Cvičenie č. 1
Formálne jazyky a operácie na nich

Peter Kostolányi

21. septembra 2022

Označenia

- Symbolom $\mathbb{N}$ označujeme množinu všetkých prirodzených čísel *s nulou*, t.j. $\mathbb{N} = \{0, 1, 2, \ldots\}$.
- Pre každé $n \in \mathbb{N}$ označujeme symbolom $[n]$ množinu $[n] = \{1, \ldots, n\}$.
- Symboly $\subseteq, \supseteq$ označujú reláciu *neostrej* inklúzie medzi množinami.
- Symboly $\subsetneq, \supsetneq$ označujú *ostrú* množinovú inklúziu.
- Symboly $\nsubseteq, \nsupseteq$ označujú *komplement neostrej inklúzie.* Pre množiny X, Y je teda $X \nsubseteq Y$ práve vtedy, keď neplatí $X \subseteq Y$.
- Symboly $\subset, \supset$ *nebudeme používať.* V literatúre môžu v závislosti od autora a oblasti označovať ako ostré, tak aj neostré inklúzie.
- *Rozdiel množín* X a Y na tomto predmete zapisujeme ako $X - Y$.
- *Komplement* – čiže doplnok – množiny X v rámci daného univerza označujeme X^C.
- *Potenčnú množinu* – čiže množinu všetkých podmnožín – množiny X označujeme 2^X.

1 Slová a operácie na slovách

Abeceda je ľubovoľná neprázdna konečná množina. Jej prvky, ktorými môžu byť ľubovoľné objekty, nazývame *symboly* (*znaky*) alebo *písmená.* Symboly tak definujeme výhradne vo vzťahu k abecede – ak nie je daná abeceda, nemá zmysel ani pojem symbolu.

Slovo nad abecedou Σ je *konečná* postupnosť symbolov z abecedy Σ. Pri zápise takýchto postupností nepoužívame čiarky ani zátvorky a namiesto $(a_1, a_2, \ldots, a_n)$ píšeme iba $a_1 a_2 \ldots a_n$. Príkladom slova nad dvojprvkovou abecedou $\{a, b\}$ tak môže byť napríklad $w = ababb$.

Prázdne slovo nad abecedou Σ – špeciálny prípad slova – je prázdna postupnosť prvkov abecedy Σ. V rámci tohto predmetu budeme prázdne slovo vždy označovať ε. V literatúre možno natrafiť aj na iné označenia, napríklad λ, 1, 1_{Σ^*} alebo e.

Ak pre dvojicu abecied Σ, Γ platí $\Sigma \subseteq \Gamma$, považujeme každé slovo nad abecedou Σ súčasne aj za slovo nad abecedou Γ.[1] To nám napríklad umožňuje hovoriť len o slove $abab$ bez upresnenia, či ho chápeme ako slovo nad abecedou $\{a, b\}$ alebo, povedzme, ako slovo nad abecedou $\{a, b, c\}$. Podobne pre nás existuje iba jedno prázdne slovo ε, ktoré považujeme za slovo *nad ľubovoľnou abecedou.*

Symboly abecedy Σ *a jednopísmenové slová nad* Σ *stotožňujeme* – to je v skutočnosti dané už naším spôsobom zápisu slov, v ktorom napríklad symbol a a slovo a nemožno odlíšiť.

Poznámka 1. Častou chybou u študentov býva zahrnutie prázdneho slova ε do abecedy. Treba si uvedomiť, že ε slúži iba ako naše označenie pre prázdnu *postupnosť* symbolov z abecedy, ktorá sama o sebe *symbolom nie je*, a teda *nie je ani prvkom abecedy.* Zdrojom tohto nedorozumenia je drobný nesúlad medzi pojmom „symbol" v zmysle definovanom vyššie a pojmom „symbol" používaným v bežnej reči, ktorému sa ani v nasledujúcom nevyhneme.

[1] Každú postupnosť prvkov Σ totiž v takom prípade bežne považujeme aj za postupnosť prvkov Γ – napriek tomu, že sa obe postupnosti líšia v koobore.

Dĺžka slova w, označovaná $|w|$, je definovaná ako dĺžka príslušnej postupnosti znakov. Ak teda $w = a_1a_2 \dots a_n$, kde $a_1, a_2, \dots, a_n$ sú symboly[2] z nejakej abecedy Σ, je $|w| = n$.

Okamžite vidieť, že pre všetky slová w je $|w| \geq 0$, pričom $|w| = 0$ práve vtedy, keď $w = \varepsilon$.

Počet výskytov symbolu c v slove w označujeme $\#_c(w)$; v literatúre sa možno stretnúť aj s notáciou $|w|_c$. *Formálne*: pre každdú abecedu Σ, všetky slová $w = a_1a_2 \dots a_n$ s $a_1, a_2, \dots, a_n \in \Sigma$ a ľubovoľné $c \in \Sigma$ kladieme $\#_c(w) = |\{i \in [n] \mid a_i = c\}|$.

Množinu všetkých slov – alebo v terminológii, ktorú zavedieme neskôr: **jazyk všetkých slov** – nad abecedou Σ označujeme Σ^*. Množinu všetkých *neprázdnych* slov nad Σ – slovo je neprázdne, ak nie je prázdne – označujeme Σ^+ a množinu všetkých slov nad Σ dĺžky $k \in \mathbb{N}$ označujeme Σ^k. Neskôr uvidíme, že abeceda je špeciálnym prípadom jazyka a uvedené označenia možno chápať aj ako aplikácie určitých operácií na jazykoch.

Môžeme teraz definovať prvú, asi najprirodzenejšiu a zďaleka najvýznamnejšiu spomedzi operácií na slovách, ktorou je ich *zreťazenie.*

Zreťazenie slov $u = a_1 \dots a_n$ a $v = b_1 \dots b_m$, kde $a_1, \dots, a_n$ a $b_1, \dots, b_m$ sú písmená, je slovo $u \cdot v = a_1 \dots a_nb_1 \dots b_m$. Namiesto $u \cdot v$ často píšeme len uv.

Ak je u slovom nad abecedou Σ_1 a v slovom nad abecedou Σ_2, je uv slovom nad abecedou $\Sigma_1 \cup \Sigma_2$. V prípade, že obidve slová u, v uvažujeme nad rovnakou abecedou Σ, je aj uv slovom nad tou istou abecedou. Zreťazenie je teda pre ľubovoľnú abecedu Σ *binárnou operáciou* na Σ^*. Ľahko možno nahliadnuť, že je táto operácia asociatívna a prázdne slovo ε je vzhľadom na ňu neutrálne. To možno vyjadriť konštatovaním, že $(\Sigma^*, \cdot, \varepsilon)$ je *monoid*, čiže niečo ako „grupa bez inverzných prvkov“.
Pre všetky dvojice slov u, v ďalej zrejme $|uv| = |u| + |v|$.

Mocnina slova nad abecedou Σ je definovaná na základe operácie zreťazenia rovnako, ako býva zvykom pri ľubovoľnej multiplikatívnej operácii. Pre $w \in \Sigma^*$ teda kladieme $w^0 = \varepsilon$ a $w^{n+1} = w^n w$ pre všetky $n \in \mathbb{N}$. Všimnime si, že nultá mocnina je definovaná prirodzeným spôsobom ako neutrálny prvok vzhľadom na operáciu zreťazenia, aj vďaka čomu pre všetky $m, n \in \mathbb{N}$ platí $w^m w^n = w^{m+n}$.

Reverz alebo **zrkadlový obraz** slova w nad abecedou Σ definujeme ako slovo w^R získané zo slova w jeho „čítaním sprava doľava“. Ak je teda $w = a_1a_2 \dots a_n$ pre $a_1, a_2, \dots, a_n \in \Sigma$, je reverzom slova w slovo $w^R = a_na_{n-1} \dots a_1$. Napríklad $(ab)^R = ba$, $(abba)^R = abba$, $(aababb)^R = bbabaa$ a $\varepsilon^R = \varepsilon$. Slová w spĺňajúce $w = w^R$, napríklad $abba$ alebo ε, nazývame **palindrómy**.

Operácie mocniny a reverzu majú, pomerne prirodzene, vyššiu precedenciu, než operácia zreťazenia. Napríklad teda $aab^2 = aabb$ a $aab^R = aab$, kým $(aab)^2 = aabaab$ a $(aab)^R = baa$.

Podslovo – alebo **faktor** – slova $w = a_1a_2 \dots a_n$ nad abecedou Σ je ľubovoľná *súvislá* podpostupnosť postupnosti $a_1a_2 \dots a_n$. Presnejšie ide o slovo $x \in \Sigma^*$ také, že pre nejaké $u, v \in \Sigma^*$ je $uxv = w$. Všimnime si, že *prázdne slovo je podslovom ľubovoľného slova.*

Prefix slova w je podslovo „začínajúce na začiatku slova w“ a **sufix** je podslovo „končiace na konci slova w“. To exaktne sformulujeme tak, že *prefixom* slova $w \in \Sigma^*$ nazveme slovo x také, že pre nejaké $v \in \Sigma^*$ je $xv = w$ a *sufixom* slova $w \in \Sigma^*$ slovo x také, že pre nejaké $u \in \Sigma^*$ je $ux = w$. *Prázdne slovo je tak prefixom aj sufixom ľubovoľného slova.*

[2] Takýto predpoklad znamená, že $a_1, \dots, a_n$ nadobúdajú hodnoty v Σ, pričom tieto hodnoty *nemusia* byť po dvoch rôzne. Ak napríklad $\Sigma = \{a, b\}$, tak pre $i = 1, \dots, n$ môže byť $a_i = a$ alebo $a_i = b$.

Všetky podslová slova abb teda napríklad sú $\varepsilon, a, b, ab, bb$ a abb. Prefixmi tohto slova sú ε, a, ab a abb a jeho sufixmi sú slová ε, b, bb a abb.

Poznámka 2. Niektoré z uvedených definícií nie sú v literatúre úplne ustálené. Tak je tomu napríklad pri pojme abecedy, kde sa občas možno stretnúť aj s definíciami nevyžadujúcimi jej konečnosť alebo neprázdnosť. U niektorých autorov sa tiež možno stretnúť s pojmom „podslovo" chápaným ako ľubovoľná (nie nutne súvislá) podpostupnosť, pričom pre súvislú podpostupnosť sa používa výhradne termín „faktor". Podobne aj viaceré ďalšie pojmy, s ktorými na tomto predmete budeme pracovať neskôr, bývajú často definované v rôznych – väčšinou v určitom zmysle ekvivalentných – obmenách, ktoré sú predovšetkým otázkou zamýšľaného použitia a v neposlednom rade aj vkusu autora.

V rámci tohto predmetu sa, prirodzene, budeme držať výhradne definícií z prednášky. Pri štúdiu literatúry je však nutné s podobnými drobnými rozdielmi v definíciách počítať.

2 Jazyky

Jazyk nad abecedou Σ je ľubovoľná množina slov nad abecedou Σ. To znamená, že ide o ľubovoľnú množinu $L \subseteq \Sigma^*$. Jazyk je *konečný*, ak ide o konečnú množinu; v opačnom prípade hovoríme o *nekonečnom* jazyku. Zaujímavé pre nás budú predovšetkým nekonečné jazyky.

Množina všetkých jazykov nad abecedou Σ je evidentne daná ako 2^{Σ^*}.

Príklad 1. Uveďme zopár príkladov jazykov:

- *Prázdny jazyk* je z definície prázdna množina $\emptyset$ neobsahujúca žiadne slovo. Je dôležité si uvedomiť, že ide o *iný objekt*, než jazyk $\{\varepsilon\}$, ktorý obsahuje práve jedno – prázdne – slovo.
- $L_1 = \{\varepsilon, a, abb, baa\}$ je konečný jazyk nad abecedou $\{a, b\}$.
- $L_2 = \{a, b\}^*$ je jazyk všetkých slov nad abecedou $\{a, b\}$.
- $L_3 = \{w \in \{a, b\}^* \mid w = w^R\}$ je jazyk všetkých palindrómov nad abecedou $\{a, b\}$.
- $L_4 = \{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$ je jazyk tých slov nad abecedou $\{a, b\}$, ktoré obsahujú rovnaký počet výskytov oboch symbolov.
- $L_5 = \{a^n b^n \mid n \in \mathbb{N}\}$ je ďalším príkladom jazyka nad abecedou $\{a, b\}$.
- Každá abeceda je konečný jazyk nad sebou samým.
- Množina všetkých dekadických zápisov prvočísel je jazykom nad abecedou $\{0, 1, \ldots, 9\}$.
- Všetky korektné programy v jazyku Java tvoria jazyk nad „abecedou Unicode".
- Všetky PDF dokumenty, chápané ako binárne súbory, tvoria jazyk nad abecedou $\{0, 1\}$.

Podobne ako pri slovách možno každý jazyk nad abecedou Σ chápať aj ako jazyk nad ľubovoľnou abecedou Γ takou, že $\Gamma \supseteq \Sigma$; napríklad jazyk L_1 z uvedeného príkladu môžeme súčasne chápať aj ako jazyk nad abecedou $\{a, b, c\}$. To motivuje nasledujúce označenie.

Abecedu Σ_L definujeme pre každý jazyk L rôzny od $\emptyset$ a $\{\varepsilon\}$ ako *najmenšiu abecedu* Σ takú, že L je jazykom nad Σ. Pre neprázdne slovo w píšeme namiesto $\Sigma_{\{w\}}$ len Σ_w – ide potom o množinu všetkých písmen vyskytujúcich sa v slove w. Abecedy $\Sigma_\emptyset$ a Σ_ε nemožno definovať, pretože abeceda musí byť podľa našej definície neprázdna. V literatúre sa namiesto Σ_L a Σ_w často používajú označenia $\mathrm{alph}(L)$ resp. $\mathrm{alph}(w)$, ktoré sú však občas používané aj v inom význame.

Poznámka 3. Skutočnosť, že slovo w obsahuje aspoň jeden výskyt písmena c, teda správne vyjadríme ako $c \in \Sigma_w$. Zápis $c \in w$, často používaný niektorými študentmi, *je vonkoncom nekorektný.*

3 Operácie na jazykoch

Množinové operácie. Keďže je jazyk definovaný ako množina, má zmysel zaoberať sa množinovými operáciami aplikovanými na jazyky. Možno teda uvažovať predovšetkým:

a) *Zjednotenie* $L_1 \cup L_2$, *prienik* $L_1 \cap L_2$ a *rozdiel* $L_1 - L_2$ jazykov L_1 a L_2.[3]

b) *Komplement* L^C jazyka L. Podobne ako aj inde v matematike je komplement nutné uvažovať v rámci nejakého univerza. To v teórii formálnych jazykov znamená, že pri práci s komplementom je, prísne vzaté, vždy potrebné uviesť abecedu Σ, nad ktorou komplementovaný jazyk L uvažujeme – potom je $L^C = \Sigma^* - L$. Použitie notácie L^C bez uvedenia univerza v literatúre väčšinou znamená, že univerzum je pre dané účely buď nepodstatné, alebo zrejmé z kontextu.

Zreťazenie jazykov L_1 a L_2 je jazyk $L_1 \cdot L_2 = \{uv \mid u \in L_1; v \in L_2\}$. Ide teda o jazyk zreťazení uv všetkých dvojíc slov u, v takých, že $u \in L_1$ a $v \in L_2$.

Príklad 2. Uvažujme *konečné* jazyky $L_1 = \{\varepsilon, a, ab, abb\}$ a $L_2 = \{b, ab, bb\}$. Jazyk $L_1 \cdot L_2$ potom možno získať nasledujúcim postupom:

1. Zreťazenia jednotlivých dvojíc slov zapíšeme do tabuľky.

	b	ab	bb
ε	b	ab	bb
a	ab	aab	abb
ab	abb	$abab$	$abbb$
abb	$abbb$	$abbab$	$abbbb$

2. Po odstránení duplikátov dostávame $L_1 \cdot L_2 = \{b, ab, bb, aab, abb, abab, abbb, abbab, abbbb\}$.

Zreťazenie má vyššiu precedenciu ako zjednotenie a prienik – napríklad výraz $L_1 \cup L_2 \cdot L_3$ teda treba čítať ako $L_1 \cup (L_2 \cdot L_3)$.

Poznámka 4. Na množine všetkých jazykov nad danou abecedou Σ zohráva zjednotenie úlohu aditívnej operácie a zreťazenie zohráva úlohu multiplikatívnej operácie. Prázdny jazyk $\emptyset$ je evidentne neutrálny vzhľadom na zjednotenie – pre všetky jazyky L je $L \cup \emptyset = \emptyset \cup L = L$ – a agresívny vzhľadom na zreťazenie – pre všetky L je $L \cdot \emptyset = \emptyset \cdot L = \emptyset$. Jazyk $\{\varepsilon\}$ je naopak neutrálny vzhľadom na zreťazenie – pre všetky L je $L \cdot \{\varepsilon\} = \{\varepsilon\} \cdot L = L$. Tieto dva jazyky sú teda *fundamentálne* odlišné: kým prázdny jazyk $\emptyset$ zohráva úlohu nuly, jazyk $\{\varepsilon\}$ zohráva úlohu jednotky.

Mocnina jazyka L je, podobne ako mocnina slova, definovaná prirodzeným induktívnym predpisom: $L^0 = \{\varepsilon\}$ a $L^{n+1} = L^n \cdot L$ pre všetky $n \in \mathbb{N}$. Dôležitá je nasledujúca charakterizácia jazyka L^n, podľa ktorej tento jazyk pozostáva z práve všetkých zreťazení n-tíc slov z jazyka L.

Tvrdenie 1. *Nech L je jazyk a $n \in \mathbb{N}$. Potom $L^n = \{w_1 \dots w_n \mid w_1, \dots, w_n \in L\}$.*

Dôkaz. Indukciou vzhľadom na n.

1. Pre $n = 0$ je z definície $L^0 = \{\varepsilon\} = \{w_1 \dots w_0 \mid w_1, \dots, w_0 \in L\}$.

2. Predpokladajme platnosť tvrdenia pre $n = k$. Dokážeme jeho platnosť aj pre $n = k + 1$. Skutočne:

$$L^{k+1} = L^k \cdot L = \{w_1 \dots w_k \mid w_1, \dots, w_k \in L\} \cdot L = \{w_1 \dots w_{k+1} \mid w_1, \dots, w_{k+1} \in L\},$$

kde prvá rovnosť je z definície mocniny jazyka, druhá z indukčného predpokladu a posledná z definície zreťazenia. □

[3] Pri množinovom rozdiele má uprednostňovanie notácie $L_1 - L_2$ pred $L_1 \setminus L_2$ korene u tých autorov, ktorí druhé z označení používajú pre tzv. ľavý kvocient – čo je operácia, s ktorou sa v rámci týchto cvičení stretneme neskôr. Iná skupina autorov zas ľavý kvocient označuje ako $L_1^{-1}L_2$; pre množinový rozdiel je potom možné „bezpečne" používať zvyčajný operátor „\".

Poznámka 5. Častou chybou u študentov býva zamieňanie si n-tej mocniny jazyka L s jazykom n-tých mocnín slov z jazyka L. Čitateľ určite ľahko dokáže, že vo všeobecnosti $L^n \neq \{w^n \mid w \in L\}$.

Poznámka 6. Bezprostredne z definície mocniny jazyka vyplýva aj rovnosť $\emptyset^0 = \{\varepsilon\}$. Ide tu o určitú analógiu (dobre odôvodnenej) konvencie $0^0 = 1$, ktorá sa používa v niektorých oblastiach matematiky.

Iterácia – alebo **(Kleeneho) uzáver**, či hovorovo **(Kleeneho) hviezdička** – jazyka L je jazyk

$$L^* = \bigcup_{k=0}^{\infty} L^k$$

Kladná iterácia – alebo **kladný (Kleeneho) uzáver**, či hovorovo **(Kleeneho) plus** – jazyka L je jazyk

$$L^+ = \bigcup_{k=1}^{\infty} L^k.$$

Nasledujúca dôležitá charakterizácia jazykov L^* a L^+ je dôsledkom tvrdenia 1.

Tvrdenie 2. *Uvažujme ľubovoľný jazyk L. Potom je $L^* = \{w_1 \ldots w_k \mid k \in \mathbb{N};\ w_1, \ldots, w_k \in L\}$ a $L^+ = \{w_1 \ldots w_k \mid k \in \mathbb{N} - \{0\};\ w_1, \ldots, w_k \in L\}$.*

Dôkaz. Rozpísaním jazykov L^k podľa tvrdenia 1 dostávame pre iteráciu

$$L^* = \bigcup_{k=0}^{\infty} L^k = \bigcup_{k=0}^{\infty} \{w_1 \ldots w_k \mid w_1, \ldots, w_k \in L\} = \{w_1 \ldots w_k \mid k \in \mathbb{N};\ w_1, \ldots, w_k \in L\}$$

a pre kladnú iteráciu

$$L^+ = \bigcup_{k=1}^{\infty} L^k = \bigcup_{k=1}^{\infty} \{w_1 \ldots w_k \mid w_1, \ldots, w_k \in L\} = \{w_1 \ldots w_k \mid k \in \mathbb{N} - \{0\};\ w_1, \ldots, w_k \in L\},$$

čo bolo treba dokázať. □

Poznámka 7. Nech Σ je abeceda. Z tvrdení 1 a 2 vyplýva konzistentnosť označení Σ^k (pre $k \in \mathbb{N}$), Σ^* a Σ^+ z oddielu 1 s operáciami mocniny, iterácie a kladnej iterácie jazyka.

Reverz jazyka L je jazyk $L^R = \{w^R \mid w \in L\}$. Platí teda napríklad $\{\varepsilon, abb, bab\}^R = \{\varepsilon, bba, bab\}$, či $\{a^n b^n \mid n \in \mathbb{N}\}^R = \{b^n a^n \mid n \in \mathbb{N}\}$.

Operácie mocniny, iterácie, kladnej iterácie a reverzu jazyka majú vyššiu precedenciu, než zreťazenie. Napríklad teda $L_1 \cdot L_2^* = L_1 \cdot (L_2^*)$ a podobne pre zvyšné spomenuté operácie.

Niektoré notačné skratky. Za účelom uľahčenia zápisu sa obyčajne používa konvencia, podľa ktorej slovo w, použité ako argument zreťazenia s jazykom, iterácie, či kladnej iterácie, interpretujeme ako jednoprvkový jazyk $\{w\}$. Namiesto $\{w\} \cdot L$ tak napríklad môžeme písať iba wL. Namiesto $\{w\}^*$ a $\{w\}^+$ píšeme obvykle iba w^* a w^+ – treba si však uvedomiť, že použitím týchto operátorov získame *jazyk*. Napríklad teda $a^* c b^* = \{a^m c b^n \mid m, n \in \mathbb{N}\}$.

4 Riešené úlohy

Dokazovanie rovnosti $L_1 = L_2$ dvoch jazykov je často najvýhodnejšie rozdeliť na dve časti a postupne dokázať obidve inklúzie $L_1 \subseteq L_2$ a $L_1 \supseteq L_2$. Pri dokazovaní inklúzie $L_1 \subseteq L_2$ potom obyčajne uvažujeme ľubovoľné slovo $w \in L_1$ a snažíme sa ukázať, že toto slovo musí patriť aj do jazyka L_2; pri inklúzii $L_1 \supseteq L_2$ postupujeme opačne. (Takýto postup, hoci je spravidla najspoľahlivejší, samozrejme nemusí byť vždy najjednoduchší alebo najelegantnejší.)

Úloha 1. Nech Σ je ľubovoľná abeceda. Dokážte, že $\Sigma \cdot \Sigma^* = \Sigma^+$.

Riešenie. Dokážeme postupne obe inklúzie medzi jazykmi $\Sigma \cdot \Sigma^*$ a Σ^+.

- $\subseteq$: Nech $w \in \Sigma \cdot \Sigma^*$. Potom $w = av$, kde $a \in \Sigma \subseteq \Sigma^*$ a $v \in \Sigma^*$. V dôsledku toho $w \in \Sigma^*$. Keďže $a \in \Sigma$, je $|w| = |av| \geq |a| = 1$. Slovo w je teda neprázdne, takže musí patriť do Σ^+.
- $\supseteq$: Nech $w \in \Sigma^+$. Vďaka tvrdeniu 2 existuje $k \in \mathbb{N} - \{0\}$ také, že $w = a_1 a_2 \ldots a_k$ pre nejaké $a_1, a_2, \ldots, a_k \in \Sigma$. Z toho $w = a_1 v$ pre $a_1 \in \Sigma$ a $v = a_2 \ldots a_k \in \Sigma^*$, takže $w \in \Sigma \cdot \Sigma^*$. □

Úloha 2. Nech L je ľubovoľný jazyk. Dokážte, že $L \cdot L^* = L^+$.

Riešenie. Dokážeme postupne obe inklúzie medzi jazykmi $L \cdot L^*$ a L^+.

- $\subseteq$: Nech $w \in L \cdot L^*$. Potom $w = uv$ pre nejaké $u \in L$ a $v \in L^*$. Z tvrdenia 2 navyše $v = v_1 \ldots v_k$ pre nejaké $k \in \mathbb{N}$ a $v_1, \ldots, v_k \in L$. Preto je slovo $w = uv_1 \ldots v_k$ zreťazením nenulového počtu slov $u, v_1, \ldots, v_k \in L$, a teda $w \in L^+$ podľa tvrdenia 2.
- $\supseteq$: Nech $w \in L^+$. Vďaka tvrdeniu 2 existuje $k \in \mathbb{N} - \{0\}$ také, že $w = w_1 w_2 \ldots w_k$ pre nejaké $w_1, w_2, \ldots, w_k \in L$. To znamená, že $w = w_1(w_2 \ldots w_k)$, kde $w_1 \in L$ a $w_2 \ldots w_k \in L^*$ podľa tvrdenia 2. Preto $w \in L \cdot L^*$. □

Úloha 3. Uvažujme dvojicu jazykov

$$L_1 = \{w \in \{a,b\}^* \mid \#_a(w) \leq \#_b(w)\},$$
$$L_2 = \{w \in \{a,b\}^* \mid \#_a(w) \geq \#_b(w)\}.$$

Nájdite jazyk $L_1 \cdot L_2$.

Riešenie. Dokážeme, že $L_1 \cdot L_2 = \{a,b\}^*$.

- $\subseteq$: Nech $w \in L_1 \cdot L_2$. Potom $w = uv$, kde $u \in L_1 \subseteq \{a,b\}^*$ a $v \in L_2 \subseteq \{a,b\}^*$. Preto aj $w = uv \in \{a,b\}^*$.
- $\supseteq$: Nech $w \in \{a,b\}^*$. Potom buď $\#_a(w) \leq \#_b(w)$, alebo $\#_a(w) \geq \#_b(w)$. Ak $\#_a(w) \leq \#_b(w)$, tak $w = w\varepsilon$, pričom $w \in L_1$ a $\varepsilon \in L_2$. Ak $\#_a(w) \geq \#_b(w)$, tak $w = \varepsilon w$, pričom $\varepsilon \in L_1$ a $w \in L_2$. V oboch prípadoch tak dostávame $w \in L_1 \cdot L_2$. □

Úloha 4. Uvažujme jazyk $L = \{w \in \{a,b\}^* \mid |w| \geq 4\}$. Nájdite jazyk L^*.

Riešenie. Dokážeme, že $L^* = L \cup \{\varepsilon\}$.

- $\subseteq$: Nech $w \in L^*$. Potom existuje $k \in \mathbb{N}$ a slová $w_1, \ldots, w_k \in L$ tak, že $w = w_1 \ldots w_k$. Ak $k = 0$, tak $w = \varepsilon \in L \cup \{\varepsilon\}$. Ak $k > 0$, tak $|w| \geq |w_1| \geq 4$, keďže $w_1 \in L$. Preto $w \in L \subseteq L \cup \{\varepsilon\}$.
- $\supseteq$: Nech $w \in L \cup \{\varepsilon\}$. Potom $w \in L = L^1 \subseteq L^*$, alebo $w \in \{\varepsilon\} = L^0 \subseteq L^*$. □

Porovnať dvojicu jazykov L, L' znamená zistiť, či platia inklúzie $L \subseteq L'$ a $L \supseteq L'$. Pri riešení úloh tohto typu je vždy potrebné overiť obidve inklúzie – nestačí sa teda napríklad uspokojiť s konštatovaním, že sa daná dvojica jazykov nemusí rovnať.

Často pritom býva úlohou porovnať jazyky typu $F(L_1, \ldots, L_k)$ a $G(L_1, \ldots, L_k)$, kde $L_1, \ldots, L_k$ sú *nejaké* (ľubovoľné) jazyky a F, G sú operácie na jazykoch (v nasledujúcich dvoch úlohách je $k = 2$). V takom prípade je potrebné dvojicu jazykov $F(L_1, \ldots, L_k)$ a $G(L_1, \ldots, L_k)$ porovnať *vo všeobecnosti* – to znamená zistiť, či jednotlivé inklúzie platia, nech sú jazyky $L_1, \ldots, L_k$ zvolené akokoľvek. Na dôkaz niektorej z inklúzií teda nemožno o jazykoch $L_1, \ldots, L_k$ predpokladať nič špeciálne. Za účelom vyvrátenia inklúzie naopak stačí nájsť konkrétne jazyky $L_1, \ldots, L_k$, pre ktoré inklúzia neplatí.

Úloha 5. Nech L_1, L_2 sú jazyky. Porovnajte jazyky $(L_1 \cup L_2)^*$ a $L_1^* \cup L_2^*$.

Riešenie. Dokážeme, že $(L_1 \cup L_2)^* \supseteq L_1^* \cup L_2^*$, pričom opačná inklúzia vo všeobecnosti neplatí.

- $\not\subseteq$: Nech $L_1 = \{a\}$ a $L_2 = \{b\}$. Potom $ab \in (L_1 \cup L_2)^*$, ale $ab \notin L_1^* \cup L_2^*$.
- $\supseteq$: Nech $w \in L_1^* \cup L_2^*$. Potom buď $w \in L_1^*$, alebo $w \in L_2^*$.

  Ak $w \in L_1^*$, podľa tvrdenia 2 existuje $k \in \mathbb{N}$ a slová $w_1, \ldots, w_k \in L_1$ tak, že $w = w_1 \ldots w_k$. Slová $w_1, \ldots, w_k$ sú súčasne aj v $L_1 \cup L_2$; vďaka tvrdeniu 2 tak $w = w_1 \ldots w_k \in (L_1 \cup L_2)^*$.

  Podobne ak $w \in L_2^*$, tak $w = w_1 \ldots w_k$, kde $k \in \mathbb{N}$ a $w_1, \ldots, w_k \in L_2$. Slová $w_1, \ldots, w_k$ sú zároveň aj v $L_1 \cup L_2$, a teda $w = w_1 \ldots w_k \in (L_1 \cup L_2)^*$. □

Úloha 6. Nech L_1, L_2 sú jazyky. Porovnajte jazyky $(L_1 \cup L_2)^*$ a $L_1^* \cdot (L_2 \cdot L_1^*)^*$.

Než pristúpime k riešeniu tejto úlohy, pokúsme sa získať trochu intuície. Jazyk $(L_1 \cup L_2)^*$ obsahuje zreťazenia nejakého – prípadne aj nulového – počtu slov, ktoré sú vždy buď z L_1 alebo z L_2. Jazyk $L_1^* \cdot (L_2 \cdot L_1^*)^*$ tiež obsahuje takéto slová, ale zdanlivo požaduje trochu viac: zreťazenie má začínať niekoľkými slovami z L_1 a následne pokračovať niekoľkými „úsekmi", kde každý z nich obsahuje najprv práve jedno slovo z L_2 a za ním niekoľko slov z L_1; „niekoľko" tu vždy môže byť aj nula.

Istotne teda bude $(L_1 \cup L_2)^* \supseteq L_1^* \cdot (L_2 \cdot L_1^*)^*$. Platnosť opačnej inklúzie závisí od toho, či je *každé* zreťazenie $w_1 \ldots w_k$ slov $w_1, \ldots, w_k$ z L_1 alebo z L_2 súčasne aj zreťazením druhého uvedeného typu. Vyznačme ale v zreťazení $w_1 \ldots w_k$ tie spomedzi slov $w_1, \ldots, w_k$, ktoré patria do L_2 a rozdelme ho na „úseky" tak, aby každý z nich začínal označeným slovom a spolu s ním obsahoval aj všetky bezprostredne nasledujúce *neoznačené* slová (každý „úsek" sa teda vždy bude končiť buď pred nasledujúcim označeným slovom, alebo na konci slova). Z tohto pohľadu už ľahko vidieť, že je naše zreťazenie v skutočnosti aj zreťazením druhého typu. Platí teda aj inklúzia $(L_1 \cup L_2)^* \subseteq L_1^* \cdot (L_2 \cdot L_1^*)^*$ a vo výsledku tak budeme dokazovať rovnosť oboch jazykov. Samotné riešenie úlohy bude pozostávať z náležitej *formalizácie* uvedených pozorovaní.

Riešenie. Dokážeme, že $(L_1 \cup L_2)^* = L_1^* \cdot (L_2 \cdot L_1^*)^*$.

- $\supseteq$: Nech $w \in L_1^* \cdot (L_2 \cdot L_1^*)^*$. Z definície zreťazenia jazykov je $w = uv$, kde $u \in L_1^*$ a $v \in (L_2 \cdot L_1^*)^*$. Z tvrdenia 2 je $u = u_1 \ldots u_m$ pre nejaké $m \in \mathbb{N}$ a $u_1, \ldots, u_m \in L_1 \subseteq (L_1 \cup L_2)$. Podobne $v = v_1 \ldots v_n$ pre nejaké $n \in \mathbb{N}$ a $v_1, \ldots, v_n \in L_2 \cdot L_1^*$. Z definície zreťazenia teraz pre $i = 1, \ldots, n$ dostávame $v_i = x_i y_i$, kde $x_i \in L_2 \subseteq (L_1 \cup L_2)$ a $y_i \in L_1^*$. Ďalšou aplikáciou tvrdenia 2 napokon pre $i = 1, \ldots, n$ zisťujeme, že $y_i = z_{i,1} \ldots z_{i,s_i}$, kde $s_i \in \mathbb{N}$ a $z_{i,1}, \ldots, z_{i,s_i} \in L_1 \subseteq (L_1 \cup L_2)$. V dôsledku toho

  $$\begin{aligned} w = uv = u_1 \ldots u_m v_1 \ldots v_n = u_1 \ldots u_m (x_1 y_1) \ldots (x_n y_n) = \\ = u_1 \ldots u_m (x_1 z_{1,1} \ldots z_{1,s_1}) \ldots (x_n z_{n,1} \ldots z_{n,s_n}) \in (L_1 \cup L_2)^*, \end{aligned}$$

  opäť podľa tvrdenia 2.

- $\subseteq$: Nech $w \in (L_1 \cup L_2)^*$. Potom podľa tvrdenia 2 existuje $k \in \mathbb{N}$ a slová $w_1, \ldots, w_k \in L_1 \cup L_2$ tak, že $w = w_1 \ldots w_k$. Nech $1 \leq i_1 < \ldots < i_s \leq k$ sú indexy také, že $w_{i_1}, \ldots, w_{i_s}$ sú *práve všetky* slová spomedzi $w_1, \ldots, w_k$, ktoré sú v L_2. Slová

  $$u_0 = w_1 w_2 \ldots w_{i_1 - 1}, \quad u_1 = w_{i_1+1} w_{i_1+2} \ldots w_{i_2-1}, \quad \ldots, \quad u_s = w_{i_s+1} w_{i_s+2} \ldots w_k$$

  potom podľa tvrdenia 2 musia byť v L_1^* – niektoré z nich môžu byť aj prázdne. Zrejme

  $$w = u_0 w_{i_1} u_1 w_{i_2} u_2 \ldots w_{i_s} u_s = u_0 \left((w_{i_1} u_1)(w_{i_2} u_2) \ldots (w_{i_s} u_s) \right). \tag{1}$$

  Pre $j = 1, \ldots, s$ je $w_{i_j} u_j \in L_2 \cdot L_1^*$ a podľa tvrdenia 2 tak $(w_{i_1} u_1)(w_{i_2} u_2) \ldots (w_{i_s} u_s) \in (L_2 \cdot L_1^*)^*$. Keďže $u_0 \in L_1^*$, z definície zreťazenia jazykov a (1) napokon dostávame $w \in L_1^* \cdot (L_2 \cdot L_1^*)^*$. □