

Formális nyelvek és fordítóprogramok

1. Nyelvek és nyelvtanok I.

Bodó Zalán

Babeş–Bolyai Tudományegyetem
Matematika és Informatika Kar
Magyar Matematika és Informatika Intézet



Alapfogalmak

- ▶ **ábécé:** szimbólumok nem üres véges halmaza; pl.
 $\Sigma = \{a, b, c\}$, $\Sigma = \{0, 1\}$
- ▶ **szó:** az ábécé felett alkotott szimbólumsorozat; $a_1 a_2 \dots a_n$
egy szó, azaz $a_1 a_2 \dots a_n \in \Sigma^*$ vagy $a_1, \dots, a_n \in \Sigma$, $n \geq 0$; a
szó hossza n
- ▶ **a szó hossza:** $w = a_1 \dots a_n$, akkor $|w| = n$
- ▶ **üres szó:** ε vagy λ ; ha $w = \varepsilon$, akkor $|w| = 0$
- ▶ $\Sigma^* = \{a_1 a_2 \dots a_n \mid n \geq 0, a_1, a_2, \dots, a_n \in \Sigma\}$
- ▶ $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$
- ▶ $\Sigma^n = \{a_1 \dots a_m \mid a_1, \dots, a_m \in \Sigma, m = n\}$
 - ▶ $\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \dots$
 - ▶ $\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \dots$
- ▶ **konkatenáció** (szorzat, összefűzés):
 $u = a_1 a_2 \dots a_m, v = b_1 b_2 \dots a_n \Rightarrow uv = a_1 \dots a_m b_1 \dots b_n$;
 $|uv| = |u| + |v|$
- ▶ **szavak hatványozása:** $\forall u \in \Sigma^*$:

$$u^0 = \varepsilon$$

$$u^n = u^{n-1} u$$

- ▶ **szó tükörképe:** $u = a_1 a_2 \dots a_n$ tükörképe
 $u^{-1} = a_n a_{n-1} \dots a_1$
 - ▶ $(u^{-1})^{-1} = u$
 - ▶ $(uv)^{-1} = v^{-1} u^{-1}$
- ▶ **részszó:** v részszoja u -nak, ha $\exists p, q \in \Sigma^*$ úgy, hogy $u = pvq$
- ▶ **valódi részszó:** ha érvényes a fenti és $pq \neq \varepsilon$
- ▶ **kezdőszelet** (prefixum): a fenti esetben p prefixuma u -nak
- ▶ **végszelet** (szuffixum): a fenti esetben q szuffixuma u -nak
- ▶ **formális nyelv:**
 - ▶ L -t, $L \subseteq \Sigma^*$, a Σ ábécé feletti (formális) nyelvnek nevezzük
 - ▶ $L = \emptyset$ az **üres nyelv**
 - ▶ $L = \{\varepsilon\}$ az **üres szóból álló nyelv**

Műveletek nyelvekkel

Legyen L , L_1 és L_2 egy-egy Σ fölötti nyelv.

- ▶ **egyesítés:** $L_1 \cup L_2 = \{u \in \Sigma^* \mid u \in L_1 \text{ vagy } u \in L_2\}$
- ▶ **metszet:** $L_1 \cap L_2 = \{u \in \Sigma^* \mid u \in L_1 \text{ és } u \in L_2\}$
- ▶ **különbség:** $L_1 \setminus L_2 = \{u \in \Sigma^* \mid u \in L_1 \text{ és } u \notin L_2\}$
- ▶ **komplementum:** $\bar{L} = \Sigma^* \setminus L$
- ▶ **szorzat:** $L_1 L_2 = \{uv \mid u \in L_1 \text{ és } v \in L_2\}$
- ▶ **nyelv hatványa:** $L^0 = \{\varepsilon\}$, $L^n = L^{n-1}L$, $n \geq 1$
- ▶ **nyelv iteráltja:** $L^* = \bigcup_{i=0}^{\infty} L^i$, $L^+ = L^* \setminus \{\varepsilon\}$
- ▶ **tükrözés:** $L^{-1} = \{u^{-1} \mid u \in L\}$
- ▶ **reguláris műveletek:** egyesítés, szorzás, iteráció

Nyelvek, nyelvtanok

nyelvek megadása:

- ▶ **felsorolással:** $L_1 = \{0, 1\}$, $L_2 = \{\varepsilon, a, aa, aaa, \dots\}$
- ▶ **tulajdonság segítségével:** $L_3 = \{a^n b^n \mid n \geq 0\}$,
 $L_4 = \{uu^{-1} \mid u \in \{0, 1\}^*\}$, $L_5 = \{u \in \{a, b\}^* \mid n_a(u) = n_b(u)\}$
- ▶ **nyelvtannal** ([generatív] grammatikával)

Def. Generatív nyelvtan

Generatív nyelvtannak nevezzük a $G = (N, T, P, S)$ négyest, ahol

- ▶ N a **nemterminális szimbólumok** halmaza,
- ▶ T a **terminális szimbólumok** halmaza,
- ▶ $P \subseteq (N \cup T)^* N (N \cup T)^* \times (N \cup T)^*$ a **helyettesítési szabályok** véges halmaza,
- ▶ $S \in N$ a grammatika **kezdőszimbóluma**.

- ▶ nemterminális szimbólumok: a szavak generálásában játszanak szerepet, nem szerepelnek magukban a szavakban; általában nagybetűkkel jelöljük.
Pl. $N = \{A, B, C\}$
- ▶ terminális szimbólumok: a szavakban szerepelnek; a szavak ábécéje.
Pl. $T = \{0, 1\}$
- ▶ helyettesítési szabályok: $u \rightarrow v$, ahol $u, v \in (N \cup T)^*$, és u legalább egy darab nemterminálist tartalmaz(!)
Pl. $P = \{S \rightarrow a|aS|aA, A \rightarrow b|bA|bS\}$
- ▶ kezdőszimbólum: általában S -sel jelöljük.

Levezetések

- ▶ **közvetlen levezetés:** $u \Rightarrow v$, ha $u = p_1pp_2$, $v = p_1qp_2$ és $(p \rightarrow q) \in P$
- ▶ $\Rightarrow$ helyett használhatjuk a $\vdash$ vagy $\models$ jelek egyikét
- ▶ közvetlen levezetés a G nyelvtannal: $\Rightarrow_G$
- ▶ $\Rightarrow$ reflexív és tranzitív lezártja: $\Rightarrow^*$
- ▶ $\Rightarrow$ tranzitív lezártja: $\Rightarrow^+$
- ▶ **levezetés:** $\Rightarrow^*$

Def. Grammatika által generált nyelv

A $G = (N, T, P, S)$ grammatika által generált nyelv

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

Def. Mondat, mondatforma

Ha $w \in T^*$, akkor azt **mondat**nak nevezzük;

ha $w \in (N \cup T)^*$, akkor **mondatformá**nak nevezzük.

1. Példa

Legyen $G = (N, T, P, S)$, ahol $N = \{S\}$, $T = \{a, b\}$,
 $P = \{S \rightarrow aSb, S \rightarrow ab\}$. Könnyű belátni, hogy a grammatika
által generált nyelv $L(G) = \{a^n b^n \mid n \geq 1\}$, mivel

$$S \Rightarrow aSb \Rightarrow a^2 Sb^2 \Rightarrow \dots a^n b^n$$

Def. Ekvivalens nyelvtanok

G_1 és G_2 **ekvivalensek**, vagyis $G_1 \cong G_2$ vagy $G_1 \equiv G_2$, ha $L(G_1) = L(G_2)$, vagyis a grammatikák által generált nyelvek megegyeznek.

2. Példa

Legyen

$$G_1 = (N_1, T, S_1, P_1):$$

$$N_1 = \{S_1\}$$

$$T = \{a, b\}$$

$$P_1 = \{S_1 \rightarrow aS_1b \mid ab\}$$

A két grammatika ekvivalens.

$$G_2 = (N_2, T, S_2, P_2):$$

$$N_2 = \{S_2, A\}$$

$$P_2 = \{S_2 \rightarrow aA \mid ab,$$

$$A \rightarrow S_2b\}$$

3. Példa

Legyen

$$G_1 = (N_1, T, S_1, P_1):$$

$$N_1 = \{S_1\}$$

$$T = \{a\}$$

$$P_1 = \{S_1 \rightarrow aS_1 \mid \varepsilon\}$$

A két grammatika ekvivalens.

$$G_2 = (N_2, T, S_2, P_2):$$

$$N_2 = \{S_2\}$$

$$P_2 = \{S_2 \rightarrow S_2a \mid \varepsilon\}$$

4. Példa

Az alábbi $G_1 = (N_1, T, P_1, S_1)$ és $G_2 = (N_2, T, P_2, S_2)$ nyelvtanok ekvivalensek, mert mindkettő az $\{a^n b^n c^n \mid n \geq 1\}$ nyelvet generálja.

$$N_1 = \{S_1, X, Y\}, \quad T_1 = \{a, b, c\},$$
$$P_1 = \{S_1 \rightarrow abc \mid aXbc, Xb \rightarrow bX, Xc \rightarrow Ybcc, bY \rightarrow Yb, aY \rightarrow aaX \mid aa\}$$
$$N_2 = \{S_2, A, B, C\}$$
$$P_2 = \{S_2 \rightarrow aS_2BC \mid aBC, CB \rightarrow BC, aB \rightarrow ab, bB \rightarrow bb, bC \rightarrow bc, cC \rightarrow cc\}.$$

Tétel

Létezik olyan formális nyelv, amit nem lehet nyelvtannal megadni.

BIZONYÍTÁS. A bizonyításhoz a nyelvtanokat a $\{0, 1\}$ ábécé felett kódoljuk. Adott $G = (N, T, P, S)$ nyelvtan esetében legyen $N = \{S_1, \dots, S_n\}$, $T = \{a_1, \dots, a_m\}$ és $S = S_1$. A kódolás pedig legyen a következő:

S_i kódja $10 \underbrace{11 \dots 11}_{i\text{-szer}} 01$

a_j kódja $100 \underbrace{11 \dots 11}_{j\text{-szer}} 001$

A szimbólumok kódját a kódolásban 000 választja el egymástól, a nyíl kódja 0000, a szabályokat pedig a 00000 szimbólumsorozattal választjuk el.

Mivel a szabályok meghatározzák a többi halmazt (feltéve, hogy például mindig azzal a szabállyal kezdjük a felsorolást, amelyik bal oldalán a kezdőszimbólum van), elegendő csak azokat kódolni.

Például legyen a következő nyelvtan:

$$G = (\{S\}, \{a, b\}, \{S \rightarrow aSb \mid ab\}, S)$$

$$\begin{array}{c|l} S & 10101 \\ a & 1001001 \\ b & 10011001 \end{array}$$

A nyelvtan kódja:

$$\underbrace{10101}_{10101} \underbrace{0000}_{0000} \underbrace{1001001}_{1001001} \underbrace{000}_{000} \underbrace{10101}_{10101} \underbrace{000}_{000} \underbrace{10011001}_{10011001} \underbrace{00000}_{00000}$$

$$\underbrace{10101}_{10101} \underbrace{0000}_{0000} \underbrace{1001001}_{1001001} \underbrace{000}_{000} \underbrace{10011001}_{10011001}$$

$\Rightarrow$ a nyelvtanok *sorrendbe rendezhetők és felsorolhatók*:

$G_1, G_2, \dots, G_k, \dots \Rightarrow$ ezen nyelvtanok halmazának számossága *megszámlálhatóan végtelen*.

Legyen $\mathcal{L}_T = \{L \mid L \subseteq T^*\}$, vagyis $\mathcal{L}_T = \mathcal{P}(T^*)$. T^* megszámlálhatóan végtelen, mert szavai sorrendbe írhatók. Legyen ez a sorrend $s_0, s_1, \dots$, ahol $s_0 = \varepsilon$. Ekkor minden $L \in \mathcal{L}_T$ nyelvnek megfeleltetünk egy végtelen $b_0, b_1, \dots$ bináris sorozatot a következőképpen:

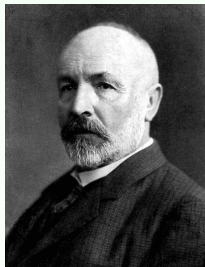
$$b_i = \begin{cases} 1, & \text{ha } s_i \in L \\ 0, & \text{ha } s_i \notin L \end{cases}, i = 0, 1, \dots$$

Minden ilyen sorozat tekinthető úgy, mint egy 1-nél kisebb kettes számrendszerbeli valós szám – a tizedesvesszőt az első számjegy elé képzeljük. Ez fordítva is igaz.

⇒ végtelen bináris sorozatok számossága, és így $\mathcal{L}_T$ számossága megegyezik a $[0, 1]$ intervallum *kontinuum* számosságával.

⇒ a formális grammatikával megadott, T terminális ábécével rendelkező nyelvek halmazának számossága kisebb, mint az összes T ábécé fölött értelmezett nyelv halmaza. $\square$

Georg Cantor



Georg Cantor (Georg Ferdinand Ludwig Philipp Cantor, 1845–1918)
orosz–német matematikus, a
halmazelmélet tudományág
megalkotója.

Hozzá kapcsolódó
fogalmak/tételek/...:

- ▶ $\mathbb{N}$, $\mathbb{N}_0$ (természetes számok
halmazának számossága)
- ▶ Cantor átlós eljárása (*Cantor's
diagonal argument*)
- ▶ kontinuum-hipotézis,
kontinuum-számosság $\mathfrak{c}$

Egy kis halmazelmélet

- ▶ $\mathbb{N}$ számossága: $\aleph_0$
- ▶ $\mathbb{R}$ számossága, bármely (a, b) valós intervallum számossága: $\mathfrak{c}$
- ▶ Cantor: $\mathfrak{c} > \aleph_0$
- ▶ Cantor–Bernstein–Schröder-tétel: $\mathfrak{c} = |\mathcal{P}(\mathbb{N})| = 2^{\aleph_0}$