Formális nyelvek és a fordítóprogramok alapjai

5. előadás

Előadó: Nagy Sára, mesteroktató
Algoritmusok és Alkalmazásaik Tanszék

Emlékeztető

Chomsky féle hierarchia:

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

Pontosabban valódi tartalmazás van

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

Megjegyzés: A következő tételek szükséges feltételeket fogalmaznak meg a 3-as típusú nyelvekre. Vannak nyelvek, amelyek bizonyíthatóan nem teljesítik a feltételeket, de 2-es típusú grammatikával generálhatók.

Szükséges feltétel 3-as típusú nyelvekre

Tétel: (Kis Bar-Hillel lemma)

Minden $L \in \mathcal{L}_3$ nyelvhez van olyan $n \geq 1$ nyelvfüggő konstans, hogy $\forall\ u \in L$, ahol $\ell(u) \geq n$ szó esetén van u-nak olyan $u = xyz$ felbontása, amelyre

- $\ell\,(xy) \leq n$,
- $y \neq \varepsilon$,
- $\forall i \geq 0$ egész esetén $xy^iz \in L$.

Kis Bar-Hillel lemma

Bizonyítás vázlata:

Ha $L \in \mathcal{L}_3$, akkor adható hozzá egy minimális véges determinisztikus automata.

Tegyük fel, hogy ennek az automatának *n* állapota van. Ha tekintünk egy legalább n hosszú szót, amit az automata elfogad, akkor a szó legalább n hosszú prefixszének olvasásakor legalább egy állapotot kétszer kellett érinteni.

Ez gráfos ábrázolásban azt jelenti, hogy a szó elemzése közben tettünk egy kört. Ezt a kört tetszőleges sokszor ismételhetjük. Így a kör során érintett részszót tetszőleges sokszor „bepumpálhatjuk” a szóba úgy, hogy az L nyelvhez tartozó szót kapunk.

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

Van olyan nyelv, ami nem 3-as típusú.

Megmutatjuk, hogy $\mathbf{L=\{a^k b^k \mid k \geq 0\}} \notin \mathcal{L}_3$.

Tegyük fel indirekt, hogy $\exists n \geq 1$ a lemma szerint.

Legyen $u= a^k b^k$, ahol $k>n$.

Ekkor léteznie kellene egy $u=xyz$ felbontásnak, ahol $\ell(xy) \leq n$ és $y \neq \varepsilon$.

De $k>n$ miatt y csak ‚a' betűket tartalmazhat.

Legyen $y=a^j$, ahol $j \geq 1$. Ekkor $a^{k+j}b^k$ szónak is benne kéne lenni a nyelvben, de az nem igaz.

Mivel L-re nem teljesülnek a lemma feltételei, így $L \notin \mathcal{L}_3$.

Nyelv maradéknyelvei

Definíció:

Legyen L egy T ábácé felett értelmezett nyelv.

Az L nyelv egy $p \in T^*$ szóra értelmezett maradéknyelve a következő:

$L_p := \{ u \in T^* \mid pu \in L \}$

Nyelv maradéknyelvei: $L_p:=\{ u \in T^* \mid pu \in L\}$

Példa:

Legyen R = **a(a|b)*b** és L az R kifejezésnek megfelelő nyelv. Ennek néhány maradéknyelve:

L_a=**(a|b)*b**

L_{aaaa}=**(a|b)*b**

L_{abb}=**(a|b)*b|ε**

$L_{ba}=\emptyset$

L_ε= **a(a|b)*b**

Myhill-Nerode tétel

Tétel:

$L \in \mathcal{L}_3$ akkor és csak akkor, ha az L-hez tartozó maradéknyelvek száma véges, azaz $|\{L_p | p \in T^*\}| < \infty$.

Megjegyzés: A szavakon egy osztályozást végzünk az adott nyelvtől függően.

Myhill-Nerode tétel

Bizonyítás vázlata:

1. Ha véges sok maradéknyelv van, akkor a maradéknyelvek segítségével megkonstruálható egy A véges determinisztikus automata, amire belátható, hogy L(A)=L. Az automata pedig átírható 3-as típusú grammatikává.

2. Ha L 3-as típusú nyelv, akkor adható hozzá 3-as típusú grammatika, ami átírható véges determinisztikus automatává. Belátható, hogy az egyes állapotokhoz rendelhető egy-egy maradéknyelv. Az így kapott maradék nyelvek között még lehetnek ekvivalensek, ha az állapotok is ekvivalensek. Mivel Q véges halmaz, így a maradéknyelvek száma is véges.

VDA előállítása maradéknyelvekből

Határozzuk meg a szavak hossza szerint haladva a lehetséges maradék nyelveket!

Legyen $p_1, p_2, \dots, p_n$ az egyes maradék nyelvek egy-egy reprezentáns szava!

Feleltessük meg az állapotokat a maradék nyelveknek, azaz

legyen $Q:=\{L_{p_i} \mid n \geq i \geq 1\}$ és

$\delta(L_p, a):=L_{pa}\ \forall a \in T;$

$q_0:=L_{\varepsilon};$

$F:=\{L_p \mid \varepsilon \in L_p\}.$

VDA előállítása maradéknyelvekből

Megjegyzés: Az így kapott automata minimális.

Példa: R = **a(a|b)*b** és L az R kifejezésnek megfelelő nyelv. Maradéknyelvei:

L_ε= **a(a|b)*b**, L_a=**(a|b)*b**, L_b=∅,

L_{aa}=**(a|b)*b** = L_a

L_{ab}=**(a|b)*b|ε**

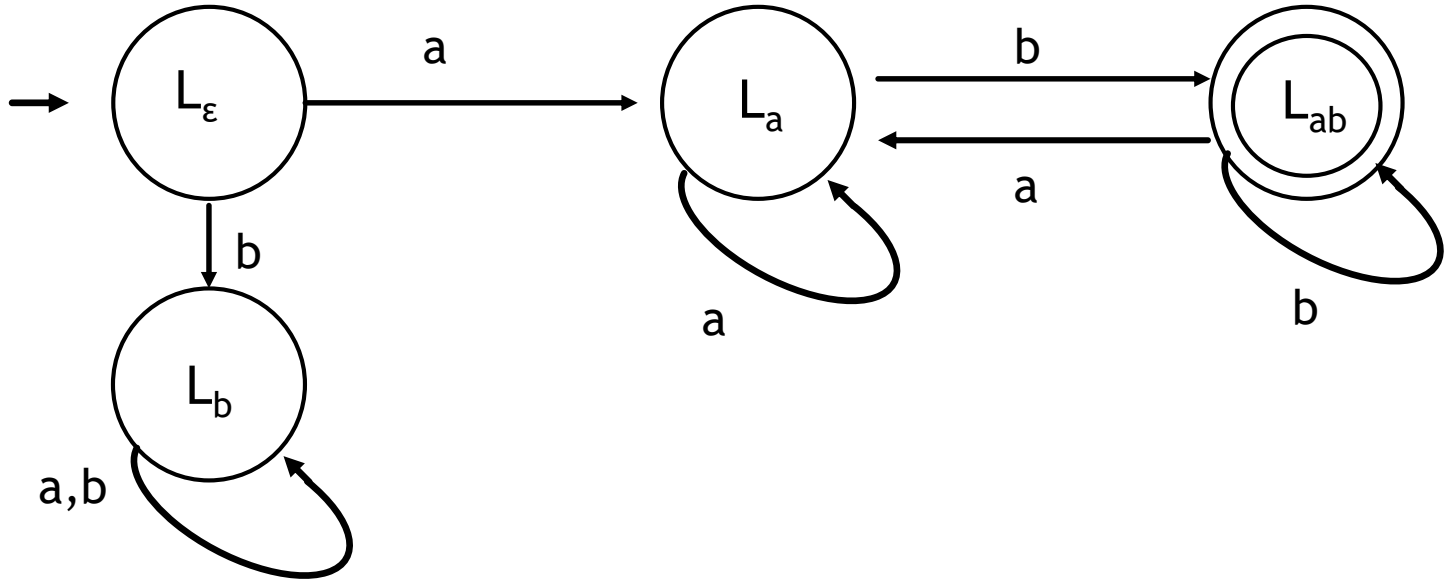
L_{ba}=∅ = L_{bb} = L_b

(L_{aaa}=**(a|b)*b** = L_a, L_{aab}=**(a|b)*b|ε** = L_{ab})

L_{aba}=**(a|b)*b** = L_a, L_{abb}=**(a|b)*b|ε** = L_{ab}

VDA előállítása maradéknyelvekből

A kapott automata:



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

Van olyan nyelv, ami nem 3-as típusú.

Megmutatjuk, hogy **$L=\{a^k b^k \mid k \geq 0\} \notin \mathcal{L}_3$.**

$L_\varepsilon = \{a^k b^k \mid k \geq 0\}$

$L_a = \{a^{k-1} b^k \mid k \geq 1\}$ és természetesen $L_\varepsilon \neq L_a$

általában $L_{a^i} \neq L_{a^j}$, ha $i \neq j$.

Mivel i és j tetszőleges természetes számok, így a maradék nyelvek száma végtelen, azaz végtelen állapotú automata kellene a nyelvhez.

Mivel L-re nem teljesül a Myhill-Nerode tétel, így $L \notin \mathcal{L}_3$.

Emlékeztető

G=(N,T,P,S) grammatika **2-es típusú**, ha szabályai

$$\mathbf{A} \rightarrow \mathbf{u} \text{ alakúak, ahol } A \in N,\ u \in (N \cup T)^*$$

Ezeket nevezzük ***környezetfüggetlen*** grammatikáknak.

Ilyenekkel írható le a programozási nyelvek ***szintaxisa***.

Emlékeztető

G=(N,T,P,S) négyest nevezzük grammatikának

Nyelvtan által generált nyelv:

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

Szóprobléma:

Adott G grammatika és adott $u \in T^*$ szó estén eldöntendő, hogy igaz-e, hogy $u \in L(G)$?

Programozási nyelvek szintaxisa

Gyakran Backus-Naur formában (BNF) adják meg.

Példa:

<kifejezés> ::= <tag> | <tag> + <kifejezés>

<tag> :: = <faktor> | <faktor> * <tag>

<faktor> ::= i | (<kifejezés>)

Példák kifejezésekre:

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

Backus-Naur forma (BNF)

A BNF lényegében egy 2-es típusú grammatika.

- Szabályok véges halmaza, ahol a szabályok bal- és jobb oldalát a

  ::= jel választja el.

- A bal oldalon egy fogalom (egy nemterminális) szerepel.

  < *fogalom* **>** (A < > jelek közé tetszőleges szöveg írható.)

- A jobb oldalon a bal oldal kifejtése szerepel. Ha több alternatíva is van, akkor az alternatívákat | jel választja el.
- A terminálisokat nem kell semmilyen jel közé tenni.
- Egy alternatíva terminálisok és nem terminálisok sorozata.

Szóprobléma

Példa:

<kifejezés> ::= <tag> | <tag> + <kifejezés>

<tag> :: = <faktor> | <faktor> * <tag>

<faktor> ::= i | (<kifejezés>)

Szintaktikusan helyes-e az i+i*i kifejezés?

Ha levezethető a <kifejezés> fogalmából, akkor igen.

<kifejezés> ⇒ <tag> + <kifejezés>⇒ <tag> + <tag> ⇒ <faktor> + <tag>

⇒ <faktor> + <faktor> * <tag>⇒ **i** + <faktor> * <tag> ⇒ **i** + **i** * <tag>

⇒ **i** + **i** * <faktor> ⇒ **i** + **i** * **i**

Levezetési fa (szintaxisfa)

Definició:

Legyen G = (N,T,P,S) tetszőleges 2-es típusú grammatika.

A **t** nemüres fát G feletti levezetési (szintaxis) fának nevezzük, ha

- pontjai $T \cup N \cup \{\varepsilon\}$ elemeivel vannak címkézve;
- belső pontjai N elemeivel vannak címkézve;
- ha egy belső pont címkéje A, a közvetlen leszármazottjainak címkéi pedig balról jobbra olvasva $X_1, X_2, \ldots, X_k$, akkor $A \rightarrow X_1X_2\ldots X_k \in P$.
- az ε -nal címkézett pontoknak nincs testvére.

Levezetési fa (szintaxisfa)

Tétel:

Ha adott G grammatika esetén **u∈ L(G)** akkor és csak akkor, ha u-hoz **megadható egy szintaxisfa.**

Megjegyzés: Az u-hoz tartozó szintaxisfa gyökere S és a leveleit balról jobbra összeolvasva az u szót kapjuk.

Példa

G = ({S,A,B}, {i,+,*,(,)}, P, S) egy grammatika, ahol a szabályok a következők:

P : $S \rightarrow A \mid A + S$

$A \rightarrow B \mid B * A$

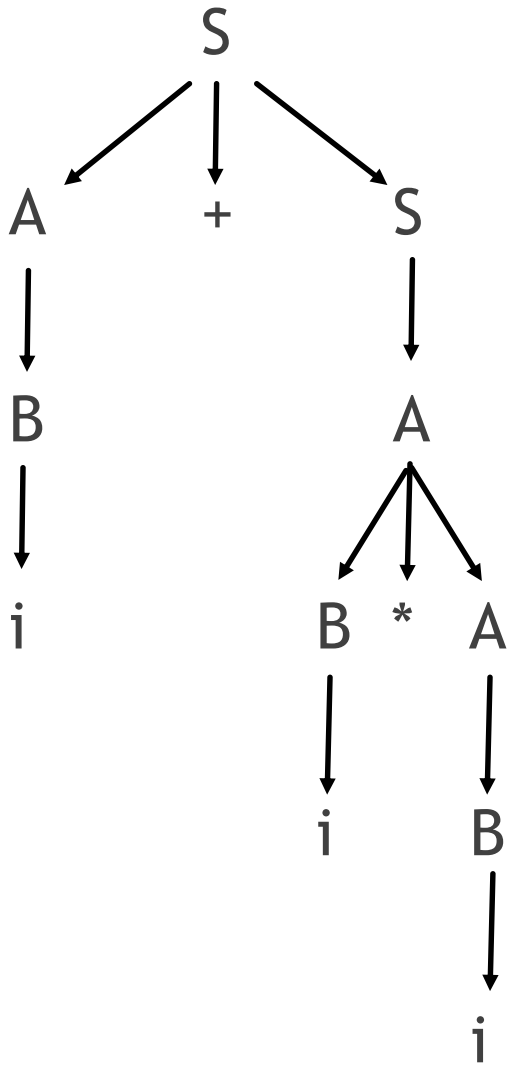
$B \rightarrow i \mid (S)$

u = i + i * i $u \in L(G)$, ha megadható hozzá levezetési fa.

Megjegyzés: A fenti G grammatika ekvivalens a korábban megadott kifejezés leírásával.

Szintaxisfa

u = i+i*i



$S \rightarrow A \mid A + S$

$A \rightarrow B \mid B * A$

$B \rightarrow i \mid (S)$

Szintaxisfa

Állítás:

Minden szintaxisfához megadható egy levezetés és fordítva.

Legbal levezetés: A legbal levezetés olyan levezetés, hogy ha a levezetés folyamán a mondatforma i. betűjén helyettesítés történik, akkor a korábbi pozíciókat (1., … , i-1.) a levezetés a további lépesekben már nem érinti, azok változatlanul maradnak.

Szintaxisfa

Egy G 2-es típusú grammatika **egyértelmű**, ha minden u $\in$ L(G) szóhoz egyetlen szintaxisfa tartozik.

Ellenpélda:

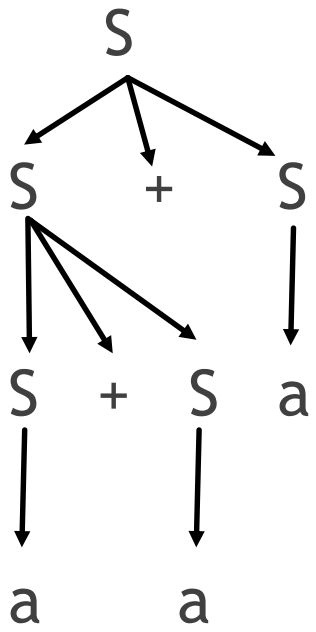
$S \rightarrow a \mid S + S$ u = a + a + a

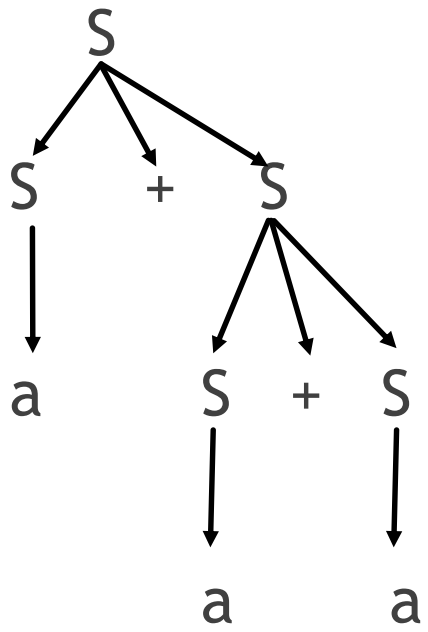
Szintaxisfa

Ellenpélda:

$S \rightarrow a \mid S + S$

$u = a + a + a$





Emlékeztető

Chomsky féle hierarchia:

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

G=(N,T,P,S) grammatika 2-es típusú, ha szabályai

$\mathbf{A} \rightarrow \mathbf{u}$ alakúak, ahol $A \in N$, $u \in (N \cup T)^*$

Ezeket nevezzük ***környezetfüggetlen*** grammatikáknak. Ilyenekkel írható le a programozási nyelvek szintaxisa.

Emlékeztető

Tétel: (ε-mentesítés)

Minden G=(N,T, P, S) környezetfüggetlen grammatikához meg lehet konstruálni egy olyan G'= (N', T, P', S') környezetfüggetlen grammatikát, amelyre igaz, hogy

- egyetlen szabály jobboldala sem az üresszó (ε);
- kivéve, ha az üresszó benne van a G által generált nyelvben, mert akkor
  S' → ε megengedett, de ekkor S' nem fordul elő egyetlen szabály jobboldalán sem

2-típusú grammatikák normál formája

Definíció:

Egy G=(N,T,P,S) környezetfüggetlen grammatikát **Chomsky normálformájúnak** mondunk, ha szabályai

- **A** $\rightarrow$ **a**, ahol A $\in$ N és a $\in$ T vagy
- **A** $\rightarrow$ **BC** alakúak, ahol A,B,C $\in$ N.
- S $\rightarrow$ ε, de ekkor S nem fordul elő egyetlen szabály jobboldalán sem.

Chomsky normál forma

Tétel:

Minden **környezetfüggetlen** grammatikához megkonstruálható egy vele ekvivalens **Chomsky normálformájú** grammatika.

Megjegyzés:

- A 2-es típusú grammatikák Chomsky normálformára hozásának algoritmusa nem a tananyag része.
- Chomsky normálformájú grammatikákhoz megadható olyan elemző program, amely $O(n^3)$ időben eldönti a szóproblémát (Cocke-Younger-Kasami algoritmus).
- Bizonyos állítások bizonyítását elég elvégezni a normálformájú grammatikákra.

2-es típusú grammatikák redukálása

A grammatikák transzformálása közben keletkezhetnek olyan szabályok, amelyek egyetlen szó levezetésében sem használhatóak.

A grammatikában lehetnek olyan nemterminálisok, amelyekből

1. nem lehet csupa nem terminálisból álló sorozatot előállítani; (zsákutcák)
2. nem érhetők el a kezdőszimbólumból.

Hasznos/ nem hasznos nemterminálisok

Definició: ***Aktív*** nemterminálisok halmaza egy adott G=(N,T,P,S) környezetfüggetlen grammatika esetén:

$$A := \{ X \in N \mid X \underset{G}{\Rightarrow}^* u \text{ és } u \in T^* \}.$$

Inaktív (zsákutca) nemterminálisok: $N \setminus A$.

Definició: ***Elérhető*** nemterminálisok halmaza:

$$R := \{ X \in N \mid S \underset{G}{\Rightarrow}^* uXw \text{ és } u,w \in (T \cup N)^* \}.$$

Nemelérhető nemterminálisok: $N \setminus R$.

Hasznos/ nem hasznos nemterminálisok

Definició: Egy nemterminálist **hasznos**nak mondunk, ha aktív és elérhető.

Definició: Egy környezetfüggetlen grammatika **redukált**, ha minden nemterminálisa hasznos, azaz a grammatika zsákutcamentes és összefüggő.

Tétel: Minden 2-es típusú grammatikához megkonstruálható egy vele ekvivalens redukált grammatika.

Hasznos/ nem hasznos nemterminálisok

Tétel: Minden 2-es típusú grammatikához megkonstruálható egy vele ekvivalens redukált grammatika.

Bizonyítás:

1. Zsákutcák meghatározása és minden olyan szabály elhagyása, amiben inaktív nemterminálisok szerepelnek.
2. Az S-ből nem elérhető nemterminálisokhoz tartozó szabályok elhagyása, azaz a grammatika összefüggővé tétele.

Redukált grammatika meghatározása

Bizonyítás folytatása:

$A_1 = \{ X \in N \mid X \rightarrow u \in P \text{ és } u \in T^* \}$

$A_{i+1} = A_i \cup \{ X \in N \mid X \rightarrow w \in P \text{ és } w \in (A_i \cup T)^* \}$, ahol $i \geq 1$.

$\exists$ k úgy, hogy $\forall$ m>k esetén $A_k = A_m$.

Ekkor A_k a grammatika aktív nemterminálisainak halmaza.

Az $N \setminus A_k$ inaktív (zsákutca) nemterminálisokat elhagyjuk a grammatikából és minden olyan szabályt is, amiben szerepelnek.

Redukált grammatika meghatározása

Bizonyítás folytatása:

$R_1 = \{ S \}$

$R_{i+1} = R_i \cup \{ Y \in N \mid X \rightarrow uYw \in P, X \in R_i , u,w \in (N \cup T)^* \}$, ahol $i \geq 1$.

$\exists$ k úgy, hogy $\forall$ m>k esetén $R_k = R_m$.

Ekkor R_k a grammatika elérhető nemterminálisainak halmaza.

Az $N \setminus R_k$ nem elérhető nemterminálisokat elhagyjuk a grammatikából és minden olyan szabályt is, amiben szerepelnek.

Az így kapott grammatika redukált.

Példa: grammatika redukálása

S →AB | aaC

A → AS | aDa

B → aaS | bAD

C → aAD | ab

D → bA

Aktív nemterminálisok meghatározása:

$A_1 = \{ C \}$, $A_2 = \{ C \} \cup \{ S \}$, $A_3 = \{ C, S \} \cup \{ B \}$, $A_4 = A_3$

Inaktív nem terminálisok, amelyek elhagyhatók a grammatikából:

A, D

Példa: grammatika redukálása

Inaktív nem terminálisok, amelyek elhagyhatók a grammatikából:

A, D

S →~~AB~~ | aaC

~~A → AS | aDa~~

B → aaS | ~~bAD~~

C → ~~aAD~~ | ab

~~D → bA~~

Példa: grammatika redukálása

Csak aktív nem terminálisokat tartalmazó ekvivalens grammatika:

$S \rightarrow aaC$

$B \rightarrow aaS$

$C \rightarrow ab$

Elérhető nem terminálisok:

$R_1 = \{ S \}$, $R_2 = \{ S \} \cup \{ C \}$, $R_4 = R_3$

Nem elérhető: B

Redukált grammatika:

$S \rightarrow aaC$

$C \rightarrow ab$

Köszönöm a figyelmet!