

Környezetfüggetlen nyelvtanok

Környezetfüggetlen nyelvtanok

A környezetfüggetlen nyelvtanokat Noam Chomsky vezette be az 1950-es években a természetes nyelvek struktúrájának leírására

- Erre viszont csak korlátozottan alkalmasak

Mire jók akkor?

- Programozási nyelvek szintaxisának megadására (Backus—Naur-forma, BNF)

A bal oldalon
van amit
definiálunk

Egyszerű kifejezések definíciója BNF-fel:

```
<kifejezés> ::= <kifejezés> "+" <kifejezés> |  
                <kifejezés> "*" <kifejezés> |  
                "(" <kifejezés> ")" | <konstans>  
<konstans> ::= "a" | "b" | "c"
```

A jobb oldalon
megmondjuk,
miből állhat egy
kifejezés
(a | jel „vagy”-
ként funkcionál,
mint a Unix
regkifejknél)

Hogyan ellenőrizhető, hogy az $(a + b) * c$ helyes-e?

```
<kifejezés>  
⇒ <kifejezés> * <kifejezés>  
⇒ (<kifejezés>) * <kifejezés>  
⇒ (<kifejezés> + <kifejezés>) * <kifejezés>  
⇒3 (<konstans> + <konstans>) * <konstans>  
⇒3 (a + b) * c
```

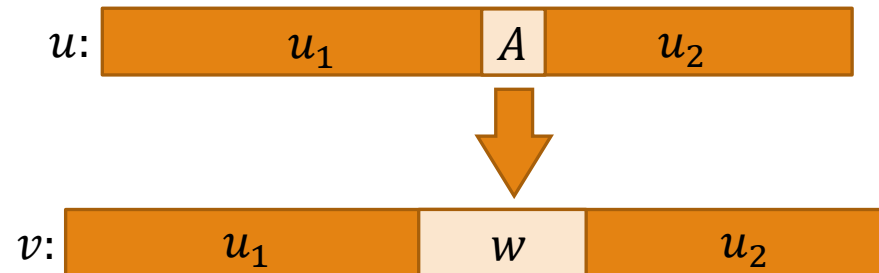
Környezetfüggetlen nyelvtanok

Környezetfüggetlen (röviden: CF) nyelvtan egy $G = (N, T, R, S)$ rendszer, ahol

- N véges, nemüres halmaz: a **nemterminálisok** halmaza
- T véges, nemüres halmaz: a **terminálisok** halmaza, $N \cap T = \emptyset$
- $S \in N$ a **kezdőszimbólum**
- R : $A \rightarrow w$ alakú **szabályok** véges halmaza, ahol $A \in N, w \in (N \cup T)^*$
 - A továbbiakban **G szimbólumai** alatt az $(N \cup T)$ halmazt értjük

Tetszőleges $u, v \in (N \cup T)^*$ szavakra, u -ból **közvetlenül deriválható** (vagy levezethető) v (jele: $u \Rightarrow v$) ha u, v felbontható $u = u_1 A u_2$ és $v = u_1 w u_2$ alakban úgy, hogy $A \rightarrow w \in R$

- Szemléletesen:



Környezetfüggetlen nyelvtanok

Egy u_0, u_1, \dots, u_n ($n \geq 0$) sorozatot a v szó u -ból való **derivációjának** (levezetésének) nevezzük ha

- $u = u_0 \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \dots \Rightarrow u_{n-1} \Rightarrow u_n = v$

Ez esetben azt mondjuk, hogy a v **deriválható** vagy **levezethető** u -ból, jele: $u \Rightarrow^* v$

- Mondatformának** nevezzük egy u szót, ha $S \Rightarrow^* u$

A G által **generált nyelv**: azon T -feletti szavak halmaza, melyek levezethetők S -ből

- Ezt a nyelvet **$L(G)$** jelöljük

Egy L nyelvet környezetfüggetlennek (röviden: CF) nevezzük, ha van olyan G CF nyelvtan, melyre $L = L(G)$

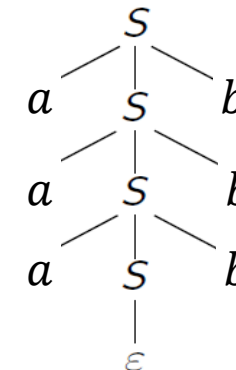
$$N = \{S\}, T = \{a, b\}$$

$$R = \{S \rightarrow \varepsilon, S \rightarrow aSb\}$$

Kezdőszimbólum: S

$$(R = \{S \rightarrow aSb | \varepsilon\}),$$

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb \quad \text{deriváció}$$



derivációs fa
(pontos def. később)

$$\text{A generált nyelv: } \{a^n b^n \mid n \geq 0\}$$

Környezetfüggetlen nyelvtanok – Példa

Az ilyen nyelvtanok egyik legjellemzőbb tulajdonsága az, hogy a helyes zárójelezéseknek megfelelő struktúrák modellezhetők vele

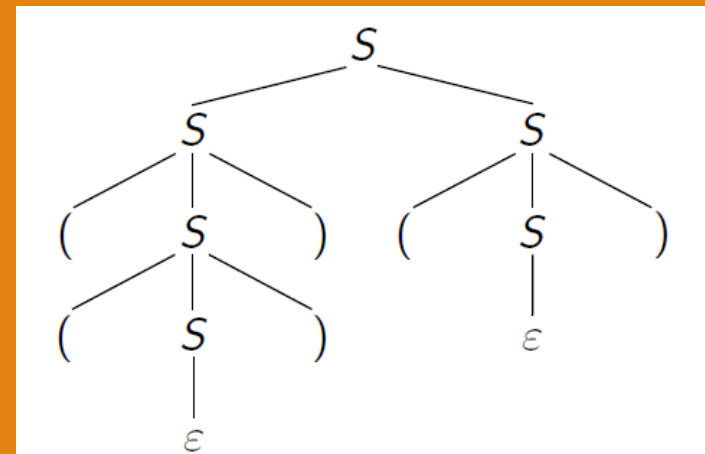
Például a helyes zárójelezések nyelve a következő nyelvtannal generálható:

- A terminális jelek: (és)
- A nemterminálisok: S
- Kezdő nemterminális: S
- A szabályok: $S \rightarrow SS \mid (S) \mid \varepsilon$

Az $((()))$ szó egy levezetése:

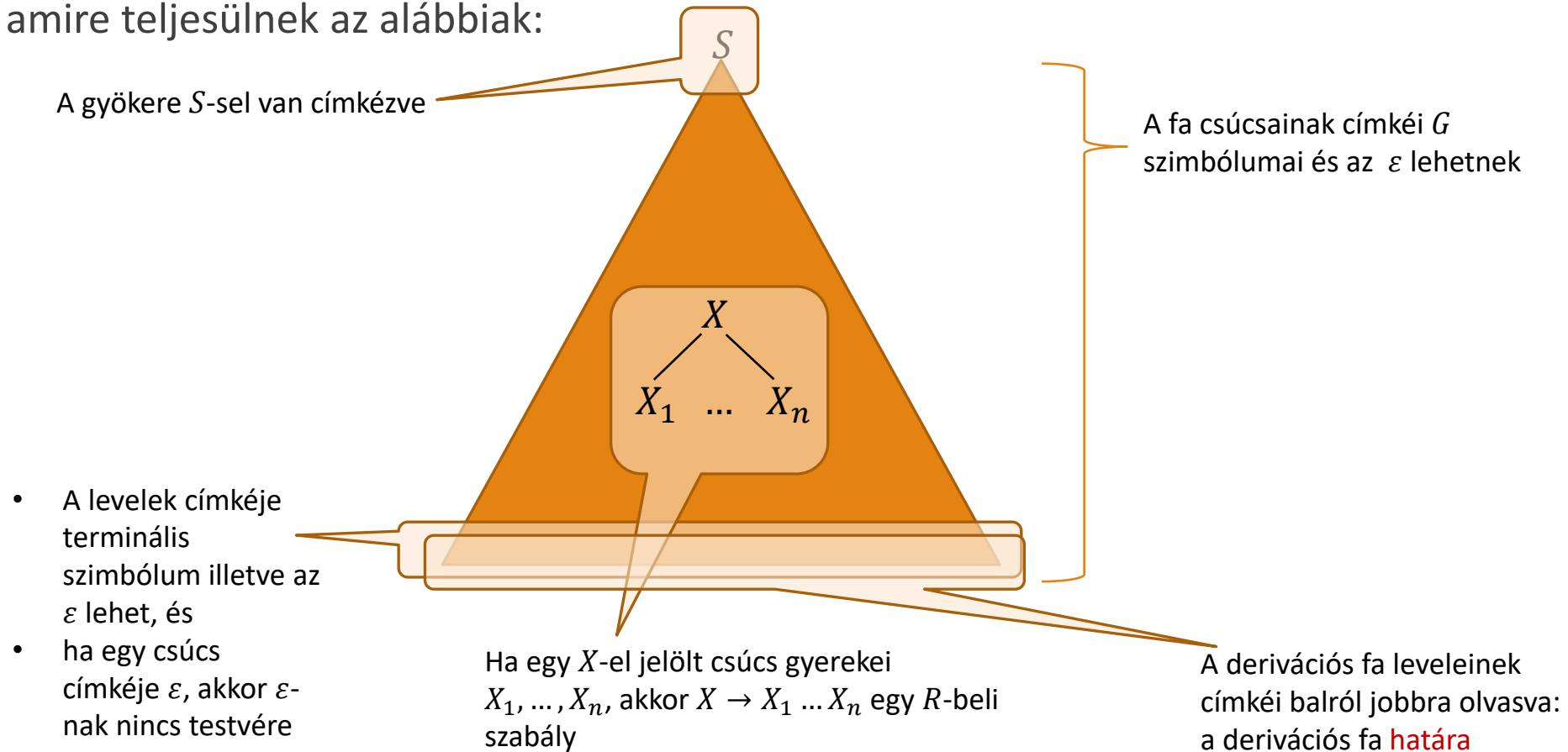
- $S \Rightarrow SS \Rightarrow (S)S$
 $\Rightarrow ((S))S \Rightarrow (())S$
 $\Rightarrow (())(S) \Rightarrow (())()$

Derivációs fa



Derivációs fák

Legyen $G = (N, T, R, S)$ egy környezetfüggetlen nyelvtan. Egy G -feletti **derivációs fa** egy olyan fa, amire teljesülnek az alábbiak:



A szóprobléma

Egy $G = (N, T, S, R)$ CF nyelvtan által generált nyelv szavai és G derivációs fái között szoros kapcsolat van:

- Tetszőleges $u \in T^*$ szóra:
- u levezethető G -ben $\Leftrightarrow G$ -nek van olyan derivációs fája, melynek a határa u

Egy $L \subseteq \Sigma^*$ nyelv **szóproblémája** alatt a következőt értjük:

- Legyen $w \in \Sigma^*$
- Döntsük el, hogy $w \in L$ teljesül-e

Azt a programot, ami megoldja a szóproblémát hívjuk **elemzőnek**

Ha egy L nyelv egy G CF nyelvtannal adott, akkor a fentiek alapján azt, hogy $w \in L$ teljesül-e eldönthetjük úgy is, hogy megpróbálunk egy olyan derivációs fát felépíteni, melynek határa w

Ha a derivációs fa még nincs készen, azaz van olyan levele, ami egy nemterminális, akkor általában a következő kérdésekre kell választ adni:

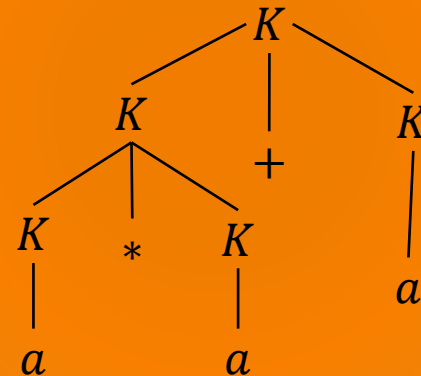
1. Melyik nemterminálist terjesszük ki
2. Melyik szabályt alkalmazzuk erre a nemterminálisra

Derivációs fák – Példa

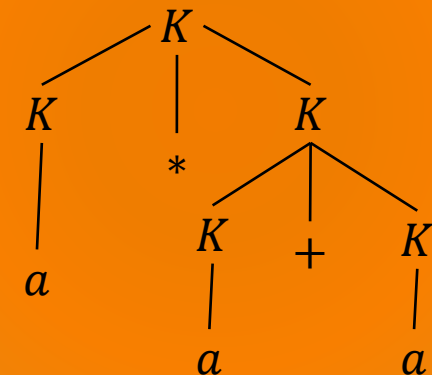
Az egyszerű kifejezések korábbi – BNF-fel megadott – leírása CF nyelvtannal megadva (nevezzük ezt a nyelvtant CFEXP1-nek):

- Terminális jelek: $+, *, (,), a$ (az egyszerűség kedvéért a továbbiakban a konstansok jelölésére csak az a terminálist használjuk)
- Nemterminálisok: K , ez lesz a kezdő is
- Szabályok:
 - $K \rightarrow K + K \mid K * K \mid (K) \mid a$
- Az $a * a + a$ kifejezés néhány levezetése:
 1. $K \Rightarrow K + K \Rightarrow K * K + K \Rightarrow K * a + K$
 $\Rightarrow a * a + K \Rightarrow a * a + a$
 2. $K \Rightarrow K * K \Rightarrow a * K \Rightarrow a * K + K$
 $\Rightarrow a * a + K \Rightarrow a * a + a$
 3. $K \Rightarrow K * K \Rightarrow K * K + K \Rightarrow a * K + K$
 $\Rightarrow a * a + K \Rightarrow a * a + a$

Az első levezetés
derivációs fája:



A második és
harmadiké:



A harmadik csak annyiban különbözik a másodiktól, hogy más sorrendben írja át a K -kat a -ra (ezért egyezik meg a derivációs fájuk)

Derivációs fák – Baloldali levezetés

A derivációs fából könnyen kiolvasható, hogy melyik nemterminális melyik szabály jobb oldalával lett kiterjesztve (előző megjegyzés 2. pont), de az nem, hogy ez milyen sorrendben történt (1. pont)

G egy levezetését **baloldalinak** nevezzük ha minden lépésben az aktuális mondatforma legelső nemterminálisa van átírva (a **jobboldali** levezetések hasonlóan definiálhatók)

A baloldali deriváció jelölése: $u_0 \Rightarrow_l u_1 \Rightarrow_l \dots \Rightarrow_l u_n$ vagy $u_0 \Rightarrow_l^* u_n$

CFEXP1 levezetései közül csak a 2. baloldali levezetés

Egy G CF nyelvtan derivációs fái és $L(G)$ szavainak baloldali levezetései között kölcsönösen egyértelmű kapcsolat van

Egyértelmű CF nyelvtanok

A G CF nyelvtant **egyértelműnek** nevezzük, ha minden $u \in L(G)$ szónak **pontosan** egy baloldali levezetése (azaz derivációs fája) van

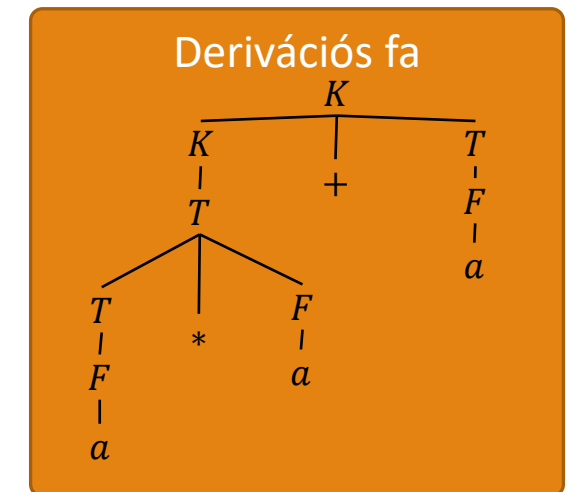
Az „A derivációs fa felépítése” megjegyzésben írtak 1. pontja világos, ha baloldali levezetéseket tekintünk

A megfelelő szabály kiválasztását (2-es pont) segítheti az, hogy a nyelvtan egyértelmű

Ezért célszerű adott feladatra egyértelmű nyelvtant megadni

CFEXP1 **nem egyértelmű**, de az általa generált nyelvhez **lehet egyértelmű** nyelvtant adni:

- A **terminális jelek**: $+, *, a, (,)$, a **nemterminálisok**: K, T, F , **kezdő nemterminális**: K
- A **szabályok**:
 - $K \rightarrow K + T \mid T; T \rightarrow T * F \mid F; F \rightarrow (K) \mid a$
- Az $a * a + a$ kifejezés **egyetlen baloldali** levezetése:
 - $K \Rightarrow_l K + T \Rightarrow_l T + T \Rightarrow_l T * F + T \Rightarrow_l F * F + T \Rightarrow_l a * F + T \Rightarrow_l a * a + T \Rightarrow_l a * a + F \Rightarrow_l a * a + a$



Egyértelmű CF nyelvtanok

Egy CF nyelvet **egyértelműnek** nevezünk ha generálható egyértelmű CF nyelvtannal

Nem minden CF nyelv egyértelmű

Ráadásul nem lehet algoritmust adni annak **eldöntésére**, hogy egy CF nyelvtan egyértelmű-e (lásd később)

Egy **nem egyértelmű** CF nyelv: $L = \{a^i b^j c^k \mid i, j, k \geq 1, i = j \text{ vagy } j = k\}$

L generálására alapvetően nincs más mód, mint a következő szabályokkal rendelkező G nyelvtan használata (a nagybetűk a nemterminálisok, a kisbetűk a terminálisok, S a kezdő):

$S \rightarrow AB$

$S \rightarrow CD$

$A \rightarrow aAb \mid ab$

$B \rightarrow cB \mid c$

$C \rightarrow aC \mid a$

$D \rightarrow bDc \mid bc$

Ezt használjuk ha olyan szavakat akarunk levezetni, melyekben ugyanannyi a van mint b

Ezt pedig akkor ha olyanokat, melyekben ugyanannyi b van mint c

De akkor azok a szavak, melyekben ugyanannyi a , b és c van levezethetők akkor is ha az első és akkor is ha a második szabállyal kezdjük a levezetést. Azaz a nyelvtan nem egyértelmű

Jobblineáris nyelvtanok

- Egy $G = (N, T, S, R)$ CF nyelvtant **jobblineárisnak** nevezünk ha minden szabálya $A \rightarrow uB$ vagy $A \rightarrow u$ alakú, ahol $A, B \in N$ és $u \in T^*$
- Egy L nyelvet **jobblineárisnak** nevezünk ha van olyan jobblineáris G nyelvtan, melyre $L = L(G)$

A felismerhető nyelvek megegyeznek a jobblineáris nyelvekkel

Bizonyítás

Egyik irány: Legyen $M = (Q, \Sigma, \delta, q_0, F)$ egy véges automata

- $L(M)$ generálható azzal a $G = (Q, \Sigma, R, q_0)$ jobblineáris nyelvtannal, melyre

$$R = \{q \rightarrow aq' \mid q \xrightarrow{a} q' \text{ az } M \text{ egy átmenete}\} \cup \{q \rightarrow \varepsilon \mid q \in F\}$$

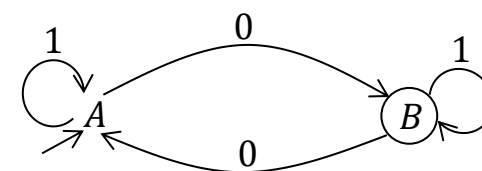
Másik irány: legyen $G = (N, T, S, R)$ egy jobblineáris nyelvtan

- feltehető, hogy R szabályai $A \rightarrow aB$ vagy $A \rightarrow \varepsilon$ alakúak, ahol $a \in T, B \in N$
 - Pl. egy $A \rightarrow auB$ alakú szabály, ahol $a \in T, u \in T^+, B \in N$, lecserélhető $A \rightarrow aA', A' \rightarrow uB$ szabályokra, ahol A' egy új nemterminális. Ezt a gondolatmenetet alkalmazva elérhető, hogy minden szabály a fenti alakú legyen
- $L(G)$ felismerhető azzal az $M = (N, T, \delta, S, F)$ NVA-val, melyre tetszőleges $A \in N, a \in T$ esetén

$A \xrightarrow{a} B$ pontosan akkor egy átmenete az M -nek ha $A \rightarrow aB \in R$,
és $A \in F$ pontosan akkor ha $A \rightarrow \varepsilon \in R$

Az

$L = \{u \in \{0,1\}^* \mid |u|_0 \text{ páratlan}\}$
nyelv felismerhető az alábbi **véges automatával**:



és generálható az alábbi **jobblineáris nyelvtannal**:

- Nemterminálisok: A, B
- Terminálisok: $0, 1$
- Kezdőszimbólum: A
- Szabályok: $A \rightarrow 1A, A \rightarrow 0B, B \rightarrow 1B, B \rightarrow 0A, B \rightarrow \varepsilon$

Környezetfüggetlen nyelvek – zártsági tulajdonságok

Következmény: Minden reguláris nyelv környezetfüggetlen

A környezetfüggetlen nyelvek zártak a reguláris műveletekre

Bizonyítás

Legyen $G_1 = (N_1, T, S_1, R_1)$ egy L_1 nyelvet, $G_2 = (N_2, T, S_2, R_2)$ pedig egy L_2 nyelvet generáló CF nyelvtan

- Feltehetjük, hogy N_1 és N_2 diszjunktak
- Legyen S egy új nemterminális
- Az $L_1 \cup L_2$ nyelvet generáló CF nyelvtan:

$$G = (N_1 \cup N_2 \cup \{S\}, T, R_1 \cup R_2 \cup \{S \rightarrow S_1, S \rightarrow S_2\}, S)$$

- Az $L_1 L_2$ nyelvet generáló CF nyelvtan:

$$G = (N_1 \cup N_2 \cup \{S\}, T, R_1 \cup R_2 \cup \{S \rightarrow S_1 S_2\}, S)$$

- A L_1^* nyelvet generáló CF nyelvtan :

$$G = (N_1 \cup \{S\}, T, R_1 \cup \{S \rightarrow SS_1, S \rightarrow \varepsilon\}, S)$$

Környezetfüggetlen nyelvek – zártsági tulajdonságok

A környezetfüggetlen nyelvek nem zártak a metszetképzésre és a komplementerképzésre

Bizonyítás

Elég megmutatni az állítást a metszetképzésre, abból az egyesítésre való zártság miatt adódik az állítás a komplementerképzésre

Legyen $L_1 = \{a^n b^n c^m \mid n, m \geq 0\}$ és $L_2 = \{a^m b^n c^n \mid n, m \geq 0\}$

Világos, hogy $L_1 \cap L_2 = \{a^n b^n c^n \mid n \geq 0\}$

Ugyanakkor $\{a^n b^n c^n \mid n \geq 0\}$ nem környezetfüggetlen (ennek bizonyítása hamarosan)

A környezetfüggetlen nyelvek korlátai

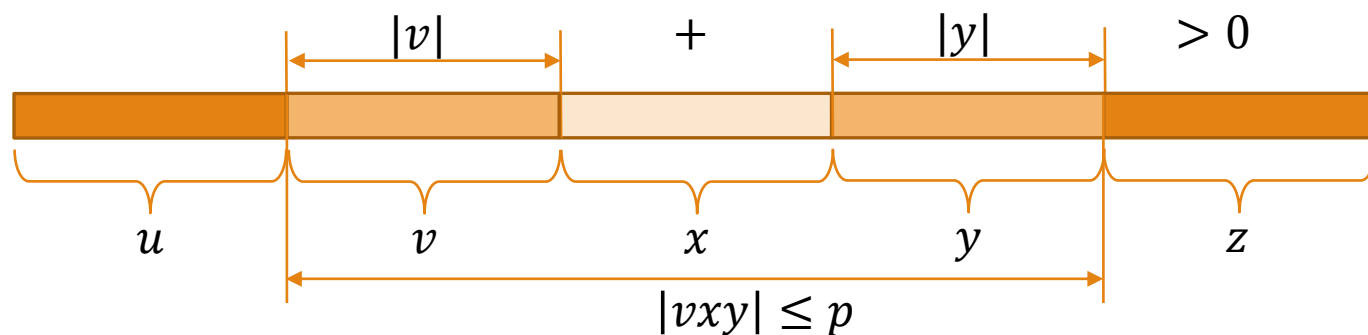
Ha egy G környezetfüggetlen nyelvtannak egy t derivációs fája „elég magas”, akkor t -ben lesz olyan út, melyen egy nemterminális legalább kétszer fordul elő

CF pumpáló lemma

Legyen $G = (N, T, R, S)$ egy L nyelvet generáló CF nyelvtan

Legyen $p = k^{|N|+1}$, ahol k az R leghosszabb szabályában a jobb oldal hossza

Legyen w egy olyan legalább p hosszú szó, ami levezethető G -ben. Ekkor w felírható



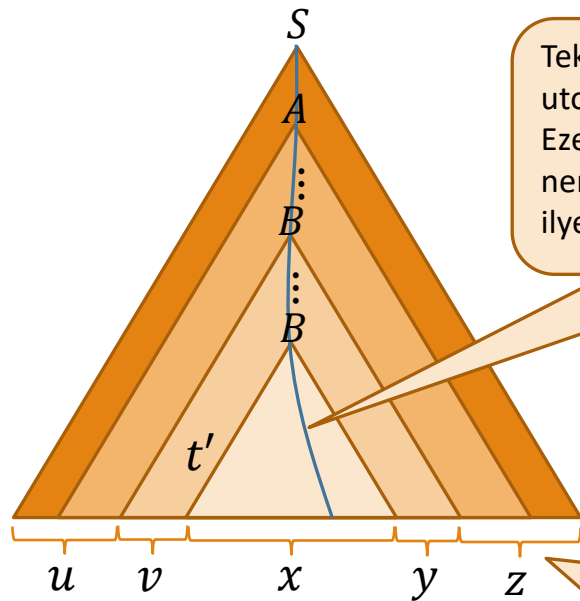
alakban úgy, hogy

1. $|vy| > 0$,
2. $|vxy| \leq p$,
3. $uv^i xy^i z \in L$ minden $i \geq 0$ számra

A környezetfüggetlen nyelvek pumpáló lemmája

Bizonyítás

- Vegyük a w szó egy **minimális magasságú** t derivációs fáját
 - t **magassága** a leghosszabb úton lévő élek száma; jele: $|t|$
- A t határának hossza legfeljebb $k^{|t|}$
- Mivel w hossza legalább $k^{|N|+1}$, $|t|$ legalább $|N| + 1$
- Legyen h a t -ben az egyik leghosszabb út, ekkor t felírható a következő alakban:

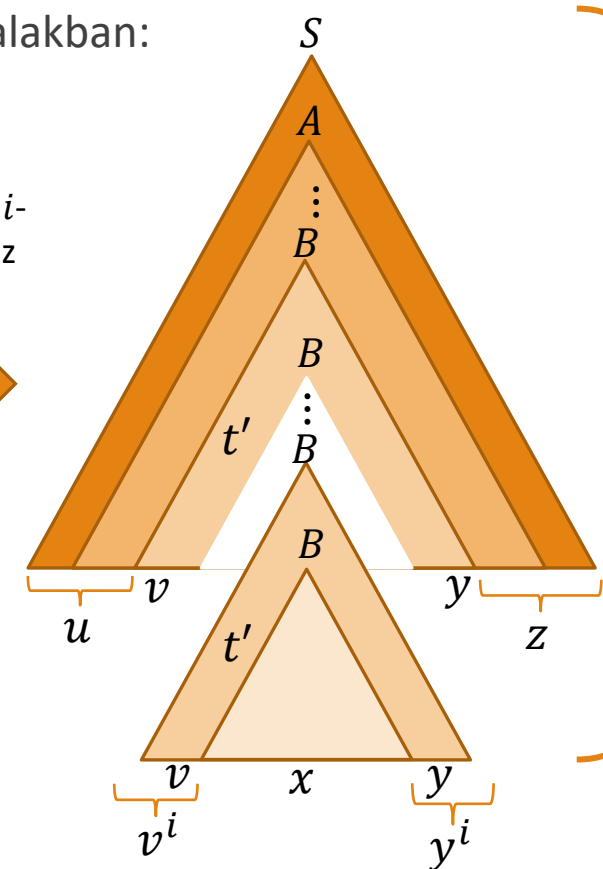


Tekintsük a h utat, és vegyük a h -n az utolsó $|N| + 1$ nemtermináliszt. Ezek között kell legyen egy ismétlődő nemterminális, vegyük a legelső ilyen, legyen ez B .

w ezen felbontására teljesülnek a lemma feltételei:

1. vy nem lehet üres, mert akkor készíthetnénk egy alacsonyabb derivációs fát melynek határa ugyanúgy w
2. vxy hossza legfeljebb p mert az őt levezető der. fa magassága legfeljebb $|N| + 1$

A t' -vel jelölt részfa i -szeres iterálása (azaz „egymás alá írása”)



3. Ez is egy derivációs fa, melynek határa L -beli

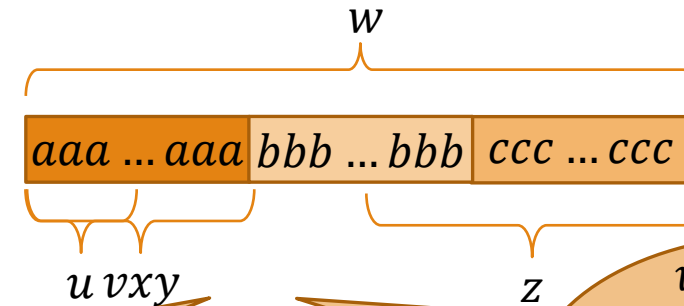
Az $L = \{a^n b^n c^n \mid n \geq 0\}$ nyelv nem környezetfüggetlen

Hogyan lehet ezt megmutatni?

Tegyük fel, hogy mégis van olyan G CF nyelvtan ami L -et generálja és legyen p az előző állításban definiált szám

Legyen $w = a^p b^p c^p$, ez levezethető G -ben

Mivel $|w| \geq p$, a w -nek van egy ilyen felbontása:



vy legalább
1 betűt
tartalmaz

vxy hossza
max. p , ezért
legfeljebb
kétféle betű
lehet benne

Legyen $w' = u \underbrace{v^2}_{v^2} x \underbrace{y^2}_{y^2} z$

Azt tanultuk, hogy G w' -t is tudja generálni, de abban valamelyik betűből kevesebb van mint a többiből!

Hol a hiba? Hát ott, hogy feltettük, hogy a G CF nyelvtan képes generálni az L -et generálni

Tehát L nem CF nyelv!