

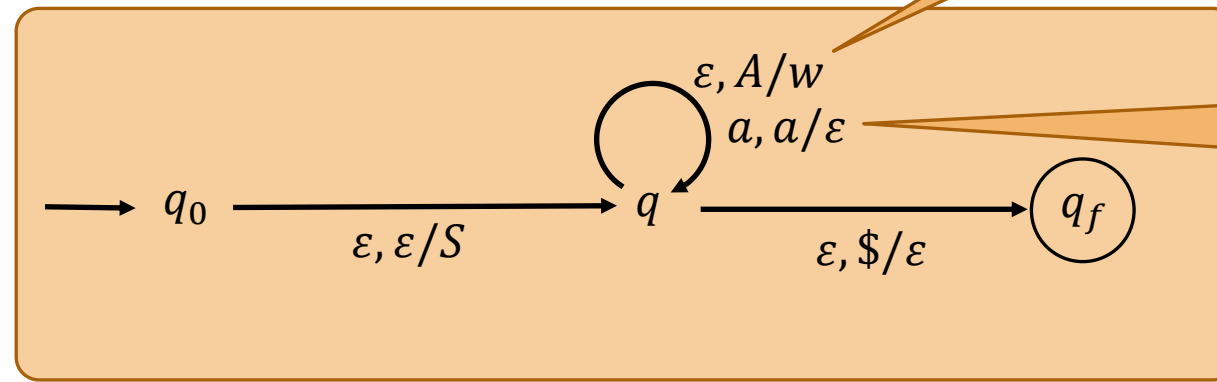
Veremautomaták

A környezetfüggetlen nyelvek osztálya megegyezik a veremautomatával felismerhető nyelvek osztályával

Bizonyítás (csak az egyik irány, vázlat)

Legyen $G = (N, T, R, S)$ egy környezetfüggetlen nyelvtan

A következő veremautomata $L(G)$ -t ismeri fel
(bemenő jelek: T , veremjelek: $N \cup T \cup \{\$, \}$, verem alja: $\$$):



Ezt az átmenetet akkor vesszük fel ha van $A \rightarrow w$ szabály R -ben

Ezt az átmenetet meg minden a terminálusra felvesszük

CF nyelvtanhoz ekvivalens veremautomata konstrukciója – Példa

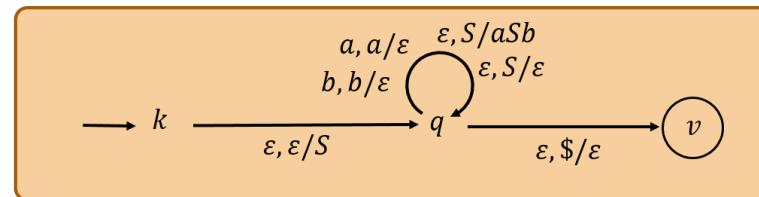
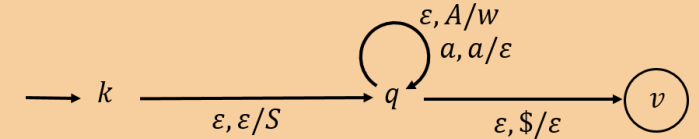
Tekintsük az alábbi $\{a^n b^n \mid n \geq 0\}$ nyelvet generáló környezetfüggetlen nyelvtant:

$G = (N, T, R, S)$, ahol

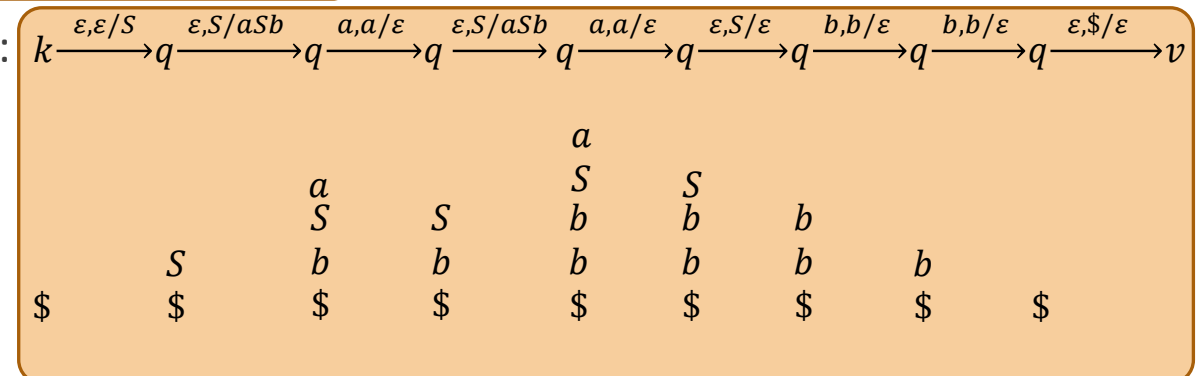
- $N = \{S\}, T = \{a, b\}, R = \{S \rightarrow aSb, S \rightarrow \varepsilon\}$

Akkor az általános konstrukció alapján a következő $L(G)$ nyelvet generáló (ekvivalens) veremautomatát kapjuk:

Az általános konstrukció



Ennek egy k -ból induló sikeres futása az $aabb$ szón:



Determinisztikus veremautomata

Determinisztikus az $M = (Q, \Sigma, \Gamma, \delta, q_0, F)$ veremautomata, ha az átmenetekre az alábbiak teljesülnek:

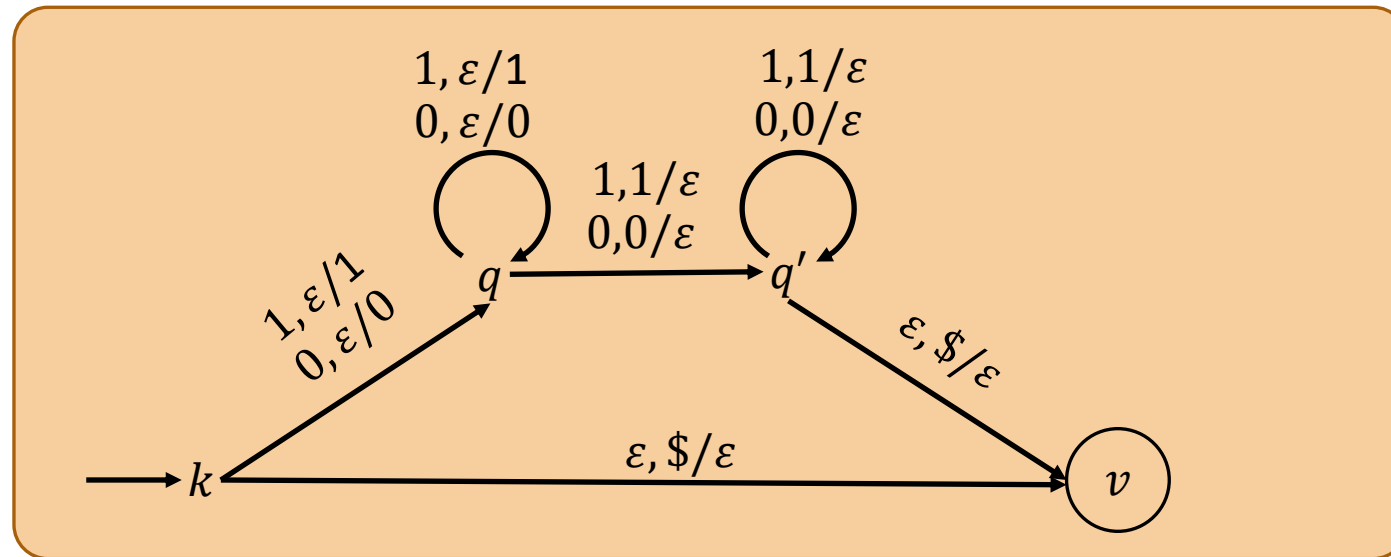
- Ha M -nek van egy $q \xrightarrow{a,b/\gamma} p$ átmenete, ahol a, b, γ tetszőlegesek, akkor **nincs** $q \xrightarrow{a,b/\gamma'} p'$ átmenete semmilyen $\gamma' \neq \gamma$ vagy $p' \neq p$ esetén
- Ha M -nek van egy $q \xrightarrow{\varepsilon,b/\gamma} p$ átmenete, ahol b, γ tetszőlegesek, akkor **nincs** $q \xrightarrow{a,b/\gamma'} p'$ átmenete (γ' tetszőleges) semmilyen $a \in \Sigma$ betűre és p' állapotra
- Ha M -nek van egy $q \xrightarrow{a,\varepsilon/\gamma} p$ átmenete, ahol a, γ tetszőlegesek, akkor **nincs** $q \xrightarrow{a,b/\gamma'} p'$ átmenete (γ' tetszőleges) semmilyen $b \in \Gamma$ betűre és p' állapotra

Nem minden környezetfüggetlen nyelv ismerhető fel **determinisztikus** veremautomatával

Az $\{uu^{-1} \mid u \in \{0,1\}^*\}$ nyelv például környezetfüggetlen (következő fólia), de nem ismerhető fel determinisztikus veremautomatával

Determinisztikus veremautomaták

Az $\{uu^{-1} \mid u \in \{0,1\}^*\}$ nyelvet felismerő nemdeterminisztikus veremautomata:



CF nyelvek elemzése

- A továbbiakban legyen $G = (N, T, R, S)$ egy CF nyelvtan, w pedig egy T -feletti szó
- A $w \in L(G)$ -t (azaz a szóproblémát) eldöntő **elemző** (parser) megkapja w -t mint betűk sorozatát (hasonlóan a VA-khoz meg a veremautomatákhoz) és leellenőrzi, hogy **w levezethető-e S -ből**
- Ezt a folyamatot hívjuk **elemzésnek**
- Az elemzés során az elemző a w egy **derivációs fáját** (azaz egy olyan derivációs fát aminek a határa w) próbálja felépíteni
 - Ha a derivációs fát a **gyökerétől kezdve** a levelek felé haladva építi fel, akkor **top-down elemzésről** beszélünk
 - Ha a derivációs fát a **levelektől kezdve** a gyökér felé haladva építi fel, akkor **bottom-up elemzésről** beszélünk

Top-down elemzés

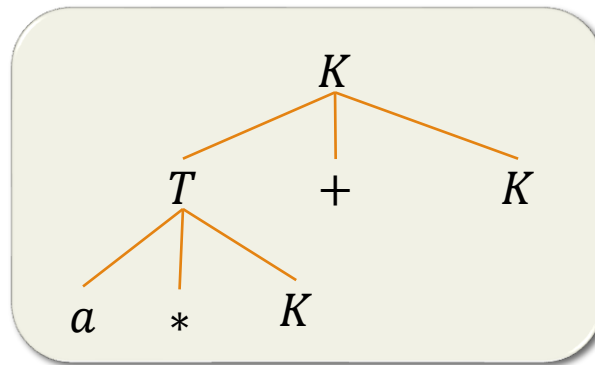
Az általános (visszalépéses) top-down elemzés **alapötlete**:

- w **baloldali levezetését** szimulálva felépítjük a w egy t **derivációs fáját** az alábbiak szerint
- Kezdetben legyen t egy **S -sel címkézett** csúcs
- Amíg van nemterminális t -ben
 - Kiválasztjuk a t **balról legelső** olyan ℓ levelét ami nemterminálissal címkézett, legyen ez a nemterminális A
 - Kiválasztjuk a G egy **A baloldalú** szabályát (*)
 - Legyen a szabály **jobboldala** $X_1 \dots X_k$
 - Minden $j = 1, \dots, k$ -ra, létrehozunk ℓ **egy új gyermekét** és címkézzük X_j -vel
 - Ha t **határának** maximális hosszú terminális prefixe **nem illeszkedik** w -re, akkor **visszalépés** történik: fordított sorrendben **töröljük** a t -beli nemterminálisok kifejtéseit addig amíg (*)-ban nem tudunk egy másik szabályt választani
 - Ha már nem tudunk visszalépni, akkor w **nem vezethető** le G -ben
- Ha t határa megegyezik w -vel, akkor w **levezetető** G -ben

Top-down elemzés – Példa

Tekintsük az aritmetikai kifejezések korábban látott nyelvét és az azt generáló alábbi egyértelmű nyelvtant:

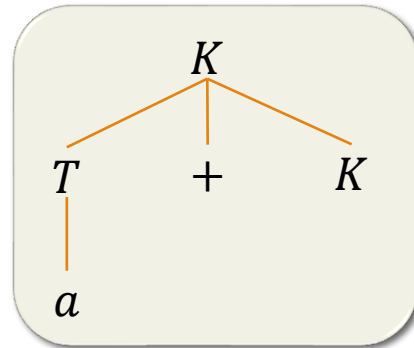
- A **terminális jelek**: $+, *, a, (,)$, a **nemterminálisok**: K, T , **kezdőszimbólum**: K
- A **szabályok**: $K \rightarrow T + K \mid T$; $T \rightarrow a * T \mid a \mid (K)$
- Az $w = a + a$ kifejezés elemzése (a t derivációs fa felépítése):
 - Legyen t egy K -val címkézett csúcs
 - Válasszuk az K -t és a $K \rightarrow T + K$ szabályt
 - Válasszuk T -t és az r : $T \rightarrow a * T$ szabályt
 - Ekkor a derivációs fa:



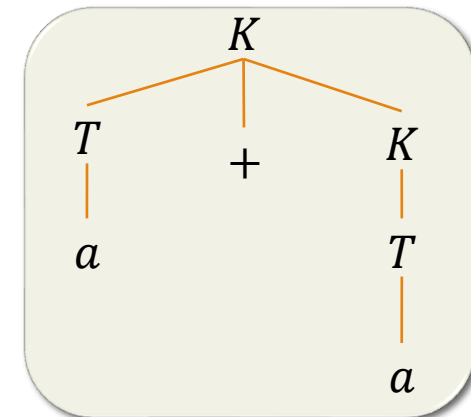
- a egyezik a w első betűjével, de $*$ nem a másodikkal, zsákutca!

Top-down elemzés – Példa

- Visszalépés: r helyett válasszuk a $T \rightarrow a$ szabályt
- A derivációs fa most így néz ki:



- a egyezik a w első betűjével, és most már a soron következő $+$ is egyezik a w második betűjével
- Ezután csak a K -t lehet helyettesíteni
 - Ha most a $K \rightarrow T + K$ szabályt választjuk, akkor később megint visszalépésre lesz szükség
 - Válasszuk helyette a $K \rightarrow T$ -t
 - És végül a T -t és $T \rightarrow a$ -t
- Készen vagyunk, sikerült levezetni az $a + a$ -t, a derivációs fája:



Prediktív elemzők

A **prediktív elemzők** képesek a visszalépések megelőzésére

A bemenet **végét** egy $N \cup T$ -n kívüli szimbólum jelöli; mi a $\$$ -t fogjuk használni

A bemenet aktuálisan olvasott betűjét **lookahead** szimbólumnak nevezzük, ami alapján mindig egyértelmű lesz, hogy milyen szabályt kell alkalmazni:

- Tegyük fel, hogy egy elemzés során a **lookahead az α betű** és tudjuk, hogy egy $A\gamma$ alakú mondatformából kell levezetni egy olyan szót, ami α -val kezdődik
- Tegyük fel, hogy A -ra az **$A \rightarrow \alpha \mid \beta$ szabályaink vannak** ($A \in N, \alpha, \beta, \gamma \in (N \cup \Sigma)^*$)
- **Melyik szabályt érdemes választani** az $A\gamma$ -beli első A átírására?
- Ha például az α megjelenik egy α -ból levezethető szó első betűjeként, de ugyanez nem igaz a β esetén és β -ból nem vezethető le az üres szó, akkor egyértelmű, hogy az $A \rightarrow \alpha$ szabályt kell választani
- Hiszen az $A \rightarrow \beta$ választása esetén egyrészt β -ból nem vezethető le olyan szó, aminek az első betűje egyezik a lookahead-del, azaz az α -val
- Másrészt β nem is törölhető, azaz az sem jó, ha γ -ból esetleg levezethető egy α -val kezdődő szó

Az ilyen és ehhez hasonló feltételek ellenőrzésében segítenek a következő fólián definiált ***first és follow* halmazok**

A $first(\alpha)$ és $follow(A)$ halmazok

Tetszőleges $\alpha \in (N \cup T)^*$ szóra és $a \in T \cup \{\varepsilon\}$ szimbólumra $a \in first(\alpha)$ akkor és csak akkor

- ha $a \in T$ és $\alpha \Rightarrow^* a\beta$ valamely $\beta \in (N \cup T)^*$ szóra vagy
- ha $a = \varepsilon$ és $\alpha \Rightarrow^* \varepsilon$

Tetszőleges $A \in N$ és $a \in T$ esetén $a \in follow(A)$ akkor és csak akkor

- ha $S \Rightarrow^* \alpha A a \beta$ valamely $\alpha, \beta \in (N \cup T)^*$ szavakra

Egy G CF nyelvtant $LL(1)$ -nek nevezünk ha tetszőleges két különböző $A \rightarrow \alpha \mid \beta$

szabályra a következők teljesülnek:

- $first(\alpha) \cap first(\beta) = \emptyset$
- ha $\varepsilon \in first(\beta)$ akkor $first(\alpha) \cap follow(A) = \emptyset$
- ha $\varepsilon \in first(\alpha)$ akkor $first(\beta) \cap follow(A) = \emptyset$

Mivel nem minden CF nyelv egyértelmű, de az $LL(1)$ nyelvtanok egyértelműek, kapjuk, hogy nem minden CF nyelvhez adható őt generáló $LL(1)$ nyelvtan

$LL(1)$ nyelvtanok

Tekintsük egy korábbi példa nyelvtanunk szabályait:

- $K \rightarrow T + K \mid T; T \rightarrow a * T \mid a \mid (K)$
 - $first(a * T) = first(a) = \{a\}$, $first(K) = first(T) = \{a, (\}$, $first(T + K) = \{a, (\}$
- **Nem $LL(1)$** mert $first(T + K) \cap first(T) = \{a, (\} \neq \emptyset$

Egy ekvivalens nyelvtan:

- $K \rightarrow TK', K' \rightarrow +TK' \mid \varepsilon, T \rightarrow FT', T' \rightarrow * FT' \mid \varepsilon, F \rightarrow (K) \mid a$
 - $first(K) = first(T) = first(F) = \{a, (\}$, $first(K') = \{+, \varepsilon\}$, $first(T') = \{*, \varepsilon\}$
 - $follow(K) = follow(K') = \{), \$ \}$ (notice: K a kezdő, ezért $\$ \in follow(K)$)
 - $follow(T) = follow(T') = \{+,), \$ \}$
 - $follow(F) = \{*, +,), \$ \}$
- Ez a nyelvtan viszont **már $LL(1)$** , vegyük például a $T' \rightarrow * FT' \mid \varepsilon$ szabályokat
 - $first(* FT') \cap first(\varepsilon) = \{*\} \cap \{\varepsilon\} = \emptyset$ és $first(* FT') \cap follow(T') = \{*\} \cap \{+,), \$ \} = \emptyset$

Az elemzőtábla

$LL(1)$ elemzőknek nevezzük az $LL(1)$ nyelvtanokra konstruálható prediktív elemzőket

Ezek működése az **elemzőtáblán** alapul

Egy G $LL(1)$ nyelvtan M **elemzőtáblájának** sorai G nemterminálisaival, oszlopai pedig G terminálisaival címkézettek, továbbá

M minden cellája egy szabályhalmaz, amit a következőképpen számolunk ki

for all $A \rightarrow \alpha$ szabályra **do**

for all $a \in first(\alpha)$ terminálisra írjuk be $A \rightarrow \alpha$ -t $M[A, a]$ -ba

if $\epsilon \in first(\alpha)$ **then** minden $a \in follow(A)$ -ra írjuk be $A \rightarrow \alpha$ -t $M[A, a]$ -ba

Az $LL(1)$ jelölésben

- az első L : a bementet beolvasása balról jobbra történik
- a második L : baloldali levezetést szimulál
- 1: egy darab lookahead szimbólum kerül beolvasásra

Lehet definiálni $LL(k)$ elemzőket is tetszőleges k -ra, de a gyakorlatban általában elegendő az $LL(1)$

Ha G egy $LL(1)$ nyelvtan, akkor az elemzőtáblájának minden cellája legfeljebb egy szabályt tartalmaz

Az elemzőtábla – Példa

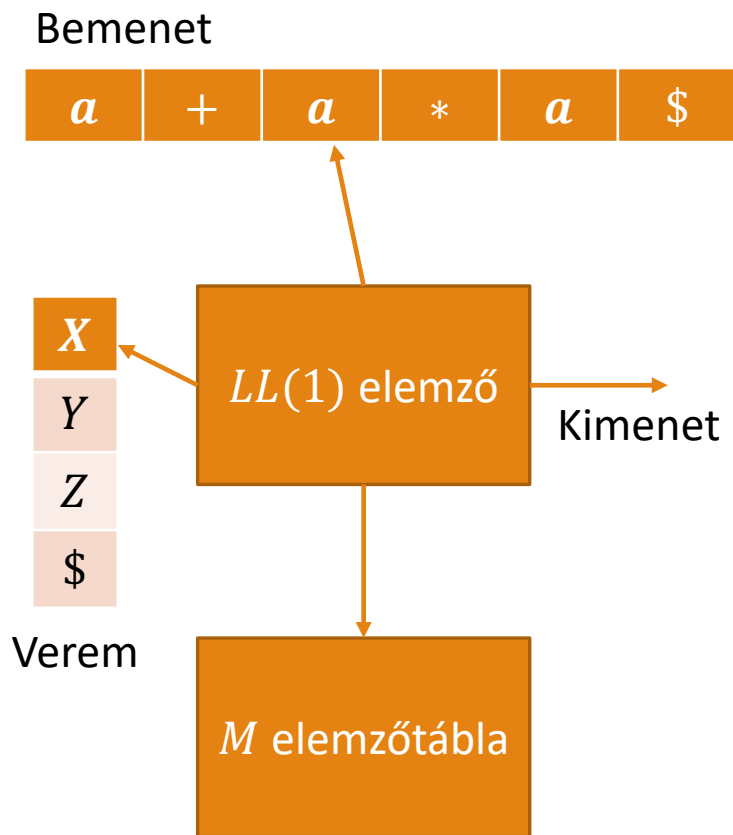
Tekintsük újra a $K \rightarrow TK'$, $K' \rightarrow +TK' \mid \varepsilon$,
 $T \rightarrow FT'$, $T' \rightarrow * FT' \mid \varepsilon$, $F \rightarrow (K) \mid a$ szabályokat

M	a	$+$	$*$	$($	$)$	$\$$
K	$K \rightarrow TK'$			$K \rightarrow TK'$		
K'		$K' \rightarrow +TK'$			$K' \rightarrow \varepsilon$	$K' \rightarrow \varepsilon$
T	$T \rightarrow FT'$			$T \rightarrow FT'$		
T'		$T' \rightarrow \varepsilon$	$T' \rightarrow * FT'$		$T' \rightarrow \varepsilon$	$T' \rightarrow \varepsilon$
F	$F \rightarrow a$			$F \rightarrow (K)$		

A	$follow(A)$
K	$), \$$
K'	$), \$$
T	$+,), \$$
T'	$+,), \$$
F	$+, *,), \$$

α	$first(\alpha)$
TK'	$(, a$
$+TK'$	$+$
FT'	$(, a$
$* FT'$	$*$
(K)	$($
a	a

Az $LL(1)$ elemző felépítése és működése



Bemenet: egy w szó és a G nyelvtan M elemzőtáblája

Kimenet: ha $w \in L(G)$, akkor a w egy **baloldali levezetése**, egyébként **hibaüzenet**

Módszer: **kezdetben** az elemző a következő konfigurációban van:

- $w\$$ van a bemeneten és a veremben az $S\$$ szó van (a $\$$ van legalul)
- Az i mutató a bemenet **első betűjére** (azaz a lookahead szimbólumra) mutat

Legyen X a **verem tetején** lévő szimbólum

while $X \neq \$$

if X a **lookahead** szimbólum **then** töröljük a verem tetejét és növeljük i -t eggyel

else if X egy **terminális** **then** kimenetre: hibaüzenet; **stop**

else if $M[X, a]$ **üres** **then** kimenetre: hibaüzenet; **stop**

else if $M[X, a] = X \rightarrow Y_1 \dots Y_k$ **then do**

kimenetre: $X \rightarrow Y_1 \dots Y_k$;

töröljük a verem tetejét;

a verem tetejére írjuk az $Y_1 \dots Y_k$ szót % Y_k kerül legalulra

Legyen X a **verem tetején** lévő szimbólum

if X nem a lookahead szimbólum **then** kimenetre: hibaüzenet

Az $LL(1)$ elemző működése – Példa

Tekintsük az alábbi elemzőtáblát és a $w = a + a * a$ bemenetet

M	a	$+$	$*$	$($	$)$	$\$$
K	$K \rightarrow TK'$			$K \rightarrow TK'$		
K'		$K' \rightarrow +TK'$			$K' \rightarrow \varepsilon$	$K' \rightarrow \varepsilon$
T	$T \rightarrow FT'$			$T \rightarrow FT'$		
T'		$T' \rightarrow \varepsilon$	$T' \rightarrow * FT'$		$T' \rightarrow \varepsilon$	$T' \rightarrow \varepsilon$
F	$F \rightarrow a$			$F \rightarrow (K)$		

A w szó $LL(1)$ elemzése a következő:

Verem	Bemenet	Művelet
$K\$$	$a + a * a\$$	
$TK'\$$	$a + a * a\$$	Output: $K \rightarrow TK'$
$FT'K'\$$	$a + a * a\$$	Output: $T \rightarrow FT'$
$aT'K'\$$	$a + a * a\$$	Output: $F \rightarrow a$
$T'K'\$$	$+a * a\$$	Olvasás: a
$K'\$$	$+a * a\$$	Output: $T' \rightarrow \varepsilon$
$+TK'\$$	$+a * a\$$	Output: $TK' \rightarrow +TK'$
$TK'\$$	$a * a\$$	Olvasás: $+$
$FT'K'\$$	$a * a\$$	Output: $T \rightarrow FT'$
$aT'K'\$$	$a * a\$$	Output: $F \rightarrow a$
$T'K'\$$	$* a\$$	Olvasás: a
$* FT'K'\$$	$* a\$$	Output: $T \rightarrow * FT'$
$FT'K'\$$	$a\$$	Olvasás: $*$
$aT'K'\$$	$a\$$	Output: $F \rightarrow a$
$T'K'\$$	$\$$	Olvasás: a
$K'\$$	$\$$	Output: $T' \rightarrow \varepsilon$
$\$$	$\$$	Output: $E' \rightarrow \varepsilon$

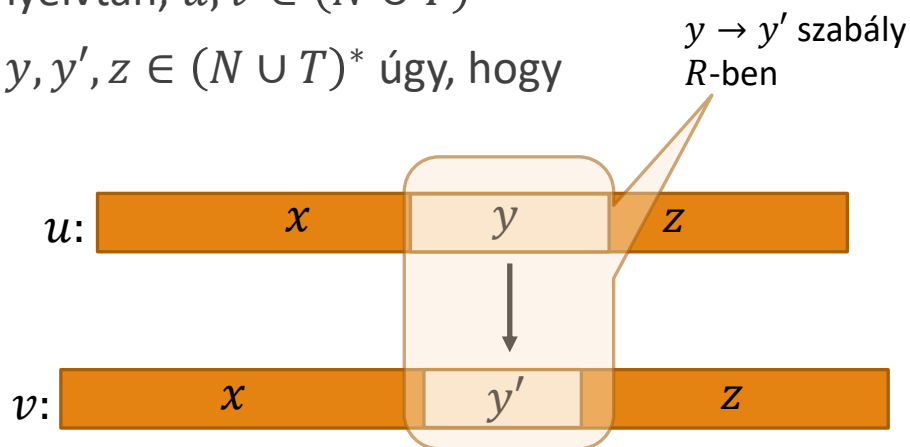
Általános nyelvtanok

(Általános) nyelvtan egy olyan $G = (N, T, R, S)$ rendszer, ahol

- N, T, S ugyanazok, mint környezetfüggetlen nyelvtan esetén,
- R pedig $u \rightarrow v$ alakú szabályok véges halmaza, ahol $u, v \in (N \cup T)^*$ és u tartalmaz legalább egy nemterminálist

Legyen $G = (N, T, R, S)$ egy nyelvtan, $u, v \in (N \cup T)^*$

- $u \Rightarrow v$ ha létezik olyan $x, y, y', z \in (N \cup T)^*$ úgy, hogy



$$G : S \rightarrow aSBc \mid \varepsilon$$

$$cB \rightarrow Bc$$

$$aB \rightarrow ab$$

$$bB \rightarrow bb$$

$$L(G) = \{a^n b^n c^n \mid n \geq 0\}$$

- $u \Rightarrow^* v$ és $L(G)$ (azaz a G által generált nyelv) definíciója ugyanaz, mint a környezetfüggetlen nyelvtanok esetében

Nyelvtanok

Környezetfüggőnek nevezünk egy $G = (N, T, R, S)$ nyelvtant, ha R minden eleme $uXv \rightarrow uwv$ alakú, ahol $u, v, w \in (N \cup T)^*$, $X \in N$, $w \neq \varepsilon$ + KES (kivéve az $S \rightarrow \varepsilon$ szabályt, de akkor ...)

$G' : S_0 \rightarrow S \mid \varepsilon$
 $S \rightarrow aSBC \mid aBC$
 $CB \rightarrow XB$
 $XB \rightarrow XC$
 $XC \rightarrow BC$
 $aB \rightarrow ab$
 $bB \rightarrow bb$
 $C \rightarrow c$

$$L(G) = \{a^n b^n c^n \mid n \geq 0\}$$

Emlékeztető: Jobblineárisnak nevezünk egy $G = (N, T, R, S)$ nyelvtant, ha R minden eleme $A \rightarrow uB$ vagy $A \rightarrow u$ alakú, ahol $A, B \in N$, $u \in T^*$

Chomsky-féle hierarchia

Jobblineáris nyelvtan: **3-as típusú** nyelvtan

Környezetfüggetlen nyelvtan: **2-es típusú** nyelvtan

Környezetfüggő nyelvtan: **1-es típusú** nyelvtan

Általános nyelvtan: **0-ás típusú** nyelvtan

\mathcal{L}_i : az i -típusú nyelvtanokkal ($i = 0,1,2,3$) generálható nyelvek osztálya

\mathcal{L}_3 pontosan a reguláris nyelvek osztálya

