

Syntaktická analýza, Top-down prístup, LL(1)

Ing. Viliam Hromada, PhD.

C-510
Ústav informatiky a matematiky
FEI STU

viliam.hromada@stuba.sk



Príklad

Gramatika $G = (\{ \langle \text{program} \rangle, \langle \text{príkaz} \rangle, \langle \text{príkazy} \rangle \}, \{ \text{begin}, \text{end}, p, ; \}, P, \langle \text{program} \rangle)$, pravidlá P :

$\langle \text{program} \rangle \rightarrow \text{begin} \langle \text{príkazy} \rangle \text{end}$

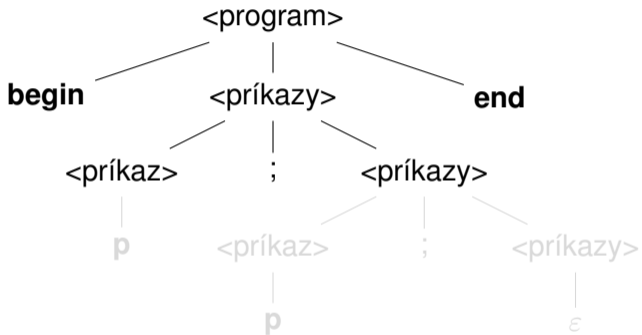
$\langle \text{príkazy} \rangle \rightarrow \langle \text{príkaz} \rangle ; \langle \text{príkazy} \rangle$

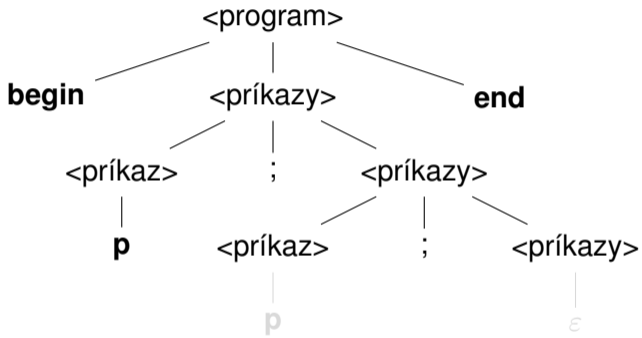
$\langle \text{príkazy} \rangle \rightarrow \varepsilon$

$\langle \text{príkaz} \rangle \rightarrow p$

Uvažujme odvodenie vety **begin p; p; end** od koreňa (počiatočného neterminálu) k listom (terminálom) tak, že vždy nahradíme prvý neterminál zľava (ľavá derivácia):







Top-down analýza

- Uvedený postup hľadania derivačného stromu od koreňa k listom, pričom sa vlastne vytvára ľavá derivácia vstupného slova, sa dá modelovať pomocou nedeterministického zásobníkového automatu.

Teoretický model top-down analyzátor

Top-down syntaktický analyzátor pre bezkontextovú gramatiku $G = (N, T, P, S)$ bude (nedeterministický) zásobníkový automat, ktorý pracuje na základe nasledovných princípov:

- Automat v zásobníku simuluje ľavé odvedenie slova w , t.j. zásobníkové symboly sú terminály a neterminály gramatiky G .
- Počiatočný zásobníkový symbol bude začiatkový neterminál S gramatiky G .



Teoretický model top-down analyzátor (pokr.)

- Kým sa automat nezasekne opakuje nasledovné činnosti:
 1. **Expanzia** - ak sa na vrchu zásobníka nachádza neterminál $A \in N$, algoritmus nedeterministicky vyberie jedno z A -pravidiel $A \rightarrow \alpha$ a neterminál A nahradí pravou stranou α .
 2. **Porovnanie** - ak sa na vrchu zásobníka nachádza terminál $a \in T$, porovná sa so symbolom na vstupe t ; ak sa zhodujú, automat odstráni symbol a z vrchu zásobníka a prejde na ďalší vstupný symbol; ak sa nezhodujú, automat sa zasekne, slovo neakceptuje (=syntaktická chyba).
- Automat akceptuje vstupné slovo (=syntakticky správne), ak spracuje celý vstup a skončí s prázdnyim zásobníkom.
- Ak automat vždy pošle na výstup pravidlo, podľa ktorého vykonal expanziu, získa sa postupnosť pravidiel, ktorá dáva ľavú deriváciu slova w v gramatike G .



Syntaktická analýza top-down I

Vstup: Bezkontextová gramatika $G = (N, T, P, S)$, reťazec w

Výstup: Zistenie, či $w \in L(G)$ a postupnosť pravidiel ľavého odvodu w

- 1: vlož do zásobníka S
- 2: **pokiaľ** zásobník nie je prázdny **rob**
- 3: **ak** na vrchu zásobníka je $A \in N$ **potom**
- 4: odstráň z vrchu zásobníka A
- 5: vlož do zásobníka pravú stranu α niektorého z pravidiel $A \rightarrow \alpha$
- 6: pošli na výstup pravidlo $A \rightarrow \alpha$
- 7: **koniec ak**
- 8: **ak** na vrchu zásobníka je $a \in T$ **potom**
- 9: **ak** a je totožné s aktuálnym vstupným symbolom t **potom**
- 10: odstráň a zo zásobníka
- 11: prejdi na ďalší vstupný symbol t
- 12: **inak**
- 13: neakceptuj vstupné slovo a skončí



Syntaktická analýza top-down II

- 14: **koniec ak**
- 15: **koniec ak**
- 16: **koniec pokiaľ**
- 17: **ak** bolo spracované celé vstupné slovo **potom**
- 18: akceptuj vstupné slovo
- 19: **inak**
- 20: neakceptuj vstupné slovo
- 21: **koniec ak**



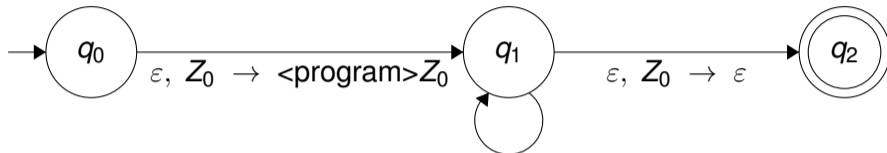
Príklad

Uvažujme gramatiku zo slajdu č. 6 a vetu **begin p ; p ; end**.

r.	Zvyšok vstupu	Zásobník <input type="checkbox"/>	Akcia
1	begin p ; p ; end	<program>	E1
2	begin p ; p ; end	begin <príkazy> end	P
3	p ; p ; end	<príkazy> end	E2
4	p ; p ; end	<príkaz>;<príkazy> end	E4
5	p ; p ; end	p ;<príkazy> end	P
6	; p ; end	;<príkazy> end	P
7	p ; end	<príkazy> end	E2
8	p ; end	<príkaz>;<príkazy> end	E4
9	p ; end	p ; <príkazy> end	P
10	; end	; <príkazy> end	P
11	end	<príkazy> end	E3
12	end	end	P
13	ϵ	ϵ	A



Ekvivalentný ZA



$\epsilon, \langle \text{program} \rangle \rightarrow \mathbf{begin} \langle \text{prikazy} \rangle \mathbf{end}$

$\epsilon, \langle \text{prikazy} \rangle \rightarrow \langle \text{prikaz} \rangle ; \langle \text{prikazy} \rangle$

$\epsilon, \langle \text{prikazy} \rangle \rightarrow \epsilon$

$\epsilon, \langle \text{prikaz} \rangle \rightarrow \mathbf{p}$

$\mathbf{begin}, \mathbf{begin} \rightarrow \epsilon$

$\mathbf{end}, \mathbf{end} \rightarrow \epsilon$

$\mathbf{p}, \mathbf{p} \rightarrow \epsilon$

$;, ; \rightarrow \epsilon$



Top-down analýza $\{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$

Gramatika pre uvedený jazyk:

1. $S \rightarrow SaSbS$
2. $S \rightarrow SbSaS$
3. $S \rightarrow \varepsilon$



Príklad - výpočet *abaabb*

Uvažujme gramatiku zo slajdu č. 16 a reťazec *abaabb*. Potom akceptujúci výpočet ZA bude (zápis vo forme tabuľky):

r.	Zvyšok vstupu	Zásobník □	Akcia
1	<i>abaabb</i>	S	E1
2	<i>abaabb</i>	SaSbS	E3
3	<i>abaabb</i>	aSbS	P
4	<i>baabb</i>	SbS	E3
5	<i>baabb</i>	bS	P
6	<i>aabb</i>	S	E1
7	<i>aabb</i>	SaSbS	E3
8	<i>aabb</i>	aSbS	P
9	<i>abb</i>	SbS	E1
10	<i>abb</i>	SaSbSbS	E3
11	<i>abb</i>	aSbSbS	P
12	<i>bb</i>	SbSbS	E3



Príklad - výpočet *abaabb*

r.	Zvyšok vstupu	Zásobník □	Akcia
13	<i>bb</i>	<i>bSbS</i>	P
14	<i>b</i>	<i>SbS</i>	E3
15	<i>b</i>	<i>bS</i>	P
16	ϵ	<i>S</i>	E3
17	ϵ	ϵ	A

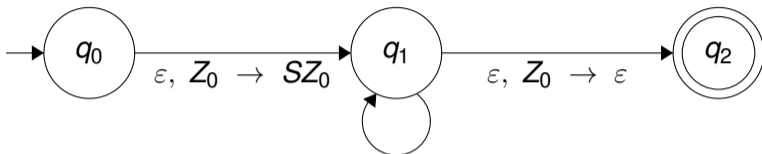


Príklad - výpočet *abb*

r.	Zvyšok vstupu	Zásobník □	Akcia
1	<i>abb</i>	<i>S</i>	E1
2	<i>abb</i>	<i>SaSbS</i>	E3
3	<i>abb</i>	<i>aSbS</i>	P
4	<i>bb</i>	<i>SbS</i>	E3
5	<i>bb</i>	<i>bS</i>	P
6	<i>b</i>	<i>S</i>	E2
7	<i>b</i>	<i>SbSaS</i>	E3
8	<i>b</i>	<i>bSaS</i>	P
9	ϵ	<i>SaS</i>	E3
10	ϵ	<i>aS</i>	ERROR

Pre *abb* by aj všetky iné možné výpočty skončili s chybou - slovo *abb* teda nepatrí do $L(G)$.





- $\epsilon, S \rightarrow SaSbS$
- $\epsilon, S \rightarrow SbSaS$
- $\epsilon, S \rightarrow \epsilon$
- $a, a \rightarrow \epsilon$
- $b, b \rightarrow \epsilon$



Vlastnosti teoretického modelu

- Predstavený model je použiteľný s ľubovoľnou bezkontextovou gramatikou. Dokáže nájsť derivačný strom pre slová generované aj nejednoznačnými, aj nedeterministickými gramatikami.
- **Nedeterminizmus** teoretického modelu je však prekážkou pri jeho praktickej použiteľnosti, pretože pri hľadaní derivácie / spracovaní slova je potrebné sledovať všetky možné vetvy výpočtu.
 - V príklade na slajde 14 je nedeterminizmus na riadkoch 3,7,11, kde sa vyberalo z 2 pravidiel na expanziu neterminálu <príkazy>.
- Ako teda vytvoriť prakticky použiteľný (t.j. deterministický) top-down parser?



Syntaktický analyzátor jazyka LL(1)

Aby bol syntaktický analyzátor použiteľný v praxi, je potrebné, aby vedel **deterministicky** spracovať vstupné reťazce s **lineárnou** časovou zložitou.

T.j. musíme odstrániť nedeterminizmus z teoretického modelu.

Daný model analyzátoru obsahuje jediný nedeterminizmus - keď je potrebné expandovať neterminál A na vrchu zásobníka a príslušná gramatika obsahuje viacero pravidiel s neterminálom A na ľavej strane.

V praxi sa nedeterminizmus odstráni tak, že automat si príslušné pravidlo **vyberie** na základe **aktuálneho vstupného symbolu** t .

Takýto analyzátor má názov LL(1) = left-left-1. T.j. vstupný reťazec sa spracúva **zľava**, vytvára sa **ľavé** odvodenie a na rozhodovanie sa používa 1 vstupný symbol.



Syntaktický analyzátor jazyka LL(1)

Definícia

Nech $G = (N, T, P, S)$ je bezkontextová gramatika. **Rozkladová tabuľka syntaktického analyzátoru LL(1)** $RT[A, t]$ je zobrazenie $N \times T_\epsilon \rightarrow P$ popisujúce, ktoré pravidlo sa použije na expanziu neterminálu A nachádzajúceho sa na vrchu zásobníka, ak aktuálny vstupný symbol je t (ak $t = \epsilon$, potom bol prečítaný celý vstupný reťazec).

Rozkladová tabuľka je zvyčajne neúplné zobrazenie - ak neexistuje v tabuľke záznam pre aktuálny neterminál A a symbol t , znamená to **syntaktickú chybu**.



Syntaktický analyzátor jazyka LL(1) I

Vstup: Bezkontextová gramatika $G = (N, T, P, S)$, reťazec w , rozkladová tabuľka syntaktického analyzátoru LL(1) RT

Výstup: Zistenie, či $w \in L(G)$ a postupnosť pravidiel ľavého odvodu w

- 1: vlož do zásobníka S
- 2: $t \leftarrow$ prvá lexéma zo vstupu
- 3: **pokiaľ** zásobník nie je prázdny **rob**
- 4: $X \leftarrow$ symbol z vrchu zásobníka a zároveň ho odstráň zo zásobníka
- 5: **ak** $X \in N$ a zároveň $RT[X, t] = X \rightarrow \alpha$ **potom**
- 6: vlož do zásobníka α aby ľavý krajný symbol bol na vrchu
- 7: pošli na výstup pravidlo $X \rightarrow \alpha$
- 8: **inak**
- 9: **ak** $X \in T$ a zároveň $X = t$ **potom**
- 10: $t \leftarrow$ ďalšia lexéma zo vstupu
- 11: **inak**
- 12: spracuj syntaktickú chybu



Syntaktický analyzátor jazyka LL(1) II

- 13: **koniec ak**
- 14: **koniec ak**
- 15: **koniec pokiaľ**
- 16: akceptuj vstupné slovo



Ako vytvoriť rozkladovú tabuľku - idea

- Ak je na vrchu zásobníka neterminál A a na vstupe terminál t , pýtame sa, za akých podmienok použiť pravidlo $A \rightarrow \alpha$, aby sa automat nezasekol.
- Automat sa zasekne v momente, keď sa vstupný terminál nezhoduje s terminálom na vrchu zásobníka.
- Preto pri výbere pravidla do RT postupujeme tak, aby sme mali zabezpečené, že po expanzii neterminálu A zvoleným pravidlom $A \rightarrow \alpha$ bude vstupný symbol t rovný prvému terminálu, ktorý sa môže zjaviť na vrchu zásobníka.
- Ak expandujeme podľa pravidla $A \rightarrow \alpha$, tak na vrchu zásobníka sa môže zjaviť len terminál patriaci do množiny:
 - $FIRST(\alpha) - \{\varepsilon\}$
 - ak $\varepsilon \in FIRST(\alpha)$ tak $FOLLOW(A)$.



Ako vytvoriť rozkladovú tabuľku - množiny *PREDICT*

Definícia

Nech $G = (N, T, P, S)$ je redukovaná gramatika a $A \rightarrow \alpha \in P$. Potom:

$$\begin{aligned} PREDICT(A \rightarrow \alpha) &= \\ &= \begin{cases} FIRST(\alpha), & ak \ \varepsilon \notin FIRST(\alpha) \\ (FIRST(\alpha) - \{\varepsilon\}) \cup FOLLOW(A), & ak \ \varepsilon \in FIRST(\alpha) \end{cases} \end{aligned}$$

T.j. množina *PREDICT* signalizuje, ktoré pravidlo použiť na expanziu A .
Samozrejme, ak sa máme **jednoznačne** rozhodnúť, musí platiť: Ak $A \rightarrow \alpha_i$ a $A \rightarrow \alpha_j$ sú 2 rôzne A -pravidlá, potom:

$$PREDICT(A \rightarrow \alpha_i) \cap PREDICT(A \rightarrow \alpha_j) = \emptyset$$



Definícia

Redukovaná gramatika $G = (N, T, P, S)$ v ktorej všetky dvojice pravidiel s rovnakou ľavou stranou majú prázdny prienik množín *PREDICT*, sa nazýva **LL(1) gramatika**.

Ak je teda gramatika typu *LL(1)*, potom *RT* zostrojíme nasledovne:

Ak $t \in \text{PREDICT}(A \rightarrow \alpha)$, potom $RT[A, t]$ obsahuje pravidlo $A \rightarrow \alpha$.

Ak je gramatika typu *LL(1)*, každá položka rozkladovej tabuľky obsahuje maximálne jedno pravidlo a rozhodovanie o použití pravidla pri expanzii je jednoznačné.



Príklad zostrojenia RT

Zostrojte RT pre gramatiku:

1. $\langle \text{program} \rangle \rightarrow \mathbf{begin} \langle \text{príkazy} \rangle \mathbf{end}$
2. $\langle \text{príkazy} \rangle \rightarrow \langle \text{príkaz} \rangle ; \langle \text{príkazy} \rangle$
3. $\langle \text{príkazy} \rangle \rightarrow \varepsilon$
4. $\langle \text{príkaz} \rangle \rightarrow \mathbf{p}$

Najprv zostrojíme množiny *FIRST* a *FOLLOW*:

FIRST	$\langle \text{program} \rangle$	$\langle \text{príkazy} \rangle$	$\langle \text{príkaz} \rangle$	begin	end	;	p
Celkom	begin	p, ε	p	begin	end	;	p

FOLLOW	$\langle \text{program} \rangle$	$\langle \text{príkazy} \rangle$	$\langle \text{príkaz} \rangle$
Celkom:	ε	end	;



Príklad zostrojenia RT (pokr.)

Výsledné množiny *PREDICT* pre pravidlá:

- $PREDICT(1.) = FIRST(\mathbf{begin}\langle\text{príkazy}\rangle\mathbf{end}) = \{\mathbf{begin}\}$.
- $PREDICT(2.) = FIRST(\langle\text{príkaz}\rangle;\langle\text{príkazy}\rangle) = \{\mathbf{p}\}$
- $PREDICT(3.) = FOLLOW(\langle\text{príkazy}\rangle) = \{\mathbf{end}\}$
- $PREDICT(4.) = FIRST(\mathbf{p}) = \{\mathbf{p}\}$.

T.j. rozkladová tabuľka:

RT	begin	end	;	p	ϵ
$\langle\text{program}\rangle$	1.				
$\langle\text{príkazy}\rangle$		3.		2.	
$\langle\text{príkaz}\rangle$				4.	

Keďže každá položka obsahuje max. 1 pravidlo, gramatika je naozaj LL(1) gramatika.



Príklad zostrojenia RT

Zostrojte RT pre gramatiku:

1. $S \rightarrow SaSbS$
2. $S \rightarrow SbSaS$
3. $S \rightarrow \varepsilon$

Najprv zostrojíme množiny *FIRST* a *FOLLOW*:

FIRST	S	a	b
Celkom	$\{\varepsilon, a, b\}$	$\{a\}$	$\{b\}$

FOLLOW	S
Celkom:	$\{\varepsilon, a, b\}$



Príklad zostrojenja RT (pokr.)

Výsledné množiny *PREDICT* pre pravidlá:

- $PREDICT(1.) = FIRST(SaSbS) = \{a, b\}$.
- $PREDICT(2.) = FIRST(SbSaS) = \{a, b\}$
- $PREDICT(3.) = (FIRST(\varepsilon) \setminus \{\varepsilon\}) \cup FOLLOW(S) =$
 $= FOLLOW(S) = \{a, b, \varepsilon\}$

T.j. rozkladová tabuľka:

RT	<i>a</i>	<i>b</i>	ε
<i>S</i>	1., 2., 3.	1., 2., 3.	3.

Keďže existujú položky s viac ako 1 pravidlom, gramatika **nie je** LL(1) gramatika.



Poznámky

- LL(1) analyzátory/jazyky sa dajú rozšíriť na rozhodovanie o expanzii na základe k vstupných symbolov, $k \geq 1$. Vtedy hovoríme o $LL(k)$ gramatikách/analyzátoroch.
- Problémom je, že pre niektoré bezkontextové gramatiky sa nedá zostrojiť ani $LL(1)$ analyzátor, ani $LL(k)$ analyzátor.
- V niektorých prípadoch je riešením pokúsiť sa nájsť ekvivalentnú gramatiku pre daný jazyk, ktorá $LL(1)$ je.



Príklad

Gramatika, ktorá generuje jazyk $\{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$:

1. $S \rightarrow SaSbS$
2. $S \rightarrow SbSaS$
3. $S \rightarrow \varepsilon$

nie je $LL(1)$

Ten istý jazyk, t.j. $\{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$, sa však predsa len dá generovať *LL*(1) gramatikou, konkrétne:

1. $S \rightarrow aBS \mid bAS$
2. $A \rightarrow a \mid bAA$
3. $B \rightarrow b \mid aBB$

