

# Formális nyelvek és a fordítóprogramok alapjai

Előadók: Nagy Sára, mesteroktató

Algoritmusok és Alkalmazásai Tanszék

Dévai Gergely, adjunktus

Programozási Nyelvek és Fordítóprogramok Tanszék

# Elérhetőség:

Nagy Sára

2.608-as szoba, 8489-es mellék

Fogadóóra: hétfő, csütörtök 16-17 óráig

(2019/20. 2. félévben)

E-mail: [saci@inf.elte.hu](mailto:saci@inf.elte.hu)

# A formális nyelvek rész tartalma:

- A formális nyelvek és automaták alapvető fogalmainak megismertetése, a közöttük fennálló összefüggések bemutatása.
- A formális nyelvek elmélete szimbólumsorozatok halmazainak véges, tömör leírásának különféle módszereit adja. Módszereket ad annak eldöntésére, hogy egy szimbólumsorozat hozzátartozik-e egy adott formális nyelvhez vagy sem.

# Kapcsolódó tudományágak

- ▶ Fordítóprogramok
- ▶ Természetes nyelvek gépi feldolgozása
- ▶ Képfeldolgozás
- ▶ Molekuláris számítás (DNS számítás)
- ▶ stb.

# Irodalom:

- ▶ A. Salomaa, Formal Languages, Academic Press, 1973.
- ▶ Révész György, Bevezetés a formális nyelvek elméletébe, Tankönyvkiadó, 1977.
- ▶ Bach Iván: Formális nyelvek, Typotex, 2001.
- ▶ Fülöp Zoltán, Formális nyelvek és szintaktikus elemzésük, Polygon, Szeged, 2004.
- ▶ Hunyadvári László, Manhertz Tamás, Automaták és formális nyelvek, Elektronikus előadásjegyzet, 2006.
- ▶ Csima Judit, Friedl Katalin: Nyelvek és automaták, BMGE jegyzet, 2013.

# Alapfogalmak és jelölések

**Ábécé:** Ábécének nevezzük a jelek egy nem üres véges halmazát. Jele:  $V$

**Betű:** Az ábécé elemeit betűknek hívjuk. ( $a \in V$ )

**Szó:** A  $V$  ábécé elemeinek egy tetszőleges véges sorozatát a  $V$  ábécé feletti szónak nevezzük. (Ha  $V$  nem lényeges vagy egyértelmű, akkor szóról beszélünk.)

Ha  $u$  egy tetszőleges szó, akkor  $\ell(u)$  jelöli a szó hosszát.

$0 \leq \ell(u) < \infty$  ;  $\ell(\varepsilon) = 0$ , ahol  $\varepsilon$  az üres szó.

# Alapfogalmak és jelölések

A  $V$  ábécé feletti szavak halmazát  $V^*$ -gal jelöljük. A nem üres szavakét  $V^+$ -szal.

**Nyelv:**  $V^*$  valamely részhalmazát a  $V$  ábécé feletti nyelvnek nevezzük. Jele:  $L$

$$L \subseteq V^*$$

**Nyelvosztály (nyelvcsalád):** Nyelvek valamely összességét nyelvosztálynak hívjuk.

# Példák nyelvekre

$$V_1 = \{a, b, c\}$$

$$L_1 = \{ aab, b, acc, bab, a, bb \}$$

$$L_2 = \{ u \in V_1^* \mid u\text{-ban pontosan egy darab ,a' betű van.} \}$$

Lexikografikusan felsorolva:

$$L_2 = \{ a, ab, ac, ba, ca, abb, abc, acb, acc, bab, bac, bba, bca, cab, cac, cba, cca, \dots \}$$

---

$$V_2 = \{!, \#, @\}$$

$$L_3 = \{ u \in V_2^* \mid u\text{-ban nincs @ betű.} \}$$

Lexikografikusan felsorolva:

$$L_3 = \{ \varepsilon, !, \#, !!, !\#, \#!, ##, !!!, !!\#, \dots \}$$

---

$$L_4 = \{ \varepsilon \} \text{ az üres szót tartalmazó nyelv.}$$

$$L_5 = \emptyset \text{ az üres nyelv, ami egy üres halmaz.}$$



# Műveletek szavakon

## Két szó konkatenációja:

Legyenek  $u = t_1 \dots t_k$  és  $v = q_1 \dots q_m$  szavak egy  $V^*$  felett értelmezve.

Ekkor a két szó konkatenáltja az

$$uv := t_1 \dots t_k q_1 \dots q_m$$

(A két szó egymás utáni leírásával kapott szó.)

# Műveletek szavakon

$V^*$  zárt a konkatenáció műveletére.

$V^*$  a konkatenációra nézve egységelemes félcsoportot alkot.

Asszociatív:  $u, v, w \in V^*$

$$(uv)w = u(vw)$$

Egységelem:  $\varepsilon$  (üres szó) és  $v \in V^*$

$$\varepsilon v = v = v\varepsilon$$

# Műveletek szavakon

## Szó hatványa:

Legyen  $u$  egy tetszőleges szó.

Nemnegatív egész hatványai:

$$u^0 := \varepsilon$$

$$u^1 := u$$

$$u^n := u^{n-1}u \quad , \text{ ahol } n \geq 1.$$

# Műveletek szavakon

## Szó megfordítása:

Ha  $u$  egy tetszőleges szó és  $u = t_1 \dots t_k$   
és  $v = t_k \dots t_1$ , akkor  $v$  az  $u$  fordítottja,  
másképpen mondva a tükörképe.

Jelölése:  $u^{-1} = v$

# További alapfogalmak

- ▶ **Részszó:** A  $v$   $u$ -nak részszoja, ha léteznek olyan  $w_1; w_2$  szavak, hogy  $u = w_1vw_2$ .
- ▶ **Szó prefixe:** A  $v$  az  $u$  szó prefixe, ha van olyan  $w$  szó, hogy  $u = vw$ .  
Valódi prefix, ha  $v \neq \varepsilon$  és  $v \neq u$ .
- ▶ **Szó suffixe:** A  $v$  az  $u$  szó suffixe, ha van olyan  $w$  szó, hogy  $u = wv$ .  
Valódi suffix, ha  $v \neq \varepsilon$  és  $v \neq u$ .

# Példák

Legyen  $V=\{a,b,c\}$ .

$u=$ **aab**babbc

$v=$ **aab** //  $v$  prefixsze az  $u$ -nak

$u=$ aabb**abbc**

$w=$ **abbc** //  $w$  suffixe  $u$ -nak

$u=$ aa**bb**abbc

$z=$ **bb** //  $z$  részszoava  $u$ -nak, két helyen is

# Műveletek nyelveken

## Két nyelv uniója:

Legyenek  $L_1$  és  $L_2$  nyelvek  $V^*$  felettiek.

Ekkor az  $L_1$  és  $L_2$  nyelvek unióján az

$L_1 \cup L_2 := \{u \in V^* \mid u \in L_1 \text{ vagy } u \in L_2\}$   
nyelvet értjük.

# Műveletek nyelveken

Az unió kommutatív, asszociatív és egységelemes művelet.

Az egységelem az üres nyelv (üres halmaz).

Jele:  $\emptyset$

$$\emptyset \cup L = L = L \cup \emptyset$$



# Műveletek nyelveken

## Két nyelv metszete:

Legyenek  $L_1$  és  $L_2$  nyelvek  $V^*$  feletti.

Ekkor az  $L_1$  és  $L_2$  nyelvek metszetén az

$L_1 \cap L_2 := \{u \in V^* \mid u \in L_1 \text{ és } u \in L_2\}$   
nyelvet értjük.

# Műveletek nyelveken

Az  $L \subseteq V^*$  nyelv **komplementere** a  $V$  ábécére vonatkozóan:

$$\bar{L} := V^* \setminus L$$

,azaz minden olyan szó, ami nem tartozik az  $L$  nyelvbe.

$$\bar{L} \cap L = \emptyset \quad \text{és} \quad \bar{L} \cup L = V^*$$

# Műveletek nyelveken

Az  $L \subseteq V^*$  nyelv **tükörképe** az a nyelv, amely a szavainak megfordítottját tartalmazza.

Jele:  $L^{-1}$

$$L^{-1} := \{u \in V^* \mid u^{-1} \in L\}$$

Példa:

$$L = \{abc, aabb, acaa, baab\}$$

$$L^{-1} = \{cba, aaca, baab, bbaa\}$$

# Műveletek nyelveken

## Két nyelv konkatenációja:

Legyenek  $L_1$  és  $L_2$  nyelvek. Ekkor az  $L_1$  és  $L_2$  nyelvek konkatenációján az

$L_1 L_2 := \{uv \mid u \in L_1 \text{ és } v \in L_2\}$  nyelvet értjük.

A nyelvek halmaza a konkatenációra nézve egység elemes félcsoport alkot.

Egység elem:  $\{\varepsilon\}$   
(az üres szót tartalmazó nyelv)

$$\{\varepsilon\}L = L = L\{\varepsilon\}$$

$$\emptyset L = \emptyset = L\emptyset$$

# Pédák

Legyen  $L_1 = \{a, ab, bb\}$  és  $L_2 = \{b, ba\}$  .

$$\begin{aligned} L_1 L_2 &= \{ ab, ab, bb \} \cup \{ aba, abba, bbba \} \\ &= \{ ab, aba, abb, bbb, abba, bbba \} \end{aligned}$$

---

Legyen  $L_1 = \{a, ab\}$  és  $L_2 = \{\epsilon, b\}$  .

$$L_1 L_2 = \{ a\epsilon, ab\epsilon \} \cup \{ ab, abb \} = \{ a, ab, abb \}$$

---

Legyen  $L = \{a, ab, b\}$ .

$$\begin{aligned} LL &= \{ aa, aba, ba \} \cup \{ aab, abab, bab \} \cup \{ ab, abb, bb \} \\ &= \{ aa, ab, ba, bb, aab, aba, abb, bab, abab \} \end{aligned}$$

# Műveletek nyelveken

## Nyelv hatványa:

Legyen  $L$  egy tetszőleges nyelv.

Nemnegatív egész hatványai:

$$L^0 := \{\varepsilon\}$$

$$L^1 := L$$

$$L^n := L^{n-1}L, \text{ ahol } n \geq 1.$$

# Műveletek nyelveken

## Nyelv lezártja (iteráltja):

Legyen  $L$  egy tetszőleges nyelv.

$L^* := L^0 \cup L^1 \cup L^2 \cup \dots$  az  $L$  nyelv lezártja.

Másképpen:

$$L^* := \bigcup_{i \geq 0} L^i \quad \text{valamint} \quad L^+ = \bigcup_{i \geq 1} L^i$$

# Pédák

Legyen  $L = \{a, ab\}$  .

$L^* = \{\varepsilon\} \cup \{a, ab\} \cup \{aa, aab, aba, abab\} \cup \dots$

$(ab)^3 \in L^*$  , de  $ab^3 \notin L^*$ .      //  $(ab)^3 = ababab$ ;  $ab^3 = abbb$

$\{ a^n b \mid n > 0 \} \subseteq L^*$ , de  $a^0 b \notin L^*$ .    //  $a^0 b = \varepsilon b = b$

$\varepsilon \in L^*$ , de ebben a konkrét esetben  $\varepsilon \notin L^+$  .



# Példák

Legyen  $L = \{\epsilon, a, ab\}$  .

$L^* = \{\epsilon\} \cup \{\epsilon, a, ab\} \cup \{\epsilon, a, aa, ab, aab, aba, abab\} \cup \dots$

$(ab)^3 \in L^*$  , de  $ab^3 \notin L^*$ .      //  $(ab)^3 = ababab$ ;  $ab^3 = abbb$

$\{ a^n b \mid n > 0 \} \subseteq L^*$ , de  $a^0 b \notin L^*$ .    //  $a^0 b = \epsilon b = b$

$\epsilon \in L^*$ , de ebben a konkrét esetben  $\epsilon \in L^+$  -nak is.

Ebben az esetben  $L^* \setminus \{\epsilon\} \neq L^+$ .

# Műveletek nyelveken

Az alábbi három nyelvi műveletet **reguláris műveletnek** nevezzük:

- unió,
- konkatenáció,
- lezárás.

# Nyelvek megadási módjai

- logikai formulával
- strukturális rekurzióval
- algoritmussal
- matematikai gépekkel
- produkciós rendszerekkel

# Programozási nyelvek szintaxisa

Gyakran Backus-Naur formában (BNF) adják meg.

Példa:

$$\langle \text{kifejezés} \rangle ::= \langle \text{tag} \rangle \mid \langle \text{tag} \rangle + \langle \text{kifejezés} \rangle$$
$$\langle \text{tag} \rangle ::= \langle \text{faktor} \rangle \mid \langle \text{faktor} \rangle * \langle \text{tag} \rangle$$
$$\langle \text{faktor} \rangle ::= i \mid ( \langle \text{kifejezés} \rangle )$$

Példák kifejezésekre:

$i+i*i$ ,  $(i+i)*i$ ,  $i*i*i*i$ ,  $((i))$ ,  $i$

$V=\{+,*,(,),i\}$  felett értelmezett szavak

# Generatív grammatika (nyelvten)

Definíció: Grammatikának a következő négyest nevezzük:

$G=(N,T,P,S)$

- $N$  a nemterminális ábácé,
- $T$  a terminálisok ábécéje,
- $P$  az átírási szabályok véges halmaza,
- $S$  a kezdőszimbólum.

# Generatív grammatika (nyelvtan)

- ▶  $N$  és  $T$  diszjunkt halmazok, azaz  $N \cap T = \emptyset$ .
- ▶  $S \in N$ , kezdőszimbólum.
- ▶ A szabályok  $p \rightarrow q$  alakúak, ahol  $p, q \in (N \cup T)^*$  és  $p$  jelöli a szabály baloldalát,  $q$  a jobboldalát,  $\rightarrow$  a két oldalt elválasztó jel.
- ▶ A szabályok baloldala kötelezően tartalmaz legalább egy nemterminális szimbólumot.
- ▶  $(N \cup T)^*$  elemeit *mondatformáknak* nevezzük.

# Generatív grammatika (nyelvten)

Példa:

$G = (\{S\}, \{a,b\}, \{S \rightarrow aSb, S \rightarrow ab\}, S)$  egy grammatika.

Ez a grammatika az  $L = \{ a^n b^n \mid n \geq 1 \}$  nyelvet definiálja.

Levezetés:

$$S \xRightarrow[G]{ } aSb \xRightarrow[G]{ } aaSbb \xRightarrow[G]{ } aaaSbbb \xRightarrow[G]{ } aaaabbbb$$

$$S \xRightarrow[G]{*} a^4b^4$$

# Közvetlen levezetés

Legyen  $G = (N, T, P, S)$  egy adott grammatika.

Legyen  $u, v \in (N \cup T)^*$ .

Azt mondjuk, hogy a  $v$  mondatforma **közvetlenül** levezethető az  $u$  mondatformából, ha létezik  $u_1, u_2 \in (N \cup T)^*$  és  $x \rightarrow y \in P$  úgy, hogy  $u = u_1xu_2$  és  $v = u_1yu_2$ .

Jelölése:  $u \Rightarrow_G v$



# Példa

$u = ab\textcolor{red}{Bca}aAcb$

$v = ab\textcolor{red}{caBA}aAcb$

$Bca \rightarrow caBA \in P$

Az  $u$  modatformából közvetlenül levezethető a  $v$  mondatforma a megadott szabály segítségével.

$\underline{ab}\textcolor{red}{Bca}\underline{aAcb} \xRightarrow{G} \underline{ab}\textcolor{red}{caBA}\underline{aAcb}$ , ahol  $u_1=ab$  és  $u_2=aAcb$

## Közvetett levezetés

Legyen  $G = (N, T, P, S)$  egy adott grammatika.

Legyen  $u, v \in (N \cup T)^*$ .

Azt mondjuk, hogy a  $v$  mondatforma **közvetetten** levezethető az  $u$  mondatformából, ha létezik olyan  $k \geq 0$  szám és  $x_0, \dots, x_k \in (N \cup T)^*$ , hogy  $u = x_0$  és  $v = x_k$  és  $\forall i \in [0, k-1]: x_i \xRightarrow[G]{} x_{i+1}$ .

Jelölése:  $u \xRightarrow[G]{*} v$

# Grammatika által generált nyelv

Minden olyan szó, amely közvetetten levezethető a kezdőszimbólumból.

$$L(G) := \{ u \in T^* \mid S \xRightarrow[G]{*} u \}$$

# Példa:

$G = (\{S,A,B\}, \{a,b\}, P, S)$  egy grammatika.

$P: S \rightarrow ASB$

$S \rightarrow AB$

$AB \rightarrow BA$  //csere szabály

$A \rightarrow a$

$B \rightarrow b$

$L(G) = \{ u \in \{a,b\}^* \mid \ell_a(u) = \ell_b(u) \geq 1 \}$  nyelvet definiálja.

(Ugyanannyi ,a' és ,b' betű van a szavakban.)

Levezetés (példa egy szó levezetésére):

$$S \xRightarrow[G]{*} A^n B^n \xRightarrow[G]{} A^{n-1} B A B^{n-1} \xRightarrow[G]{*} B A^{n-1} A B^{n-1} \xRightarrow[G]{*} b a^n b^{n-1}$$

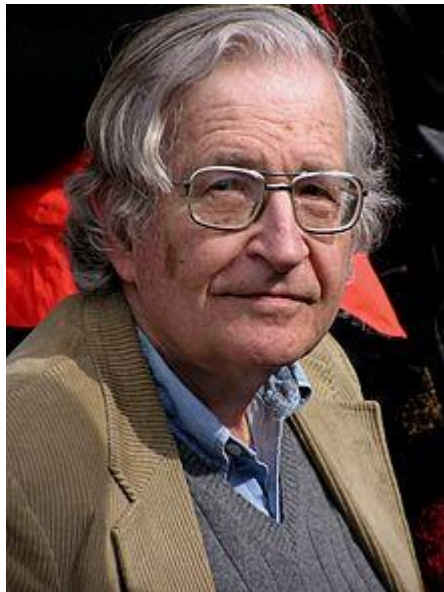
# Ekvivalencia

A  $G_1$  es  $G_2$  nyelvtanok **ekvivalensek**, ha

$L(G_1) = L(G_2)$ , azaz ugyanazt a nyelvet generálják.

Gyengén ekvivalensek, ha  $L(G_1) \setminus \{\varepsilon\} = L(G_2) \setminus \{\varepsilon\}$ .

# Noam Chomsky (született: 1928)



Noam Chomsky amerikai nyelvész, a Massachusetts Institute of Technology professzora, a generatív nyelvtan elméletének megalkotója, filozófus, politikai aktivista, előadó és lektor. Kidolgozója a róla elnevezett Chomsky-hierarchiának. (Wikipédia)

*Köszönöm a figyelmet!*