Formális nyelvek és a fordítóprogramok alapjai

2. előadás

Előadó: Nagy Sára, mesteroktató
Algoritmusok és Alkalmazásaik Tanszék

Emlékeztető:

V - ábécé, jelek nem üres véges halmaza;

V^* - az adott jelkészlet felett értelmezett összes szó;

$L \subseteq V^*$ - formális nyelv, szavak halmaza.

Nyelv megadása szabályrendszerrel

Definíció: Grammatikának (nyelvtannak) 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.

Grammatika: G=(N,T,P,S)

- N és T diszjunkt halmazok, azaz $N \cap T = \emptyset$.
- $S \in N$, kezdőszimbólum.
- A szabályok $p \rightarrow q$ alakúak, ahol $p \in (N \cup T)^* N (N \cup T)^*$, $q \in (N \cup T)^*$ 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 *mondatformá*knak nevezzük.

Grammatika által generált nyelv

Minden olyan szó, amely közvetetten levezethető a kezdőszimbólumból.

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

Generatív grammatika (nyelvtan)

Példa:

G = ({S}, {a,b}, {S → aSb, S → ab}, S) egy grammatika.

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

Levezetés:

$S \underset{G}{\Rightarrow} aSb \underset{G}{\Rightarrow} aaSbb \underset{G}{\Rightarrow} aaaSbbb \underset{G}{\Rightarrow} aaaabbbb$

$S \underset{G}{\Rightarrow}^* a^4b^4$

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 $x \to y \in P$ úgy, hogy $u = u_1 x u_2$ és $v = u_1 y u_2$.

Jelölése: $u \underset{G}{\Rightarrow} v$

Példa

u = abBcaaAcb

v = abcaBAaAcb

$Bca \longrightarrow caBA \in P$

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

$\underline{ab}Bca\underline{aAcb} \underset{G}{\Rightarrow} \underline{ab}caBA\underline{aAcb}$, 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 $x_0, \ldots, x_k \in (N \cup T)^*$, hogy

$u = x_0$ és $v = x_k$ és $\forall\, i \in [0,k-1]$: $x_i \underset{G}{\Rightarrow} x_{i+1}$.

Jelölése: $u \underset{G}{\Rightarrow}^* v$

Grammatika által generált nyelv

Minden olyan szó, amely közvetetten levezethető a kezdőszimbólumból.

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

Példa:

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

P: $S \to ASB$

$S \to AB$

$AB \to BA$ //csere szabály

$A \to a$

$B \to b$

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

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

Levezetés (példa egy szó levezetésére):

$S \underset{G}{\Rightarrow}^* A^nB^n \underset{G}{\Rightarrow} A^{n-1}BAB^{n-1} \underset{G}{\Rightarrow}^* BA^{n-1}AB^{n-1} \underset{G}{\Rightarrow}^* ba^nb^{n-1}$

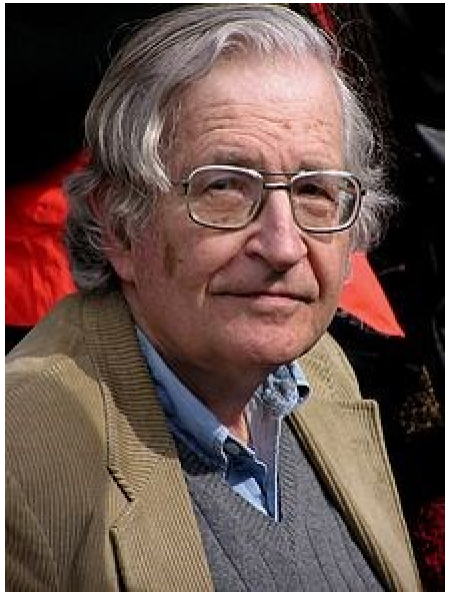
Ekvivalencia

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)

Chomsky féle grammatika típusok

Definició: A G =(N,T,P,S) grammatika i-típusú (i =0,1,2,3), ha P szabályhalmazára teljesülnek a következők:

- **i =0**: Nincs korlátozás.
- **i =1:** P minden szabálya $\mathbf{u_1Au_2} \rightarrow \mathbf{u_1vu_2}$ alakú,
  ahol $u_1,u_2,v \in (N \cup T)^*$, $A \in N$, és $v \neq \varepsilon$,
  kivéve az $S \rightarrow \varepsilon$ alakú szabályt, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem.
  (Ezt "Korlátozott ε szabály”-nak, röviden: KES szabálynak hívjuk.)
- **i =2**: P minden szabálya $\mathbf{A} \rightarrow \mathbf{v}$ alakú, ahol $A \in N$, $v \in (N \cup T)^*$.
- **i =3**: P minden szabálya vagy $\mathbf{A} \rightarrow \mathbf{uB}$ vagy $\mathbf{A} \rightarrow \mathbf{u}$ alakú,
  ahol $A,B \in N$ és $u \in T^*$.

Chomsky féle grammatika típusok

Jelölje $\mathcal{G}_i$ az i-típusú grammatikák halmazát.

A grammatikák alakjából következik, hogy

$\mathcal{G}_i \subseteq \mathcal{G}_0$, ahol i=1,2,3.

$\mathcal{G}_3 \subseteq \mathcal{G}_2$

Nyelvek típusai

Egy L nyelvet i-típusúnak nevezünk ($i \in \{0,1,2,3\}$), ha létezik olyan i-típusú grammatika, ami az L nyelvet generálja.

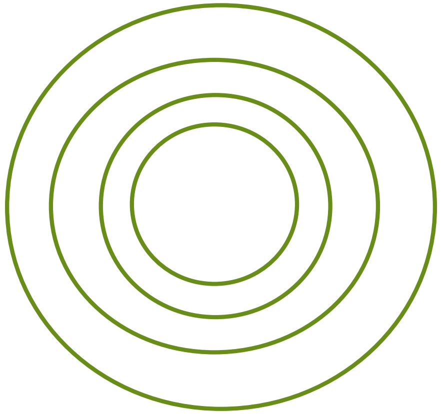
Jelölje $\mathcal{L}_i$ az i-típusú nyelvek halmazát. (Nyelvcsalád.)

Chomsky féle hierarchia

$$\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_1 \subseteq \mathcal{L}_0$$

Pontosabban valódi tartalmazás van

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



Chomsky féle grammatika típusok

Típus	Alaptípus szabályai	Speciális alakok szabályai	Normál forma szabályai
0.	Nincs korlátozás.	$p \to q$, ahol $p \in N^+$, $q \in (N \cup T)^*$	
1.	$u_1 A u_2 \to u_1 v u_2$, ahol $u_1, u_2, v \in (N \cup T)^*$, $A \in N$, és $v \neq \varepsilon$, kivéve az $S \to \varepsilon$, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem. (környezetfüggő grammatika)	$p \to q$, ahol $l(p) \leq l(q)$ kivéve az $S \to \varepsilon$, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem. (hosszúság nemcsökkentő grammatika)	Kuroda normál forma $A \to a$ vagy $A \to B$ vagy $A \to BC$ vagy $AB \to CD$ alakúak a szabályok, ahol $a \in T$ és $A,B,C,D \in N$, kivéve az $S \to \varepsilon$, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem.
2.	$A \to v$, ahol $v \in (N \cup T)^*$, $A \in N$ (környezetfüggetlen grammatika)	$A \to v$, ahol $v \in (N \cup T)^*$, $A \in N$ és $v \neq \varepsilon$, kivéve az $S \to \varepsilon$, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem.	Chomsky normál forma $A \to a$ vagy $A \to BC$ alakúak a szabályok, ahol $a \in T$ és $A,B,C \in N$, kivéve az $S \to \varepsilon$, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem.
3.	$A \to uB$ vagy $A \to u$, ahol $u \in T^*$, $A,B \in N$ (reguláris grammatika)	$A \to aB$ vagy $A \to a$, ahol $a \in T$, és $A,B \in N$, kivéve az $S \to \varepsilon$, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem.	3-as normál forma $A \to aB$ vagy $A \to \varepsilon$, ahol $a \in T$, és $A,B \in N$.

Példa

$L=\{ u \in \{a,b\}^* \mid \ell_a(u)=\ell_b(u) \geq 0 \}$

$G_1 = (\{S', S, A, B\}, \{a,b\}, P, S')$ egy grammatika.

P_1: $S' \to \varepsilon$	KES
$S' \to S$	0,1,2,**3**-as típusú szabály
$S \to ASB$	0,1,**2**-es típusú szabály
$S \to AB$	0,1,**2**-es típusú szabály
$AB \to BA$	**0**-s típusú szabály
$A \to a$	0,1,2,**3**-as típusú szabály
$B \to b$	0,1,2,**3**-as típusú szabály

A G_1 grammatika 0-s típusú. $G_1 \in \mathcal{G}_0$

Példa

$L=\{ u \in \{a,b\}^* \mid \ell_a(u)=\ell_b(u) \geq 0 \}$

$G_2 = (\{S,A,B\}, \{a,b\}, P, S)$ egy grammatika.

P_2: $S \to aSbS$

$S \to bSaS$

$S \to \varepsilon$

Állítás: $L=L(G_2)$.

Bizonyítás: Mivel ‚a' és ‚b' együtt generálódik, ezért $L(G_2) \subseteq L$.

Kérdés, hogy minden L beli szó generálható-e, azaz $L \subseteq L(G_2)$?

Példa folytatása

Kérdés, hogy minden L beli szó generálható-e, azaz $L \subseteq L(G_2)$?

Minden szó hossza páros, azaz $\ell(u) \geq 2*k$, ahol $k \in \mathbb{N}$.

A tartalmazás k szerinti teljes indukcióval bizonyítható.

k=1 esetén látható, hogy $S \underset{G}{\Rightarrow} aSbS \underset{G}{\Rightarrow^*} ab$ és $S \underset{G}{\Rightarrow} bSaS \underset{G}{\Rightarrow^*} ba$

Tegyük fel, hogy a 2k hosszú szavakat le tudjuk vezetni. Legyen $\ell(u)=2*(k+1)$.

Ha az u szó ‚a'-val kezdődik, akkor biztosan van egy ‚b' párja, azaz

u=vw, ahol v az első olyan prefix, hogy $v \in L$, azaz $\ell_a(v)=\ell_b(v)$ és v=axb.

Ekkor $x,w \in L$ és $2*k \geq \ell(x) \geq 0$ és $2*k \geq \ell(w) \geq 0$.

Így $S \underset{G}{\Rightarrow} aSbS \underset{G}{\Rightarrow^*} axbS \underset{G}{\Rightarrow^*} axbw$, ahol teljes indukciót alkalmazhatunk x és w levezetésére.

Ha ‚b'-vel kezdődne a szó, akkor hasonlóan járunk el.

Példa

$L=\{ u \in \{a,b\}^* \mid \ell_a(u)=\ell_b(u) \geq 0 \}$

G_2 = ({S,A,B}, {a,b}, P, S) egy grammatika.

P_2: $S \to aSbS$ 0,1,**2**-es típusú szabály

$S \to bSaS$ 0,1,**2**-es típusú szabály

$S \to \varepsilon$ 0,2,**3**-as típusú szabály (nem KES)

A G_2 grammatika 2-es típusú. $G_2 \in \mathcal{G}_2$

$L(G_1)=L(G_2)$, azaz a két grammatika ekvivalens.

Az L nyelv 2-es típusú, mert van hozzá 2-es típusú grammatika.

$L \in \mathcal{L}_2$

Lehet, hogy $L \in \mathcal{L}_3$?

Válasz: nem. Bizonyítás később.

Chomsky féle grammatika típusok

Definició: A $G = (N,T,P,S)$ grammatika i-típusú $(i = 0,1,2,3)$,ha P szabályhalmazára teljesülnek a következők:

- **i =0**: Nincs korlátozás.
- **i =1:** P minden szabálya $\mathbf{u_1Au_2 \rightarrow u_1vu_2}$ alakú,
  ahol $u_1,u_2,v \in (N \cup T)^*$, $A \in N$, és $v \neq \varepsilon$,

kivéve az $S \rightarrow \varepsilon$ alakú szabályt, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem (röviden: KES)

- **i =2**: P minden szabálya $\mathbf{A \rightarrow v}$ alakú, ahol $A \in N$, $v \in (N \cup T)^*$.
- **i =3**: P minden szabálya vagy $\mathbf{A \rightarrow uB}$ vagy $\mathbf{A \rightarrow u}$ alakú,
  ahol $A,B \in N$ és $u \in T^*$.

Chomsky féle hierarchia

$$\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_1 \subseteq \mathcal{L}_0$$

Azonban

$$\mathcal{G}_3 \subseteq \mathcal{G}_2 \not\subseteq \mathcal{G}_1 \subseteq \mathcal{G}_0$$

Ha a 2-es típusú szabályoknál is kikötnénk, hogy v≠ε, akkor igaz lenne a tartalmazás, és akkor triviálisan igaz lenne a nyelvcsaládokra is tartalmazás.

Nyelvtani transzformáció

A nyelvtani transzformáció olyan eljárás, amely egy G grammatikából egy másik G' grammatikát készít.

Ekvivalens transzformációról beszélünk, ha minden G grammatikára és az ő G' transzformáltjára igaz, hogy **L(G)=L(G')**.

ε-mentesítés

Tétel:

Minden $G=(N,T,P,S)$ környezetfüggetlen (2-es típusú) grammatikához megkonstruálható egy vele ekvivalens $G'=(N',T,P',S')$ környezetfüggetlen grammatika úgy, hogy P'-ben **nincs $A\rightarrow\varepsilon$** alakú szabály, kivéve, ha $\varepsilon\in L(G)$, mert akkor $S'\rightarrow\varepsilon\in P'$, de ekkor S' nem szerepelhet szabály jobboldalán.

ε-mentesítés

Első lépésben meghatározzuk, hogy mely nemterminálisokból vezethető le az üres szó.

$H := \{ A \in N \mid A \underset{G}{\Rightarrow}^{*} \varepsilon \}$

Ehhez definiáljuk a H_i ($i \geq 1$) halmazokat:

$H_1 := \{ A \in N \mid \exists A \rightarrow \varepsilon \in P \}$

$H_{i+1} := H_i \cup \{ A \in N \mid \exists A \rightarrow w \in P \text{ és } w \in H_i^{*} \}$

$H_1 \subseteq H_2 \subseteq \ldots \subseteq H_k = H_{k+1}$ $\exists k$ és legyen $H := H_k$.

ε-mentesítés

Ekkor látható,
ha $A \in N$ és $A \underset{G}{\Rightarrow}^{*} \varepsilon$, akkor, és csak akkor, ha $A \in H$.

Ennek következménye, hogy

$\varepsilon \in L(G)$, akkor, és csak akkor, ha $S \in H$.

ε-mentesítés

Második lépésben átalakítjuk H ismeretében a grammatika szabályait a kellő alakúra.

$S \notin H$ estén:

$A \rightarrow v' \in P'$, akkor, és csak akkor, ha $v' \neq \varepsilon$ és $\exists A \rightarrow v \in P$ úgy,hogy v'-t v-ből úgy kapjuk, hogy elhagyunk nulla vagy több H-beli nemterminálist v-ből.

ε-mentesítés

$S \in H$ estén:

A korábbi szabályokhoz hozzá vesszük még a következő két szabályt:

$$S' \to \varepsilon \text{ és } S' \to S$$

,ahol $S' \notin N$ a G' grammatika új kezdőszimbóluma.

Megjegyzés: Az átalakítás megőrzi a 2. és 3. típust.

Példa

$L=\{ u \in \{a,b\}^* \mid \ell_a(u)=\ell_b(u) \geq 0 \}$

$G = (\{S\}, \{a,b\}, P, S)$

P: $S \rightarrow aSbS$

$S \rightarrow bSaS$

$S \rightarrow \varepsilon$

és $L(G)=L$. (Ezt korábban bizonyítottuk.)

Példa

$L=\{ u \in \{a,b\}^* \mid \ell_a(u)=\ell_b(u) \}$, azaz ugyanannyi „a” és „b” van a szavakban.

$G = (\{S\}, \{a,b\}, P, S)$ $G' = (\{S',S\}, \{a,b\}, P', S')$

P: $S \rightarrow aSbS$ P': $S \rightarrow aSbS$

$S \rightarrow abS$

$S \rightarrow aSb$

$S \rightarrow ab$

$S \rightarrow bSaS$ $S \rightarrow bSaS$

$S \rightarrow baS$

$S \rightarrow bSa$

$S \rightarrow ba$

$S \rightarrow \varepsilon$ $S' \rightarrow \varepsilon$, $S' \rightarrow S$

és $L(G)=L(G')$.

Példa

u=abba szó levezetése:

$S \underset{G}{\Rightarrow} aSbS \underset{G}{\Rightarrow} aSbbSaS \underset{G}{\Rightarrow} abbSaS \underset{G}{\Rightarrow} abbaS \underset{G}{\Rightarrow} abba$

$S' \underset{G'}{\Rightarrow} S \underset{G'}{\Rightarrow} abS \underset{G'}{\Rightarrow} abba$

Példa

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

P: $S \rightarrow aAS$

$S \rightarrow AaB$

$S \rightarrow AB$

$A \rightarrow BB$

$B \rightarrow bA$

$B \rightarrow \varepsilon$

$H_1=\{B\}$

$H_2= H_1 \cup \{A\} =\{A,B\}$

$H_3= H_2 \cup \{S\} = \{A,B,S\} = N$

$H=H_3$

$G' = (\{S',S,A,B\}, \{a,b\}, P', S')$

P': $S \rightarrow aAS \mid aS \mid aA \mid a$

$S \rightarrow AaB \mid aB \mid Aa \mid a$

$S \rightarrow AB \mid B \mid A$

$A \rightarrow BB \mid B$

$B \rightarrow bA \mid b$

$S' \rightarrow \varepsilon$

$S' \rightarrow S$

Köszönöm a figyelmet!