Formális Nyelvek - 1. Előadás

Csuhaj Varjú Erzsébet

Algoritmusok és Alkalmazásaik Tanszék
Informatikai Kar
Eötvös Loránd Tudományegyetem
H-1117 Budapest
Pázmány Péter sétány 1/c

E-mail: csuhaj@inf.elte.hu

A kurzus célja, hogy megismerkedjünk a formális nyelvek és automaták elméletének, a számítástudomány egyik tradícionális ágának alapjaival.

Irodalom:

1. György E. Révész, Introduction to Formal Languages, McGraw-Hill Book Company, 1983.

További irodalom:

2. A. Salomaa, Formal Languages, Acedemic Press, 1973.

3. K. Krithivasan, Rama, R., Introduction to Formal Languages, Automata Theory and Computation, Pearson, 2009.

4. J. E. Hopcroft, Rajeev Motwani, J.D. Ullman, Introduction to Automata Theory, Languages, and Computation. Second Edition. Addison-Wesley (2001).

Magyar nyelvű irodalom:

1. Révész György, Bevezetés a formális nyelvek elméletébe, Tankönyvkiadó, 1977.

2. Fülöp Zoltán, Formális nyelvek és szintaktikus elemzésük, Polygon, Szeged, 2004.

Tudnivalók

Az előadások alapjául az irodalomjegyzék [1], [2] és [3] eleme szolgál. Az előadásvázlatok (a slide-ok) minden lényeges információt tartalmaznak és kizárólag tanulási célokra használhatók.

A kurzus vizsgával zárul, a vizsga során az előadásokon elhangzott és az előadásvázlatokon szereplő anyagot kérem számon. Az előadásokon további információk is elhangozhatnak, ezek kiegészítő jellegűek.

A vizsga előtt három héttel részletes információt adok a számonkérés alapjául szolgáló anyagról és a számonkérés módjáról.

Az előadások látogatása nem kötelező, de ajánlott. A gyakorlatok látogatása kötelező, a lehetséges hiányzások számát a gyakorlatvezetők ismertetni fogják a gyakorlatokon. A gyakorlati jegy megszerzésének feltételeit is a gyakorlatvezetők fogják meghatározni és ismertetni.

Az előadásvázlatokat .pdf file formájában az előadások utáni pénteken felteszem a honlapomra (http://people.inf.elte.hu/csuhaj), és a pontos webcímet a Neptun rendszeren keresztül közölni fogom.

Minden szerdán de. 10-12 h között fogadóórám van a Déli tömb 2.511-es hivatali szobámban, ahol az érdeklődőket szeretettel várom. Ugyancsak keressenek meg emailben (csuhaj@inf.elte.hu) vagy a fogadóórán, ha bármilyen kérdésük felmerül a tantárggyal kapcsolatban.

Jó tanulást kívánok!

Budapest, 2013. február

Csuhaj Varjú Erzsébet

A kurzus tartalmának rövid leírása

1. **Bevezetés, a formális nyelv fogalma:** alapvető fogalmak és jelölések,szavak, nyelvek, grammatikák, a grammatikák Chomsky-féle hierarchiája.

2. **Műveletek nyelveken:** definíciók, nyelvosztályok zártsági tulajdonságai.

3. **Környezetfüggetlen grammatikák és nyelvek:** redukált grammatikák, a Chomsky normálforma, levezetési fa, lineáris grammatikák, reguláris grammatikák, reguláris nyelvek, reguláris kifejezések. A generált nyelvek és nyelvosztályok tulajdonságai.

4. **Környezetfüggő- és mondatszerkezetű grammatikák:** hossz-nemcsökkentő grammatikák, Kuroda normál forma, mondatszerkezetű grammatikák normálformái. A generált nyelvek és nyelvosztályok tulajdonságai. Nyelvosztályok Chomsky-féle hierarchiája.

A kurzus tartalmának rövid leírása - folytatás

1. **Automaták és nyelvek:** véges automaták, veremautomaták, kétvermű automata, lineárisan korlátolt automata, Turing gép. Az automaták tulajdonságai, a felismert nyelvosztályok, az automaták és a grammatikák kapcsolatai.

2. **Szintaktikai elemzés:** kapcsolat szintaxis és szemantika között; környezetfüggetlen grammatikák és nyelvek egyértelműsége; $LL(k)$ és $LR(k)$ grammatikák.

A formális nyelvek és automaták elmélete - a gyökerek

A nyelv **grammatikájának** fogalma már kb. időszámításunk előtt az V. században felmerült Indiában (Panini).

Fontosabb lépések:

- Axel Thue, Emil Post, matematika, a XX. század eleje.
- W. Mc Culloch, W. Pitts, 1943, az idegrendszer modellje - a véges állapotú gép;

  S.C. Kleene, 1956, neurális háló - a véges automata.
- Noam Chomsky, 1959, matematikai model, az angol nyelv grammatikájának matematikai modellje.
- Programnyelvek, ALGOL 60, 1960

Mivel foglalkozik a formális nyelvek és automaták elmélete?

A formális nyelvek elmélete szimbólumsorozatok halmazaival foglalkozik.

Célja - többek között - véges, tömör leírását adni az ilyen halmazoknak.

Az elmélet módszereket ad formális nyelvek definiálására, a formális elemek nyelvhez való tartozásának eldöntésére, a nyelvi elemek struktúrájának felismerésére.

A szimbólum fogalmát alapfogalomnak tekintjük, ezért nem definiáljuk.

Milyen tudományágakhoz kapcsolódnak a formális nyelvek és automaták?

- A természetes nyelvek gépi feldolgozása, matematikai modellezése, matematikai nyelvészet,

- programozási nyelvek, fordítóprogramok elmélete,

- kódelmélet,

- képfeldolgozás,

- mintafelismerés,
- fejlődő rendszerek modellezése (Lindenmayer rendszerek),
- molekuláris számítástudomány (DNS számítás),
- multi-ágens rendszerek formális leírásai,
- stb.

Alapfogalmak és jelölések - I

Szimbólumok véges nemüres halmazát **ábécének** nevezzük.

Példa: $V = \{a, b, c\}$

Egy V ábécé elemeiből képzett véges sorozatokat V feletti **szavaknak** vagy **sztringeknek** mondunk.

Példa: Legyen az ábécé $V = \{a, b, c\}$ és akkor $aaabbbccc$ egy szó.

A 0 hosszúságú sorozatot **üres szónak** nevezzük és ε-nal jelöljük.

A V ábécé feletti szavak halmazát (beleértve az üres szót is) V^*-gal, a nemüres szavak halmazát V^+-**szal jelöljük.**

Alapfogalmak és jelölések - II

Legyen V egy ábécé és legyenek u, v V feletti szavak (azaz, legyen $u, v \in V^*$). Az uv szót az u és v szavak **konkatenáltjának** nevezzük.

Példa: Legyen az ábécé $V = \{a, b, c\}$, legyenek $u = abb$ és $v = cbb$ szavak. Akkor $uv = abbcbb$ az u és v konkatenáltja.

A **konkatenáció** mint művelet **asszociatív**, de **általában nem kommutatív**.

Példa: Legyen $u = ab$, $v = ba$, akkor $uv = abba$ és $vu = baab$.

Alapfogalmak és jelölések - II - folytatás

Legyen V egy ábécé. Megállapíthatjuk, hogy V^* **zárt a konkatenáció műveletére nézve** (azaz, bármely $u, v \in V^*$ esetén $uv \in V^*$ teljesül), továbbá a **konkatenáció egységelemes művelet**, ahol az egységelem ε (azaz, bármely $u \in V^*$ esetén $u\varepsilon \in V^*$ és $\varepsilon u \in V^*$).

Alapfogalmak és jelölések - III

Legyen i nemnegatív egész szám és legyen w a V ábécé feletti szó ($w \in V^*$).

A w **szó** i**-edik hatványa** alatt a w szó i példányának konkatenálját értjük és w^i-vel jelöljük.

Példa: Legyen az ábécé $V = \{a, b, c\}$, és legyen $w = abc$. Akkor $w^3 = abcabcabc$.

Konvenció alapján minden $w \in V^*$ szóra $w^0 = \varepsilon$.

Alapfogalmak és jelölések - IV

Legyen V egy ábécé és legyen w egy V feletti szó (azaz, legyen $w \in V^*$).

A w **szó hosszán** a w szót alkotó szimbólumok számát értjük (azaz, w mint sorozat hosszát) és $|w|$-vel jelöljük.

Példa: Legyen az ábécé $V = \{a, b, c\}$ és legyen $w = abcccc$. Akkor w hossza 6.

Az üres szó hossza - nyilvánvalóan - 0, azaz $|\varepsilon| = 0$.

Alapfogalmak és jelölések - V

Egy V ábécé feletti két u és v szót azonosnak nevezünk, ha mint szimbólumsorozatok egyenlőek (azaz, mint sorozatok elemről-elemre megegyeznek.)

Legyen V egy ábécé és legyenek u és v szavak V felett. Az u szót a v szó **részszavának** nevezzük, ha $v = xuy$ teljesül valamely x és y V feletti szavakra.

Az u szót a v szó **valódi részszavának** mondjuk, ha x és y közül legalább az egyik nemüres, azaz, $xy \neq \varepsilon$.

Példa: Legyen $V = \{a, b, c\}$ ábécé és legyen $v = aabbbcc$ szó. Az $u = abbbc$ szó valódi részszava v-nek.

Ha $x = \varepsilon$, akkor u-t a v szó **prefixének**, ha $y = \varepsilon$, akkor u-t a v szó **szufixének** hívjuk.

Legyen $v = aabbbcc$ szó. Az $u = aabbb$ szó prefixe, a $bbbcc$ szó szufixe v-nek.

Alapfogalmak és jelölések - VI

Legyen u egy V ábécé feletti szó. Az u szó **tükörképe** vagy **fordítottja** alatt azt a szót értjük, amelyet úgy kapunk, hogy u szimbólumait megfordított sorrendben írjuk. Az u szó tükörképét u^{-1} -gyel jelöljük.

Legyen $u = a_1 \ldots a_n$, $a_i \in V$, $1 \leq i \leq n$. Ekkor $u^{-1} = a_n \ldots a_1$.

Alapfogalmak és jelölések - VII

Legyen V egy ábécé és legyen L tetszőleges részhalmaza V^*-nak. Akkor L-et egy V feletti **nyelvnek** nevezzük.

Az **üres nyelv** - amely egyetlen szót sem tartalmaz - jelölése $\emptyset$.

Egy V ábécé feletti nyelvet **véges nyelvnek** mondunk, ha véges számú szót tartalmaz, ellenkező esetben **végtelen nyelvről** beszélünk.

Példák nyelvekre

Legyen $V = \{a, b\}$ ábécé.

Akkor $L_1 = \{a, b, \varepsilon\}$ véges nyelv, $L_2 = \{a^i b^i \mid 0 \leq i\}$ végtelen nyelv.

Példa L_2-beli szavakra: $ab, aabb, aaabbb, \ldots$

Legyen $L_3 = \{uu^{-1} | u \in V^*\}$.

Példa L_3-beli szavakra: $u = ababb$, $u^{-1} = bbaba$ és $uu^{-1} = ababbbbaba$.

Nyelvek sokféle módon előállíthatók, egyik mód a nyelvek generálása grammatikával.

Generatív grammatika - Definíció

Egy G **generatív grammatikán** (grammatikán vagy (generatív) nyelvtanon) egy (N, T, P, S) négyest értünk, ahol

- N és T diszjunkt ábécék, a **nemterminális** és a **terminális** szimbólumok ábécéi;
- $S \in N$ a **kezdőszimbólum** (axióma),
- P **véges halmaza** (x, y) rendezett pároknak, ahol $x, y \in (N \cup T)^*$ és x legalább egy nemterminális szimbólumot tartalmaz.

  A P halmaz elemeit **átírási szabályoknak** (röviden szabályoknak) vagy **produkcióknak** nevezzük.

  Az (x, y) jelölés helyett használhatjuk az $x \to y$ jelölést is, ahol a $\to$ szimbólum nem eleme az $(N \cup T)$ halmaznak.

Példa Generativ Grammatikára

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol

$N = \{S\}$ a nemterminálisok ábécéje,

$T = \{a, b\}$ a terminálisok ábécéje, és

$$P = \{S \to aSb, \quad S \to ab, \\ S \to ba\}$$

a szabályok halmaza.

Közvetlen levezetési lépés - Definíció

Legyen $G = (N, T, P, S)$ egy generatív grammatika és legyen $u, v \in (N \cup T)^*$.

Azt mondjuk, hogy a v szó **közvetlenül** vagy **egy lépésben levezethető** az u szóból G-ben és ezt

$$u \Longrightarrow_G v$$

módon jelöljük, ha $u = u_1 x u_2$, $v = u_1 y u_2$, $u_1, u_2 \in (N \cup T)^*$ és $x \to y \in P$.

Példa közvetlen levezetésre

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S\}$ a nemterminálisok ábécéje, $T = \{a, b\}$ a terminálisok ábécéje és $P = \{S \to aSb, S \to ab, S \to ba\}$ a szabályok halmaza.

Legyen $u = aaaSbbb$.

Akkor $v = aaaaSbbbb$ közvetlenül (egy lépésben) levezethető u-ból, azaz

$$u \Longrightarrow_G v,$$

ugyanis $u_1 = aaa$, $u_2 = bbb$, $x = S$, $y = aSb$ és $S \to aSb \in P$.

Levezetés - Definíció

Legyen $G = (N, T, P, S)$ egy generatív grammatika és legyen $u, v \in (N \cup T)^*$.

Azt mondjuk, hogy a v szó k **lépésben levezethető** az u szóból G-ben, $k \geq 1$, ha létezik olyan $u_1, \ldots, u_{k+1} \in (N \cup T)^*$ szavakból álló sorozat, amelyre $u = u_1$, $v = u_{k+1}$, valamint $u_i \Longrightarrow_G u_{i+1}$, $1 \leq i \leq k$ teljesül.

A v szó **levezethető** az u szóból G-ben, ha vagy $u = v$, vagy létezik olyan $k \geq 1$ szám, hogy a v szó az u szóból k lépésben levezethető.

Levezetés

Legyen $G = (N, T, P, S)$ egy tetszőleges generatív grammatika és legyen $u, v \in (N \cup T)^*$.

Azt mondjuk, hogy a v szó levezethető az u szóból G-ben és ezt

$$u \Longrightarrow_G^* v$$

módon jelöljük, ha vagy $u = v$ vagy valamely $z \in (N \cup T)^*$ szóra fennáll, hogy $u \Longrightarrow_G^* z$ és $z \Longrightarrow_G v$ teljesül.

$\Longrightarrow^*$ a $\Longrightarrow$ reláció reflexív tranzitív lezártját jelöli.

A $\Longrightarrow$ reláció tranzítív lezártját $\Longrightarrow^+$-val jelöljük.

A kezdőszimbólumból levezethető sztringeket **mondatformának** nevezzük.

A generált nyelv - Definíció

Legyen $G = (N, T, P, S)$ egy tetszőleges generatív grammatika. A G **grammatika által generált** $L(G)$ **nyelv** alatt az

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

szavakból álló halmazt értjük.

Azaz, a G grammatika által generált nyelv a T^* halmaz azon elemei, amelyek levezethetők a G grammatika S kezdőszimbólumából.

Példa

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S\}$, $T = \{a, b\}$ és $P = \{S \to aSb, S \to ab, S \to ba\}$.

Akkor $L(G) = \{a^n abb^n, a^n bab^n | n \geq 0\}$.

Példa egy levezetésre:

$S \Longrightarrow_G aSb \Longrightarrow_G aaSbb \Longrightarrow_G aababb.$

Példa

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S, X, Y\}$, $T = \{a, b, c\}$. Legyen

$$\begin{aligned} P = \{ & S \to abc, \quad S \to aXbc, \\ & Xb \to bX, \quad Xc \to Ybcc, \\ & bY \to Yb, \quad aY \to aaX, \quad aY \to aa\}. \end{aligned}$$

Akkor $L(G) = \{a^n b^n c^n | n \geq 1\}$.

Példa egy levezetésre:

$$S \Longrightarrow aXbc \Longrightarrow_G abXc \Longrightarrow_G abYbcc \Longrightarrow_G aYbbcc \Longrightarrow_G aabbcc.$$

Ekvivalens Grammatikák és Nyelvek

Két generatív grammatikát (gyengén) **ekvivalensnek** nevezünk, ha ugyanazt a nyelvet generálják.

Két nyelvet **gyengén ekvivalensnek** mondunk, ha legfeljebb az üres szóban különböznek.

A Chomsky-féle hierarchia

A $G = (N, T, P, S)$ **generatív grammatikát** i**-típusúnak mondjuk**, $i = 0, 1, 2, 3$, ha P szabályhalmazára teljesülnek a következők:

- $i = 0$: Nincs korlátozás.
- $i = 1$: P minden szabálya $u_1 A u_2 \to u_1 v u_2$ alakú, ahol $u_1, u_2, v \in (N \cup T)^*$, $A \in N$, és $v \neq \varepsilon$, kivéve az $S \to \varepsilon$ alakú szabályt, feltéve, hogy P-ben ilyen szabály létezik. Ha P tartalmazza az $S \to \varepsilon$ szabályt, akkor S nem fordul elő P egyetlen szabályának jobboldalán sem.
- $i = 2$: P minden szabálya $A \to v$ alakú, ahol $A \in N$ és $v \in (N \cup T)^*$.
- $i = 3$: P minden szabálya vagy $A \to uB$ vagy $A \to u$, alakú, ahol $A, B \in N$ és $u \in T^*$.

Példa

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S\}$, $T = \{a, b\}$ és $P = \{S \to aSb, S \to ab, S \to ba\}$.

Ez a grammatika 2-típusú (környezetfüggetlen).

A Chomsky-féle hierarchia - folytatás

Legyen $i = 0, 1, 2, 3$. Egy L nyelvet **i-típusúnak mondunk**, ha i-típusú grammatikával generálható.

Az i-típusú nyelvek osztályát $\mathcal{L}_i$-vel jelöljük.

A 0-típusú grammatikát **mondatszerkezetű grammatikának**, az 1-típusú grammatikát **környezetfüggő grammatikának**, a 2-típusú grammatikát **környezetfüggetlen grammatikának** is nevezzük. A 3-típusú grammatikát **reguláris** vagy **véges állapotú** grammatikának is mondjuk.

A 0,1,2,3-típusú nyelvek osztályait rendre **rekurzíven felsorolható**, **környezetfüggő**, **környezetfüggetlen**, valamint **reguláris nyelvosztálynak** is mondjuk.

A Chomsky-féle hierarchia - folytatás

Nyilvánvaló, hogy $\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_0$ és $\mathcal{L}_1 \subseteq \mathcal{L}_0$.

A későbbiekben megmutatjuk, hogy

$$\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0.$$

Megjegyzés: Könnyen észrevehetjük, hogy a $\mathcal{L}_2$ és a $\mathcal{L}_1$ nyelvosztályok közötti, a tartalmazásra vonatkozó reláció nem azonnal látható a megfelelő grammatikák definíciójából.

Hivatkozás:

György E. Révész, Introduction to Formal Languages, McGraw-Hill Book Company, 1983, Chapter 1.