# Formális nyelvek - 6. 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

## Reguláris kifejezések

**Motiváció:**

Ismeretes, hogy **minden véges nyelv reguláris**. Tudjuk továbbá, hogy az $\mathcal{L}_3$ **nyelvosztály** (a reguláris nyelvek osztálya) **zárt az unió, a konkatenáció és az iteráció lezártja** műveletekre nézve.

Következésképpen, kiindulva véges számú véges nyelvből és az előzőekben felsorolt, ún. reguláris műveleteket véges sokszor alkalmazva reguláris nyelvet kapunk.

Kérdés az, hogy vajon ezzel az **eljárással minden reguláris nyelvet elő tudunk-e állítani**, azaz, ez a módszer elégséges-e az $\mathcal{L}_3$ nyelvosztály leírására?

# Reguláris kifejezések

**Definíció**

Legyenek $V$ és $V' = \{\varepsilon, \cdot, +, *, (, )\}$ diszjunkt ábécék. A $V$ ábécé feletti reguláris kifejezéseket rekurzív módon a következőképpen definiáljuk:

1. $\varepsilon$ reguláris kifejezés $V$ felett.
2. Minden $a \in V$ reguláris kifejezés $V$ felett.
3. Ha $R$ reguláris kifejezés $V$ felett, akkor $(R)^*$ is reguláris kifejezés $V$ felett.
4. Ha $Q$ és $R$ reguláris kifejezések $V$ felett, akkor $(Q) \cdot (R)$ és $(Q) + (R)$ is reguláris kifejezések $V$ felett.

A $*$, $\cdot$ és $+$ szimbólumok rendre az iteráció lezártjára, a konkatenációra és az unióra utalnak, azt jelölik.

## Reguláris kifejezések - megjegyzések

Minden reguláris kifejezés jelöl (meghatároz) valamely reguláris nyelvet.

Például $\varepsilon$ a $\{\varepsilon\}$ nyelvet, $a$ az $\{a\}$ nyelvet, $a + b$ az $\{a\} \cup \{b\} = \{a, b\}$ nyelvet és $a \cdot b$ az $\{a\}\{b\} = \{ab\}$ nyelvet jelöli.

A reguláris kifejezés a szintaxis, az, hogy hogyan értelmezzük, a szemantika.

# Példák

Legyen $V = \{a, b\}$. Az alábbi reguláris kifejezések mellett az általuk jelölt nyelv található.

**Megjegyzés:**

A zárójelek egyrésze elhagyható, ha a műveleteken precedenciát definiálunk. A szokásos sorrend $*, \cdot, +$.

- $a^*$ ugyanaz, mint $(a)^*$ és az $\{a\}^*$ nyelvet jelöli.
- $(a+b)^*$ ugyanaz, mint $((a)+(b))^*$ és az $\{a,b\}^*$ nyelvet jelöli.
- $a^* \cdot b$ ugyanaz, mint $((a)^*) \cdot (b)$ és az $\{a\}^* b$ nyelvet jelöli.
- $b + ab^*$ ugyanaz, mint $(b) + ((a) \cdot (b)^*)$ és a $\{b\} \cup \{a\}\{b\}^*$ nyelvet jelöli.
- $(a+b) \cdot a^*$ ugyanaz, mint $((a)+(b)) \cdot ((a)^*)$ és az $\{a,b\}\{a\}^*$ nyelvet jelöli.

## Axiómák reguláris kifejezésekre

Könnyen látható, hogy $\{a,b\}\{a\}^* = \{a\}\{a\}^* \cup \{b\}\{a\}^*$. Így

$$(a+b)a^* = a \cdot a^* + b \cdot a^*,$$

azaz, a két reguláris kifejezés ugyanazt a nyelvet jelöli.

Legyenek $P, Q, R$ reguláris kifejezések. Akkor $P$, $Q$ és $R$ helyébe reguláris kifejezéseket írva fennállnak az alábbi egyenlőségek.

$$P + (Q + R) = (P + Q) + R \qquad P \cdot (Q \cdot R) = (P \cdot Q) \cdot R$$

$$P + Q = Q + P \qquad P \cdot (Q + R) = P \cdot Q + P \cdot R$$

$$(P + Q) \cdot R = P \cdot R + Q \cdot R \qquad P^* = \varepsilon + P \cdot P^*$$

$$\varepsilon \cdot P = P \cdot \varepsilon = P \qquad P^* = (\varepsilon + P)^*$$

## Axiómák reguláris kifejezésekre - folytatás

Sok érdekes, reguláris kifejezések között fennálló egyenlőséget megkapunk, ha a fenti axiómákban a $P$, $Q$, $R$ reguláris kifejezéseket reguláris kifejezésekkel helyettesítünk.

Azonban sem az előbbi axiómákból, sem más véges axiómahalmazból nem kaphatjuk meg az összes reguláris kifejezést kizárólag helyettesítés segítségével.

Még egy további inferencia szabályra van szükségünk, nevezetesen, ha

$P = R + P \cdot Q$ és $\varepsilon \notin Q$, akkor $P = R \cdot Q^*$.

## Axiómák reguláris kifejezésekre - folytatás

Ha $P = R + P \cdot Q$ és $\varepsilon \notin Q$, akkor $P = R \cdot Q^*$.

Vegyük észre a hasonlóságot a $p = r + pq$ egyenlet megoldásával, amely $p = r(1-q)^{-1}$, ha $q \neq 1$-re. Továbbá, ha $|q| \leq 1$, akkor $(1-q)^{-1} = 1 + q + q^2 + q^3 + ...$, amely megfelel $Q^*$-nak.

## Axiómák reguláris kifejezésekre - folytatás

A teljesség biztosítása céljából még hozzáadjuk az $\emptyset$ szimbólumot a reguláris kifejezések halmazához, amely az üres nyelvet jelöli.

Ebben az esetben nincs szükségünk a $\varepsilon$ szimbólumra, mivel $\emptyset^* = \{\varepsilon\}$.

Így, a definícióban helyettesíthetjük a $\varepsilon$ szimbólumot az $\emptyset$ szimbólummal.

Ekkor helyettesítjük $\varepsilon$-t a megelőző axióma rendszerben $(\emptyset)^*$-gal és még egy további axiómát tekintünk:

$$\emptyset \cdot P = P \cdot \emptyset = \emptyset.$$

Az axiómák és a két inferencia szabály, nevezetesen a helyettesítés és a fenti feltételes egyenlőség elégséges ahhoz, hogy levezessünk minden érvényes egyenlőséget reguláris kifejezések között.

## Reguláris kifejezések versus reguláris nyelvek

**Tétel**

Minden reguláris kifejezés egy reguláris (3-típusú) nyelvet jelöl, és megfordítva, minden reguláris nyelvhez megadható egy, ezen nyelvet jelölő reguláris kifejezés.

## Bizonyításvázlat

1) Az állítás első fele a megelőző diszkusszióból következik.

2) Megmutatjuk, hogy minden $L$ reguláris nyelvhez, amelyet a $G = (N, T, P, S)$ normálformában adott reguláris grammatika generál, meg tudunk konstruálni egy reguláris kifejezést, amely az $L$ nyelvet jelöli.

Legyen $N = \{A_1, \ldots, A_n\}$, $n \geq 1$, $S = A_1$. ($G$ minden szabálya vagy $A_i \to aA_j$ vagy $A_i \to \varepsilon$ alakú, ahol $a \in T$, $1 \leq i, j \leq n$.)

Azt mondjuk, hogy az $A_i \Longrightarrow^* uA_j$ $(u \in T^*)$ levezetés **érinti** az $A_m$ nemterminálist, ha $A_m$ előfordul valamely közbülső mondatformában $A_i$ és $uA_j$ között a levezetésben.

Az $A_i \Longrightarrow^* uA_j$ levezetést $k$-**megszorítottnak** nevezzük, ha $0 \leq m \leq k$ teljesül minden $A_m$ nemterminálisra, amely a levezetésben előfordul.

## Bizonyításvázlat - folytatás

Definiáljuk a következő halmazokat:

$E^k_{i,j} = \{u \in T^* \mid$ létezik $A_i \Longrightarrow^* uA_j$ $k$-megszorított levezetés $\}$.

$k$**-szerinti indukcióval** bizonyítjuk, hogy az $E^k_{i,j}$ nyelvhez létezik őt jelölő reguláris kifejezés $i, j, k$-ra, ahol $0 \leq i, j, k \leq n$.

**Bázis:** $i \neq j$ esetén az $E^0_{i,j}$ halmaz vagy üres vagy $T$-beli betűkből áll. ($a \in E^0_{i,j}$, akkor és csak akkor, ha $A_i \to aA_j \in F$.) Ha $i = j$, akkor $E^0_{i,i}$ tartalmazza $\varepsilon$-t és nulla vagy több elemét $T$-nek, így $E^0_{i,j}$ reguláris kifejezéssel jelölhető.

**Indukciós lépés:** tegyük fel, hogy rögzített $k$-ra, $0 < k \leq n$, az $E^{k-1}_{i,j}$ nyelvek mindegyike jelölhető reguláris kifejezéssel. Akkor minden $i, j, k$-ra fennáll, hogy

$$E^k_{i,j} = E^{k-1}_{i,j} + E^{k-1}_{i,k} \cdot (E^{k-1}_{k,k})^* \cdot E^{k-1}_{k,j}.$$

Ekkor az indukciós feltevés alapján $E^k_{i,j}$ szintén jelölhető reguláris kifejezéssel.

Legyen $I_\varepsilon$ azon $i$ indexek halmaza, amelyekre $A_i \to \varepsilon$. Akkor $L(G) = \cup_{i \in I_\varepsilon} E^n_{1,i}$, azaz, $L$ reguláris kifejezéssel jelölhető.

# Helyettesítés

## Definíció

Legyen $V$ egy ábécé, valamint legyen minden $a \in V$-re $V_a$ ábécé és $s(a) \subseteq V_a^*$. Minden $u = a_1 a_2 \dots a_n \in V^*$ szóra definiáljuk az $u$ szó $s$ helyettesítését a következőképpen:

$$s(u) = s(a_1)s(a_2)\dots s(a_n).$$

Legyen továbbá $s(\varepsilon) = \varepsilon$. Az $s$ helyettesítés kiterjeszthető bármely $L \subseteq V^*$ nyelvre a következő módon: $s(L) = \{w \mid w \in s(u), u \in L\}$.

## Reguláris nyelvek zártsága a helyettesítésre nézve

Reguláris kifejezéseket használva könnyen látható, hogy az $\mathcal{L}_3$ nyelvosztály zárt a helyettesítésre nézve. A reguláris kifejezések halmaza nyilvánvalóan zárt a kifejezés minden betűjének valamely reguláris kifejezéssel való helyettesítésére nézve. (Lásd a megelőző diszkussziót).

Megjegyzés: A helyettesítés a homomorfizmus általánosítása.

**Irodalom:**

György E. Révész, Introduction to Formal Languages, McGraw-Hill Book Company, 1983, Chapter 3.5

## Áttekintés - 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 $u_1Au_2 \to u_1vu_2$ 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 $u \in T^*$.

# Áttekintés - normálformák

- Minden $G = (N, T, P, S)$ generatív grammatikához meg tudunk konstruálni egy vele ekvivalens és azonos típusú $G' = (N', T, P', S)$ generatív grammatikát úgy, hogy $P'$ egyetlen szabályának baloldalán sem fordul elő terminális szimbólum.

- Minden $G = (N, T, P, S)$ **környezetfüggetlen grammatikához** meg tudunk konstruálni egy **vele ekvivalens** $G' = (N', T, P', S')$ **környezetfüggetlen grammatikát** úgy, hogy
  - $G'$ minden szabályának jobboldala **nemüres szó**,
  - **kivéve** azt az esetet, ha az **üres szó benne van a** $G$ **által generált nyelvben**,
  - mely esetben $S' \to \varepsilon$ az **egyetlen** olyan szabály, melynek jobboldala az üres szó és ekkor $S'$ **nem fordul elő** a $G'$ egyetlen szabályának jobboldalán sem.

# Áttekintés - normálformák

A $G = (N, T, P, S)$ környezetfüggetlen grammatikát **Chomsky-normálformájúnak** mondjuk, ha minden egyes szabálya vagy

1. $X \to a$, ahol $X \in N$, $a \in T$, vagy

2. $X \to YZ$, ahol $X, Y, Z \in N$ alakú.

**Tétel:**

Minden $\varepsilon$-mentes $G = (N, T, P, S)$ környezetfüggetlen grammatikához meg tudunk konstruálni egy vele **ekvivalens** $G' = (N', T, P', S)$ **Chomsky-normálformájú** környezetfüggetlen grammatikát.

# Áttekintés - normálformák

**Definíció**

Egy környezetfüggetlen grammatika redukált, ha minden nemterminálisa aktív és elérhető.

**Tétel**

Minden környezetfüggetlen grammatikához meg tudunk konstruálni egy vele ekvivalens redukált környezetfüggetlen grammatikát.

## Áttekintés - normálformák

**Tétel**

Minden 3-típusú, azaz reguláris nyelv generálható egy olyan grammatikával, amelynek szabályai vagy

1. $X \to aY$, ahol $X, Y \in N$ és $a \in T$ vagy

2. $X \to \varepsilon$ alakúak, ahol $X \in N$.

## Áttekintés - zártsági tulajdonságok

Az unió, a konkatenáció, valamint az iteráció lezárása műveleteket együttesen **reguláris** műveleteknek nevezzük.

**Tétel:**

Az $\mathcal{L}_i$, $i = 0, 1, 2, 3$ nyelvosztályok mindegyike zárt a reguláris műveletekre nézve.

## Áttekintés - zártsági tulajdonságok

Az $\mathcal{L}_3$ nyelvosztály zárt a tükrözés és a helyettesítés műveletére nézve.

Az $\mathcal{L}_2$ nyelvosztály nem zárt a metszet műveletére nézve.

## Áttekintés - eldönthetőségi tulajdonságok

Eldönthető, hogy egy környezetfüggetlen grammatika által generált nyelv

- tartalmazza-e az üres szót,
- az üres nyelv-e,
- véges nyelv-e,
- végtelen nyelv-e.

## Áttekintés - környezetfüggetlen nyelv szavainak alakja

### Bar-Hillel lemma

Minden $L$ környezetfüggetlen nyelvhez meg tudunk adni két $p$ és $q$ természetes számot úgy, hogy minden olyan szó $L$-ben, amely hosszabb, mint $p$

$$uxwyz$$

alakú, ahol $|xwy| \leq q$, $xy \neq \varepsilon$, és minden

$$ux^iwy^iv$$

szó is benne van az $L$ nyelvben minden $i \geq 0$ egész számra ($u, x, w, y, v \in T^*$).

## Áttekintés - reguláris nyelvek előállíthatósága

Minden reguláris nyelv generálható bal-lineáris grammatikával, jobb-lineáris grammatikával, illetve leírható reguláris kifejezéssel.

## Valódi tartalmazások a Chomsky-hierarchiában

A reguláris nyelvek osztály valódi részosztálya a lineáris nyelvek osztályának, a lineáris nyelvek osztálya valódi részosztálya a környezetfüggetlen nyelvek osztályának.