A számításelmélet alapjai I.

10. előadás

előadó: Tichler Krisztián
ktichler@inf.elte.hu

Levezetési fa

A környezetfüggetlen grammatikák levezetéseit gyökeres irányított fákkal is leírhatjuk.

Definíció

$G = \langle N, T, P, S \rangle$ környezetfüggetlen grammatika feletti **levezetési fának** nevezünk egy fát, ha teljesülnek a következők:

- A levezetési fa gyökerének címkéje S.
- Minden további csúcs címkéje $N \cup T \cup \{\varepsilon\}$ valamely eleme.
- Ha egy csúcs címkéje X és gyerekeinek címkéi balról jobbra olvasva rendre $X_1, \ldots, X_m$, $m \geq 1$, akkor $X \to X_1 \cdots X_m \in P$.
- Minden levél címkéje a $T \cup \{\varepsilon\}$ halmaz valamely eleme, az ε-nal címkézett csúcsoknak nincs testvére.

Definíció

Egy levezetési fa **határa**: levélcímkéi balról jobbra összeolvasva (konkatenálva).

Levezetési fa

Nyilván minden terminális szót generáló levezetéshez hozzá tudunk rendelni egy levezetési fát, melynek határa éppen az adott szó.

Igaz-e fordítva? Nem, maga a levezetési fa általában nem adja meg a levezetés során alkalmazott szabályok sorrendjét.

Megjegyzés: A fára tekinthetünk úgy, mint egy aciklikus irányított gráfra. Ilyenkor a csúcshalmaz elemeinek vannak topolgikus sorrendjei, azaz olyan sorrendjei, ahol nincs későbbi csúcsból korábbiba él (pl. a szélességi bejárás ilyen). Éppen ezek a topológikus sorrendek felelnek meg a levezetési fához tartozó levezetéseknek.

Definíció

Egy 2-típusú grammatika két levezetését **lényegében azonosnak** nevezzük, ha ugyanazt a levezetési fát határozzák meg.

Levezetési fa

A lényegében azonos (azaz ugyanahhoz a levezési fához tartozó) levezetések közül kitűntetünk egyet.

Definíció

Egy 2-típusú grammatika egy levezetése **baloldali levezetés**, ha minden levezetési lépésben az aktuális mondatforma balról legelső nemterminálisra történt a grammatika egy szabályának alkalmazása.

Így minden levezetési fához pontosan egy baloldali levezetés tartozik.

Definíció

Egy G környezetfüggetlen grammatika **egyértelmű**, ha minden $L(G)$-beli szónak egyetlen baloldali levezetése van.

Ami ezzel ekvivalens:

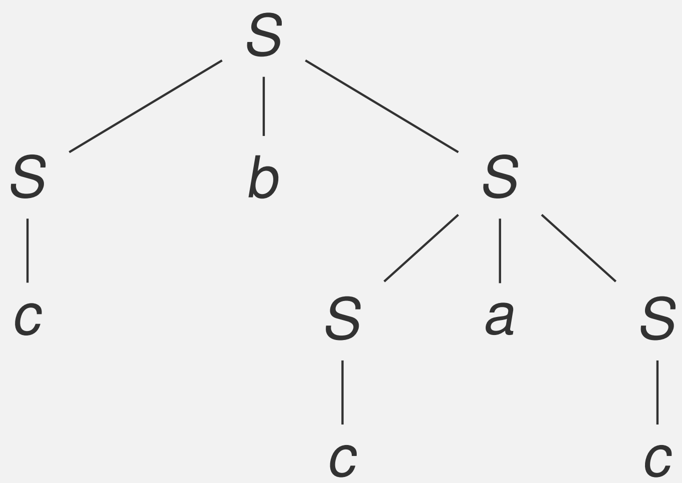
minden $u \in L(G)$-nek egyetlen levezetési fája van.

Levezetési fa

Példa: Legyen $G = \langle\{S\}, \{a, b, c\}, \{S \to SaS \,|\, SbS \,|c\}, S\rangle$.

$L(G)$ olyan páratlan hosszú szavakból áll, melynek páratlan pozícióin c, páros pozícióin a vagy b áll.

Tekintsük a következő levezetési fát:



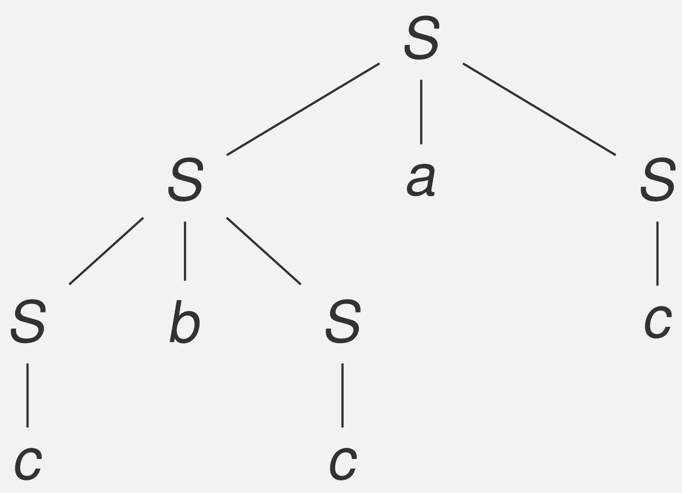
$S \Rightarrow SbS \Rightarrow cbS \Rightarrow cbSaS \Rightarrow cbcaS \Rightarrow cbcac$ és

$S \Rightarrow SbS \Rightarrow SbSaS \Rightarrow SbcaS \Rightarrow cbcaS \Rightarrow cbcac$

ugyanannak a szónak 2 levezetése, de ugyanaz a levezetési fa tartozik hozzájuk, így a két levezetés lényegében azonos. Az első levezetés az ezen fához tartozó baloldali levezetés.

Levezetési fa

Tekintsük most ezt a levezetési fát:



$$S \Rightarrow SaS \Rightarrow SbSaS \Rightarrow cbSaS \Rightarrow cbcaS \Rightarrow cbcac$$

ennek a szónak egy harmadik levezetése. Ehhez már más levezetési fa tartozik (a határ mindkét esetben *cbcac*). Ez a levezetés ebben a fában szintén baloldali levezetés.

A *G* grammatika nem egyértelmű, mert nem minden szónak van egyetlen baloldali levezetése.

Nagy Bar-Hillel lemma

Tétel (Nagy 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 felírható $uxwyz$ alakban, ahol $|xwy| \leq q, xy \neq \varepsilon$, továbbá, ekkor minden ux^iwy^iz, $i \geq 0$ alakú szó is benne van az L nyelvben $(u, x, w, y, z \in T^*)$.

Megjegyzés A tétel gyakran használt másik elnevezése: "pumpálási lemma".

Nagy Bar-Hillel lemma

Bizonyítás:

Legyen L környezetfüggetlen nyelv, amelyet a $G = \langle N, T, P, S \rangle$ Chomsky-normálformájú grammatika generál.

Tegyük fel, hogy G nemterminálisainak száma n és legyen $p = 2^{n-1}$, valamint legyen $q = 2^n$.

Mivel a G Chomsky normálformájú, ezért a levezetési fa bináris fa.

Minden ág utolsó lépése egy „$A \to a$" ($a \in T, A \in N$) típusú átírás.

Nevezzük **redukált levezetésnek** az ezen szabályokat a végén már nem alkalmazó levezetéseket, így egy **redukált levezetési fát** kapunk, amelyik szintén bináris.

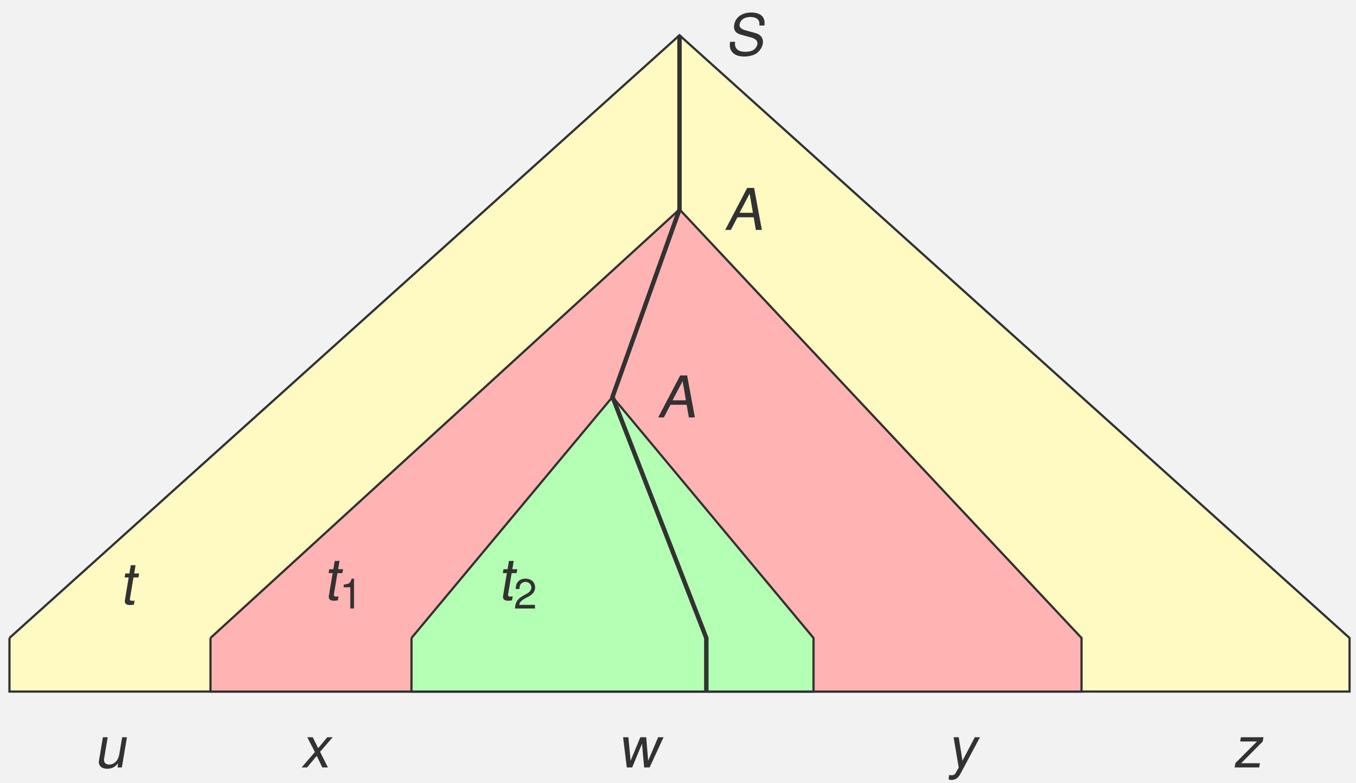
Nagy Bar-Hillel lemma

Ha a levezetési fában a leghosszabb út hossza k, akkor ez a redukált levezetési fában $k-1$ hosszúságnak felel meg. Tehát a redukált levezetési fa leveleinek száma legfeljebb 2^{k-1}, és ez a becslés az eredeti levezetési fára is igaz.

Tehát minden $|\beta| > p$ szó minden levezetési fájában a leghosszabb út hossza nagyobb, mint n. Ez az n-nél hosszabb út legalább $n+2$ csúcsot tartalmaz, a megfelelő redukált levezetés legalább $n+1$-et.

A skatulya-elv szerint így lennie kell olyan nemterminálisnak, amely ezen az úton legalább kétszer előfordul. Ha több ilyen van, akkor válasszuk az úton alulról az első ismétlődő párt. Legyen ez a nemterminális A.

Nagy Bar-Hillel lemma



Ekkor léteznek $u, x, w, y, z \in T^*$ amelyekre $S \Rightarrow^* uAz$, $A \Rightarrow^* xAy$, $A \Rightarrow^* w$, továbbá $xy \neq \varepsilon$, mivel G ε- és láncszabály mentes.

Ekkor az első levezetést 1-szer, a másodikat i-szer ($i \geq 0$), a harmadikat újra egyszer alkalmazva $S \Rightarrow^* ux^iwy^iz$ adódik.
($i = 1$-re β-t kapjuk meg.)

Nagy Bar-Hillel lemma

A leghosszabb útnak az A utolsó előtti előfordulása gyökerű t_1 részfába eső része nyilván t_1-ben is leghosszabb út.

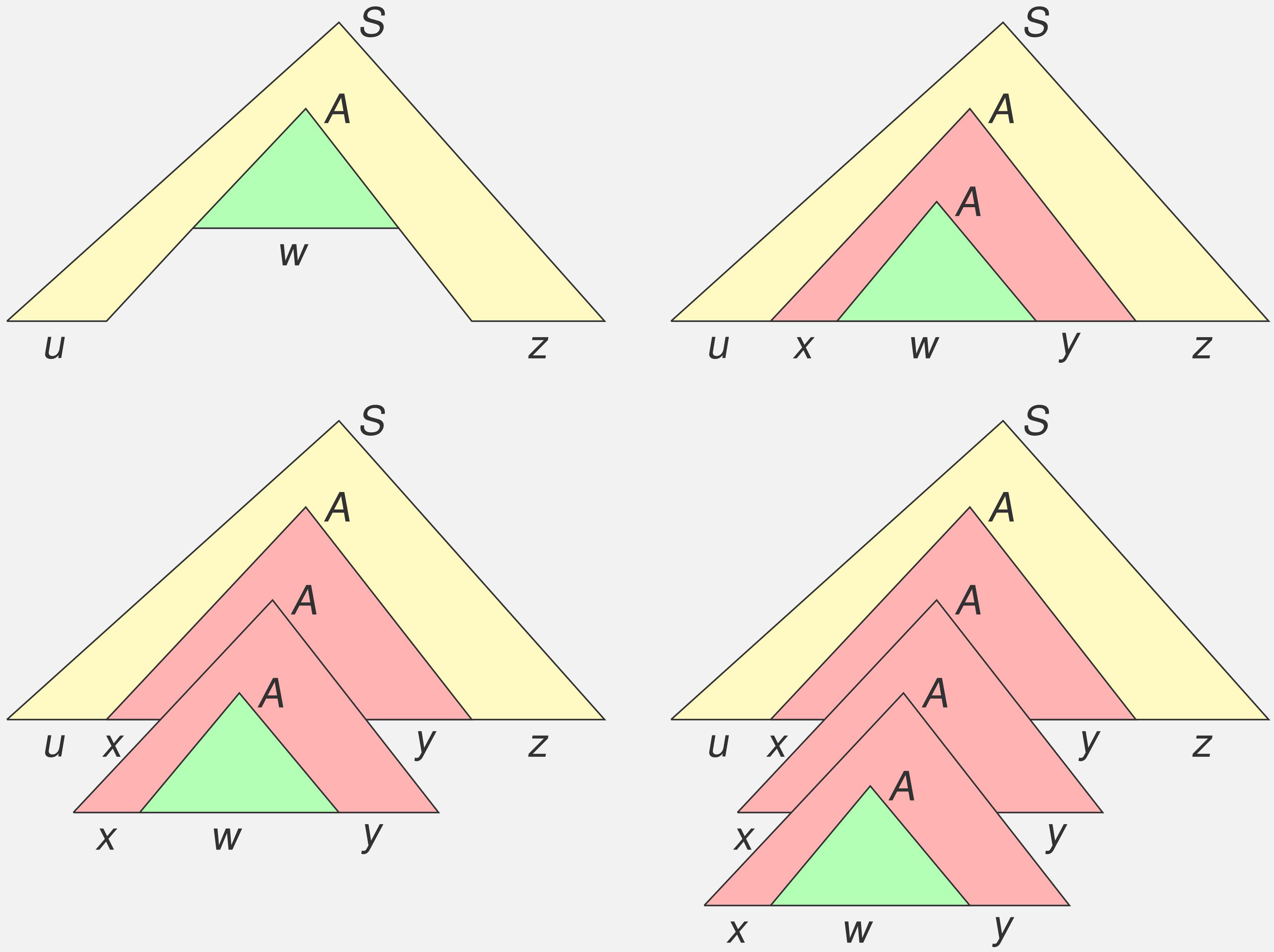
Továbbá a levezetés ezen az úton A-n kívül minden más nemterminálist legfeljebb egyszer tartalmazhat, mivel A alulról az első ismétlődő nemterminális.

Így a leghosszabb út t_1-be eső része legfeljebb $n+1$ élt, a redukált része legfeljebb n élt tartalmaz.

Tehát t_1 redukált részében minden út legfeljebb n hosszú így t_1 redukált határa legfeljebb 2^n hosszú. Mivel ezután már csak $A \to a$ alakú szabályokat alkalmazunk, ez a teljes t_1 határára is igaz, tehát az $|xwy| \leq q$ feltétel is teljesül. $\square$

Megjegyzés: p és q értéke csak a nyelvtől függ, magától a pumpált szótól nem.

Pumpálás



Nagy Bar-Hillel lemma

1. Példa: $L = \{a^n b^n c^n \mid n \in \mathbb{N}\} \notin \mathcal{L}_2$, azaz létezik nem környezetfüggetlen nyelv.

Megoldás: Legyenek p, q a Nagy Bar-Hillel lemma szerint létező megfelelő konstansok. Legyen továbbá $w = a^k b^k c^k$, valamely $k > \max\{p, q\}$-ra. Ekkor nyilván $w \in L$ és $|w| > p$.

Tekintsük w-nek egy olyan $w = uxvyz$ felbontását, melyre $u, x, v, y, z \in \{a, b, c\}^*$, $|xvy| \leq q$, $|xy| > 0$.

Tekintsük most az $uvz = ux^0vy^0z$ szót. Mivel xy hossza legalább 1 és az $\{a, b, c\}$-ből a k nagyságára vonatkozó feltétel miatt legalább egy betűt nem tartalmaz, így uvz, amely valamelyik betűből pontosan k darabot, egy másikból pedig kevesebb, mint k-t tartalmaz nem lehet L-nek eleme.

Tehát a nyelv nem teljesíti a Nagy Bar-Hillel lemma feltételeit, vagyis L nem környezetfüggetlen.

Nagy Bar-Hillel lemma

2. Példa: $L = \{a^{n^2} \mid n \in \mathbb{N}\} \notin \mathcal{L}_2$

Megoldás: Indirekt, tegyük fel, hogy $L \in \mathcal{L}_2$. Ekkor a Nagy Bar-Hillel lemma szerint léteznek a nyelvfüggő p és q konstansok. Legyen $M = \max\{p, q\}$. Tekintsük az $u = a^{M^2}$ szót.

Mivel $|u| > M \geq p$, ezért a Nagy Bar-Hillel lemma szerint létezik az u-nak $u = xyzvw$ felbontása, ahol $K := |yv| > 0$, $|yzv| \leq q \leq M$.

$xy^izv^iw = a^{M^2+(i-1)K}$ $(i \in \mathbb{N})$. A kitevők egy K differenciájú számtani sorozatot alkotnak. Mivel egy nem nulla differenciájú számtani sorozatban biztosan van nem négyzetszám (az egymást követő négyzetszámok távolsága végtelenhez tartva nő), ezért a Nagy Bar-Hillel lemma állítása L-re nem teljesül, tehát $L \notin \mathcal{L}_2$.

Nem reguláris műveletek zártsági kérdései

Következmény $\mathcal{L}_2$ nem zárt a metszetre, komplementerre, különbségre, szimmetrikus differenciára.

Bizonyítás:

$\{a^n b^n c^n \mid n \in \mathbb{N}\} = \{a^k b^n c^n \mid k, n \in \mathbb{N}\} \cap \{a^n b^n c^k \mid k, n \in \mathbb{N}\}$.

Előbb láttuk, hogy a baloldali nyelv nem $\mathcal{L}_2$-beli, a jobboldali kettő azonban könnyen láthatóan $\mathcal{L}_2$-beli. (Pl. $S \to aS \mid A,\ A \to bAc \mid \varepsilon$)

Tegyük fel, hogy $\mathcal{L}_2$ zárt a komplementerre. A De Morgan-azonosság alapján $L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$. Mivel L_2 az unióra zárt, ezért a komplementerre való zártságból következne a metszetre való zártsága, amiről az imént kimutattuk, hogy nem teljesül.

Ha a különbségre, illetve a szimmetrikus differenciára zárt lenne, akkor tetszőleges L mellett tartalmazná $T^* \setminus L$-et, illetve $T^* \triangle L$-et (ahol T L ábécéje), de ezek mindegyike $\overline{L}$, így zárt lenne a komplementerre is. $\square$

Reguláris és környezetfüggetlen nyelv metszete

Két környezetfüggetlen nyelv metszete nem feltétlen környezetfüggetlen. Vajon egy környezetfüggetlen és egy reguláris nyelv metszete környezetfüggetlen-e?

Tétel: Legyen $L \in \mathcal{L}_2$ és $L' \in \mathcal{L}_3$. Ekkor $L \cap L' \in \mathcal{L}_2$.

Bizonyítás:

Mivel $L \in \mathcal{L}_2$, ezért létezik egy $A = \langle Z, Q, T, \delta, z_0, q_0, F \rangle$ veremautomata, amelyre $L(A) = L$. Mivel $L' \in \mathcal{L}_3$, ezért létezik egy $A' = \langle Q', T, \delta', q_0', F' \rangle$ véges determinisztikus automata, amelyre $L(A') = L'$.

Konstruálunk egy $A'' = \langle Z, Q'', T, \delta'', z_0, q_0'', F'' \rangle$ veremautomatát melyre $L(A'') = L \cap L'$.

Reguláris és környezetfüggetlen nyelv metszete

Valójában A'' nem lesz más, mint a két gép direkt szorzata:

$$Q'' := \{ [q, q'] \mid q \in Q \ \wedge \ q' \in Q' \}. \quad q'' := \ [q_0, q_0'].$$

$$(z', [r, r']) :\in \delta''(z, [q, q'], a) \iff (z'r) \in \delta(z, q, a) \ \wedge \delta'(q', a) = r'.$$
$$(q, r \in Q, q', r' \in Q', z \in Z, z' \in Z^*, a \in T)$$

$$(z', [r, r']) :\in \delta''(z, [q, q'], \varepsilon) \iff (z'r) \in \delta(z, q, \varepsilon) \ \wedge r' = q.'$$
$$(q, r \in Q, q', r' \in Q', z \in Z, z' \in Z^*)$$

$$F'' := \{ [q, q'] \mid q \in F \ \wedge \ q' \in F' \}.$$

A'' tehát úgy működik, mint A, csak közben az input minden betűjének feldolgozására lép az A' géppel is. F'' definíciója szerint A'' azon L-beli szavakat fogadja el, amelyekre A' is F'-beli állapotba jut. $\square$

Algoritmikus eldönthetetlen problémák

A 2-es típusú grammatikák esetében **vannak algoritmikusan eldönthetetlen problémák**, azaz olyan eldöntési („igen”/„nem” kimenetű) problémák, amelyekre nem létezik olyan mindig termináló algoritmus, amelyik éppen a probléma „igen”-példányaira ad igen választ. Ilyenek például (bizonyítás nélkül):

1. **Input:** G_1, G_2 2-es típusú grammatikák
   Output: $L(G_1) \stackrel{?}{=} L(G_2)$
2. **Input:** G_1, G_2 2-es típusú grammatikák
   Output: $L(G_1) \stackrel{?}{\subseteq} L(G_2)$
3. **Input:** G_1, G_2 2-es típusú grammatikák
   Output: $L(G_1) \cap L(G_2) \stackrel{?}{=} \emptyset$
4. **Input:** $G = \langle N, T, P, S \rangle$ 2-es típusú grammatika
   Output: $L(G) \stackrel{?}{=} T^*$
5. **Input:** G 2-es típusú grammatika
   Output: G egyértelmű grammatika-e.

Algoritmikusan eldönthető problémák

Szóprobléma:

Input: $G = \langle N, T, P, S \rangle$ 2-es típusú grammatika; $u \in T^*$.

Output: $u \overset{?}{\in} L(G)$.

Az előző előadáson láttuk, hogy a szóprobléma a CYK algoritmussal hatékonyan (polinomiális műveletigénnyel) eldönthető.

Ebből következően az alábbi algoritmikus problémák szintén (polinomiális műveletigénnyel) eldönthetők

Input: G 2-es típusú grammatika és egy L véges nyelv.

Output: $L \overset{?}{\subseteq} L(G)$.

Input: G 2-es típusú grammatika és egy L véges nyelv.

Output: $L \cap L(G) \overset{?}{=} \emptyset$.

Algoritmikusan eldönthető problémák

Tétel: Eldönthető, hogy egy környezetfüggetlen grammatika végtelen nyelvet generál-e vagy sem.

Bizonyítás: Legyenek p és q a Nagy Bar-Hillel lemma alapján létező konstansok.

Állítás: $L(G)$ akkor és csak akkor végtelen, ha létezik $\beta \in L(G)$, melyre $p < |\beta| \leq p + q$.

$(\Rightarrow)$ Ha van ilyen szó, akkor a Nagy Bar-Hillel lemma alapján $L(G)$ végtelen nyelv.

$(\Leftarrow)$ Ha $L(G)$ végtelen, akkor van p-nél hosszabb α szava. Ekkor a Nagy Bar-Hillel lemma alapján léteznek olyan u, x, v, y, z szavak, amelyekre $\alpha = uxvyz$, ahol $|xvy| \leq q$, $xy \neq \varepsilon$ és $\alpha' = uvz \in L(G)$ ($i = 0$ iteráció).

Algoritmikusan eldönthető problémák

Mivel $1 \leq |xy| = |\alpha| - |\alpha'| \leq |xvy| \leq q$, ezért ha α-t α'-re cseréljük, akkor legalább 1-gyel, de legfeljebb q-val csökken a szó hossza.

Ezt iterálva kapni fogunk egy olyan $\beta \in L(G)$ szót, amelyre $p < |\beta| \leq p + q$.

A Chomsky normálformára hozás algoritmikus, így a Nagy Bar-Hillel lemma p és q konstansai algoritmikusan kiszámíthatóak.

Egy Chomsky normálformújú grammatika p-nél hosszabb, de legfeljebb $p + q$ hosszú határral rendelkező levezetési fáiból véges sok van (binárisak, legfeljebb $2(p + q) - 1$ csúcsuk van és minden egyes csúcs címkéje legfeljebb $|T| + |N|$ féle lehet).

Ezek a fák tehát algoritmikusan, sorra előállíthatóak és ellenőrizhetőek. $\square$

Algoritmikusan eldönthető problémák

Tétel: Eldönthető, hogy egy környezetfüggetlen grammatika az üres nyelvet generálja-e vagy sem.

Bizonyítás: Legyenek p és q a Nagy Bar-Hillel lemma alapján létező konstansok.

Állítás: $L(G)$ akkor és csak akkor nem üres, ha létezik legfeljebb p hosszú szava.

Ehhez elég azt kell látni, hogy ha $L(G)$-nak van p-nél hosszabb szava, akkor legfeljebb p hosszú szava is van. Az előző bizonyításhoz hasonlóan amíg a szó hossza p-nél nagyobb tudunk adni egy nála rövidebbet (vegyük mindig a 0. iterációt), iteráljuk ezt addig, amíg egy legfeljebb p hosszú szót nem kapunk.

A Chomsky normálformára hozás algoritmikus, így a Nagy Bar-Hillel lemma p és q konstansai algoritmikusan kiszámíthatóak.

Egy Chomsky normálformájú grammatika legfeljebb p hosszú határral rendelkező levezetési fáiból véges sok van, ezek sorra algoritmikusan előállíthatóak és ellenőrizhetőek. $\square$