Cvičenie 1 - Formálne jazyky a gramatiky

Ing. Viliam Hromada, PhD.

A-413
Ústav informatiky a matematiky
FEI STU

viliam.hromada@stuba.sk

Úloha č. 1

Nech abeceda $\Sigma = \{a, b, c\}$, nech reťazce nad touto abecedou sú $x = aa, y = \varepsilon, z = abc$.

1. Napíšte obrátený reťazec x^R, y^R, z^R a ich dĺžky
2. Napíšte výsledok zreťazenia $xx, xy, xz, yx, yy, yz, zx, zy, zz$ a ich dĺžky
3. Napíšte x^i, y^i, z^i, pre $i \in \{0, 1, 2, 3\}$ a ich dĺžky
4. Vymenujte reťazce patriace do Σ^*, Σ^+ dĺžok aspoň 2.

Úloha č. 1 - odpovede

1. $x^R = aa, y^R = \varepsilon, z^R = cba$. $|x^R| = |x| = 2, |y^R| = |y| = 0, |z^R| = |z| = 3$.
2. $xx = aaaa, xy = aa, xz = aaabc, yx = aa, yy = \varepsilon, yz = abc, zx = abcaa, zy = abc, zz = abcabc$.

   Zrejme $|uv| = |u| + |v|$, t.j.:
   $|xx| = 2|x| = 4, |xy| = |x| + |y| = 2, |xz| = |x| + |z| = 2 + 3 = 5, |yx| = 2, |yy| = |y| + |y| = 0, |yz| = 3, |zx| = 5, |zy| = 3, |zz| = 6$.
3. $x^0 = \varepsilon, x^1 = aa, x^2 = aaaa, x^3 = aaaaaa, y^0 = \varepsilon, y^1 = \varepsilon, y^2 = \varepsilon, y^3 = \varepsilon, z^0 = \varepsilon, z^1 = abc, z^2 = abcabc, z^3 = abcabcabc$

   Zrejme $|u^i| = i|u|$, t.j. $|x^0| = 0|x| = 0, |x^2| = 2|x| = 4, |x^3| = 3|x| = 6, \ldots, |y^2| = 2|y| = 0, \ldots, |z^3| = 3|z| = 9$
4. $\Sigma^* = \{\varepsilon, a, b, c, aa, ab, ac, ba, bb, bc, ca, cb, cc, \ldots\}$

   $\Sigma^+ = \{a, b, c, aa, ab, ac, ba, bb, bc, ca, cb, cc, \ldots\} = \Sigma^* \setminus \{\varepsilon\}$.

Úloha č. 2

Uvažujme abecedu $\Sigma = \{a, b, c\}$ a nasledovné jazyky nad touto abecedou:

- $L_1 = \{c\}$
- $L_2 = \{aaa, aba, aab, abb, baa, bab, bba, bbb\}$
- $L_3 = \{aw \mid w \in \{a, b, c\}^*\}$
- $L_4 = \{wcw^R \mid w \in \{a, b\}^*\}$

1. Ktoré jazyky L_1, L_2, L_3, L_4 sú konečné a ktoré sú nekonečné? Popíšte slovne jazyky L_3, L_4
2. Uveďte jazyk, ktorý vznikne ako zreťazenie $L_1L_1, L_1L_2, L_2L_1, L_2L_2$
3. Popíšte jazyk $L_1^2, L_1^3, L_1^*, L_1^+, L_2^*, L_3^C, L_4^2$
4. Popíšte jazyk $L_2 \cap L_3, L_2 \cap L_4, L_3 \cap L_4, L_4 \setminus L_1, L_1 \cup L_1^C$

Úloha č. 2 - odpovede

1. L_1, L_2 sú konečné jazyky, pretože $|L_1| = 1, |L_2| = 8$. Jazyky L_3, L_4 sú nekonečné.
   $L_3 = \{a, aa, ab, ac, aaa, aab, aac, aba, abb, abc, aca, acb, acc, aaaa, ...\}$ t.j. všetky reťazce začínajúce a

   $L_4 = \{c, aca, bcb, aacaa, abcba, bacab, bbcbb, aaacaaa, aabcbaa, abacaba, ...\}$ t.j. všetky **palindrómy** nepárnej dlžky, ktoré majú uprostred symbol c a ostatné symboly môžu byť len a a b.
2. $L_1L_1 = \{cc\}$, $L_1L_2 = \{caaa, caba, caab, cabb, cbaa, cbab, cbba, cbbb\}$,
   $L_2L_1 = \{aaac, abac, aabc, abbc, baac, babc, bbac, bbbc\}$,
   $L_2L_2 = \{aaaaaa, aaaaba, aaaaab, ..., aaabbb, abaaaa, abaaba, abaaab, ..., ababbb, ..., bbbaaa, ..., bbbbbb\}$

Úloha č. 2 - odpovede

3. $L_1^2 = L_1 L_1 = \{cc\}, L_1^3 = L_1 L_1^2 = \{ccc\}, L_1^* = \{c\}^* = \{\varepsilon, c, cc, ccc, cccc, ...\} = \{c^i \mid i \in \{0, 1, 2, 3, ...\}\}, L_1^+ = \{c, cc, ccc, cccc, ...\} = \{c^i \mid i \in \{1, 2, 3, ...\}\}$
   $L_2^* = \{\varepsilon, aaa, aba, ..., bbb, aaaaaa, ..., aaaaaaaaa, ..., abbbaabbabbb, ...\}$

   $L_3^C = \Sigma^* \setminus L_3$, t.j. všetky reťazce **nezačínajúce** a-čkom, t.j.
   $L_3^C = \{\varepsilon\} \cup \{bw \mid w \in \{a, b, c\}^*\} \cup \{cw \mid w \in \{a, b, c\}^*\} =$
   $L_3^C = \{\varepsilon, b, c, ba, bb, bc, ca, cb, cc, baa, ...\}$

   $L_4^2 = L_4 L_4 = \{cc, caca, acac, cbcb, bcbc, acaaca, acabcb, ...\} = \{xy \mid x = wcw^R, y = ucu^R, w \in \{a, b\}^*, u \in \{a, b\}^*\}$.

Úloha č. 2 - odpovede

4. $L_2 \cap L_3 = \{aaa, aba, aab, abb\}$

   $L_2 \cap L_4 = \emptyset$

   $L_3 \cap L_4 = \{aca, aacaa, abcba, aaacaaa, abacaba, abbcbba, ...\} =$ $\{wcw^R \mid w = ax, x \in \{a, b\}^*\}$, t.j. $L_3 \cap L_4$ sú palindrómy nepárnej dĺžky majúce uprostred c, pričom ostatné symboly môžu byť len a a b a navyše začínajú a-čkom.

   $L_4 \setminus L_1 = \{aca, bcb, aacaa, ...\}$ $L_3 \cap L_4$ sú palindrómy nepárnej dĺžky aspoň 3 majúce uprostred c, pričom ostatné symboly môžu byť len a a b

   $L_1 \cup L_1^C = \Sigma^*$.

Úloha č. 3

Je daná gramatika $G = (N, T, P, S)$, kde $N = \{S, A\}$, $T = \{0, 1\}$, počiatočný neterminál je S, s pravidlami P:

$S \rightarrow 0A$

$A \rightarrow 0A|1A|0|1$

a) určte typ gramatiky (regulárna, bezkontextová, kontextová, frázová)

b) nájdite odvodenie slov 0101, 0111, 1000 v danej gramatike.

c) Určte, aký jazyk gramatika generuje (slovne, formálnym zápisom).

Úloha č. 3 - odpovede

a) Gramatika je **regulárna** (technicky je aj bezkontextová, kontextová a frázová)

b) $S \Rightarrow 0A \Rightarrow 01A \Rightarrow 010A \Rightarrow 0101$, teda $S \Rightarrow^* 0101$.

b) $S \Rightarrow 0A \Rightarrow 01A \Rightarrow 011A \Rightarrow 0111$, teda $S \Rightarrow^* 0111$.

b) $S \Rightarrow 0A \Rightarrow \ldots \Rightarrow 1000$ sa zjavne nedá, teda $S \not\Rightarrow^* 1000$

c) Vidíme, že gramatika generuje také reťazce z núl a jednotiek, ktoré začínajú nulou a sú dĺžky aspoň 2. Preto by sme mohli písať, že jazyk generovaný gramatikou G je:

$$L(G) = \{0w \mid w \in \{0,1\}^+\}$$

teda, že na začiaku obsahujú 0 a za ňou je reťazec z núl a jednotiek dĺžky aspoň 1, teda dokopy tvoria reťazce z 0 a 1 začínajúce nulou dĺžky aspoň 2.

Úloha č. 4

Je daná gramatika popisujúca syntax podmieneného príkazu $G = (N, T, P, S)$ s pravidlami P:

1. <príkaz> $\rightarrow$ **p1**
2. <príkaz> $\rightarrow$ **p2**
3. <príkaz> $\rightarrow$ **if**<podmienka>**then**<príkaz><else časť>
4. <else časť> $\rightarrow$ **else**<príkaz>
5. <else časť> $\rightarrow$ ε
6. <podmienka> $\rightarrow$ **cond**

Neterminály $N = \{$<príkaz>, <else časť>, <podmienka>$\}$ terminály $T = \{$**p1**, **p2**, **if**, **then**, **else**, **cond**$\}$, $S =$ <príkaz>

Úloha č. 4

Zistite, či nasledovné "fragmenty kódu" spĺňajú uvedenú syntax (t.j. či **majú v gramatike odvodenie**):

1. **p1**
2. **p1 p2**
3. **if cond then p1**
4. **if cond then p1 else**
5. **if cond then p1 else if cond then p1 else p2**

Úloha č. 4 - odpovede

Riešenie:

1. <príkaz> $\Rightarrow$ **p1**
2. <príkaz> $\not\Rightarrow^*$ **p1 p2**
3. <príkaz> $\Rightarrow$ **if**<podmienka>**then**<príkaz><else časť> $\Rightarrow$ **if cond then** <príkaz><else časť> $\Rightarrow$ **if cond then p1** <else časť> $\Rightarrow$ **if cond then p1**
4. <príkaz> $\Rightarrow$ **if**<podmienka>**then**<príkaz><else časť> $\Rightarrow$ **if cond then** <príkaz><else časť> $\Rightarrow$ **if cond then p1** <else časť> $\Rightarrow$ **if cond then p1 else**<príkaz> $\not\Rightarrow$ **if cond then p1 else**

   (pretože z neterminálu <príkaz> nie je možné odvodiť len ε)

Úloha č. 4 - odpovede

Riešenie pre **if cond then p1 else if cond then p1 else p2**:

<príkaz> $\Rightarrow$ **if**<podmienka>**then**<príkaz><else časť> $\Rightarrow$
if cond then <príkaz><else časť> $\Rightarrow$ **if cond then p1** <else časť> $\Rightarrow$
if cond then p1 else<príkaz> $\Rightarrow$
if cond then p1 else if<podmienka>**then**<príkaz><else časť> $\Rightarrow$
if cond then p1 else if cond then<príkaz><else časť> $\Rightarrow$
if cond then p1 else if cond then p1 <else časť> $\Rightarrow$
if cond then p1 else if cond then p1 else <príkaz> $\Rightarrow$
if cond then p1 else if cond then p1 else p2

Úloha č. 5

Je daná gramatika $G = (N, T, P, S)$, kde $N = \{X, Y, Z\}$, $T = \{a, b, c\}$, počiatočný neterminál je $S = X$, s pravidlami P:

$X \to aYb$

$aY \to bYb$

$Y \to Zc$

$Z \to aZ \mid b$

a) určte typ gramatiky (regulárna, bezkontextová, kontextová, frázová)

b) nájdite odvodenie slova *bbcbb* v danej gramatike.

c) Určte, aký jazyk gramatika generuje (slovne, formálnym zápisom).

Úloha č. 5 - odpovede

a) Gramatika je **kontextová** (technicky je aj frázová)

b) $X \Rightarrow aYb \Rightarrow bYbb \Rightarrow bZcbb \Rightarrow bbcbb$, teda $X \Rightarrow^* bbcbb$.

c) Vidíme, že gramatika generuje reťazce 2 typov (podľa toho, či sa použije pravidlo $aY \to bYb$):

- Ak sa pravidlo nepoužije, potom reťazce majú prefix a, suffix cb a medzi nimi podreťazec tvaru $a^* b$.
- Ak sa pravidlo použije, potom reťazce majú prefix b, suffix cbb a medzi nimi podreťazec $a^* b$.

$$L(G) = \{aa^* bcb\} \cup \{ba^* bcbb\}$$

Úloha č. 6

Uvažujte abecedu $A = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, -\}$. Navrhnite gramatiku, ktorá generuje celočíselné konštanty (t.j čísla z množiny $\mathbb{Z}$), t.j. reťazce:

- 0, 1, 2, 3, 4, ..., 9, 10, 11, 12, 13, ...
- $-1, -2, -3, -4, \ldots -9, -10, \ldots$

Úloha č. 6 - riešenie

Určite terminálne symboly gramatiky budú vlastne abeceda, t.j. $T = A = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, -\}$. Navrhneme sústavu pravidiel, ktoré tieto čísla generujú:

- $S \to -A|0|1|2|3|4|5|6|7|8|9|1B|2B|3B|4B|5B|6B|7B|8B|9B$
- $A \to 1|2|3|4|5|6|7|8|9|1B|2B|3B|4B|5B|6B|7B|8B|9B$
- $B \to 0|1|2|3|4|5|6|7|8|9|1B|2B|3B|4B|5B|6B|7B|8B|9B|0B$

Teda neterminály $N = \{S, A, B\}$ a S je počiatočný neterminál. Všimnite si, že naša gramatika:

- Negeneruje také čísla, ktoré by začínali nulami - s výnimkou len nuly samotnej.
- Negeneruje napr. -0, čiže nula je vždy bez znamienka.
- Je regulárna!

Úloha č. 7

Nájdite gramatiku, ktorá generuje také reťazce zo symbolov $\{a, b\}$ (t.j. nad abecedou $A = \{a, b\}$), v ktorých je rovnaký počet symbolov a a b, t.j. gramatiku, ktorá generuje jazyk:

$$L = \{w \in \{a, b\}^* \mid \sharp_a(w) = \sharp_b(w)\},$$

kde $\sharp_a(w)$ označuje počet symbolov a v reťazci w (analogicky $\sharp_b(w)$)

Úloha č. 7 - zamyslenie

Aké reťazce patria do jazyka L?

- $\{\varepsilon\}$, lebo$\sharp_a(\varepsilon) = \sharp_b(\varepsilon) = 0$
- $\{ab, ba\}$, lebo pre ne $\sharp_a(ab) = \sharp_b(ab) = 1$, resp. $\sharp_a(ba) = \sharp_b(ba) = 1$
- $\{aabb, abab, abba, baba, baab, bbaa\}$, lebo sa v nich a aj b vyskytuje 2krát
- ...

Aké reťazce nepatria do jazyka L?

- $\{a, aab, aba, baa, aa, aaab, aaba, ...\}$ lebo je v nich viac a než b
- $\{b, bba, bab, abb, bb, bbba, bbab, ...\}$ lebo je v nich viac b než a

Úloha č. 7 - pokus o riešenie

Prvý pokus o riešenie - nech gramatika G by bola:

- $S \to aS \mid bS \mid \varepsilon$

Táto gramatika určite dokáže generovať také reťazce, v ktorých je počet a a b rovnaký. Totižto, táto gramatika generuje **všetky reťazce** z písmen $\{a, b\}$ a určite platí, že $L \subseteq L(G)$.

Problém je, že generuje aj reťazce **navyše**, napr. a alebo aaa, teda také, ktoré **nepatria do** L, lebo v nich nie je rovnaký počet a a b. Teda $L(G) \not\subseteq L$. Táto gramatika **nie je správna.**

Úloha č. 7 - pokus o riešenie

Druhý pokus o riešenie - nech gramatika G by bola:

1. $S \rightarrow abS$
2. $S \rightarrow \varepsilon$

Táto gramatika funguje tak, že každá aplikácia pravidla č. 1 vyrobí vo výslednom reťazci časť *ab*, a teda určite **všetko**, čo G odvodí má tú vlastnosť, že počet *a* a *b* je rovnaký, teda $L(G) \subseteq L$.

Problém je, že **nie každý reťazec** s rovnakým počtom *a* a *b* má v danej gramatike odvodenie, napr. *abba*, sa v nej odvodiť nedá! Preto $L \not\subseteq L(G)$. Táto gramatika **nie je správna.**

Úloha č. 7 - pokus o riešenie

Tretí pokus o riešenie - nech gramatika G by bola:

- $S \rightarrow SaSbS \mid SbSaS \mid \varepsilon$

Táto gramatika má nasledovnú logiku. Uvažujme reťazec s rovnakým počtom a a b. Potom v ňom nastane jedna z 2 situácií:

- Alebo v ňom **musia existovať**, také a a b, že a je pred b a navyše platí, že aj v podreťazci pred a je rovnaký počet a, b; v podreťazci medzi a, b je rovnaký počet a, b; v podreťazci za b je rovnaký počet a, b.

  abbaaabbba

- Alebo v ňom **musia existovať**, také b a a, že b je pred a a navyše platí, že aj v podreťazci pred b je rovnaký počet a, b; v podreťazci medzi b, a je rovnaký počet a, b; v podreťazci za a je rovnaký počet a, b.

  bababa

Úloha č. 7 - pokus o riešenie

Pre gramatiku G

- $S \rightarrow SaSbS \mid SbSaS \mid \varepsilon$

teda platí:

- Určite všetko čo generuje sú reťazce, ktoré majú rovnaký počet a a b, teda $L(G) \subseteq L$.
- Nebudeme to dokazovať, ale platí aj, že **každý** reťazec s rovnakým počtom a a b má v gramatike odvodenie, teda $L \subseteq L(G)$.
- A teda $L = L(G)$, čiže **naša gramatika je správna.**
- Naša gramatika je **bezkontextová.**

Pre úplnosť v našej gramatike G: $N = \{S\}$, $T = \{a, b\}$.

Úloha č. 8

Príklad: Nájdite gramatiku generujúcu všetky "správne" ozátvorkované výrazy, t.j.:

$$L = \{\varepsilon, (), (()), ()(), ((())), (())(), ()()(), ...\}.$$

t.j. ku každej ľavej zátvorke existuje pravá zátvorka.

Musí platiť, že $L(G) = L$, t.j.:

- $L(G) \subseteq L$, (gramatika negeneruje "**nič navyše**")
- $L \subseteq L(G)$. (každé slovo z jazyka L sa dá v gramatike **odvodiť**)

Ak platia oba body, z toho logicky vyplýva, že $L(G) = L$.

Príklad (pokr.) Nech gramatika je napr.: $G = (\{S, A\}, \{(,)\}, P, S)$.
Pravidlá:

- $S \to AS$
- $S \to \varepsilon$
- $A \to (S)$.

Je zrejmé, že gramatika generuje správne ozátvorkované výrazy, t.j. $L(G) \subseteq L$ a **nič navyše**. Otázkou ale zostáva, či generuje **všetky** správne ozátvorkované výrazy, resp. či sa dá každý správne ozátvorkovaný výraz vygenerovať pomocou danej gramatiky (t.j. či $L \subseteq L(G)$).

Príklad (pokr.) Dôkaz matematickou indukciou: Uvažujme "stupeň vnorenia" l:

- $l = 0$: ε
- $l = 1$: $(), ()(), ...$
- $l = 2$: $(()), (())(), ...$

1. Pre $l = 0$ je triviálne ukázať, že gramatika takéto výrazy generuje.
2. Predpokladajme, že gramatika generuje výrazy pre $l \leq k$. Majme výraz w so stupňom vnorenia $k + 1$. Musíme dokázať, že aj takýto výraz gramatika vie vygenerovať.

Príklad (pokr.) Výraz w si rozdelíme na správne-ozátvorkované segmenty: $w = w_1 w_2 \ldots w_n$, z ktorých aspoň 1 má stupeň vnorenia $k+1$. Predpokladajme, že je to w_1. Slovo w vieme odvodiť:

$$S \Rightarrow AS \Rightarrow \ldots \Rightarrow \underbrace{A \ldots A}_{n} \Rightarrow (S)\underbrace{A \ldots A}_{n-1} \Rightarrow \ldots \Rightarrow w_1 \underbrace{A \ldots A}_{n-1} \Rightarrow \ldots \Rightarrow w_1 w_2 \ldots w_n.$$

Odstránením "vonkajšej" zátvorky (keď sme použili pravidlo $A \rightarrow (S)$) sme dostali výraz so stupňom vnorenia k, ktorý podľa indukčného predpokladu odvodiť vieme. Vieme teda odvodiť aj výraz so stupňom vnorenia $k+1$.

Samozrejme, toto nemusí byť **jediná** gramatika, ktorá daný jazyk generuje.

Úloha č. 9

Nájdite gramatiku pre jazyk $L = \{a^n b^n c^n \mid n \in \{1, 2, 3, 4, ...\}\}$.

Úloha č. 9 - zamyslenie

Reťazce patriace od jazyka L:

1. abc
2. $aabbcc$
3. $aaabbbccc$
4. $aaaabbbbcccc$

Teda reťazce majú **rovnaký počet** a, b, c a zároveň sú symboly **usporiadané**, t.j. najprv a, potom b a nakoniec c. Gramatika teda musí zabezpečiť, aby sa pri generovaní **sledoval počet symbolov**, aby bola zaručená ich rovnosť po vygenerovaní reťazca terminálov. Navyše musí zaručiť ich usporiadanie.

Úloha č. 9 - pokus o riešenie

Skúsme navrhnúť nasledovné pravidlá:

1. $S \rightarrow aSBC$
2. $S \rightarrow aBC$

Naša gramatika vie teraz generovať takúto vetnú formu:

$$S \Rightarrow^n a^n(BC)^n = a^nBC(BC)^{n-1}$$

Ak by sme v nej nahradili každé B terminálom b a C terminálom c, dostali by sme určite reťazec s rovnakým počtom a, b, c, avšak **nie usporiadaný**! Preto využijeme **možnosti kontextových gramatík**, ktoré dovoľujú meniť symboly vo vetných formách.

Úloha č. 9 - pokus o riešenie

Doplňme pravidlo č. 3:

1. $S \to aSBC$
2. $S \to aBC$
3. $CB \to BC$

Toto pravidlo umožní "preusporiadať" neterminály B, C vo vetnej forme tak, aby sme vedeli dostať:

$$S \Rightarrow^n a^n(BC)^n = a^nBC(BC)^{n-1} \Rightarrow^* a^nB^nC^n$$

Teraz môžeme pristúpiť k prepisovaniu neterminálov na terminály - ale spravíme to špeciálnym spôsobom.

Úloha č. 9 - pokus o riešenie

1. $S \to aSBC$
2. $S \to aBC$
3. $CB \to BC$
4. $aB \to ab$
5. $bB \to bb$
6. $bC \to bc$
7. $cC \to cc$

Teraz už vieme každú požadovanú deriváciu úspešne dokončiť:

$$S \Rightarrow^n a^n(BC)^n = a^nBC(BC)^{n-1} \Rightarrow^* a^nB^nC^n \Rightarrow^n a^nb^nC^n \Rightarrow^n a^nb^nc^n$$

Pre úplnosť v gramatike $N = \{S, A, B, C\}$, $T = \{a, b, c\}$, S je počiatočný neterminál

Úloha č. 9 - upozornenie

POZOR!!! V predchádzajúcom slajde **nesmieme len pridať pravidlá** $B \to b$, $C \to c$, t.j.

1. $S \to aSBC$
2. $S \to aBC$
3. $CB \to BC$
4. $B \to b$
5. $C \to c$

Lebo by sme vedeli urobiť napríklad deriváciu:

$$S \Rightarrow aSBC \Rightarrow aaBCBC \Rightarrow^* aabcbc$$

čo nie je reťazec patriaci do daného jazyka L.