manual FLEX, Manual YACC, Manual Bison, Compiler Construction using Flex and Bison

Curs 1

Alfabet, cuvânt, mulţime de cuvinte

- Alfabet: V o mulţime finită (elemente lui $V =$ simboluri)
  - Simbolurile le vom nota a, b, etc.

Alfabet, cuvânt, mulţime de cuvinte

- Alfabet: V o mulţime finită (elemente lui $V =$ simboluri)
  - Simbolurile le vom nota a, b, etc.
- Cuvânt: şir finit de simboluri
  - cuvintele le vom nota u, v, w...
  - cuvântul nul notat ϵ sau λ

Alfabet, cuvânt, mulţime de cuvinte

- Alfabet: V o mulţime finită (elemente lui $V =$ simboluri)
  - Simbolurile le vom nota a, b, etc.
- Cuvânt: şir finit de simboluri
  - cuvintele le vom nota u, v, w...
  - cuvântul nul notat ϵ sau λ
- Lungimea unui cuvânt u: numarul simbolurilor sale. Notaţie: $|u|$
  $|\epsilon| = 0$

Alfabet, cuvânt, mulţime de cuvinte

- Alfabet: V o mulţime finită (elemente lui $V =$ simboluri)
  - Simbolurile le vom nota a, b, etc.
- Cuvânt: şir finit de simboluri
  - cuvintele le vom nota u, v, w...
  - cuvântul nul notat ϵ sau λ
- Lungimea unui cuvânt u: numarul simbolurilor sale. Notaţie: $|u|$
  $|\epsilon| = 0$
- V^* - mulţimea tuturor cuvintelor peste alfabetul V, inclusiv ϵ
  - $\{0, 1\}^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, 001, ...\}$

Alfabet, cuvânt, mulţime de cuvinte

- Alfabet: V o mulţime finită (elemente lui $V =$ simboluri)
  - Simbolurile le vom nota a, b, etc.
- Cuvânt: şir finit de simboluri
  - cuvintele le vom nota u, v, w...
  - cuvântul nul notat ϵ sau λ
- Lungimea unui cuvânt u: numarul simbolurilor sale. Notaţie: $|u|$
  $|\epsilon| = 0$
- V^* - mulţimea tuturor cuvintelor peste alfabetul V, inclusiv ϵ
  - $\{0, 1\}^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, 001, ...\}$
- V^+ - mulţimea tuturor cuvintelor nenule peste alfabetul V
  - $\{0, 1\}^+ = \{0, 1, 00, 01, 10, 11, 000, 001, ...\}$

- Concatenarea a două cuvinte x, y: cuvântul $x \cdot y$ obţinut din simbolurile lui x, în ordinea în care apar, urmate de cele ale lui y de asemenea în ordinea în care apar:
  - $x = 0100, y = 100, x \cdot y = 0100100$
  - $x = 000, y = \epsilon, x \cdot y = 000$

- Concatenarea a două cuvinte x, y: cuvântul $x \cdot y$ obţinut din simbolurile lui x, în ordinea în care apar, urmate de cele ale lui y de asemenea în ordinea în care apar:
  - $x = 0100, y = 100, x \cdot y = 0100100$
  - $x = 000, y = \epsilon, x \cdot y = 000$
- Concatenarea este asociativă
- ϵ este element neutru

- Concatenarea a două cuvinte x, y: cuvântul $x \cdot y$ obţinut din simbolurile lui x, în ordinea în care apar, urmate de cele ale lui y de asemenea în ordinea în care apar:
  - $x = 0100, y = 100, x \cdot y = 0100100$
  - $x = 000, y = \epsilon, x \cdot y = 000$
- Concatenarea este asociativă
- ϵ este element neutru
- $(V^*, \cdot)$ este monoid, se numeşte monoidul liber generat de V

- Concatenarea a două cuvinte x, y: cuvântul $x \cdot y$ obţinut din simbolurile lui x, în ordinea în care apar, urmate de cele ale lui y de asemenea în ordinea în care apar:
  - $x = 0100, y = 100, x \cdot y = 0100100$
  - $x = 000, y = \epsilon, x \cdot y = 000$
- Concatenarea este asociativă
- ϵ este element neutru
- $(V^*, \cdot)$ este monoid, se numeşte monoidul liber generat de V
- Cuvântul v este un prefix al cuvântului u dacă $\exists w \in V^*, u = vw$; dacă $w \in V^+$, atunci v este un prefix propriu

- Concatenarea a două cuvinte x, y: cuvântul $x \cdot y$ obţinut din simbolurile lui x, în ordinea în care apar, urmate de cele ale lui y de asemenea în ordinea în care apar:
  - $x = 0100, y = 100, x \cdot y = 0100100$
  - $x = 000, y = \epsilon, x \cdot y = 000$
- Concatenarea este asociativă
- ϵ este element neutru
- $(V^*, \cdot)$ este monoid, se numeşte monoidul liber generat de V
- Cuvântul v este un prefix al cuvântului u dacă $\exists w \in V^*, u = vw$; dacă $w \in V^+$, atunci v este un prefix propriu
- Cuvântul v este un sufix al cuvântului u dacă $\exists w \in V^*, u = wv$; dacă $w \in V^+$, atunci v este un sufix propriu

Limbaj formal

- Fie V un alfabet. O submulţime $L \subseteq V^*$ este un limbaj (formal) peste alfabetul V (sau V-limbaj) dacă L are o descriere finită.
- O descriere poate fi:

Limbaj formal

- Fie V un alfabet. O submulţime $L \subseteq V^*$ este un limbaj (formal) peste alfabetul V (sau V-limbaj) dacă L are o descriere finită.
- O descriere poate fi:
  - neformală (în limbaj natural):
    - mulţimea cuvintelor peste alfabetul $\{0, 1\}$ care conţin un număr par de 0 .
    - $L = \{x \in V^+ : |x| \text{este par}\}$.
    - $\{a^n b^n | n \in N\}$.
    - $\{w \in \{0, 1\}^* | w \text{ se termină in } 00\}$.

Limbaj formal

- Fie V un alfabet. O submulţime $L \subseteq V^*$ este un limbaj (formal) peste alfabetul V (sau V-limbaj) dacă L are o descriere finită.
- O descriere poate fi:
  - neformală (în limbaj natural):
    - mulţimea cuvintelor peste alfabetul $\{0, 1\}$ care conţin un număr par de 0 .
    - $L = \{x \in V^+ : |x| \text{este par}\}$.
    - $\{a^n b^n | n \in N\}$.
    - $\{w \in \{0, 1\}^* | w \text{ se termină in } 00\}$.
  - formală (descriere matematică):
    - o descriere inductivă a cuvintelor
    - o descriere generativă a cuvintelor (gramatică generativă)
    - o descriere a unei metode de recunoaştere a cuvintelor din limbaj (automat finit, automat pushdown, etc.)

Operaţii cu limbaje

- Operaţiile cu mulţimi (reuniune, interseţie etc)
- Produs de limbaje: $L_1 \cdot L_2 = \{uv | u \in L_1, v \in L_2\}$
- Iteraţia (produsul Kleene): $L^* = \bigcup_{n \geq 0} L^n$, unde:
  - $L^0 = \{\epsilon\}$
  - $L^{n+1} = L^n L$
- L^R = $\{w^R | w \in L\}$; dacă $w = a_1 a_2 \dots a_n$, atunci $w^R = a_n \dots a_2 a_1$

Curs 1

Gramatici

Definiţie 1

O gramatică este un sistem $G = (N, T, S, P)$*, unde:*

1. N *şi* T *sunt două alfabete disjuncte*
   - N *este mulţimea neterminalilor*
   - T *este mulţimea terminalilor*
2. $S \in N$ *este simbolul de start (neterminalul iniţial)*
3. P *este o mulţine finită de reguli (producţii) de forma* $x \to y$*, unde* $x, y \in (N \cup T)^*$ *şi* x *conţine cel puţin un neterminal.*

Derivare

Fie $G = (N, T, S, P)$ o gramatică şi $u, v \in (N \cup T)^*$.
Spunem că v este derivat direct (într-un pas) de la u prin aplicarea regulii $x \to y$, şi notăm $u \Rightarrow v$, dacă $\exists p, q \in (N \cup T)^*$ astfel încât $u = pxq$ şi $v = pyq$

Derivare

Fie $G = (N, T, S, P)$ o gramatică şi $u, v \in (N \cup T)^*$.
Spunem că v este derivat direct (într-un pas) de la u prin aplicarea regulii $x \to y$, şi notăm $u \Rightarrow v$, dacă $\exists p, q \in (N \cup T)^*$ astfel încât $u = pxq$ şi $v = pyq$

- Dacă $u_1 \Rightarrow u_2 \ldots \Rightarrow u_n$, $n > 1$, spunem că u_n este derivat din u_1 în G şi notăm $u_1 \Rightarrow^+ u_n$.

Derivare

Fie $G = (N, T, S, P)$ o gramatică şi $u, v \in (N \cup T)^*$.
Spunem că v este derivat direct (într-un pas) de la u prin aplicarea regulii $x \to y$, şi notăm $u \Rightarrow v$, dacă $\exists p, q \in (N \cup T)^*$ astfel încât $u = pxq$ şi $v = pyq$

- Dacă $u_1 \Rightarrow u_2 \ldots \Rightarrow u_n$, $n > 1$, spunem că u_n este derivat din u_1 în G şi notăm $u_1 \Rightarrow^+ u_n$.
- Scriem $u \Rightarrow^* v$ dacă $u \Rightarrow^+ v$ sau $u = v$.

Limbaj generat

Definiţie 2

Limbajul generat de gramatica G este:

$$L(G) = \{w \in T^* | S \Rightarrow^+ w\}$$

Limbaj generat

Definiţie 2

Limbajul generat de gramatica G este:

$$L(G) = \{w \in T^* | S \Rightarrow^+ w\}$$

Definiţie 3

Două gramatici G_1 şi G_2 sunt echivalente dacă $L(G_1) = L(G_2)$

Exemplu

- $L = \{a^n b^n | n \geq 1\}$
- Definiţia inductivă:
  - $ab \in L$
  - Daca $X \in L$, atunci $aXb \in L$
  - Nici un alt cuvânt nu face parte din L

Exemplu

- $L = \{a^n b^n | n \geq 1\}$
- Definiţia inductivă:
  - $ab \in L$
  - Daca $X \in L$, atunci $aXb \in L$
  - Nici un alt cuvânt nu face parte din L
- Definiţia generativă:
  - $G = (\{X\}, \{a, b\}, X, P)$, unde $P = \{X \to aXb, X \to ab\}$
  - Derivarea cuvântului a^3b^3:

    $X \Rightarrow aXb \Rightarrow aaXbb \Rightarrow aaabbb$

Exemplu

- $L = \{a^n b^n c^n | n \geq 1\}$
- $G = (N, T, S, P)$, $N = \{S, X\}$, $T = \{a, b, c\}$, P constă din:
  - (1) $S \rightarrow abc$
  - (2) $S \rightarrow aSXc$
  - (3) $cX \rightarrow Xc$
  - (4) $bX \rightarrow bb$

- Derivarea cuvântului $a^3b^3c^3$: $S \Rightarrow^{(2)} a\underline{S}Xc \Rightarrow^{(2)} aa\underline{S}XcXc \Rightarrow^{(1)} aaab\underline{cX}cXc \Rightarrow^{(3)} aaa\underline{bX}ccXc \Rightarrow^{(4)} aaabbc\underline{cX}c \Rightarrow^{(3)} aaabb\underline{cX}cc \Rightarrow^{(3)} aaab\underline{bX}ccc \Rightarrow^{(4)} aaabbbccc = a^3b^3c^3$

Curs 1

Clasificarea gramaticilor

1. **Gramatici de tip 0 (generale)**

   Nu există restricţii asupra regulilor

Clasificarea gramaticilor

1. **Gramatici de tip 0 (generale)**

   Nu există restricţii asupra regulilor

2. **Gramatici de tip 1 (dependente de context)**

   reguli de forma $pxq \rightarrow pyq$ unde $x \in N, y \neq \epsilon, p, q \in (N \cup T)^*$
   $S \rightarrow \epsilon$, caz în care S nu apare în dreapta producţiilor

Clasificarea gramaticilor

1. **Gramatici de tip 0 (generale)**

   Nu există restricţii asupra regulilor

2. **Gramatici de tip 1 (dependente de context)**

   reguli de forma $pxq \rightarrow pyq$ unde $x \in N, y \neq \epsilon, p, q \in (N \cup T)^*$
   $S \rightarrow \epsilon$, caz în care S nu apare în dreapta producţiilor

3. **Gramatici de tip 2 (independente de context)**

   reguli de forma $A \rightarrow y$ unde $A \in N$ şi $y \in (N \cup T)^*$

Clasificarea gramaticilor

1. **Gramatici de tip 0 (generale)**

   Nu există restricţii asupra regulilor

2. **Gramatici de tip 1 (dependente de context)**

   reguli de forma $pxq \rightarrow pyq$ unde $x \in N, y \neq \epsilon, p, q \in (N \cup T)^*$
   $S \rightarrow \epsilon$, caz în care S nu apare în dreapta producţiilor

3. **Gramatici de tip 2 (independente de context)**

   reguli de forma $A \rightarrow y$ unde $A \in N$ şi $y \in (N \cup T)^*$

4. **Gramatici de tip 3 (regulate)**

   $A \rightarrow u$ sau $A \rightarrow uB$ unde $A, B \in N$ şi $u \in T^*$.

Exemple

Ce tip au urmatoarele gramatici?

- $G = (N, T, S, P)$, $N = \{S, A, B\}$, $T = \{a, b, c\}$, P:
  $(1) S \rightarrow aaAc$
  $(2) aAc \rightarrow aAbBc$
  $(3) bB \rightarrow bBc$
  $(4) Bc \rightarrow Abc$
  $(5) A \rightarrow a$
- $G = (N, T, S, P)$, $N = \{S, X\}$, $T = \{a, b, c\}$, P:
  $(1) S \rightarrow abc$
  $(2) S \rightarrow aSXc$
  $(3) cX \rightarrow Xc$
  $(4) bX \rightarrow bb$

Exemple

- Fie $G = (\{E\}, \{a, +, -, (,)\}, E, \{E \to a, E \to (E + E), E \to (E - E)\})$.
    - Ce tip are gramatica G ?
    - Construiţi derivări din E pentru cuvintele $(a + a)$ şi $((a + a) - a)$
    - Cuvântul $(a + a - a)$ poate fi derivat din E?
    - Descrieţi limbajul $L(G)$
- Fie $G = (\{A, B\}, \{a, b\}, A, \{A \to aA, A \to B, B \to bB, B \to \epsilon\})$
    - Ce tip are gramatica G ?
    - Descrieţi limbajul $L(G)$

Ierarhia lui Chomsky

- Un limbaj L este de tipul j dacă există o gramatică G de tipul j astfel încât $L(G) = L$, unde $j \in \{0, 1, 2, 3\}$.
- Vom nota cu $\mathcal{L}_j$ clasa limbajelor de tipul j, unde $j \in \{0, 1, 2, 3\}$.
- Ierarhia lui Chomsky: $\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0$
- Incluziunile sunt stricte:
  - orice limbaj de tip $j + 1$ este şi de tip $j \in \{0, 1, 2\}$
  - există limbaje de tip j care nu sunt de tip $j + 1$, $j \in \{0, 1, 2\}$

Proprietăţi

- Fiecare din familiile $\mathcal{L}_j$ cu $0 \leq j \leq 3$ conţine toate limbajele finite.
- Fiecare din familiile $\mathcal{L}_j$ cu $0 \leq j \leq 3$ este închisă la operaţia de reuniune:

$$L_1, L_2 \in \mathcal{L}_j \Longrightarrow L_1 \cup L_2 \in \mathcal{L}_j,$$

$\forall j : 0 \leq j \leq 3$

Curs 1

Gramatici de tip 3

- O gramatică $G = (N, T, S, P)$ este de tip 3, dacă regulile sale sunt de forma $A \rightarrow u$ sau $A \rightarrow uB$ unde $A, B \in N$ şi $u \in T^*$.
- Exemplu: $G = (\{D\}, \{0, 1, \ldots, 9\}, D, P)$
  Unde P este:
  $D \rightarrow 0D|1D|2D|\ldots|9D$
  $D \rightarrow 0|1|\ldots|9$

Exemple

- Fie gramatica $G = (\{A, B\}, \{l, d\}, A, P)$
  unde P este:
  $A \rightarrow lB, \quad B \rightarrow lB|dB|\epsilon$ (l = literă, d = cifră)

Exemple

- Fie gramatica $G = (\{A, B\}, \{l, d\}, A, P)$
  unde P este:
  $A \to lB, \quad B \to lB|dB|\epsilon$ (l = literă, d = cifră)
  L(G): mulţimea identificatorilor

Exemple

- Fie gramatica $G = (\{A, B\}, \{l, d\}, A, P)$
  unde P este:
  $A \rightarrow lB, \quad B \rightarrow lB|dB|\epsilon$ (l = literă, d = cifră)
  L(G): mulţimea identificatorilor

- Fie gramatica $G = (\{A, B\}, \{+, -, d\}, A, P)$
  unde P este:
  $A \rightarrow +dB|-dB|dB, \quad B \rightarrow dB|\epsilon$ (d = cifră)

Exemple

- Fie gramatica $G = (\{A, B\}, \{l, d\}, A, P)$
  unde P este:
  $A \to lB, \quad B \to lB|dB|\epsilon$ (l = literă, d = cifră)
  L(G): mulţimea identificatorilor
- Fie gramatica $G = (\{A, B\}, \{+, -, d\}, A, P)$
  unde P este:
  $A \to +dB| - dB|dB, \quad B \to dB|\epsilon$ (d = cifră)
  L(G): mulţimea constantelor întregi

Forma normală

- O gramatică de tip 3 este în formă normală dacă regulile sale sunt de forma $A \rightarrow a$ sau $A \rightarrow aB$, unde $a \in T$, şi eventual $S \rightarrow \epsilon$ (caz în care S nu apare în dreapta regulilor).

- Pentru orice gramtică de tip 3 există o gramatică echivalentă în formă normală
  - Se poate arăta că pot fi eliminate regulile de forma $A \rightarrow B$ (redenumiri) şi cele de forma $A \rightarrow \epsilon$ (reguli de ştergere), cu excepţia, eventual a regulii $S \rightarrow \epsilon$
  - Orice regulă de forma $A \rightarrow a_1 a_2 \ldots a_n$ se înlocuieşte cu $A \rightarrow a_1 B_1$, $B_1 \rightarrow a_2 B_2$,..., $B_{n-2} \rightarrow a_{n-1} B_{n-1}$, $B_{n-1} \rightarrow a_n$ $n > 1, B1, \ldots, B_{n-1}$ fiind neterminali noi
  - Orice regulă de forma $A \rightarrow a_1 a_2 \ldots a_n B$ se înlocuieşte cu $A \rightarrow a_1 B_1$, $B_1 \rightarrow a_2 B_2$,...,$B_{n-2} \rightarrow a_{n-1} B_{n-1}$, $B_{n-1} \rightarrow a_n B$, $n > 1, B_1, \ldots, B_{n-1}$ fiind neterminali noi
  - Transformarile care se fac nu modifică limbajul generat de gramatică

Fie L, L_1, L_2 limbaje regulate: există gramaticile G, G_1, G_2 de tip 3 astfel ca $L = L(G), L_1 = L(G_1)$ şi $L_2 = L(G_2)$.

Atunci, următoarele limbaje sunt de asemenea regulate:

1. $L_1 \cup L_2$
2. $L_1 \cdot L_2$
3. L^*
4. L^R
5. $L_1 \cap L_2$
6. $L_1 \setminus L_2$

Închiderea la reuniune

Fie $G_1 = (N_1, T_1, S_1, P_1)$ şi $G_2 = (N_2, T_2, S_2, P_2)$ gramatici de tip 3 cu $L_1 = L(G_1)$, $L_2 = L(G_2)$.
Presupunem $N_1 \cap N_2 = \emptyset$ şi gramaticile în forma normală

Închiderea la reuniune: se arată că $L_1 \cup L_2 \in \mathcal{L}_3$:
Gramatica $G = (N_1 \cup N_2 \cup \{S\}, T_1 \cup T_2, S, P_1 \cup P_2 \cup \{S \to S_1, S \to S_2\})$ este de tip 3 şi generează limbajul $L_1 \cup L_2$

Închiderea la produs

Gramatica $G = (N_1 \cup N_2, T_1 \cup T_2, S_1, P)$ unde P constă din:

- regulile de forma $A \to aB$ din P_1
- reguli $A \to aS_2$ pentru orice regulă $A \to a$ din P_1
- toate regulile din P_2

este de tip 3 şi generează limbajul L_1L_2

Închiderea la iteraţie

Fie $G = (N, T, S, P)$ de tip 3 care generează L ($L = L(G)$). Presupunem că simbolul de start S nu apare în partea dreaptă a vreunei reguli.

Gramatica $G' = (N, T, S, P')$ unde P' constă din

- reguli $A \to aB$ din P
- reguli $A \to aS$, pentru orice regulă $A \to a$ din P
- regula $S \to \epsilon$

este de tip 3 şi generează L^*