LIMBAJE FORMALE ŞI AUTOMATE

Curs + Seminar 13 aprilie /19 mai 2024: ora 9, Amf Spiru Haret /202

Conf.dr. Monica TĂTĂRÂM smtataram@yahoo.com tataram@fmi.unibuc.ro

- 1. Programa
- 2. Descrierea cursului
- 3. Evaluare
- 4. Bibliografie



1. Ierarhia Chomsky

- ✓ Alfabet, cuvant, operatii cu cuvinte
- ✓ Limbaj, operatii cu limbaje
- ✓ Gramatica; exemple
- ✓ Clasificarea gramaticilor generative; ierarhia lui Chomsky;

2. Automate finite

- √ Variantele deterministe si nedeterministe; echivalenta lor
- Clasa limbajelor definite de automatele finite: proprietati de inchidere, probleme de decizie, lema de pompare
- ✓ Expresii regulate; legatura cu automatele finite;

3. Limbaje independente de context

- ✓ Gramatici independente de context; forma normala Chomsky.
- ✓ Automate pushdown; echivalenta cu gramaticile independente de context.
- ✓ Proprietati de inchidere, probleme de decizie, lema de pompare.

- 1. Programa
- 2. Descrierea cursului
- 3. Evaluare
- 4. Bibliografie

- Clasificarea problemelor:
 - ✓ informale,
 - ✓ formale,
 - de analiză,
 - de sinteză.

- Formalizam
- Masuram
- Calculam
 - ✓ Instante ale problemei
 - Clasa de probleme

- # Programe de calculator.
- **#** Algoritm / Algoritmi
- **#** Problema formala de sinteza
- **#** Problema formala de analiza
- #Problema informala

Probleme

- dificile pt om dar usoare pt calculator:
 - prelucrari banale pt volume mari de date;
- usoare pt om dar dificile pt calculator:
 - ✓ probleme dificil de formalizat:

(datele de intrare pot fi incomplete sau inconsistente; metodele de rezolvare clasice sunt inaplicabile)

ex.: recunoașterea vorbirii, recunoașterea imaginilor, recunoașterea caracterelor, politicile de fidelizare bancara etc.)

- dificile pt om si pt calculator:
 - probleme caracterizate prin spaţiu mare de căutare: probleme NP dificile: SAT, TSP).

(nu se cunosc algoritmi care să permită rezolvarea lor într-un interval de timp a cărui dimensiune să depindă polinomial de dimensiunea problemei) ex.: probleme de optimizare liniara/combinatorială de dimensiuni mari, probleme de planificare.

- # Cat de USOR / DIFICIL rezolvam?
- # Ce PUTEM / NU PUTEM rezolva?
- **#Cum REZOLVAM?**



- Clasificarea problemelor:
 - ✓ rezolvabile [algoritmic]
 - efectiv (practic)
 - in principiu (teoretic)
 - ✓ nerezolvabile;

Informatica teoretica:

- problema -> functie -> algoritm?
 - exista probleme nerezolvabile?
 - exista probleme teoretic rezolvabile dar practic nerezolvabile?

Exemple?

- algoritmul: descriere, caracteristici;
 - ✓ problema opririi,
 - ✓ problema corespondentei Post;
 - problema turnurilor din Hanoi (Benares),
 - problema satisfiabilitatii formulelor logice;

Progresele tehnologice →

extinderea studiului din sfera teoretica in cea practica =>

- 3 domenii ale teoriei calculului:
- ✓ teoria limbajelor formale (gramaticilor şi automatelor),
- ✓ teoria calculabilitatii propriu-zise,
- ✓ teoria complexitatii.

Raspunsul la intrebare

diverse forme:

1950: TEORIA LIMBAJELOR FORMALE (GRAMATICILOR ŞI AUTOMATELOR:



Avram Noam Chomsky (7 dec 1928 -),
Arto SALOMAA, Grzegorz ROZENBERG,
Solomon MARCUS, Gh. PAUN,
Victor MITRANA, Alex. MATEESCU.

Raspunsul la intrebare → diverse forme:

1930: TEORIA CALCULABILITATII:



Alan Mathison TURING (23 iunie 1912 – 7 iunie 1954),

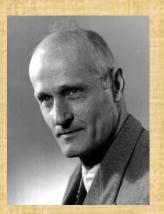
Alonzo CHURCH, Stephen Cole KLEENE,

Kurt GÖDEL, Emile Leon POST, N.N. MARKOV.









Raspunsul la intrebare → diverse forme:





1950: TEORIA COMPLEXITATII:

1953: Andrzej Grzegorczyk (22 aug. 1922 – 20 mar. 2014)

1960: Manuel BLUM (26 apr. 1938 -)

1960: Ray SOLOMONOFF (25 iulie 1926 – 7 dec. 2009) şi

1963: Andrei N. KOLMOGOROV (25 apr. 1903 – 20 oct. 1987)

1966: Gregory CHAITIN (15 nov. 1947 -)

1970: Stepan COOK (14 dec 1939 -) şi

1971: Leonid A. LEVIN (2 nov. 1948 -)







Cursul ∈

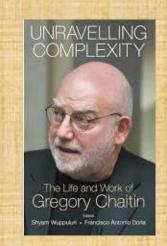
Semiotica (Teoria comunicarii), Lingvistica, Theortical Computer Science:



"Theory without practice is useless; practice without theory is blind" *Roger BACON* (1219/20 – c. 1292)

"To me, you understand something only if you can program it. (You, not someone else!) . . . programming something forces you to understand it better, it forces you to really understand it, since you are explaining it to a machine."

Gregory CHAITIN (1947 -).



Semiotica =

- = stiinta care studiaza semnele şi simbolurile din perspectiva filosofica,
- = se ocupa cu studiul proceselor prin care se formeaza şi se ataseaza un inteles semnelor care ne inconjoara;

Strans legata de Teoria comunicarii datorita relatiei triadice (cf. Charles Sanders PEIRCE) dintre

- ✓ semn
- ✓obiectul (reprezentantul) sau
- ✓ interpretantul sau

Se compune din 3 subdomenii:

✓ sintactica:

studiaza relatia dintre semnele care compun o anumita structura formala

✓ semantica:

studiaza relatia dintre semne şi lucrurile (obiecte, procese, fenomene) pe care le desemneaza (i.e. "intelesul" semnelor)

pragmatica:

studiaza relatia dintre semne și utilizatorii lor. >

Semiotica (cont.)

Studiaza diverse sisteme de semne (lingvistice şi non-lingvistice):

- √ iconii,
- √ indicii,
- ✓ simbolurile nonverbale;

Procesele semiotice:

- de gradul I
- ✓ de gradul II

(ex.: interfetele grafice).

Lingvistica =

- = stiinta care studiaza limbajul sub 3 aspecte:
 - √ forma limbajului (sintaxa),
 - ✓ intelesul limbajului (semantica),
 - ✓ contextul limbajului (pragmatica);

Primele studii "de lingvistica":

- ✓ legate de descrierea limbajului:
- ✓ tratatul Astadhyayi (Astaka) = o gramatica descriptiva a limbii sanskrite,
- ✓ Sanskrit = complet = perfect = limbajul zeilor,
- ✓ structura: 4 capitole cu cate 4 subcapitole fiecare.



Lingvistica (cont.)

- ✓ Autorul Astadhyayi : Panini
 - √ datele biografice: incerte:
 - ✓ unii biografi: 520 î.H., Shalatula (Attock) in actualul Pakistan 460 î.H., India;
 - ✓ alti biografi: secolele 4, 5, 6 şi chiar 7 î.H.,
- considerat fondatorul foneticii, fonolgiei si morfologiei
- considerat un precursor al teoriei limbajelor formale:
 - ✓ propune o structura generativa a gramaticii limbii sanskrite:
 - ✓ substantivele compuse, frazele etc: formate cu ajutorul unor reguli de productie pornind de la aprox. 1700 de elemente fundamentale (substantive, verbe, consoane etc.)
 - =>gramatica sansckita a fost comparata cu Elementele lui Euclid (esentiala şi aproape nemodificata pana azi).

Lingvistica (cont.)

- ✓ Panini: considerat un precursor al reprezentarilor
 - ✓ matematice: (primul care a utilizat litere pentru reprezentarea numerelor. cf. arheologului indian Pandit Bhagvanlal Indraji sec XIX)
 - ✓ pentru limbajele de programare: o notatie similara BNF şi EBNF

```
<simbol> ::= expresie unde:
```

<simbol> = neterminal (variabla)

expresie = una sau mai multe secvente de simboluri terminale şi neterminale;

::= operatorul de atribuire (inlocuire)

operatorul care indica posibilitatea de alegere intre mai multe expresii in m. drept;

<oferta-curs> ::= <titlu> <cadru_didactic-curs> <cadru_didactic_seminar> <an_universitar>
<cadru_didactic_curs>::= <prof_llie_Dima> | <prof_Radu_Sava>|<conf_lon_Dedu>.

Lingvistica (cont.)

Gramatica =

= un sistem de reguli care guverneaza limbajul folosit de o *anumita comunitate*;

se refera atat la:

- √ formarea şi compunerea morfemelor in cuvinte (morfologia),
- formarea şi compunerea cuvintelor in proprozitii şi fraze (sintaxa);

dar şi la:

✓ sunete, semnificatii, gesturi (fonologia = modul in care sunetele şi gesturile functioneaza impreuna). →

Teoria limbajelor formale (TLF) =

- = studiaza in principal aspectele sintactice ale acestui tip de limbaje (i.e. structura interna a constructelor tipice)
 - a aparut in cadrul Lingvisticii,
 - ca instrument de studiere a sintaxei limbajelor naturale,

initial: subdomeniu al Matematicii,

ulterior: subdomeniu al Informaticii Teoretice.

TLF constituie un instrument teoretic aplicat:

- i) in Logica matematica și in Fundamentele Matematicii, pt:
 - reprezentarea sintaxei sistemelor axiomatice şi formalismelor matematice;
- ii) in Informatica:
 - ✓ in <u>Teoria Limbajelor de Programare pt:</u>
 - definirea gramaticilor,
 - studierea proprietatilor limbajelor de programare;
 - ✓ in <u>Teoria Calculabilitatii pt:</u>
 - definirea formala a conceptului de calculabilitate,
 - definirea formala a notiunii de calculator,
 - definirea şi studierea proprietatilor unor modele matematice de calculabilitate (automate, gramatici);
 - ✓ in <u>Teoria Complexitatii Calculului pt:</u>
 - definirea claselor de complexitate intr-o forma care poate fi parsata de masini cu putere computationala limitata.

=> ?

- ? Limbaj sau limba?
- ? Ce fel de limbaj?
- ? Ce fel de gramatica?
- => diverse dihotomii:
 - A. limbaj formal vs limbaj informal
 - B. limbaj formal vs limbaj natural
 - C. limbaj formal vs limbaj artificial
 - D. limbaj formal vs limbaj construit.

Limbaj, limba, dialect?

Limbajul

- capacitatea vietuitoarelor de a invata şi utiliza sisteme complexe de semne in scopul comunicarii,
- limbajul uman: obiectul de studiu al lingvisticii;

Limba

- = sistem de comunicare alcătuit din sunete articulate, specific oamenilor, prin care aceștia își exprimă gândurile, sentimentele și dorințele
- este specifică unei comunități umane, istoric constituită,
- este caracterizata prin structură gramaticală, fonetică și lexicală proprie,
- este o instanta a limbajului (langage langue; language language!);
- se estimeaza ca in lume exista:
 intre 5.000 şi 8.000 de limbi (şi aprox. 20 alfabete "clasice");
 http://www.worldstandards.eu/other/alphabets/
 http://ro.wikipedia.org/wiki/Ideogram%C4%83

Dialect

- = grai, jargon;
- = ramificație teritorială a unei limbi, ale cărei trăsături caracteristice (fonetice, lexicale, gramaticale etc.) o deosebesc de limba comună a întregului popor și de alte ramificații teritoriale ale acestei limbi..

A. Limbaj formal vs limbaj informal

Limbaj informal!!

Limbaj formal = o multime de secvente de simboluri care respecta un set (finit) de reguli specifice;

Limbajele formale dispun de:

- un alfabet = o multime finita şi nevida de simboluri (litere/cuvinte/semne),
- o gramatica [generativa], [formala];
- ✓ Secventele de simboluri preluate din alfabet şi formate dupa regulile gramaticii se numesc **cuvinte** (**fraze**) ale limbajului;
- ✓ Nu intotdeauna au sens!!
 - abb ---- "Colorless green ideas sleep furiously "

(Noam CHOMSKY: Three Models for the Description of Language, in *I.R.E. Transactions* on *Information Theory*, Volume IT-2, Proceedings of the Symposium on Information Theory, Sept. 1956, pp.113-123):

26

propozitie corecta gramatical dar care nu are sens in realitatea cunoscuta;

eventual ----- Mirel este un analfabet erudit:

engleza/romana???? ----- propozitie corecta gramatical dar care nu respecta principiile logicii (exprima o contradictie). ->

A. Limbaj formal vs limbaj informal (cont.) Istoric:

- primul limbaj formal:
 - **Begriffsschrift** (i.e. ~"concept writing", "concept notation") = un sistem formal similar celui dedicat aritmeticii (care a şi constituit sursa de inspiratie pt Begriffsschrift),
 - Gottlob FREGE (1848-1925);
 - **1879**,
- prima gramatica formala:
 - sistemele de rescriere (Semi-Thue systems),
 - Axel THUE (1863-1922)
 - 1900?
 - 1914: o problema similara problemei opririi, formulata ca o problema de apartenenta a unui element (cuvant) la un semigrup finit generat.





B. Limbaj formal vs limbaj natural

Limbaj natural = un set de secvente de simboluri care respecta un set de reguli specifice;

Clasificare:

- limbaje vorbite (orale) = permit comunicarea orala, prin cuvant;
- limbaje de semne (bazate pe gestica) = permit comunicarea
 - alternativa (uneori antinomica),
 - in cadrul anumitor comunitati (persoane fara auz etc.),

Principii de functionare:

- ambele tipuri de limbaje naturale contin:
 - un sistem fonologic = determina modul in care simbolurile se combina pt a forma secvente (cuvinte şi /sau morfeme; ex.: cine, va, com, placea; cineva, complacea),
 - un sistem sintactic = determina modul in care cuvintele şi morfemele se combina pentru a forma propozitii şi fraze;
- ambele tipuri de limbaje naturale se bazeaza pe un proces de semioza :
 - → simbolurile conventionale primesc o semnificatie anume, unanim acceptata in respectiva comunitate.

B. Limbaj formal vs limbaj natural (cont.)

Caracteristicile limbajului natural (unele sunt distinctive pt limbajul uman):

- independent de canalul de comunicare (poate fi codificat prin scriere normala, braille, morse etc.),
- independent de timp şi spatiu (poate descrie evenimente care au loc in alte momente de timp şi in alte locatii decat cele curente),
- conventional,
- dependent de context,
- recursiv (o fraza este formata din alte fraze, numite propozitii etc.),
- ambiguu,
- generativ (dintr-un set finit de elemente simple se poate crea o potential infinitate de elemente complexe, inzestrate cu sens);

Origine:

Africa, aprox. 50.000 ani in urma.

C. Limbaj formal vs limbaj artificial

Limbaj artificial =

- = un limbaj ale carui lexic, fonetica și gramatica
 - au fost create pe baza unei surse unice
 - de catre un singur creator / grup de creatori
 - ca un exercitiu in sine (vocabularul şi gramatica se dezvolta mult mai rapid decat puterea de invatare a "vorbitorilor");

limbaj artificial = un oximoron:

- un sistem este un limbaj doar daca exista cel putin 2 vorbitori care il utilizeaza pt a comunica idei şi concepte;
- un sistem este artificial doar atat timp cat nimeni nu il foloseste pentru comunicarea conceptelor;
- numite și limbaje internationale auxiliare.

C. Limbaj formal vs limbaj artificial (cont.)

Exemple:

- Volapük, 1879, Johann Martin SCHLEYER: primul "auxlang" cu o oarecare recunoastere internationala,
- Esperanto, 1887, Ludwik Lejzer ZAMENHOF,
- Interlingua, 1951, International Auxiliary Language Association (data publicarii unui dictionar şi a unei gramatici a auxlag.),
- Eurolengo, 1972, Leslie JONES (combinatie de engleza şi spaniola),
- Klingon, 1985, Marc OKRAND (pt serialul Star Treck);

Caracteristici:

- scop
 - in sine, nu pentru comunicare,
 - => Esperanto: nu; Volapük: da!
- originalitate,
 - distanta cat mai mare fata de vocabularele existente,
 - => Esperanto, Eurolengo: nu; Volapük: da!
- dimensiune;
 - nu se masoara in numar de cuvinte ci in puterea expresiva a vocabularului şi gramaticii,
 - limbajele care nu dispun de sisteme de numeratie sau de reguli adecvate anumitor situatii sintactice au o putere expresiva f limitata.

D. Limbaj formal vs limbaj construit

Limbaj construit =

- = un limbaj ale carui lexic, fonetica şi gramatica
 - au fost create pe baza unei surse unice
 - de catre un singur creator / grup de creatori
 - in scopul expres al asigurarii comunicarii interumane sau a simularii acesteia,
- => limbajul construit difera de limbajul natural / artificial;

Clasificare:

- limbaje controlate = standardizare (simplificare) a unor limbaje naturale in scopul facilitarii asimilarii lor de catre vorbitorii nenativi (ex. Basic English, Simplified Technical English etc.),
- limbaje proiectate = create in scopul studierii unui domeniu stiintific: logica, filosofie, lingvistica etc.,
- limbajele artistice = limbaje construite in scopuri estetice (de fictiune, experimentale, micronationale, personale, ilare etc.),

Prezentam 2 clase speciale de limbaje construite:

- D1) limbajele pentru interactiunea cu robotii,
- D2) limbajele de programare.

D. Limbaj formal vs limbaj construit (cont.)

D1) Limbajele pentru interactiunea cu robotii:

Robot Interaction Language (ROILA, 2010)

= primul limbaj vorbit, unificat, creat special pentru comunicarea dintre oameni şi masini, in speta **roboti** (ex. Roomba, Lego Mindstorms NXT);

Caracteristici:

- usor de invatat de catre om,
- optimizat pentru recunoasterea de catre roboti (algoritmi de speech recognition),
- inlocuieste elementele limbajului natural care il fac dificil de recunoscut de catre roboti (omofoni, conjugari, declinari etc.) cu un vocabular artificial (format din cuvinte frecvent, usor şi similar pronuntate in mai multe limbi) generat de un algoritm genetic;

Proiectat in 2010 la Departamentul de Proiectare pentru Industrie, Universitatea Tehnica din Eindhoven,

Open source.

D. Limbaj formal vs limbaj construit (cont.)

Vocabularul ROILA:

provine din Basic English;

Gramatica ROILA:

regulata:

nu exista exceptii: aceeasi regula se aplica tuturor cuvintelor dintr-o categorie gramaticala;

izolanta:

nu exista prefixe sau sufixe care sa modifice sensul cuvintelor => exista cuvinte speciale care sunt plasate <u>dupa</u> un cuvant pentru a indica timpul trecut ("jifi") sau numarul plural ("tuji") etc.;

Limbajul ROILA:

- frazele au o structura unica, simpla: subiect-verb-complement,
- exista numai 5 parti de vorbire: substantive, verbe, adjective, adverbe, pronume (eu="pito", tu="bama", el, ea="mona")".

D. Limbaj formal vs limbaj construit (cont.)

D2) Limbajele de programare:



"To me, you understand something only if you can program it. (You, not someone else!) . . . programming something forces you to understand it better, it forces you to really understand it, since you are explaining it to a machine."

Gregory Chaitin

- ∈ clasa limbajelor construite,
- ∈ clasa limbajelor formale;

proiectate pentru a asigura comunicarea om-masina:

- masina = calculator / automat,
- masina = independenta / incorporata;

Limbaj de programare = un limbaj care permite reprezentarea algoritmilor intr-o forma "inteligibila" de catre calculator (inclusiv a directivelor de control al unor periferice ale calculatorului). -> 35

D. Limbaj formal vs limbaj construit (cont.)

Clasificarea limbajelor de programare:

- ✓ imperative / declarative,
- ✓ cod-masina / nivel inalt
- etc.;

Istoric

- limbajul pentru razboiul de tesut al lui Jaquard,
- limbajul pentru pianele mecanice,
- ✓ Plankalkül (Konrad ZUSE: 1943-1945, Z3),
- ✓ Short Code (John MAUCHLEY: 1949, ENIAC),
- generatiile de limbaje de nivel inalt.

- Descrierea cursului
- 2. Planul cursului
- 3. Evaluare
- 4. Bibliografie

Examen: fizic:

aprox. 10 subjecte

1 subject teorie + 1 problema,

nota finala: media aritmetica a 3 note:

- 1. examen: subjectul teoretic;
- 2. examen: problema
- 3. media notelor pentru cele 4 teme de casa

(minim nota 5 la fiecare)

bonus: activitate "in clasa".

- Descrierea cursului
- 2. Planul cursului
- Evaluare
- 4. Bibliografie

- [Aho et al., 74] Alfred AHO, John E HOPCROFT, Jeffrey D. ULLMAN: The Design and Analysis of Computer Algorithms, Addison-Wesley Publ. Co., Reading Mass., 1974, II 24220, II 37734
- [Atan. et al., 90] Adrian ATANASIU, Alexandru MATEESCU: Limbaje formale, TUB, Bucureşti, 1990, II 36078
- [Denn et al., 78] **Peter J. DENNING, Jack B. DENNIS, Joseph E. QUALITZ**: *Machines, Languages and Computation*, Prentice Hall Inc. Englewood Cliffs, NJ, 1978, **II 27675**
- [Gium., 04] **Cristian GIUMALE**: *Introducere în analiza* algoritmilor; Teorie şi aplicaţii, Editura Polirom, Iaşi, 2004 **II 40105**.

- [Vide et al., 04] Carlos MARTÍN-VIDE, Victor MITRANA, Gheorghe PĂUN (Eds.): Formal Languages and Applications, Springer-Verlag, Berlin Heidelberg, 2004, II 40157
- [Paun et al., 01] Gheorghe PĂUN, Grzegorz ROZENBERG, Aarto SALOMAA (Eds.): Current Trends in Theoretical Computer Science, World Scientific Publ. Co., Singapore, 2001, II 39698
- [Roze. et al., 97] **Grzegorz ROZENBERG, Aarto SALOMAA** (Eds.): Handbook of Formal Languages, 3 vols., Springer-Verlag, Berlin Heidelberg, 1997, **II 39019**
- [Salo, 73] Aarto SALOMAA: Formal Languages, Academic Press, 1973 II 23015.

