2. opravná písomka z Formálnych Jazykov a Automatov

myšlienky riešení ©MišoF.

Celý materiál je neoficiálny, ak tu mám chybu ja, vaše riešenie s rovnakou chybou **nie je** správne!

$\boxed{1}$ Formálne definujte *zbytočný neterminál* v bezkontextovej gramatike. (Sú ich dva druhy – nedosiahnuteľné a také, z ktorých sa nedá odvodiť žiadna vetná forma obsahujúca len terminály.) K danej bezkontextovej gramatike G zostrojte štandardnou konštrukciou ekvivalentnú *redukovanú* (t.j. bez zbytočných neterminálov) bezkontextovú gramatiku G'.

$G = (N, T, P, \sigma)$, $\quad N = \{\sigma, \alpha, \beta, \gamma, \delta, \varphi, \psi\}$, $\quad T = \{x, y\}$

$$P = \left\{ \begin{array}{llll} \sigma \to x^{47} \mid \gamma \mid \psi y, & \alpha \to \psi\gamma\beta\delta \mid \alpha \mid \varepsilon, & \beta \to xy\beta \mid xyyx \mid \sigma\psi \mid \varphi, & \\ \gamma \to x\delta \mid y\psi \mid x\sigma \mid y\varphi, & \delta \to x\delta \mid \psi y\beta, & \varphi \to x \mid y\sigma y, & \psi \to \psi^{42} \mid \delta\sigma\delta \end{array} \right\}$$

Nedosiahnuteľný neterminál v gramatike $G = (N, T, P, \sigma)$ je každý neterminál ρ taký, že neexistujú žiadne $u, v \in (N \cup T)^*$ také, že $\sigma \Rightarrow_G^* u\rho v$. *Neukončiteľný* neterminál v gramatike $G = (N, T, P, \sigma)$ je každý neterminál ρ taký, že neexistuje žiadne $u \in T^*$ také, že $\rho \Rightarrow_G^* u$. *Zbytočné* neterminály gramatiky sú práve všetky jej nedosiahnuteľné neterminály a všetky jej neukončiteľné neterminály.
Treba si uvedomiť nasledovné dve skutočnosti:

- Tým, že odstránime nedosiahnuteľné neterminály, nám nevzniknú žiadne nové neukončiteľné neterminály – keďže tie, čo sme odstránili, sa nedali dosiahnuť, nijako nemohli ovplyvniť odvodenia z ostatných neterminálov.
- Tým, že odstránime neukončiteľné neterminály nám **môžu** vzniknúť nové nedosiahnuteľné neterminály. Príklad: gramatika zo zadania :-)

Preto v štandardnej konštrukcii **najskôr** odstránime neukončiteľné neterminály a **potom** nedosiahnuteľné (čím už nič nepokazíme a môžeme skončiť).
Samotná štandardná konštrukcia pre neukončiteľné neterminály: zostrojeme postupne množiny $M_1, M_2, \ldots$, pričom v M_i sú tie neterminály, z ktorých sa dá na i krokov vyrobiť (nejaké, aspoň 1) terminálne slovo. Keď pre nejaké i dostaneme $M_i = M_{i-1}$, vieme, že nám žiaden nový už nepribudne – všetky ostatné sú neukončiteľné. Nedosiahnuteľné nájdeme analogicky (P_i je množina tých, ktoré vieme vo vetnej forme dostať po najviac i krokoch odvodenia). Detaily sú v skriptách (snáď).
Vo výslednej gramatike mali zostať neterminály σ, γ a φ.

$\boxed{2}$ Rozhodnite, či je jazyk

$$L = \left\{ uvx \;\middle|\; u, x \in \{a,b\}^* \;\wedge\; \#_a(ux) = \#_b(ux) \;\wedge\; v \in \{a,b,c\}^* \;\wedge\; v = v^R \right\}$$

bezkontextový. Ak nie, dokážte. Ak áno, zostrojte bezkontextovú gramatiku G, ktorá ho bude generovať. V tomto prípade nezabudnite na popis množiny pravidiel a myšlienku dôkazu $L = L(G)$.

Gramatika s pravidlami $\{\sigma \to \sigma a\sigma b\sigma \mid \sigma b\sigma a\sigma \mid \varepsilon\}$ generuje práve všetky slová w také, že $\#_a(w) = \#_b(w)$. (Dokázalo by sa to podobne ako v druhom príklade 1. opravnej písomky.) My by sme potrebovali, aby ešte niekde vo vetnej forme **raz** vygenerovala nejaký nový neterminál, z ktorého potom vygenerujeme palindróm (slovo v také, že $v = v^R$). Aby sme ho vygenerovali práve raz, musíme si "v neterminále σ" pamätať, či ho ešte treba vygenerovať alebo už nie. Budeme mať teda namiesto σ neterminály $\boxed{1}$ (ešte treba) a $\boxed{0}$ (už netreba).
Výsledná gramatika:

$G = (N, T, P, \boxed{1})$, $\quad N = \{\boxed{0}, \boxed{1}, \beta\}$, $\quad T = \{a, b, c\}$, $\quad P = \Big\{$

$\boxed{0} \to \boxed{0}a\boxed{0}b\boxed{0} \mid \boxed{0}b\boxed{0}a\boxed{0} \mid \varepsilon$

$\boxed{1} \to \boxed{u}a\boxed{v}b\boxed{x} \mid \boxed{u}b\boxed{v}a\boxed{x} \qquad (\forall u, v, x \in \mathbb{N};\ x + y + z = 1)$

$\boxed{1} \to \beta$

$\beta \to x\beta x \mid x \mid \varepsilon \qquad (\forall x \in \{a, b, c\}) \Big\}$

Myšlienka dôkazu. Indukciou od dĺžky odvodenia dokážeme, že z β sa dajú vygenerovať práve všetky palindrómy. Dokážeme, že pôvodná gramatika generuje práve všetky slová také, že $\#_a(w) = \#_b(w)$. Indukciou od dĺžky odvodenia dokážeme, že keď zoberieme ľubovoľnú vetnú formu našej gramatiky (ktorú sme dostali bez použitia pravidiel $\beta \to \dots$) a namiesto $\boxed{0}$, $\boxed{1}$ napíšeme σ, buď dostaneme vetnú formu pôvodnej gramatiky, alebo tam bude navyše práve 1 β. Z toho už vyplýva, že keď nám ostane jediný neterminál β (ten si môžeme nechať "na koniec", v bezkontextovej gramatike nazáleží na poradí prepisovania neterminálov), zvyšok vetnej formy obsahuje rovnako a a b. Z β potom vygenerujeme nejaký palindróm, čím dostaneme slovo z L, teda $L(G) \subseteq L$.
Ak sa v našej gramatike dá odvodiť w také, že $\#_a(w) = \#_b(w)$, tak sa dajú odvodiť aj všetky vetné formy, v ktorých terminály sú práve w a medzi každými dvoma terminálmi (aj na okrajoch) je práve jeden neterminál, pričom "súčet" neterminálov je 1 (opäť indukciou). Ak máme našou gramatikou vygenerovať slovo $uvx \in L$, vygenerujeme najskôr takúto vetnú formu, ktorej terminály budú tvoriť ux a neterminál $\boxed{1}$ bude medzi u a x. Potom všetky $\boxed{0}$ prepíšeme na ε, $\boxed{1}$ na β a z β vygenerujeme v. Preto $L(G) \supseteq L$.

$\boxed{3}$ Zapíšte formálne nasledovné výroky (použite značenie z prednášok). Dokážte alebo vyvráťte výrok d).

a) Automat A pri výpočte na slove w v prvom kroku číta a, zmení pri ňom stav z q_0 na q_1 a po nenulovom konečnom počte krokov skončí v stave q_f.
b) V gramatike G máme pravidlo "σ sa môže prepísať na $a\beta$", preto sa z vetnej formy $b\sigma$ dá na 1 krok odvodiť vetná forma $ba\beta$.
c) Ak pre jazyk L neplatí pumpovacia lema pre reg. jazyky (rozpísať znegovanú!), tak nie je regulárny.
d) *Zjednotenie konečného počtu regulárnych jazykov je regulárny jazyk, zjednotenie nekonečného počtu môže a nemusí byť regulárny jazyk.*
e) L je jazyk všetkých dobre uzátvorkovaných výrazov nad abecedou $\big\{ (,), [,] \big\}$.

a) $\exists u \in \Sigma^*; w = au, \qquad (q_0, w) \vdash_A (q_1, u) \vdash_A^+ (q_f, \varepsilon)$
b) $G = (N, T, P, \sigma),\ (\sigma \to a\beta) \in P \qquad \Rightarrow \qquad b\sigma \Rightarrow_G ba\beta$
c) AK $\forall n \in \mathbb{N}; \exists w \in L; |w| > n$ také, že $\Big(\forall u, v, x \in \Sigma^*;$ ak $w = uvx \wedge |uv| \le n \wedge |v| \ge 1,$ tak $\exists i; uv^ix \notin L\Big)$, TAK $L \notin \mathcal{R}$. Časť medzi AK a TAK je práve to, čo máte o jazyku dokázať, ak chcete použitím pumpovacej lemy dokázať, že nie je regulárny.
d) $\forall k \in \mathbb{N};\ \forall L_1, \dots, L_k \in \mathcal{R};\ \bigcup_{i=1}^k L_i \in \mathcal{R}$
$\Big(\exists L_1, L_2, \dots \in \mathcal{R};\ \bigcup_{i=1}^\infty L_i \in \mathcal{R}\Big) \ \wedge\ \Big(\exists J_1, J_2, \dots \in \mathcal{R};\ \bigcup_{i=1}^\infty J_i \notin \mathcal{R}\Big)$
Tu ešte treba spomenúť, že formulácia $\forall L_1, L_2, \dots \in \mathcal{R};\ \bigcup_{i=1}^\infty L_i \in \mathcal{R}\ \vee\ \bigcup_{i=1}^\infty L_i \in \mathcal{R}$ nehovorí to, čo chceme. Hovorí o nejakých konkrétnych jazykoch (o všetkých $\bigcup_{i=1}^\infty L_i$ pre rôzne L_i), že každý z nich buď je regulárny alebo nie. Ale predsa **každý** jazyk buď je regulárny, alebo nie je! Naše tvrdenie hovorí: Môže nastať situácia, že to zjednotenie bude regulárne. Môže nastať (zjavne iná) situácia, že to zjednotenie nie je regulárne.
e) $L = \{w \mid w = \varepsilon\ \vee\ \exists u \in L; w = (u)\ \vee\ \exists u \in L; w = [u]\ \vee\ \exists u, v \in L; w = uv\}$
Všimnite si, že táto definícia je "rekurzívna" – pri rozhodnutí, či $w \in L$ potrebujeme vedieť, či iné slová patria do L. Tieto iné slová sú ale určite kratšie ako w, preto o každom slove vieme jednoznačne rozhodnúť, či do w patrí, preto je definícia korektná. O moc jednoduchšia definícia neexistuje. (Iný pohľad: L je jazyk generovaný gramatikou s pravidlami $\{\sigma \to (\sigma) \mid [\sigma] \mid \sigma\sigma \mid \varepsilon\}$, len to bolo treba opäť formálne zapísať.) Za $L = \{w \mid w \in \Sigma^* \wedge w$ je dobre uzátvorkovaný výraz$\}$ bolo pol boda :-).

Dôkaz d). Na prednáške bolo dokázané, že ak $L_1, L_2 \in \mathcal{R}$, tak aj $L_1 \cup L_2 \in \mathcal{R}$. Indukciou sa dá dokázať, že potom to platí aj pre ľubovoľný konečný počet jazykov. Alebo ak chceme explicitne – majme jazyky $L_1, \dots, L_n \in \mathcal{R}$, pre ne automaty $A_1, \dots, A_n$, potom automat pre $\bigcup L_i$ zostrojíme tak, že zoberieme všetky tieto automaty (BUNV nech majú disjunktné stavy), pridáme nový počiatočný stav a hrany z neho na ε do počiatočných stavov A_i. Že tento automat robí čo má je zjavné.
Pre nekonečné zjednotenie je to zaujímavejšie. Výsledný jazyk môže byť regulárny. Nech $\forall i;\ L_i = \emptyset \in \mathcal{R}$, potom $\bigcup_{i=1}^\infty L_i = \emptyset \in \mathcal{R}$. Ale nemusí. Nech $\forall i;\ L_i = \{a^ib^i\} \in \mathcal{R}$, potom $\bigcup_{i=1}^\infty L_i = \{a^nb^n \mid n > 0\} \notin \mathcal{R}$.
Presnejšie výsledný jazyk môže byť úplne ľubovoľný jazyk L. Zoraďme slová L podľa dĺžky a rovnako dlhé podľa abecedy. V tomto poradí ich označme $w_1, w_2, \dots$ (slov v jazyku je spočítateľne veľa, preto to ide). Položme $L_i = \{w_i\}$. Potom $\forall i;\ L_i \in \mathcal{R}$ a zároveň $\bigcup_{i=1}^\infty L_i = L$.

4

Odpovede vyznačte krížikmi do príslušných políčok, nemusíte ich zdôvodňovať. Bodovanie: správna odpoveď +1, nesprávna -1, bez odpovede 0. Ak budete mať záporný súčet, získavate za túto úlohu 0 bodov.

	áno	**nie**
Existuje rekurzívny jazyk, ktorý je regulárny.	x	
Ku každej regulárnej gramatike G existuje ekvivalentná regulárna gramatika G', ktorej pravidlá sú podmnožinou $N \times (T \cup NN \cup \{\varepsilon\})$.		x
Pre každý homomorfizmus h a jazyk L je $h(h^{-1}(L)) \supseteq L$.		x
Jazyk $L = \{a^m b^n c^n \mid m, n \geq 0\} \cup \{b, c\}^*$ sa dá pumpovať (podľa pump. lemy pre R).	x	
$\mathcal{R}$ je najmenšia trieda jazykov, obsahujúca všetky konečné jazyky a uzavretá na $\cup$, $\cap$, $\cdot$ a h.		x
Všetky slová v Σ^* sú konečne dlhé.	x	
$\forall A \subseteq \Sigma^*;\ A.\emptyset = \emptyset$	x	
$\emptyset = \{\varepsilon\}$		x
Reg. výrazu $(ab^*)^* + (ba^*)^*$ vyhovujú všetky slová z $\{a, b\}^*$.	x	
V každej bezkontextovej gram. G je hĺbka ľub. stromu odvodenia ľub. slova w najviac $\|w\|$.		x

- Dokonca každý regulárny jazyk je zjavne rekurzívny.
- Pravidlo, ktoré by malo na pravej strane dva neterminály v reg. gramatike mať nemôžeme. Takže každé pravidlo by muselo mať na pravej strane terminál alebo ε. Ale to zjavne nestačí – keď nič iné, ako vygenerovať slová dĺžky 2 a viac? (Pre bezkontextové gramatiky by to už platilo, volá sa to Chomského normálny tvar.)
- Správne tam má byť $\subseteq$, pričom nemusí nastať rovnosť. Zoberme $L = \{a, b\}$, $h : \{a\}^* \to \{a, b\}^*$, $h(a) = a$, potom $h^{-1}(L) = \{a\}$, $h(h^{-1}(L)) = \{a\} \not\supseteq L$.
- Regulárny nie je, ale pumpovať sa dá. Slová z jazyka neobsahujúce a sa zjavne dajú pumpovať pri ľubovoľnom rozdelení, výsledok vždy padne do druhej časti. Slovo z jazyka obsahujúce a rozdelíme tak, že pumpovaná časť bude prvé a, ľubovoľným napumpovaním dostaneme slovo z prvej časti jazyka.
- Z konečných jazykov uvedenými operáciami dostaneme opäť len konečné jazyky. Takže najmenšia taká trieda sú práve všetky konečné jazyky a dosť. (Samozrejme keby medzi tými operáciami bola aj iterácia, bolo by to o inom. Kleeneho veta.)
- $\Sigma^* = \bigcup_{i \geq 0} \Sigma^i$, pričom Σ^i sú práve všetky slová dĺžky i nad abecedou Σ. Keď $w \in \Sigma^*$, tak $\exists i; w \in \Sigma^i$, a teda w je slovo konečnej dĺžky i.
- Zreťazenie je definované ako $L_1.L_2 = \{uv \mid u \in L_1 \wedge v \in L_2\}$. V našom prípade $A.\emptyset = \{uv \mid u \in A \wedge v \in \emptyset$. Keďže neexistuje žiadne $v \in \emptyset$, neexistuje ani žiadna taká dvojica u, v a teda $A.\emptyset = \emptyset$.
- Jedno je prázdna množina, druhé je množina obsajhujúca 1 prvok.
- Výrazu (ab^*) vyhovujú všetky slová začínajúce a, za ktorým je niekoľko b. Keď zoberieme ľubovoľné slovo začínajúce a, dá sa rozdeliť na niekoľko takýchto úsekov. (Rozdelíme ho pred každým a.) Preto všetky takéto slová vyhovujú reg. výrazu $(ab^*)^*$. Podobne druhému vyhovujú všetky slová začínajúce na b.
- Majme gramatiku s pravidlami $\sigma \to \beta$, $\beta \to a \mid \sigma$. V nej existuje ľubovoľne hlboký strom odvodenia slova a. *(Ako by to bolo, keby v otázke nebol ľubovoľný, ale najmenej hlboký strom odvodenia?)*

5

Rozhodnite, či je jazyk $L = \left\{ wcw \;\middle|\; w \in \{a, b\}^* \right\}$ bezkontextový. Ak nie, dokážte. Ak áno, zostrojte zásobnikový automat A, ktorý ho bude (stavom alebo prázdnou pamäťou) akceptovať. V tomto prípade nezabudnite na popis δ-funkcie a na myšlienku dôkazu $L = L(A)$, resp. $L = N(A)$.

Jeden z klasických príkladov jazykov, ktoré **nie sú** bezkontextové. Dokážeme, že preň neplatí pumpovacia lema. Sporom. Majme ľubovoľné p, q. Zoberme slovo $w = a^{p+q} b^{p+q} c a^{p+q} b^{p+q} \in L$. Zjavne $|w| > p$. Tvrdím, že toto slovo sa nedá nijako prípustne rozdeliť a napumpovať. Zjavne pre ľub. rozdelenie $w = uvxyz$ je $vy \in \{a, b\}^*$, lebo c nemôžeme pumpovať. Ak sú obe v, y na jednej strane c, dostaneme napumpovaním na 47 slovo, kde časti pred c a za c sú rôzne dlhé, a teda nepatrí do jazyka. Zostáva možnosť, kde v je pred c a

y je za c. Ale $|vxy| \leq q$, preto nutne v obsahuje samé b a y obsahuje samé a. Opäť zjavne napumpovaním na 42 dostaneme slovo, ktoré nie je z L (počty a, b pred a za znakom c sú rôzne, preto aj slová pred c a za c sú rôzne), q.e.d.

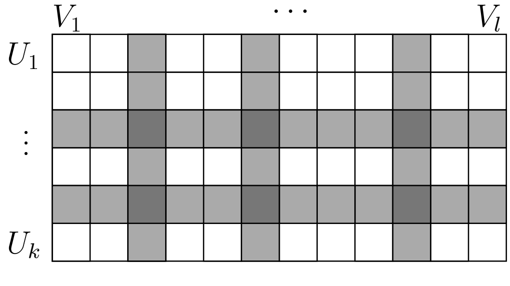
6 Vyslovte *Myhill-Nerodovu vetu* (charakterizácia regulárnych jazykov pomocou relácií ekvivalencie). Jej použitím dokážte, že trieda regulárnych jazykov je uzavretá na zjednotenie.

Veta (Myhill, Nerode). Nasledujúce tvrdenia sú ekvivalentné:

- L je regulárny jazyk.
- L je zjednotenie niektorých tried ekvivalencie nejakej sprava invariantnej relácie ekvivalencie konečného indexu.
- Relácia R_L definovaná ako $uR_Lv \iff (\forall x;\ ux \in L \iff vx \in L)$ je reláciou ekvivalencie konečného indexu.

(*Index* je počet tried ekvivalencie, *sprava invariantná* znamená, že ak slová u, v sú v tej istej triede ekvivalencie, tak aj $\forall x$; slová ux a vx sú v tej istej triede ekvivalencie.)
Intuícia za vetou: Triedy ekvivalencie zodpovedajú stavom nejakého konečného automatu pre L, pričom stavov môže samozrejme byť len konečne veľa. Jazyk je zjednotením tých tried, ktoré zodpovedajú akceptačným stavom. Keď máme regulárny jazyk, preň existuje konečný deterministický automat A, potom všetky slová zo Σ^* môžeme rozdeliť do niekoľkých množín (=tried ekvivalencie) podľa toho, v ktorom stave skončí A výpočet na w. Ak skončia výpočty na u a v v tom istom stave, výpočty na ux a vx zjavne skončia tiež v tom istom stave. Preto je naša relácia ekvivalencie sprava invariantná. A naopak, keď máme takúto reláciu, ľahko zostrojíme automat, ktorý si v stave pamätá, do ktorej triedy ekvivalencie patrí doteraz prečítaná časť slova. Totiž keďže je relácia sprava invariantná, je jednoznačne určené, ako sa trieda ekvivalencie (=stav) zmení prečítaním ďalšieho písmena.
(Toto bola vlastne ekvivalencia prvých dvoch tvrdení. Tretie hovorí o "minimálnom" konečnom automate pre jazyk – na slovách, ktoré sú v rôznych triedach takto zostrojenej R_L, musí výpočet každého automatu pre L skončiť v rôznych stavoch. Preto najmenší automat má práve toľko stavov ako je index R_L, špeciálne ak R_L nie je konečného indexu, tak žiaden automat pre L neexistuje.)

A teraz už k riešeniu. Použijeme ekvivalenciu prvých dvoch tvrdení. Majme $L_1, L_2 \in \mathcal{R}$. Keďže $L_1 \in \mathcal{R}$, existuje podľa Myhill-Nerodovej vety SIREKI[1] s triedami $U_1, \dots, U_k$, pričom L_1 je zjednotením niektorých U_i. Podobne k L_2 existuje SIREKI s triedami $V_1, \dots, V_l$, pričom L_2 je zjednotením niektorých týchto tried. Uvedomme si, že každé slovo zo Σ^* je v práve jednej z tried U_i a v práve jednej z tried V_j. Takto môžeme slová zo Σ^* rozdeliť do $k \times l$ množín podľa toho, v ktorých U_i a V_j ležia. Takto sme vlastne dostali $k \times l$ tried ekvivalencie nejakej novej relácie ekvivalencie T konečného indexu. Nie je ťažké nahliadnuť, že keďže relácie U, V boli sprava invariantné, je sprava invariantná aj relácia T. Označme triedy T ako $T_{i,j} = U_i \cap V_j$, potom ale $L_1 \cup L_2$ je práve zjednotením tých tried $T_{i,j}$, pre ktoré $U_i \subseteq L_1$ alebo $V_i \subseteq L_2$. A teda opäť podľa Myhill-Nerodovej vety je aj jazyk $L_1 \cup L_2$ regulárny, q.e.d.
Na záver obrázok, ktorý je v tomto prípade lepší ako tisíc slov.



Riadky zodpovedajú triedam ekvivalencie relácie U, stĺpce sú triedy ekvivalencie V. Vyfarbené riadky tvoria jazyk L_1, vyfarbené stĺpce jazyk L_2, všetky vyfarbené políčka dokopy tvoria ich zjednotenie. (Tmavo vyfarbené tvoria prienik.)

[1] sprava invariantná relácia ekvivalencie konečného indexu