

UNIVERZITA KOMENSKÉHO, BRATISLAVA

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

SILA A ZLOŽITOSŤ MODERNÝCH REGULÁRNYCH  
VÝRAZOV

Diplomová práca

2015

Tatiana Tóthová

UNIVERZITA KOMENSKÉHO, BRATISLAVA

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

# SILA A ZLOŽITOSŤ MODERNÝCH REGULÁRNYCH VÝRAZOV

Diplomová práca

Študijný program: Informatika  
Študijný odbor: .... Informatika  
Školiace pracovisko: Katedra Informatiky  
Školiteľ: RNDr. Michal Forišek, PhD.

Bratislava, 2015

Tatiana Tóthová



Univerzita Komenského v Bratislave  
Fakulta matematiky, fyziky a informatiky

---

## ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Tatiana Tóthová  
**Študijný program:** informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)  
**Študijný odbor:** 9.2.1. informatika  
**Typ záverečnej práce:** bakalárska  
**Jazyk záverečnej práce:** slovenský

**Názov:** Moderné regulárne výrazy

**Cieľ:** Spraviť prehľad nových konštrukcií používaných v moderných knižniciach s regulárnymi výrazmi (ako napr. look-ahead a look-behind assertions). Analyzovať tieto rozšírenia z hľadiska formálnych jazykov a prípadne tiež z hľadiska algoritmickej výpočtovej zložitosti.

**Vedúci:** RNDr. Michal Forišek, PhD.  
**Katedra:** FMFI.KI - Katedra informatiky  
**Vedúci katedry:** doc. RNDr. Daniel Olejár, PhD.  
**Dátum zadania:** 23.10.2012

**Dátum schválenia:** 24.10.2012

doc. RNDr. Daniel Olejár, PhD.  
garant študijného programu

*.....tu bude podakovanie.....*

*Tatiana Tóthová*

## Abstrakt

Tu bude abstrakt po slovensky.

**Kľúčové slová:** nejaké, kľúčové, slová

## Abstract

Here will be abstract in english...

**Key words:** some, key, words

# Obsah

<b>Úvod</b>	<b>1</b>
<b>1 Súčasný stav problematiky</b>	<b>2</b>
1.1 Základné definície . . . . .	2
1.2 Vlastnosti a sila . . . . .	6
1.3 Popisná zložitosť . . . . .	6
<b>2 Naše výsledky</b>	<b>10</b>
2.1 Vlastnosti a sila . . . . .	10
2.1.1 Sila negatívneho lookaroundu . . . . .	10
2.1.2 Vlastnosti . . . . .	14
2.1.3 Zaradenie do Chomského hierarchie . . . . .	15
2.1.4 Priestorová zložitosť . . . . .	17
2.2 Popisná zložitosť . . . . .	29
<b>Záver</b>	<b>31</b>
<b>Literatúra</b>	<b>32</b>

# Úvod

Pár slov na úvod...



# 1 Úvod do súčasného stavu problematiky

Myšlienka regulárnych výrazov bola prvýkrát spomenutá v teórii formálnych jazykov a automatov ako iný spôsob popisu regulárnych jazykov. Vtedy pozostávali z operácií zjednotenia, zretazenia a Kleeneho uzáveru. Pre ich jednoduchosť boli implementované ako nástroj na vyhľadávanie slov zo špecifikovaného jazyka. Postupom času a s inšpiráciou zo strany užívateľov k nim pribúdali ďalšie operácie. Niektoré boli len skratkou k tomu, čo sa už dalo zapísať – umožnili zapísať to isté menej znakmi – ostatné otvárali dvere k popisu dovtedy nedosiahnuteľných jazykov.

Zmes týchto operácií nazývame moderné regulárne výrazy a zaujíma nás, kam sa až dostali v popisovaní jazykov v rámci Chomského hierarchie. Túto problematiku sme z väčšej časti rozobrali v bakalárskej práci [Tó13]. V tejto práci rozbor dokončíme a pozrieme sa na ne aj z hľadiska zložitosti.

## 1.1 Základné definície

Formálna definícia je uvedená v [Tó13], tu si len neformálne uvedieme, s ktorými operáciami pracujeme a ako fungujú. Každý regulárny výraz sa skladá zo znakov a metaznakov. Znaky sú symboly, ktoré charakterizujú samé seba, teda  $L(a) = \{a\}$ . Metaznaky popisujú operáciu nad regulárnymi výrazmi. Ak potrebujeme použiť metaznak ako znak, stačí pred neho dať  $\backslash$ , teda  $L(\backslash x) = \{x\}$ . Ak metaznak vyžaduje vstup, je ním posledný znak/metaznak/uzátvorkovaný podvýraz pred ním. Metaznaky vyzerajú nasledovne:

- $\backslash$  – robí z metaznakov obyčajné znaky
- $()$  – okrúhle zátvorky slúžia na oddelovanie podvýrazov

## 1.1. ZÁKLADNÉ DEFINÍCIE KAPITOLA 1. SÚČASNÝ STAV PROBLEMATIKY

- $|$  – operácia zjednotenia (alternácia) – musí vyhovovať ľavý alebo pravý podvýraz
- $*$  – Kleeneho uzáver – predchádzajúci výraz opakuj ľubovoľný počet krát
- $+$  – opakuj 1 alebo viackrát
- $\{ \}$  – zložené zátvorky sú používané ako  $\{n, m\}$  (opakuj aspoň  $n$  a najviac  $m$ -krát) a  $\{n\} = \{n, n\}$  (opakuj  $n$ -krát)
- $[ ]$  – hranaté zátvorky – znaky vnútri tvoria množinu, z ktorej si vyberáme. Vieme použiť aj intervaly, napr.  $a-z$ ,  $A-Z$ ,  $0-9$ ,  $\dots$  a kombinovať ich. Všetky metaznaky vnútri  $[ ]$  sa považujú za normálne znaky.
- $.$  – ľubovoľný znak
- $?$  – ak samostatne: opakuj 0 alebo 1-krát  
ak za operáciou: namiesto greedy implementácie použi minimalistickú, t.j. zober čo najmenej znakov (platí pre  $*, +, ?, \{n, m\}$ )<sup>1</sup>
- $^$  – metaznak označujúci začiatok slova; špeciálnym prípadom je výraz  $[\wedge\alpha]$  (kde  $\alpha$  je nejaká množina znakov), ktorý špecifikuje ľubovoľný znak, ktorý sa v množine  $\alpha$  nenachádza
- $\$$  – metaznak označujúci koniec slova

Okrem operácií označených metaznakmi vznikli aj zložitejšie operácie, na popis ktorých treba dlhšie konštrukcie. V prvom rade sa zaviedlo **číslovanie okrúhlych zátvoriek**. Čísľuje sa zľava doprava podľa poradia ľavej (otváracej) zátvorky. Toto číslovanie sa deje automaticky pri každom behu algoritmu, teda už pri písaní výrazu ho môžeme využívať. Popíšme si teraz zložitejšie operácie:

- **komentár** ozn.  $(? \# \text{text komentáru})$  – klasický komentár, určený čitateľom kódu; nemá vplyv na výraz, algoritmus ho ignoruje
- **spätné referencie** ozn.  $\backslash k$ ,  $k \in \mathbb{N}$  – môže sa nachádzať na ľubovoľnom mieste vo výraze ZA  $k$ -tou pravou (zatváracou) zátvorkou. Odkazuje sa na  $k$ -te zátvorky a presný význam sa určuje až pri hľadaní zhody na konkrétnom vstupe. Algoritmus si zapamätá aké podslovo zo vstupu matchoval výraz vnútri  $k$ -tych zátvoriek a presne toto podslovo hľadá, keď ďalej vo výraze narazí na  $\backslash k$ .

---

<sup>1</sup>všetky spomenuté operácie „žerú“ znaky a na ich implementáciu bol použitý greedy algoritmus

## 1.1. ZÁKLADNÉ DEFINÍCIE KAPITOLA 1. SÚČASNÝ STAV PROBLEMATIKY

Predpokladáme, že číslo  $k$  je zapísané znakmi z inej abecedy ako zvyšok regexu – t.j. je jednoznačne určiteľné, kde končí zápis  $k$ .<sup>2</sup>

- **lookahead** – tzv. nazeranie dopredu. Keďže anglický výraz je kratší a zvučnejší, rozhodli sme sa ho používať.

- **pozitívny** ozn.  $(?= \alpha)$ , kde  $\alpha$  je nejaký moderný regulárny výraz

V momente, keď na lookahead narazíme si zapamätáme aktuálnu pozíciu v slove. Začneme hľadať zhodu s výrazom  $\alpha$ . Ak vyhlásil zhodu, vrátíme sa naspäť na zapamätané miesto a odtiaľ pokračujeme v slove aj regulárnym výraze akokeby tam lookahead nikdy nebol. Inak povedané lookahead nevyžiera písmenká a robí akýsi prienik. Ak neobsahuje znak pre koniec slova \$, môže skončiť kdekoľvek – je to v okamihu, keď zistil zhodu.

- **negatívny** ozn.  $(?! \alpha)$ , kde  $\alpha$  je nejaký moderný regulárny výraz

Nesmie nájsť slovo z jazyka  $L(\alpha)$ . Postup je ako pri lookaheade, ale končí úspešne len ak neexistuje žiadne podslovo začínajúce na danom mieste v slove také, že by ho  $\alpha$  akceptoval. Inak povedané, toto je pozitívny lookahead s obrátenou akceptáciou.

- **lookbehind** – tzv. nazeranie dozadu. Opäť budeme používať anglickú verziu pre jej stručnosť.

- **pozitívny** ozn.  $(?<= \alpha)$ , kde  $\alpha$  je nejaký moderný regulárny výraz

Hľadá slovo z  $L(\alpha)$  naľavo od aktuálneho miesta v slove (musí končiť na susediacom políčku vľavo od aktuálnej pozície v slove). Opäť nie je určená hranica slova, môže začínať kdekoľvek, ak nie je vynútený začiatok slova znakom  $\wedge$ .

Ak by sme chceli deterministický algoritmus, vyzeral by nasledovne: od teraz do konca vykonávania lookbehindu bude na pozícii, na ktorej v slove sme, bude teraz pre znak pre koniec slova – endmarker. Najprv vyskúšame 1 susedný symbol naľavo, či patrí do jazyka. Ak nie skúsime čítať o jeden

---

<sup>2</sup>V implementovaných regexoch vieme deterministicky určiť, kde končí číslo  $k$  – sú povolené maximálne trojciferné čísla. My chceme všeobecnejší model, bez obmedzení na veľkosť  $k$ , a práve preto je tento predpoklad oprávnený. Zároveň si to môžeme dovoliť, lebo ak máme  $k$  zapísané ako číslo v desiatkovej sústave, v teórii je jednoduché zasubstituovať hľadaný jazyk tak, aby neobsahoval znaky  $0, \dots, 9$ .

## 1.1. ZÁKLADNÉ DEFINÍCIE KAPITOLA 1. SÚČASNÝ STAV PROBLEMATIKY

znak viac (2 symboly naľavo), keď neuspejeme, opäť posunieme pomyselný začiatok slova doľava. Ak akceptujeme, môže to byť len na endmarkeri. Ak sme neakceptovali a začiatok slova nejde viac posunúť, zamietneme.

- **negatívny** ozn.  $(?<!\alpha)$ , kde  $\alpha$  je nejaký moderný regulárny výraz  
Hľadá ako pozívny lookbehind, ale akceptuje len ak neexistuje slovo z  $L(\alpha)$  vyskytujúce sa naľavo od aktuálnej pozície.

Pre lookahead a lookbehind používame spoločný názov lookaround, takisto s prívlastkom pozitívny/negatívny myslíme iba ich pozitívne/negatívne verzie.

Keďže sa týmito operáciami budeme zaoberať podrobnejšie, uvedieme pre upresnenie ich definície.

**Definícia 1.1.1** (Pozitívny lookahead).

$$L_1(? = L_2)L_3 = \{uvw \mid u \in L_1 \wedge v \in L_2 \wedge vw \in L_3\}$$

Operáciu  $(? = \dots)$  nazývame *pozitívny lookahead* alebo len lookahead.

**Definícia 1.1.2** (Negatívny lookahead).

$$L_1(? !L_2)L_3 = \{uv \mid u \in L_1 \wedge v \in L_3 \wedge \text{neexistuje také } x, y, \text{ že } v = xy \text{ a } x \in L_2\}$$

Operáciu  $(?! \dots)$  nazývame negatívny lookahead.

**Definícia 1.1.3** (Pozitívny lookbehind).

$$L_1(? \leq L_2)L_3 = \{uvw \mid uv \in L_1 \wedge v \in L_2 \wedge w \in L_3\}$$

Operáciu  $(? \leq \dots)$  nazývame *pozitívny lookbehind* alebo len lookbehind.

**Definícia 1.1.4** (Negatívny lookbehind).

$$L_1(? < !L_2)L_3 = \{uv \mid u \in L_1 \wedge v \in L_3 \wedge \text{neexistuje také } x, y, \text{ že } u = xy \text{ a } y \in L_2\}$$

Operáciu  $(? < ! \dots)$  nazývame negatívny lookbehind.

Moderné regulárne výrazy sa skladajú z **konečného počtu** znakov, metaznakov a zložitejších operácií.

Pre korektné matchovanie regulárnym výrazom potrebujeme vedieť priority operácií. Niektoré sa správajú ako znak:  $[ ]$ ,  $.$ ,  $\wedge$ ,  $\$$ ,  $\backslash k$ , pozitívny a negatívny lookahead. Ostatné kombinujú znaky s nasledovnými prioritami:

priorita	3	2	1	0
operácia	()	* + ? {}	zreťazenie	

Máme veľkú množinu operácií a budeme chcieť pracovať aj s jej podmnožinami, preto si zavedieme názvoslovie pre ich jednoznačnejšie a kratšie určenie.

*Regex* – množina operácií, pomocou ktorých vieme popísať iba regulárne jazyky; presnejšie všetky znaky a metaznaky (bez zložitejších operácií)

*Eregex* – *Regex* so spätnými referenciami

*LRegex* – *Eregex* s pozitívnym lookaroundom

*nLRegex* – *LRegex* s negatívnym lookaroundom

$\mathcal{L}_{RE}$  – trieda jazykov nad *Regex* ( $= \mathcal{R}$ )

$\mathcal{L}_{ERE}$  – trieda jazykov nad *Eregex*

$\mathcal{L}_{LERE}$  – trieda jazykov nad *LRegex*

$\mathcal{L}_{nLERE}$  – trieda jazykov nad *nle*

## 1.2 Vlastnosti a sila moderných regulárnych výrazov

**TODO!!!** opraviť referencie na vety z bakalárky

Potrebné vety z bakalárskej práce:

**Veta 1.2.1** (Veta 2.2.5.). *Regulárne jazyky sú uzavreté na lookaround.*

**Veta 1.2.2** (Veta 2.2.10.). *Trieda nad regexami s pozitívnym lookaroundom je  $\mathcal{R}$ .*

**Veta 1.2.3** (Veta 2.2.14.).  $\mathcal{L}_{LERE} \subseteq \mathcal{L}_{CS}$

## 1.3 Popisná zložitosť moderných regulárnych výrazov

V čase, keď sme hľadali články týkajúce sa popisnej zložitosti moderných regulárnych výrazov, sme nenašli nič týkajúce sa tejto problematiky. Známe boli len výsledky pre regulárne výrazy s operáciami zjednotenia, zreťazenia a Kleeneho  $*$  (RE), prípadne ešte

prieniku a komplementu (GRE). Navyše mali ešte znak pre prázdny jazyk  $\emptyset$  a prázdne slovo  $\varepsilon$ .

Spomeňme si najprv ako možno zdefinovať popisnú zložitosť [EKSwW13] [EZ]:

Nech  $E$  je regulárny výraz, potom

- $|E|$  je jeho celková dĺžka
- $rpn(E)$  je jeho celková dĺžka v reverznej polskej notácii. Táto notácia oproti klasickej neobsahuje okrúhle zátvorky a používa explicitný metaznak pre zrefazenie  $\bullet$ . Napríklad regulárny výraz  $(a|ab) * (c|d)$  má 12 znakov a v reverznej polskej notácii  $aab \bullet | * cd| \bullet$  má 10 znakov.

Uvedomme si, že dĺžka reverznej polskej notácie je rovnaká ako počet vrcholov v syntaktickom strome pre daný výraz. Preto možno vernejšie popisuje skutočnú zložitosť regulárneho výrazu (žiadne pomocné symboly, len znaky a operácie). Nakoľko však nie je veľmi prehľadná a ťažko sa v nej hľadajú chyby, nie je tak často používaná.

- $|alph(E)| = N(E)$  je počet alfabetických symbolov v  $E$
- $H(E)$  je hĺbka vzhľadom na  $*$ , počet vnorení  $*$
- $L(E)$  je dĺžka najdlhšej neopakujúcej sa cesty cez výraz
- $W(E)$  je šírka, počet zjednotených symbolov. Za touto mierou zložitosti nie je žiadna intuitívna predstava. Ako vidieť neskôr v definícii, dôvodom jej vzniku bola duálnosť k  $L$ .

V nasledujúcej tabuľke sú uvedené indukčné definície jednotlivých mier zložitosti. Zložitosť regulárneho jazyka vzhľadom na ľubovoľnú z týchto mier zložitosti je minimálna miera cez všetky regulárne výrazy pre daný regulárny jazyk.

	Alphabetical Symbol	$E \cup F$	$E \cdot F$	$E^*$
N	1	$N(E) + N(F)$	$N(E) + N(F)$	$N(E)$
H	0	$\max(H(E), H(F))$	$\max(H(E), H(F))$	$H(E) + 1$
L	1	$\max(L(E), L(F))$	$L(E) + L(F)$	$L(E)$
W	1	$W(E) + W(F)$	$\max(W(E), W(F))$	$W(E)$

Ako pri mnohých iných modeloch, aj do regulárnych výrazov vieme zakomponovať časti, ktoré nič nerobia (okrem toho, že zaberajú miesto). V rámci skúmania najjednoduchších výrazov sa prišlo k nasledujúcim definíciám [EKSwW13]:

**Definícia 1.3.1.** *Nech  $E$  je regulárny výraz nad abecedou  $\Sigma$  a nech  $L(E)$  je jazyk špecifikovaný výrazom  $E$ . Hovoríme, že  $E$  je **zmenšiteľný**, ak platí nejaká z nasledujúcich podmienok:*

1.  $E$  obsahuje  $\emptyset$  a  $|E| > 1$
2.  $E$  obsahuje podvýraz tvaru  $FG$  alebo  $GF$ , kde  $L(F) = \varepsilon$
3.  $E$  obsahuje podvýraz tvaru  $F|G$  alebo  $G|F$ , kde  $L(F) = \{\varepsilon\}$  a  $\varepsilon \in L(G)$

Inak, ak žiadna z nich neplatí,  $E$  nazývame **nezmenšiteľným**.

V originále sa používajú výrazy *collapsible* a *uncollapsible*. Táto definícia neodhalí všetky zbytočne zopakované časti, napríklad  $a|a$  je nezmenšiteľný, aj keď by sa dal zapísať jednoduchým  $a$ . Problém je, že identity výrazov nie sú konečne axiomatizovateľné (ani nad unárnou abecedou) [EKSwW13]. Teda nie je reálne určiť také pravidlá, aby sme dosiahli konečné zjednodušenie.

**Definícia 1.3.2.** *Ak  $E$  je nezmenšiteľný regulárny výraz taký, že*

1.  $E$  nemá nadbytočné  $()$ ; a zároveň
2.  $E$  neobsahuje podvýraz tvaru  $F **$

*potom vravíme, že  $E$  je **neredukovateľný** (irreducible).*

Môžeme vyvodiť, že minimálny regulárny výraz pre daný jazyk bude nezmenšiteľný a neredukovateľný, naopak to však nemusí platiť. S takýmto základom možno dokázať napríklad tvrdenie: Ak  $E$  je neredukovateľný a  $|\text{alph}(E)| \geq 1$ , potom  $|E| \leq 11|\text{alph}(E)| - 4$ .

Iné spôsoby na ohraničenie regulárnych výrazov poskytujú nasledujúce pozorovanie a veta.

**Veta 1.3.3** (Proposition 6 [EKSwW13]). *Nech  $L$  je neprázdny regulárny jazyk.*

(a) *Ak dĺžka najkratšieho slova v  $L$  je  $n$ , potom  $|\text{alph}(E)| \geq n$  pre ľubovoľný regulárny výraz  $E$ , kde  $L(E) = L$ .*

(b) Ak navyiac  $L$  je konečný a dĺžka najdlhšieho slova v  $L$  je  $n$ , potom  $|\text{alph}(E)| \geq n$  pre ľubovoľný regulárny výraz, kde  $L(E) = L$ .

Definícia non-returning NKA hovorí, že sa nevracia do počiatočného stavu, t.j. žiadne prechody nevedú smerom do  $q_0$ .

**Veta 1.3.4** (Theorem 10). *Nech  $E$  je regulárny jazyk s  $|\text{alph}(E)| = n$ . Potom existuje non-returning NKA akceptujúci  $L(E)$  s  $\leq n + 1$  stavmi a DKA akceptujúci  $L(E)$  s  $\leq 2n + 1$  stavmi.*

Rozbehnutých je viacero oblastí skúmania. Regulárne výrazy sú zaujímavé hlavne preto, že tvoria stručnejší popis jazyka a sú ekvivalentné automatom. S tým súvisí prvá oblasť. Skúma sa stavová zložitosť automatu ekvivalentného konkrétnemu výrazu a naopak tiež popisná zložitosť výrazu ekvivalentného konkrétnemu automatu. Dá sa to zhrnúť ako problémy konverzií medzi modelmi. Niektorí autori siahajú po väčšej abstrakcii a namiesto automatov uvažujú iba orientované grafy s hranami označenými symbolmi. Určia si parametre grafu a snažia sa napríklad čo najlepšie popísať cesty medzi vrcholmi.

Ďalšia oblasť zahŕňa operácie nad regulárnymi výrazmi. Jeden z problémov je napríklad vzťah popisnej zložitosti výrazu pre jazyk  $L$  a  $L^c$ . Pre automaty existuje konštrukcia pre vyrobenie komplementu jazyka. Avšak pre komplementárny regulárny výraz to nie je také jednoznačné.

Ako poslednú by sme spomenuli problém najkratšieho slova nešpecifikovaného regulárnym výrazom. Predpokladajme regulárny výraz  $E$ , kde  $|\text{alph}(E)| = n$  nad konečnou abecedou  $\Sigma$ , pričom  $L(E) \neq \Sigma^*$ . Aké dlhé môže byť najkratšie slovo NEšpecifikované výrazom  $E$ ? Najzaujímavejší výsledok je pre výraz s  $|\text{alph}(E)| = 75n + 361$ , kde najkratšie nešpecifikované slovo je dĺžky  $3(2n - 1)(n + 1) + 3$ .



## 2 Naše výsledky

### 2.1 Vlastnosti a sila moderných regulárnych výrazov

#### 2.1.1 Sila negatívneho lookaheadu

V bakalárskej práci sme zabudli ... *negatívny lookahead* **TODO!!!**

**Lema 2.1.1.** *Trieda  $\mathcal{R}$  je uzavretá na negatívny lookahead.*

*Dôkaz.* Nech  $L_1, L_2, L_3 \in \mathcal{R}$ . Ukážeme, že  $L = L_1(?! L_2)L_3 \in \mathcal{R}$ .

Keďže  $L_1, L_2, L_3$  sú regulárne, existujú DKA  $A_i = (K_i, \Sigma_i, \delta_i, q_{0i}, F_i)$  také, že  $L(A_i) = L_i, i \in \{1, 2, 3\}$ . Zostrojíme NKA  $A$  pre  $L$ .

Konštrukcia bude veľmi podobná ako pre pozitívny lookahead, keď si správne predpripravíme  $A_2$ . Negatívny lookahead sa snaží za každú cenu nájsť slovo z  $L_2$  (t.j. uspieť s  $A_2$ ) a ak sa mu to nepodarí, akceptuje. Preto musíme zaručiť, že sa  $A_2$  buď zasekne alebo prejde až do konca slova<sup>1</sup>. Ak  $A_2$  dosiahne akceptačný stav, negatívny lookahead musí zamietnuť.

Ďalšie pozorovanie nám hovorí, že negatívny lookahead akceptuje vždy nejaké slovo z komplementu  $L_2$ . Ale komplement vzhľadom na akú abecedu? V skutočnosti nekonečnú – ľubovoľný znak, ktorý nepatrí do  $\Sigma_2$ , znamená zaseknutie  $A_2$ , t.j. akceptáciu. Ak sa na to pozrieme z väčšej diaľky, uvidíme  $L_3$ , s ktorým robíme prenik.  $A_3$  musí dočítať slovo do konca a na úplne neznámom znaku sa zasekne, takže z pohľadu výslednej akceptácie slova nevedí, ak by kvôli tomu znaku zamietol už negatívny lookahead. Preto stačí urobiť komplement vzhľadom na  $\Sigma_2 \cup \Sigma_3$ .

Podme upravovať  $A_2$ , začneme vytváraním komplementu. Ak  $\Sigma_3 \setminus \Sigma_2 \neq \emptyset$ <sup>2</sup>, potom vytvoríme  $A'_2 = (K'_2, \Sigma'_2, \delta'_2, q_0, F'_2)$  tak, že pridáme nové znaky  $\Sigma'_2 = \Sigma_2 \cup \Sigma_3$ , jeden

<sup>1</sup>Ak sa zasekne, slovo z  $L_2$  tam určite nebude, lebo neexistuje akceptačný výpočet. Ak sa nezasekne, nevieme povedať o tom slove nič, pokiaľ ho neprejdeme celú.

<sup>2</sup>Inak  $A'_2 = A_2$ .

nový stav<sup>3</sup>  $K'_2 = K_2 \cup \{q_{ZLE}\}$  a prechody na nové znaky z každého stavu:

$$\begin{aligned} \forall q \in K_2 \ \forall a \in \Sigma_2 & : \delta'_2(q, a) = \delta_2(q, a) \\ \forall q \in K_2 \ \forall a \in \Sigma_3 \setminus \Sigma_2 & : \delta'_2(q, a) = q_{ZLE} \\ \forall a \in \Sigma'_2 & : \delta'_2(q_{ZLE}, a) = q_{ZLE} \end{aligned}$$

$F'_2 = F_2$ . Do všetkých stavov sme pridali prechody na nové znaky a  $q_{ZLE}$  má 1 prechod na každý znak, takže  $A'_2$  je stále deterministický. Zároveň nové znaky vedú do stavu, z ktorého sa nedá dostať do žiadneho akceptačného, z čoho vyplýva  $L(A'_2) = L_2$ .

Teraz  $A'_2$  zmeníme na  $A''_2$  tak, aby akceptoval práve vtedy, keď (!  $L_2$ ). Najprv si skonštruujeme množinu stavov, z ktorých sa vieme dostať do akceptačného stavu:  $H = \{q \in K'_2 \mid \exists w \in \Sigma_2^* \exists q_A \in F'_2 : (q, w) \vdash_{A'_2} (q_A, \varepsilon)\}$ , nasledovným spôsobom:

$$H_0 = F'_2$$

$$H_{i+1} = \{q \in K'_2 \mid \exists p \in H_i \exists a \in \Sigma'_2 : \delta(q, a) = p\}$$

Zrejme  $\exists i \in \mathbb{N} : H_{i+1} = H_i = H$ . Nás zaujíma množina  $K'_2 \setminus H$ , v ktorej sa nachádzajú všetky stavy, z ktorých sa na žiadne slovo nie je možné dostať do žiadneho akceptačného stavu.

$$A''_2 = (K''_2, \Sigma''_2, \delta''_2, q_0, F''_2) : K''_2 = K'_2 \cup \{q_A, q_Z\}, \Sigma''_2 = \Sigma'_2, F''_2 = (K'_2 \setminus H) \cup \{q_A\}$$

$$\begin{aligned} \forall q \in K'_2 \setminus F'_2 \ \forall a \in \Sigma'_2 & : \delta''_2(q, a) = \delta'_2(q, a) \\ \forall q \in K'_2 \setminus F'_2 \ \forall a \in \Sigma'_2 & : \delta''_2(q, \varepsilon) = q_A \\ \forall q \in F'_2 \ \forall a \in \Sigma'_2 & : \delta''_2(q, a) = q^4 \\ \forall a \in \Sigma'_2 & : \delta''_2(q_A, a) = q_Z \end{aligned}$$

Je dobré poznamenať, že  $A''_2$  je takmer deterministický. Chýbajú mu prechody z  $q_Z$  – môžeme ho nechať cykliť v tomto stave, ale nevadí nám ani keď sa zasekne. Nedeterministické rozhodnutie vykonáva iba jedno – pri prechode do stavu  $q_A$ . Vtedy háda, že už je na konci slova. Ak nie je,  $\delta$ -funkcia ho pošle do stavu  $q_Z$ . Podľa toho je zřejmé, že ak existuje akceptačný výpočet a dočítal slovo do konca, je práve jeden<sup>5</sup> <sup>6</sup>.

<sup>3</sup>Pre každý nový stav predpokladáme, že je jedinečný – t.j. žiaden taký v množine stavov ešte nie je.

<sup>4</sup>Tento riadok v podstate netreba,  $A''_2$  sa môže rovno zaseknúť, lebo  $A'_2$  práve akceptoval.

<sup>5</sup>Okrem prípadu, kedy dočíta slovo a je v stave z  $(K'_2 \setminus H)$  – vieme ho predĺžiť o krok na  $\varepsilon$  a skončiť v  $q_A$ . Jadro výpočtu ale zostáva rovnaké – jednoznačné.

<sup>6</sup>Zaujímavé je, že aj zamietací výpočet je pre každé slovo jednoznačný

Tvrdíme  $L_1(?=L(A_2''))L_3 = L_1(?! L_2)L_3 = L$ . Keďže  $A_1$  a  $A_3$  sa pri oboch výpočtoch budú chovať rovnako (sú nezávislé od lookaheadov), nebudeme sa nimi pri dôkaze zaoberať.

$\subseteq$ : Máme akceptačný výpočet pre  $A_2''$  na nejakom vstupe. Akceptačný stav mohol byť buď z množiny  $K_2' \setminus H$ , to znamená, že pôvodný automat  $A_2'$  sa dostal do stavu, z ktorého už nemohol nijak akceptovať<sup>7</sup>. V takom prípade  $(?! L_2)$  akceptuje. V druhom prípade mohol  $A_2''$  akceptovať pomocou  $q_A$ , čo znamená, že dočítal slovo do konca (pretože z  $q_A$  sa na ľubovoľný znak dostaneme do  $q_Z$ , v ktorom sa  $A_2''$  zasekne a nebude akceptovať) a ani raz sa nedostal do akceptačného stavu automatu  $A_2'$ . Teda aj  $(?! L_2)$  akceptuje.

$\supseteq$ : Nech  $(?! L_2)$  akceptoval, t.j. na  $A_2$  bol vykonaný nejaký neakceptujúci výpočet ( $A_2$  je deterministický, takže existuje práve jeden pre každé slovo). Rovnakú postupnosť stavov bude mať aj výpočet na  $A_2''$  (automat obsahuje všetky stavy aj  $\delta$ -funkciu z  $A_2$ , počiatočný stav je ten istý). Ak sa  $A_2$  zasekol na neznámom znaku, v  $A_2'$  na ten znak pridáme prechod do stavu  $q_{ZLE}$  a v ňom zostaneme až do konca slova. Keďže  $q_{ZLE}$  nie je v  $A_2'$  akceptačný a nedá sa z neho na žiadne slovo dostať do ľubovoľného akceptačného stavu,  $q_{ZLE} \in K_2' \setminus H$  a teda  $q_{ZLE} \in F_2''$ . Ak sa  $A_2$  nezasekol, tak dočítal slovo do konca a v postupnosti stavov jeho výpočtu sa nenachádza žiaden z množiny  $F_2$ . V tom prípade  $A_2''$  z posledného stavu prejde na  $\varepsilon$  do  $q_A$  (2.riadok definície  $\delta_2''$ ) a akceptuje.

Z práve dokázaného tvrdenia a vety 1.2.1 už vyplýva aj platnosť našej lemy.  $\square$

**Lema 2.1.2.** *Trieda  $\mathcal{R}$  je uzavretá na negatívny lookbehind.*

*Dôkaz.* Nech  $L_1, L_2, L_3 \in \mathcal{R}$ , existujú DKA  $A_i = (K_i, \Sigma_i, \delta_i, q_{0i}, F_i)$  také, že  $L(A_i) = L_i, i \in \{1, 2, 3\}$ . Ukážeme, že  $L = L_1(?! L_2)L_3 \in \mathcal{R}$ .

Nemôžeme skonštruovať  $A_2''$  také, že bude akceptovať práve vtedy, keď  $(?! L_2)$ , pretože nevieme čítať doľava. Tzn. zaručiť, že miesto, kde začína  $A_2''$  výpočet je skutočný začiatok slova. Mechanizmus lookbehindu musí byť riadený zvonku, automatom pre  $L$ . Skonštruujeme ho.  $C = (K, \Sigma, \delta, q_0, F)$  :

$$K = K_3 \cup \{(q, A) \mid q \in K_1 \wedge A \subseteq K_2\}, \Sigma = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3, q_0 = (q_{01}, \{q_{02}\}), F = F_3 \cup \delta :$$

<sup>7</sup>Sem spadá aj prípad, keď  $A_2$  prečítal neznámy znak a zasekol sa –  $A_2'$  zostáva až do konca slova v stave  $q_{ZLE}$ .

$$\begin{aligned}
& \forall q \in K_3 \ \forall a \in \Sigma_3 : \delta(q, a) \ni \delta_3(q, a) \\
& \forall q \in K_1 \ \forall A \subseteq K_2 \ \forall a \in \Sigma_1 \cap \Sigma_2 : \delta((q, A), a) \ni (p, B \cup \{q_{02}\}), \text{ kde } \delta_1(q, a) = p, \\
& \quad B = \{r \mid \exists s \in A : \delta_2(s, a) = r\} \\
& \forall q \in K_1 \ \forall A \subseteq K_2 \ \forall a \in \Sigma_1 \setminus \Sigma_2 : \delta((q, A), a) \ni (p, \{q_{02}\}), \text{ kde } \delta_1(q, a) = p \\
& \forall q \in F_1 \ \forall A : A \subseteq K_2 \wedge A \cap F_2 = \emptyset : \delta((q, A), \varepsilon) \ni q_{03}
\end{aligned}$$

Druhý riadok  $\delta$ -funkcie hovorí, že  $A_1$  a všetky rozbehnuté výpočty  $A_2$  prejdú do ďalšieho stavu a zároveň sa rozbehne nový výpočet  $A_2$ . Ak by sme hľadali jeden akceptačný výpočet  $A_2$  stačilo by tipnúť si začiatok a odtiaľ ho simulovať. Lenže simulujeme negatívny lookbehind, teda ak neexistuje akceptačný výpočet, tak my akceptujeme (prejdeme na simulovanie  $A_3$ , 4. riadok). A to vieme povedať len v prípade, že sme videli všetky možné výpočty. Keďže  $A_2$  je deterministický, pre každé slovo existuje práve jeden výpočet a zároveň sa nikdy nezasekne (na svojej abecede), teda nám stačí skontrolovať množinu stavov vtedy, keď sa  $C$  rozhodne, že  $A_1$  končí výpočet. Ak tam je, nedostane sa k simulovaniu  $A_3$  a tým ani k akceptačným stavom.

$$L(C) = L.$$

$\subseteq$ : Majme akceptačný výpočet  $C$  na  $w$ . Z definície  $F$  vyplýva, že  $A_3$  akceptoval, teda  $\exists u, v$  také, že  $w = uv$  a  $v \in L_3$ . Do  $q_{03}$  sa dá dostať len z dvojice  $q, A$  takej, že  $q \in F_1$ , teda  $u \in L_1$ , a  $a \cap F_2 = \emptyset$ , teda  $\nexists xy$  také, že  $u = xy$  a  $y \in L_2$ . To vyplýva z toho, že v každom znaku  $u$  začal jeden výpočet na  $A_2$  a žiaden z nich neakceptoval. Teda negatívny lookbehind akceptoval a  $w \in L$ .

$\supseteq$ : Nech  $w \in L$ , teda  $\exists u, v$  také, že  $w = uv$ ,  $u \in L_1, v \in L_3$  a  $\forall x, y : u = xy$  platí  $y \notin L_2$ . Pre  $u, v$  teda existujú akceptačné výpočty na  $A_1, A_3$  a pre všetky  $y$  neakceptačné na  $A_2$ . Z toho už vieme vyskladať akceptačný výpočet na  $C$ .  $\square$

**Veta 2.1.3.** *Trieda  $\mathcal{R}$  je uzavretá na negatívny lookaround.*

**Lema 2.1.4.** *Nech  $L_1, L_2, L_3, L_4 \in \mathcal{R}$ ,  $\alpha = (L_1 (?! L_2) L_3) * L_4$ . Potom  $L(\alpha) \in \mathcal{R}$ .*

*Dôkaz.* Podobne ako v dôkaze vety 2.1.1 pretransformujeme  $(?! L_2)$  na akýsi  $(?=L(A_2''))$ , kde  $A_2''$  bude akceptovať práve vtedy, keď  $(?! L_2)$ . Potom  $\beta = (L_1(?=L(A_2''))L_3) * L_4$ ,  $L(\beta) = L(\alpha) \in \mathcal{R}$  podľa vety 1.2.2.  $\square$

**Lema 2.1.5.** *Nech  $L_1, L_2, L_3, L_4 \in \mathcal{R}$ ,  $\alpha = L_4 (L_1 (? <! L_2) L_3) *$ . Potom  $L(\alpha) \in \mathcal{R}$ .*

*Dôkaz.* Použijeme konštrukciu ako v dôkaze vety 2.1.2, kde pretransformujeme  $(?<!\ L_2)$  na  $(?<=L(A_2''))$ , kde  $A_2''$  bude akceptovať práve vtedy, keď  $(?<!\ L_2$ . Z toho vznikne  $\beta = L_4(L_1(?<= L(A_2''))L_3)* = L(\alpha) \in \mathcal{R}$  podľa vety 1.2.2.  $\square$

**Veta 2.1.6.** *Trieda nad regexami s negatívnym lookaroundom je  $\mathcal{R}$ .*

**Dôsledok 2.1.7.** *Trieda nad regexami s pozitívnym a negatívnym lookaroundom je  $\mathcal{R}$ .*

### 2.1.2 Vlastnosti

Pridanie nových operácií medzi regulárne výrazy síce pridalo na sile modelu, ale mohlo pokaziť jeho uzáverové vlastnosti. Vieme, že regulárne jazyky sú uzavreté na všetky možné operácie, ktoré poznáme. Posilnenie modelu spätnými referenciami a následne lookaroundom však ohrozilo uzavretosť na základnú operáciu regulárnych výrazov – zretazenie. Pokiaľ zretazíme 2 regexy obsahujúce lookahead alebo lookbehind, tieto operácie začnú zasahovať do slova z vedľajšieho jazyka. Ak by išlo o zretazenie so zarážkou (napr.  $\alpha\#\beta$ ), na koniec každého lookaheadu v  $\alpha$  by stačilo pripojiť  $.*\#$ , prípadne znak  $\$$  nahradiť znakom  $\#$ . Podobne by sme v regexe  $\beta$  pridali na začiatok každého lookbehindu  $\#.*$  prípadne by sme vymenili znak  $\wedge$  znakom  $\#$ . Potom by tieto operácie zostali „skrotené” na území slova, ktoré danému regexu prislúcha.

Otázka nastáva, ako to spraviť, keď zarážku k dispozícii nemáme. Odpoveďou je nasledujúca veta.

**Veta 2.1.8.** *Trieda  $\mathcal{L}_{LERE}$  je uzavretá na zretazenie.*

*Dôkaz.* Nech sú le-regexy  $\alpha, \beta$ . Chceme ukázať, že jazyk  $L(\alpha)L(\beta) \in \mathcal{L}_{LERE}$ . Intuitívne nás to vedie k riešeniu  $\alpha\beta$ , čo ale nemusí byť vždy správne. Problémom sú operácie lookaround, presnejšie každý lookahead v  $\alpha$  môže zasahovať do slova z  $L(\beta)$  a takisto každý lookbehind z  $\beta$  môže zasahovať do slova z  $L(\alpha)$ . Navyše, ak lookahead obsahuje  $\$$  a lookbehind  $\wedge$ , potom zasahujú do slova z iného jazyka určite. V takom prípade môže regex  $\alpha\beta$  vynechať niektoré slová z  $L_1L_2$  a tiež pridať nejaké nevhodné slová navyše. Preto treba nájsť spôsob, ako operácie lookaroundu vhodne 'skrotiť'. Predpokladajme, že  $\alpha$  má  $k$  označených zátvoriek, potom regex pre jazyk  $L(\alpha)L(\beta)$  vyzerá nasledovne:

$$(?=(\alpha) \ ( \ \beta \ ) \$)\alpha' \setminus k + 2(?<=\wedge \setminus 1\beta') \quad (2.1)$$

$\begin{matrix} 1 & 1 & k+2 & k+2 \end{matrix}$

V  $\alpha, \beta$  treba vhodne prepísať označenie zátvoriek (po poradí).  $\alpha'$  je  $\alpha$  prepísaný tak, že pre každý lookahead:

- bez  $\$$  – na koniec pridáme  $\cdot * \backslash k + 2 \$$
- s  $\$$  – pred  $\$$  pridáme  $\backslash k + 2$

$\beta'$  je  $\beta$  prepísaný tak, že pre každý lookbehind:

- bez  $\wedge$  – na začiatok pridáme  $\wedge \backslash 1 \cdot$
- s  $\wedge$  – pred  $\wedge$  pridáme  $\wedge \backslash 1$

Čo presne robí regex 2.1? Nech  $w$  je vstupné slovo, ktoré chceme matchovať. Lookahead na začiatku ho nejako rozdelí na  $w_1$  a  $w_2$ , pričom  $w = w_1 w_2$ , tak, že  $w_1$  bude patriť do  $L(\alpha)$  a podslovo  $w_2$  do  $L(\beta)$ . Ešte musíme overiť, či  $\alpha$  matchuje  $w_1$  samostatne.

Preto sme v  $\alpha$  prepísali všetky lookaheads. V  $\alpha'$  každý z nich musí na konci slova  $w$  matchovať  $w_2$  (toto zabezpečí spätná referencia  $\backslash k + 2$  a  $\$$ ) a teda matchovanie regexu  $\alpha$  zostane výlučne na podslove  $w_1$ . Analogicky to platí pre upravené lookbehindy v  $\beta'$ .  $\square$

### 2.1.3 Zaradenie do Chomského hierarchie

**Veta 2.1.9.**  $\mathcal{L}_{LRE}$  je neporovnateľná s  $\mathcal{L}_{CF}$ .

*Dôkaz.* Majme jazyk  $L = \{ww \mid w \in \{a, b\}^*\} \in \mathcal{L}_{CF}$ , nech  $\alpha = ([ab]^*) \backslash 1$ . Zrejme  $L(\alpha) = L$ , teda  $L \in \mathcal{L}_{LRE}$ .

Ukážeme, že jazyk  $L = \{a^n b^n \mid n \in \mathbb{N}\} \notin \mathcal{L}_{LRE}$ . Sporom, nech  $L \in \mathcal{L}_{LRE}$ .

Vieme, že  $L \notin \mathcal{L}_{ERE}$ , preto musí obsahovať nejaký lookaround. Zároveň z  $L \notin \mathcal{R}$  a vety 1.2.2 vyplýva, že musí obsahovať aj spätné referencie.

Kam sa môžu spätné referencie odkazovať a kam ich potom môžeme umiestniť? Nech výraz, na ktorý ukazujú, vyrobí nejaké:

- $a^i$ , potom  $\backslash k$  musí byť v prvej polovičke slova (medzi  $a$ , inak by pokazil štruktúru slova), takže nevplýva na časť s  $b$  a teda sa zaobídeme bez nich.
- $a^i b^j$ , potom  $\backslash k$  by mohol byť len medzi  $b$ , ale tam by pokazil štruktúru slova  $a^n b^n$

- $b^j$ , potom  $\setminus k$  môže byť len medzi  $b$  a je tam zbytočný z rovnakých dôvodov, aké má prípad  $a^i$

Vidíme, že so spätnými referenciami dosiahneme rovnaký výsledok ako bez nich, čo je spor s tým, že sa vo výraze musia nachádzať (bez nich vieme urobiť len regulárny jazyk).  $\square$

**Dôsledok 2.1.10.**  $\mathcal{L}_{LRE} \subsetneq \mathcal{L}_{CS}$

**Veta 2.1.11.**  $\mathcal{L}_{nLRE} \subseteq \mathcal{L}_{CS}$

*Dôkaz.* Vieme, že  $\mathcal{L}_{LRE} \in \mathcal{L}_{CS}$  (veta 1.2.3), teda ľubovoľný regex z  $LEregex$  vieme simulovať pomocou LBA. Ukážeme, že ak pridáme operáciu negatívny lookahead/lookbehind, vieme to simulovať tiež.

Nech  $\alpha \in nLEregex$ . Potom nech  $A$  je LBA pre  $\alpha$ , ktorý ignoruje negatívny lookaround (t.j. vyrábame LBA pre regex z  $LEregex$ ). Teraz vytvoríme LBA  $B$  pre regex vnútri negatívneho lookaroundu. Z nich vytvoríme LBA  $C$  pre úplný regex  $\alpha$  tak, že  $C$  bude simulovať  $A$  a keď príde na rad negatívny lookaround zaznačí si, v akom stave je  $A$  a na ktorom políčku skončil. Skopíruje slovo na ďalšiu stopu a na nej simuluje od/do toho miesta  $B$  (podľa toho, či je to lookahead alebo lookbehind).

Teraz je to s akceptáciou náročnejšie ako pri pozitívnom lookarounde. Pokiaľ  $B$  akceptoval,  $C$  sa zasekne. Ak sa  $B$  zasekol,  $C$  sa vráti naspäť k zastavenému výpočtu  $A$  a pokračuje v ňom. Keďže slovo je konečné a  $B$  na ňom testuje regex, ktorý postupne vyjedá písmenká, určite raz príde na koniec slova. Môže sa stať, že bude skúšať viaceré možnosti – napr. skúsi pre  $*$  zobrať menej znakov, teda príde na koniec slova viackrát. Ako určíme, že skončil výpočet? Bez újmy na všeobecnosti môžeme predpokladať, že sa  $B$  nezacyklí, ale zamietne slovo na konci výpočtu. To preto, že všetkých možností na rozdelenie znakov medzi operácie  $*$ ,  $+$ ,  $?$ ,  $\{n, m\}$  je konečne veľa a keď ich systematicky skúša<sup>8</sup>, raz sa mu musia minúť. To znamená, že ak nemá v  $\delta$ -funkcii umelo vsunuté zacyklenie, nezacyklí sa. Teda ak slovo neakceptuje, tak ho určite zamietne a vtedy  $C$  môže prejsť na zastavený výpočet  $A$  a pokračovať v ňom.

<sup>8</sup>Algoritmus pre  $*$  je v praxi greedy. Ak slovo nesedí, tak sa vráti, odoberie posledný znak zožratý  $*$  a opäť skúša zvyšok výrazu, či slovo sedí. Ak stále nevyhovuje, algoritmus odoberá hviezdičky znaky dovtedy, dokým nenájde zhodu alebo odoberie všetky znaky – vtedy sa mu minuli všetky možnosti a môže slovo zamietnuť.

Podobne ako pri lookarounde aj jeho negatívna verzia môže obsahovať vnorený negatívny lookaround. Tých však môže byť iba konečne veľa, keďže každý regex musí mať konečný zápis. Čo znamená, že aj stôp bude konečne veľa a naznačeným postupom si vieme postupne vybudovať LBA, ktorý bude simulovať  $\alpha$ .  $\square$

**TODO!!!**check & rewrite

**Veta 2.1.12.**  $\mathcal{L}_{nLRE}$  je neporovnateľná s  $\mathcal{L}_{CF}$ .

*Dôkaz.* Opäť použijeme jazyk  $L = \{a^n b^n \mid n \in \mathbb{N}\}$  a budeme dokazovať sporom. Nech  $L \in \mathcal{L}_{LRE}$ .

Z vety 2.1.9 vidíme, že výraz musí obsahovať negatívny lookaround. Z jej dôkazu vidíme, že spätné referencie nepomôžu, nech sa ich pokúsime hocijako použiť. Z toho vyplýva, že ich nepoužijeme a z vety 2.1.6 zistíme, že nám zostal model schopný vyrobiť iba regulárne jazyky. Spor.  $\square$

**Dôsledok 2.1.13.**  $\mathcal{L}_{nLRE} \subsetneq \mathcal{L}_{CS}$

## 2.1.4 Priestorová zložitosť

**Veta 2.1.14.**  $\mathcal{L}_{LRE} \subseteq NSPACE(\log n)$ , kde  $n > 1$  je veľkosť vstupu.

*Dôkaz.* Ukážeme, že ľubovoľný  $\alpha \in LERegex$  vieme simulovať nedeterministickým Turingovým strojom s jednou vstupnou read-only páskou a jednou pracovnou páskou, na ktorej použijeme maximálne logaritmický počet políčok.

Celý regex si budeme uchovávať v stavoch<sup>9</sup> a pridáme doňho špeciálny znak  $\blacktriangleright$ , ktorým si budeme ukazovať, kam sme sa v regexe dopracovali – bude to akýsi smerník na znak, ktorý práve spracovávame. Teda stav bude vyzeráť takto:  $q_{\beta\blacktriangleright\xi}$  (pre lepšiu čitateľnosť budeme uvádzať namiesto stavu iba jeho dolný index:  $\beta\blacktriangleright\xi$ ), kde  $\beta\xi = \alpha$ , čiast  $\beta$  sme už namatchovali na vstup a práve sa chystáme pokračovať čiastou  $\xi$ . Ak  $\blacktriangleright$  ukazuje na znak, porovnáme ho s aktuálnym znakom na vstupnej páske. Pokiaľ sa nezhodujú, výpočet sa zasekne. Pri zhode pokračujeme až kým sa nám neminie vstup aj regex. Ak  $\blacktriangleright$  ukazuje na metaznak, potom zistíme o akú operáciu ide (ak má viac znakov, treba na to niekoľko stavov – napr. lookahead) a začneme ju vykonávať.

<sup>9</sup>Každý regex je má z definície konečnú dĺžku, to znamená aj konečný počet stavov.



Pracovná páska bude slúžiť ako úložisko smerníkov na rôzne miesta vstupnej pásky. Bude mať formát  $A_1\#A_2\#\dots\#A_m$ . Adresu nejakého políčka na vstupnej páske vieme zapísať v priestore  $\log n$ . Ak ukážeme, že  $m$  je konštanta, potom  $m \cdot \log n \in O(\log n)$ . Adresa  $A_1$  bude rezervovaná pre prvý adresný slot na aktuálnu pozíciu hlavy na vstupe, nech si ju máme odkiaľ okopírovať, keď treba. Adresy budeme využívať pri operáciách spätná referencia, lookahead a lookbehind.

Keďže celý regex vidíme pri konštrukcii Turingovho stroja, vieme si do stavov zakódovať význam a poradie adresných slotov. A celú pracovnú pásku si predpripraviť (napísať potrebný počet  $\#$ ).

Teraz skonštruujeme Turingov stroj  $M = (K, \Sigma, \delta, q_0, F)$  k regexu  $\alpha$ .

$$K = \{\beta \blacktriangleright \xi \mid \beta, \xi \text{ sú podslová } \alpha \text{ také, že } \beta\xi = \alpha\}$$

$$\Sigma = \Sigma(\alpha)^{10}, \quad q_0 = \blacktriangleright \alpha, \quad F = \{\alpha \blacktriangleright\}$$

$\delta$  : (v tvare: stav, znak čítaný na vstupnej páske)

Kvôli prehľadnosti popíšeme len hlavné kroky algoritmu.

$\delta(\beta \blacktriangleright a\xi, a) = \{(\beta a \blacktriangleright \xi, 1)\} \forall a \in \Sigma$  – Ak je v regexe znak, matchujeme ho s tým na vstupe.

$\delta(\beta \blacktriangleright (\gamma)\xi, a) = \{(\beta(\blacktriangleright \gamma)\xi, 0)\}$  – Ak sú to  $k$ -te zátvorky a regex obsahuje  $\backslash k$ , zapíšeme aktuálnu adresu ako adresu začiatku pre  $\backslash k$ .

$\delta(\beta(\gamma \blacktriangleright)\xi, a) = \{(\beta(\gamma) \blacktriangleright \xi, 0)\}$  – Ak sú to  $k$ -te zátvorky a regex obsahuje  $\backslash k$ , zapíšeme aktuálnu adresu ako adresu konca pre  $\backslash k$ . Používame poloopený interval – znak na koncovej adrese do podslova nepatrí.

$\delta(\beta(\gamma) \blacktriangleright * \xi, a) = \{(\beta(\blacktriangleright \gamma) * \xi, 0), (\beta(\gamma) * \blacktriangleright \xi, 0)\}$  – Kleeneho  $*$ : buď opakujeme alebo pokračujeme ďalej.

$\delta(\beta \blacktriangleright \backslash k \xi, a) = \{(q_{najdi\_zaciatok(k)}, 0)\}$  – Najprv si zapamätáme aktuálnu pozíciu na vstupe (ďalej označovanú ako 'aktuálna pracovná pozícia'). Stav  $q_{najdi\_zaciatok(k)}$  zodpovedá presunu hlavy na vstupnej páske na začiatočnú pozíciu podslova. Tu začneme algoritmus porovnávania podslov podľa definície spätnej referencie – aké

<sup>10</sup>T.j. všetky znaky a metaznaky použité v regexe  $\alpha$ .

podслово matchujú  $k$ -te zátvorky, také isté musí ležať aj na 'aktuálnej pracovnej pozícii'. Algoritmus bude porovnávať vždy postupne po jednom znaku od začiatkovej pozície (ľavá zátvorka) po (koniec - 1) (pravá zátvorka):

pokiaľ zaciatok != koniec:

    zapamätaj si znak

    začiatok++

    presuň hlavu na pozíciu 'aktuálna pracovná pozícia'

    porovnaj znak (ak neseď, zasekni sa)

    (aktuálna pracovná pozícia)++

    presuň hlavu na pozíciu 'začiatok'

presuň hlavu na pozíciu 'aktuálna pracovná pozícia'

Vidíme, že si dočasne potrebujeme zapamätať adresu 'aktuálna pracovná pozícia', ale po skončení tohto algoritmu ju môžeme zahodiť.

Po úspešnom zbehtutí algoritmu TS prejde do stavu  $\beta \setminus k \blacktriangleright \xi$ .

$\delta(\beta \blacktriangleright (?=\gamma)\xi, a) = \{(\blacktriangleright \gamma, 0)\}$  – Lookahead: zapamätáme si aktuálnu pozíciu na vstupe do zodpovedajúceho slotu na pracovnej páske a spracujeme regex vnútri lookaheadu. Ak uspejeme, presunieme hlavu na vstupnej páske naspäť na zapamätanú pozíciu a pokračujeme v regexe ďalej:  $\delta(\gamma \blacktriangleright, a) = \{(\beta(?=\gamma) \blacktriangleright \xi, 0)\}$ .

$\delta(\beta \blacktriangleright (?<=\gamma)\xi, a) = \{(\text{doľava}_\gamma, 0)\}$  – Lookbehind: zapamätáme si aktuálnu pozíciu na vstupe do zodpovedajúceho slotu na pracovnej páske. Toto je zarážka pre lookbehind – svoje matchovanie musí skončiť na tejto pozícii tak, že tento znak už neberie do úvahy. Nedeterministicky sa vrátíme o niekoľko políčok doľava ( $\delta(\text{doľava}_\gamma, a) = \{(\text{doľava}_\gamma, -1), (\blacktriangleright \gamma, 0)\}$ ) a skúsime matchovať regex v lookbehinde. Keď uspejeme, porovnáme aktuálnu pozíciu so zarážkou. Ak sú rôzne, Turingov stroj sa zasekne. Inak pokračuje vo výpočte ďalej, od tejto pozície:  $\delta(\gamma \blacktriangleright \text{overene}, a) = \{(\beta(?<=\gamma) \blacktriangleright \xi, 0)\}$ .

### Počet adries:

Pre **spätné referencie** potrebujeme vždy 2 adresy – na začiatok a koniec. Ak sa náhodou budú  $k$ -te zátvorky opakovať, napr. kvôli Kleeneho \*, v definícii stojí, že sa vždy berie do úvahy posledný výskyt, takže adresy prepisujeme pri každom opakovaní.

V prípade, že sa  $k$ -te zátvorky nachádzajú vnútri lookaheadu/lookbehindu, tiež máme pre ne rezervované 2 sloty. Algoritmus porovnávania potrebuje 1 ďalšiu adresu – aktuálnu pracovnú pozíciu. Po jeho dokončení adresu môžeme vymazať (tzn. v ďalšom výpočte prepísať niečím iným).

V prípade **lookaheadu a lookbehindu** spotrebujeme len 1 adresný slot a to tiež len dočasne – dokým sa operácia celá nevykoná. Potom je nám tento údaj zbytočný. Tieto operácie však môžu byť vnorené a tak v najhoršom prípade zaberú  $p \cdot \log n$  priestoru, ak ich počet je  $p$ .

Ak máme 1 rezervovaný slot pre aktuálnu adresu,  $s$  spätných referencií,  $l_a$  lookaheadov a  $l_b$  lookbehindov, najviac spotrebujeme  $(1 + 2s + 1 + l_a + l_b) \log n$  priestoru. Celý regex  $\alpha$  je konečne dlhý, teda počet operácií je konečný. Čo znamená, že  $m = 1 + 2s + 1 + l_a + l_b$  je konštanta a to sme chceli dokázať.

□

**Veta 2.1.15** (Savitch). *Nech  $S(n) \geq \log n$  je páskovo konštruovateľná, potom*

$$NSPACE(S(n)) \subseteq DSPACE(S^2(n))$$

**Dôsledok 2.1.16.**  $\mathcal{L}_{LRE} \subseteq DSPACE(\log^2 n)$ , kde  $n > 1$  je veľkosť vstupu.

**Veta 2.1.17.**  $\mathcal{L}_{nLRE} \subseteq DSPACE(\log^2 n)$ , kde  $n > 1$  je veľkosť vstupu.

*Dôkaz.* **TODO!!!**

□

**Definícia 2.1.18.** *Konfiguráciou regexu  $\alpha = r_1 \dots r_n$  nazývame dvojicu  $(r, w)$ , kde  $r \in (\lceil \alpha \rceil \cup (\alpha \rceil) \cup (r_1 \dots \lceil r_i \dots r_n)$   $w \in \Gamma^* \lceil \Gamma^*$  a symbol  $\lceil$  ukazuje, kde sa nachádzame vo výpočte v regexe a v slove.*

**Definícia 2.1.19.** *Konfiguráciu  $(r_1 \dots r_n \lceil, w_1 \dots w_m \lceil)$  nazývame **akceptačná**.*

**Definícia 2.1.20.** *Zátvorka  $($  v regexe  $\alpha$  je **indexovateľná**, ak sa bezprostredne za ňou nenachádza metaznak  $?$  a zároveň sa nenachádza vnútri negatívneho lookaroundsu.*

Je zakázané odkazovať sa spätnými referenciami na operácie formy  $(? \dots)$ , preto ich ani nechceme a nebudeme brať do úvahy v poradí zátvoriek. Z týchto operácií sa v našom modeli nachádza iba pozitívny a negatívny lookaround. Je povolené odkazovať sa na zátvorky vnútri lookaroundsu, takže sa vieme odvolať na podslovo, čo sa zhoduje

s pozitívnu formou (ak lookaround považujeme za neindexovateľné zátvorky, stačí ho prepísať do formy  $(?= \dots)$  a vieme sa referencovať na jeho obsah). Problém nastáva pri jeho negatívnej verzii – podľa definície nesmie nájsť žiadnu zhodu, inak sa výpočet zastaví. Potom po akceptácii nedefinuje žiadne podslovo, na ktoré by sme sa mohli odvolať. To isté platí o ľubovoľných zátvorkách v jeho vnútri. **TODO!!!** ale na samotné matchovanie vnútri potrebuje aj spätné referencie

**Definícia 2.1.21.** Zátvorka  $)$  v regexe  $\alpha$  je **indexovateľná**, ak  $k$  nej prislúchajúca otváracia zátvorka je indexovateľná.

**Definícia 2.1.22.** Nech regex  $\alpha \in nLRegex$  obsahuje alternáciu. Potom jeho podslovo  $\beta$  nazývame **alternovateľným**, ak je validným regexom z nle a zároveň zodpovedá jednej z týchto podmienok:

(i) **prvá**

$\beta$  je prefix  $\alpha$  alebo znak pred  $\beta$  je <sup>11</sup>

$\wedge$  za  $\beta$  nasleduje metaznak  $|$

(ii) **stredná**

znak pred  $\beta$  je  $|$

$\wedge \beta$  je dobre uzátvorkovaný výraz

$\wedge$  za  $\beta$  nasleduje  $|$

(iii) **posledná**

znak pred  $\beta$  je  $|$

$\wedge \beta$  je suffix  $\alpha$  alebo za  $\beta$  nasleduje  $)$  <sup>12</sup>

**Lema 2.1.23.** Alternácia používa iba alternovateľné regexy.

*Dôkaz.* Ukážeme indukciou na počet alternácií v regexe.

Majme regex  $\alpha \in nLRegex$ , ktorý obsahuje alternáciu. Bez újmy na všeobecnosti nech je to práve 1 alternácia (potenciálne zložená z viacerých metaznakov  $|$ ). Podľa tabuľky priorít vieme, že alternácia má najnižšiu prioritu. Regex  $\alpha$  môže byť v takýchto tvaroch:

<sup>11</sup>Z podmienky hovoriacej, že  $\beta$  musí byť validný regex, vyplýva, že prislúchajúca  $)$  sa bude nachádzať až za metaznakom  $|$

<sup>12</sup>To isté ako pri podmienke (i) – prislúchajúca  $($  sa musí nachádzať pred metaznakom  $|$  (susediacim s  $\beta$ )

1.  $\beta_1 | \dots | \beta_n$
2.  $\alpha_1(\beta_1 | \dots | \beta_n)\alpha_2$

Pre  $\alpha_1, \alpha_2, \beta_i \in nLRegex$   $i \in \{1, \dots, n\}$ . Je zrejmé, že  $\beta_1$  spĺňa podmienku prvého alternovateľného regexu,  $\beta_i$  pre  $i \in \{2, \dots, n-1\}$  spĺňajú podmienku pre stredný alternovateľný regex a  $\beta_n$  spĺňa podmienku pre posledný alternovateľný regex.

Nech platí pre regexy obsahujúce  $m-1$  alternácií, že tieto alternácie používajú iba alternovateľné regexy. Zoberme teraz regex  $\alpha \in nLRegex$  obsahujúci  $m$  alternácií. Opäť vieme  $\alpha$  rozdeliť buď spôsobom 1. alebo 2. na validné regexy z  $nLRegex$ . Každý z  $\beta_1, \dots, \beta_n, \alpha_1, \alpha_2$  obsahuje najviac  $m-1$  alternácií, teda alternácie v týchto regexoch používajú iba alternovateľné regexy. Navyše je zrejmé, že aj regexy  $\beta_1, \dots, \beta_n$  majú tvar alternovateľných regexov.  $\square$

Na základe definovania pojmu alternovateľnosti vieme jednoznačne určiť, ktoré regexy patria ku ktorej alternácii. Dôležité je, že to vieme naprogramovať do Turingovho stroja – pre každú alternáciu bude skúmať, či je ohraničená zátvorkami alebo rozdeľuje celý regex na časti (teda či nastáva prípad 1. alebo 2. z dôkazu predchádzajúcej vety).

**Definícia 2.1.24.** *Krok výpočtu* je relácia  $\vdash$  na konfiguráciách definovaná nasledovne:

$$I. \forall a \in \Sigma : (r_1 \dots [a \dots r_n, w_1 \dots [a \dots w_m) \vdash (r_1 \dots a [\dots r_n, w_1 \dots a [\dots w_m)$$

$$II. \text{Nech } ( \text{ je indexovateľná, } k\text{-ta v poradí: } (r_1 \dots [(\dots r_n, w_1 \dots [w_j \dots w_m)$$

$$(1) \vdash (r_1 \dots ([\dots r_n, w_1 \dots [w_j^k \dots w_m)$$

Ak za jej uzatváracou zátvorkou nasleduje  $*$ , t.j.  $\alpha$  je tvaru  $r_1 \dots [(\dots) * \dots r_n$ , potom

$$(2) \vdash (r_1 \dots (\dots) * [\dots r_n, w_1 \dots [w_j^{k'} \dots w_m)$$

$$III. \text{Nech } ) \text{ je indexovateľná, } k\text{-ta v poradí:}$$

$$(r_1 \dots [) \dots r_n, w_1 \dots [w_j \dots w_m) \vdash (r_1 \dots ) [\dots r_n, w_1 \dots [w_j^{k'} \dots w_m)$$

$$IV. \text{Nech podslová } \alpha_1, \alpha_2, \dots, \alpha_A \text{ regexu } \alpha \text{ sú všetkými členmi zobrazenej alternácie:}$$

$$(r_1 \dots [\alpha_1 | \alpha_2 | \dots | \alpha_A \dots r_n, w_1 \dots [w_j \dots w_m)$$

$$(1) \vdash \text{ďalší prechod v } \alpha_1$$

- $$(2) \vdash (r_1 \dots \alpha_1 | \lceil \alpha_2 | \dots | \alpha_A \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$
- $$\dots$$
- $$(A) \vdash (r_1 \dots \alpha_1 | \alpha_2 | \dots | \lceil \alpha_A \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$
- V.  $(r_1 \dots \underset{k}{(r_i \dots r_l)} \lceil \underset{k}{*} \dots r_n, w_1 \dots \underset{k}{w_a} \dots \underset{k'}{w_b} \dots \lceil w_j \dots w_m)^{13}$
- $$(1) \vdash (r_1 \dots \underset{k}{(r_i \dots r_l)} \underset{k}{*} \dots r_n, w_1 \dots \underset{k}{w_a} \dots \underset{k'}{w_b} \dots \lceil w_j \dots w_m)$$
- $$(2) \vdash (r_1 \dots \underset{k}{(\lceil r_i \dots r_l)} \underset{k}{*} \dots r_n, w_1 \dots w_a \dots w_b \dots \lceil \underset{k}{w_j} \dots w_m)$$
- VI.  $(r_1 \dots \lceil \backslash k \dots r_n, w_1 \dots \underset{k}{w_a} \dots \underset{k'}{w_b} \dots \lceil w_j \dots w_m)$
- $$\vdash (r_1 \dots \lceil \backslash k \dots r_n, w_1 \dots \mathsf{T} \underset{k}{w_a} \dots \underset{k'}{w_b} \dots \lceil w_j \dots w_m)$$
- VII.  $(r_1 \dots \lceil \backslash k \dots r_n, w_1 \dots \underset{k}{w_a} \dots \mathsf{T} w_c \dots \underset{k'}{w_b} \dots \lceil w_j \dots w_m), \text{ kde } a \leq c \leq b \text{ a zároveň } w_c = w_j^{14}$
- $$\vdash (r_1 \dots \lceil \backslash k \dots r_n, w_1 \dots \underset{k}{w_a} \dots w_c \mathsf{T} \dots \underset{k'}{w_b} \dots w_j \lceil \dots w_m)$$
- VIII.  $(r_1 \dots \lceil \backslash k \dots r_n, w_1 \dots \underset{k}{w_a} \dots \mathsf{T} \underset{k'}{w_b} \dots \lceil w_j \dots w_m)$
- $$\vdash (r_1 \dots \lceil \backslash k \dots r_n, w_1 \dots \underset{k}{w_a} \dots \underset{k'}{w_b} \dots w_j \lceil \dots w_m)$$
- IX. *Nech  $(?= \dots)$  je  $k$ -ty pozitívny lookahead v poradí:*
- $$(r_1 \dots \lceil (?= \dots) \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$
- $$\vdash (r_1 \dots (?= \lceil \dots) \dots r_n, w_1 \dots \lceil \overset{k}{\vec{w}}_j \dots w_m)$$
- X. *Nech  $\rangle$  patrí ku  $k$ -temu pozitívnemu lookaheadu v poradí:*
- $$(r_1 \dots (?= \dots \lceil) \dots r_n, w_1 \dots \overset{k}{\vec{w}}_l \dots \lceil w_j \dots w_m)$$
- $$\vdash (r_1 \dots (?= \dots) \lceil \dots r_n, w_1 \dots \lceil w_l \dots w_j \dots w_m)$$

<sup>13</sup>Podľa definície spätných referencií platí podsledné podslovo nájdené regexom v  $k$ -tych zátvorkách.

Pri tejto pracovnej pozícii v regexe je zrejmé, že nejde o prvý prechod cez tieto zátvorky a teda existuje také  $a, b$ , že  $k$  je v slove nad  $w_a$  a  $k'$  nad  $w_b$ . Ak nastane prechod (2), pôvodné horné indexy  $k, k'$  miznú a pridáva sa  $k$  nad  $w_j$ .

<sup>14</sup> $w_c$  a  $w_j$  môžu byť poschodové symboly, avšak pri tejto rovnosti poschodia ignorujeme – chceme porovnať iba písmenká v slove, prislúchajúce týmto pozíciám.

XI. Nech  $(?<=\dots)$  je  $k$ -ty pozitívny lookbehind v poradí,  $\forall L \in \{0, \dots, j-1\}$ :

$$(r_1 \dots \lceil (?<=\dots) \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$

$$\vdash (r_1 \dots (?<=\lceil \dots) \dots r_n, w_1 \dots \lceil w_{j-L} \dots \overset{k\leftarrow}{w}_j \dots w_m)$$

XII. Nech  $\lceil$  patrí ku  $k$ -temu pozitívnemu lookbehindu v poradí:

$$(r_1 \dots (?<=\dots \lceil) \dots r_n, w_1 \dots \lceil \overset{k\leftarrow}{w}_j \dots w_m)$$

$$\vdash (r_1 \dots (?<=\dots) \lceil \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$

XIII. Ak  $\nexists p \in \{j, \dots, m\} : (\lceil r_k \dots r_l, \lceil w_j \dots w_p) \vdash^* (r_k \dots r_l \lceil, w_j \dots w_p \lceil)$ , potom:

$$(r_1 \dots \lceil (?! r_k \dots r_l) \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$

$$\vdash (r_1 \dots (?! r_k \dots r_l) \lceil \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$

XIV. Ak  $\nexists p \in \{1, \dots, j-1\} : (\lceil r_k \dots r_l, \lceil w_p \dots w_{j-1}) \vdash^* (r_k \dots r_l \lceil, w_p \dots w_{j-1} \lceil)$ , potom:

$$(r_1 \dots \lceil (?<! r_k \dots r_l) \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$

$$\vdash (r_1 \dots (?<! r_k \dots r_l) \lceil \dots r_n, w_1 \dots \lceil w_j \dots w_m)$$

**Definícia 2.1.25.** *Jazyk generovaný regexom  $\alpha$  je množina*

$$L(\alpha) = \{w \mid \text{platí } (\lceil \alpha, \lceil w) \vdash^* (\alpha \lceil, w \lceil)\}$$

**Poznámka.** Postupnosť konfigurácií  $(\lceil \alpha, \lceil w) \vdash^* (\alpha \lceil, w \lceil)$  pre daný regex  $\alpha$  a slovo  $w$  nazývame **akceptačný výpočet**.

**Lema 2.1.26.** *Nech  $\alpha \in \mathcal{L}_{LRE}$  a  $w \in L(\alpha)$ . Potom existuje akceptačný výpočet, ktorý má najviac  $5 \cdot |\alpha| \cdot |w|$  konfigurácií.*

*Dôkaz.* Pri prechodoch medzi konfiguráciami I., II., III., IV., V.(1), VII., IX., XII., XIII. a XIV. sa ukazovateľ posúva vždy vpred buď v slove alebo v regexe alebo v oboch. Keby sme využívali iba tieto prechody, akceptačný výpočet má najviac  $|\alpha| + |w|$  konfigurácií.

Existujú prechody, pri ktorých sa žiaden ukazovateľ nepohne: VI. a VIII. Tieto prechody nastávajú pri spätných referenciách a jedná sa o objavenie sa alebo zmiznutie pomocného ukazovateľa. Pri každej spätnej referencii nastávajú oba javy práve jedenkrát. Zároveň počet spätných referencií je najviac  $|\alpha|$ , teda tieto 2 konfigurácie sa objavia dokopy najviac  $(2 \cdot |\alpha|)$ -krát.

Zostali konfigurácie, v ktorých ukazovateľ skáče dozadu, rozoberme si ich postupne:

V.(2) – v prípade Kleeneho  $*$  nastáva ďalšie opakovanie. Keďže  $w \in L(\alpha)$ , existuje akceptačný výpočet. Takýchto výpočtov môže byť viac, napríklad ak sa v  $\alpha$  vyskytuje regex  $(\beta)^*$  taký, že  $\varepsilon \in \beta$ . Potom existuje nekonečne veľa výpočtov, ktoré sa líšia v počte prechodov týmto regexom. Preto si vyberieme najkratší akceptačný výpočet. To znamená, že v každom prechode cez regex opakovaný operáciou  $*$  musíme nájsť zhodu aspoň s jedným písmenkom z  $w$  (t.j. ukazovateľ v slove  $w$  sa pohne aspoň o 1 pozíciu doprava). Regex opakovaný Kleeneho  $*$  je dlhý najviac  $|\alpha|-1$  znakov a opakujeme ho najviac  $w$ -krát. Z toho vidíme, že existuje akceptačný výpočet, ktorý má najviac  $|\alpha| \cdot |w|$  konfigurácií.

X., XI. – ukončenie operácie lookahead, začiatok operácie lookbehind. V tomto prípade ukazovateľ skáče dozadu v slove. Podstatné však je, že v regexe sa posúva dopredu. Takýchto prechodov vieme spraviť najviac  $|\alpha|$ .

Spolu je to najviac  $|\alpha| + |w| + 2 \cdot |\alpha| + |\alpha| \cdot |w| + |\alpha| \leq 5 \cdot |\alpha| \cdot |w|$  konfigurácií.

□

Príkladom toho, že táto lema poskytuje tesný odhad, je jazyk  $(a * b * c * d * e^*)^*$  a slovo  $edcba$ . V najkratšom akceptačnom výpočte v každom opakovaní  $*$  naozaj nájdeme zhodu práve s jedným písmenkom.

V praxi často regex dopredu nepoznáme. Vstupným údajom je text na vyhľadávanie a rovnako aj regex, ktorý znázorňuje požiadavku na vyhľadávanie. Tento problém predstavuje jazyk

$$L(\text{regex}\#\text{word}) = \{\text{word} \mid \text{word} \in L(\text{regex}) \wedge \text{regex} \in \mathcal{U}\}$$

kde  $\mathcal{U} \in \{\text{Regex}, \text{Eregex}, \text{LEregex}, \text{nLEregex}\}$ .

**Veta 2.1.27.**  $L(\text{regex}\#\text{word}) \in \text{NSPACE}(r \log w)$ , kde  $r = |\text{regex}|$ ,  $w = |\text{word}|$  a  $\text{regex} \in \text{LEregex}$ .

*Dôkaz.* **TODO!!!** Zostrojíme nedeterministický Turingov stroj  $M$ , ktorý bude akceptovať jazyk  $L(\text{regex}\#\text{word})$  a na pracovných páskach zapíše najviac  $O(r \log w)$  políčok.

Na pracovnej páske

□

**Veta 2.1.28.**  $L(\text{regex}\#\text{word}) \in \text{DSPACE}(n \log^2 n)$ , kde  $\text{regex} \in \text{LEregex}$  a  $n$  je dĺžka vstupu.



*Dôkaz.* Pre účely dôkazu budeme označovať  $w = |word|$  a  $r = |regex|$ .

Zostrojíme deterministický Turingov stroj  $M$ , ktorý bude akceptovať jazyk  $L(regex\#word)$  a na každej pracovnej páske použije najviac  $O(n \log^2 n)$  políčok.  $M$  bude mať vstupnú read-only pásku a 2 pracovné pásy.

Myšlienka je podobná dôkazu Savitchovej vety –  $M$  bude vykonávať funkciu  $TESTUJ(C_1, C_2, i)$ , ktorá zistí, či sa vieme dostať z konfigurácie  $C_1$  do konfigurácie  $C_2$  na  $i$  krokov.

Podľa lemy 2.1.26 vieme, že ak existuje akceptačný výpočet pre  $w$ , potom existuje aj akceptačný výpočet taký, ktorý má najviac  $5rw$  konfigurácií. Na základe tohto výsledku bude  $M$  zisťovať, či sa z počiatočnej konfigurácie  $C_0$  vieme dostať do akceptačnej konfigurácie  $C_a$  na  $5rw$  krokov –  $TESTUJ(C_0, C_a, 5rw)$ .

Pseudokód procedúry  $TESTUJ$ :

```

1  bool TESTUJ( $C_1, C_2, i$ )
2      if ( $C_1 == C_2$ ) then return true
3      if ( $i > 0 \wedge C_1 \vdash C_2$ ) then return true
4      if ( $i \leq 1$ ) return false
5      iteruj cez všetky konfigurácie  $C_3$ 
6          if ( $TESTUJ(C_1, C_3, \lfloor \frac{i}{2} \rfloor) \wedge TESTUJ(C_3, C_2, \lceil \frac{i}{2} \rceil)$ ) then return true
7      return false

```

Pre podrobnejší popis pseudokódu je nutné, aby sme najprv definovali tvar konfigurácie. Použijeme zápis z definície 2.1.24, v tvare  $(a_1 \dots \lceil a_i \dots a_r, b_1 \dots \lceil b_j \dots b_w \rceil)$  respektíve  $(a_1 \dots \lceil a_i \dots a_r, b_1 \dots \lceil b_j \dots b_w \rceil)$ , kde  $a_1 \dots a_r = regex$  a  $b_1 \dots b_w = word$ . Reprezentovať ich budeme vo forme 2 resp. 3 adries – pracovná pozícia v regexe ( $i$ ), pracovná pozícia v slove ( $j$ ) a adresa k spätnej referencii ( $l$ ). Namiesto poschodových symbolov si  $M$  bude pre každú konfiguráciu pamätať informáciu, ktoré zátvorky zodpovedajú ktorému podslovu slova  $word$ . Pre každé zátvorky si uloží 2 adresy – začiatok podslova a 1 políčko za koncom podslova (použijeme polootvorený interval  $\langle z, k \rangle$ ). Pre každý lookahead a lookbehind bude mať vyhradené 1 adresné miesto aby si zapamätal, kam do slova ukazoval, keď naňho narazil.

Popis pseudokódu:

**riadok 2** Ak  $C_1 = C_2$ , potom vieme prejsť z  $C_1$  do  $C_2$  na ľubovoľný počet krokov.

**riadok 3** Platí  $C_1 \neq C_2$ . Ak  $i = 0$ , nevieme prejsť do žiadnej inej konfigurácie ako  $C_1$ ,

teda vrátíme **false**. Nech  $i > 0$ . Skontrolujeme podľa definície 2.1.24, či platí  $C_1 \vdash C_2$ . Konfigurácie sú uložené na páske ako  $m$ -tice, kde  $m = 3 + 2 \cdot (\text{počet zátvoriek}) + (\text{početlookaheadov}) + (\text{početlookbehindov})$ :  $C_1 = (d_1, \dots, d_m)$ ,  $C_2 = (e_1, \dots, e_m)$ . Nech  $d_1, e_1$  sú pracovné adresy v regexe,  $d_2, e_2$  sú pracovné adresy v slove a  $d_3, e_3$  pracovné adresy pri spätných referenciách, pričom  $d_3, e_3$  môžu byť nedefinované (na prislúchajúcom mieste medzi oddeľovačmi adries nebude nič zapísané). TS M overí, či je splnená nejaká z týchto podmienok (rímske čísla zodpovedajú tým v definícii 2.1.24):

- I.  $regex[d_1] = word[d_2] \wedge (e_1, \dots, e_m) = (d_1 + 1, d_2 + 1, nedef, d_4, \dots, d_m)$
- II. (1)  $regex[d_1] = '('$   
 $\wedge regex[d_1]$  je indexovateľná  
 $\wedge$  nech  $d_k$  prislúcha k  $regex[d_1]$ ,  $4 \leq k \leq m$ :  $d_2 = d_k$  (sedí začiatok podslova)  
 $\wedge (e_1, \dots, e_m) = (d_1 + 1, d_2, nedef, d_4, \dots, d_m)$
- II. (2)  $regex[d_1] = '('$   
 $\wedge$  2 políčka pred pozíciou  $e_2$  je v regexe  $'\ast'$   
 $\wedge$  zátvorky  $regex[d_1]$  a  $regex[e_1 - 2]$  sú k sebe prislúchajúce<sup>15</sup>  
 $\wedge (e_1, \dots, e_m) = (d_1 + l, d_2, nedef, d_4, \dots, d_k, \dots, d_m)$  pre  $l \in \mathbb{N}$
- III. podobne ako II.(1) – kontrola, či sedí koniec podslova
- IV.  $regex[e_1 - 1] = '|'$   
 $\wedge$  regex medzi  $d_1$  a najbližším metaznakom  $|$  je alternovateľný  
 $\wedge$  všetky regexy ohraňované  $|$  medzi  $d_1$  a  $e_1$  sú alternovateľné  
 $\wedge (e_1, \dots, e_m) = (d_1 + l, d_2, nedef, d_4, \dots, d_m)$  pre  $l \in \mathbb{N}$
- V. (1)  $regex[d_1] = '\ast'$   $\wedge (e_1, \dots, e_m) = (d_1 + 1, d_2, nedef, d_4, \dots, d_m)$
- V. (2)  $regex[d_1] = '\ast'$   
 $\wedge regex[e_1 - 1] = '('$  a prislúcha k  $regex[d_1]$   
 $\wedge$  nech  $e_k, e_{k+1}$  prislúchajú k týmto zátvorkám, potom  $e_k = e_2$  a  $e_k \leq e_{k+1}$   
 $\wedge (e_1, \dots, e_m) = (d_1 - l, d_2, nedef, d_4, \dots, d_{k-1}, e_k, e_{k+1}, d_{k+2}, \dots, d_m)$  pre  $l \in \mathbb{N}$
- VI. na pozícii  $d_1$  je podslovo  $'\backslash k'$  pre nejaké  $k \in \mathbb{N}$ ,  $0 \leq k \leq (\text{počet indexovateľných zátvoriek})$

<sup>15</sup>To M skontroluje tak, že si overí, že medzi nimi ku každej  $'('$  existuje  $')$  – teda počet výskytov  $'('$  a  $')$  musí byť rovnaký.

- $\wedge (e_1, \dots, e_m) = (d_1, d_2, d_{k+3}, d_4, \dots, d_m)$  ( $d_{k+3}$  je adresa začiatku podslova prislúchajúca ku  $k$ -tým zátvorkám)
- VII. podobne ako I.– musí platiť  $d_3 < d_{k+3}$  a správne sa posunú adresy  $d_2, d_3$
- VIII. na pozícii  $d_1$  je podslovo ' $\backslash k$ '
- $\wedge d_3 = d_{k+3}$
- $\wedge (e_1, \dots, e_m) = (d_1, d_2, \text{ndef}, d_4, \dots, d_m)$
- IX. na pozícii  $d_1$  je podslovo ' $(?=$ '
- $\wedge$  nech  $d_l$  adresa prislúchajúca k tomuto lookaheadu:  $e_l = d_2$ , teda  $(e_1, \dots, e_m) = (d_1 + 3, d_2, \text{ndef}, d_4, \dots, d_{l-1}, d_2, d_{l+1}, \dots, d_m)$
- X.  $\text{regex}[d_1] = )'$  a prislúcha lookaheadu, ktorému patrí adresa  $d_l$
- $\wedge (e_1, \dots, e_m) = (d_1 + 1, d_l, \text{ndef}, d_4, \dots, d_{l-1}, \text{ndef}, d_{l+1}, \dots, d_m)$
- XI. podobne ako IX. – zapíšeme si aktuálnu adresu v slove
- XII.  $\text{regex}[d_1] = )'$  a prislúcha lookbeindu, ktorému patrí adresa  $d_l$
- $\wedge d_2 = d_l$
- $\wedge (e_1, \dots, e_m) = (d_1 + 1, d_2, \text{ndef}, d_4, \dots, d_{l-1}, \text{ndef}, d_{l+1}, \dots, d_m)$

Keďže regex neobsahuje negatívny lookbaround, ďalšie riadky z definície netestujeme.

**riadok 4** Po predošlých riadkoch platí  $C_1 \neq C_2 \wedge C_1 \not\leq C_2$ . Ak zároveň  $i \leq 1$ , potom sa z  $C_1$  do  $C_2$  na  $i$  krokov dostať nedokážeme. *TESTUJ* vráti **false**.

**riadok 5** M začne iterovať cez všetky možné konfigurácie – t.j. všetky možné kombinácie adries, ktoré ukazujú najďalej 1 políčko za regex/slovo. Vygenerované  $C_3$  bude mať M uložené ako lokálnu premennú.

**riadok 6** M testuje, či je  $C_3$  vo výpočte v strede medzi  $C_1$  a  $C_2$ . Ak obe volané procedúry vrátia **true**, vrátime **true**. Inak sa M vráti na **riadok 5** a vygeneruje ďalšie  $C_3$ .

**riadok 7** Neexistuje vhodná konfigurácia  $C_3$ , teda sa nevieme dostať z  $C_1$  do  $C_2$  na  $i$  krokov. Vrátime **false**.

Podľa vyššie špecifikovaných  $m$ -tíc pre konfigurácie bude úvodné nastavenie nasledovné:

$$C_0 = ( \quad 1, \quad 1, \text{ndef}, 0, 0, \dots, 0, \text{ndef}, \dots, \text{ndef} )$$

$$C_a = (r + 1, w + 1, \underbrace{ndef, 0, 0, \dots, 0}_{\text{adresy pre spätné referencie}}, \underbrace{ndef, \dots, ndef}_{\text{adresy pre lookahead a lookbehind}})$$

Turingov stroj  $M$  rekurzívne volá procedúru  $TESTUJ$ . Preto na prvej páske bude mať zásobník, kde budú uložené záznamy o jednotlivých volaniach. Druhá páska je pomocná pri vykonávaní konkrétneho volania –  $M$  potrebuje konštantný počet adries, aby mohol realizovať porovnávanie, zisťovanie príslušnosti zátvoriek, zisťovanie, koľký v poradí je daný lookahead/lookbehind/otváracia zátvorka, ...

Zrejme ak existuje akceptačný výpočet,  $M$  ho nájde. Treba ukázať, že sa pri tom na každej páske zmestí do pamäte  $O(n \log^2 n)$ . Podľa predchádzajúceho odstavca vieme, že na druhej (pomocnej) páske potrebuje  $O(\log n)$  políček, čo spĺňa podmienku. Spočítajme veľkosť potrebnú zásobníka.

Adresy vieme zapísať v logaritmickom priestore závislom od dĺžky slova, kam ukazujú. Pre regex to bude  $\log r$  a pre slovo  $\log w$ , pretože vieme adresovať od oddeľovača  $\#$ . Číslo  $i$  je najviac  $5rw$  a tiež ho vieme zapísať v logaritmickom tvare, čo zaberie  $\log(5rw) = \log 5 + \log r + \log w$  políček. Platí  $r \leq n, w \leq n$ .

Počet zátvoriek, lookaheadov a lookbehindov je v regexe najviac  $r$ , teda jedna konfigurácia bude potrebovať najviac  $\log r + 2 \log w + 2r \log w$  priestoru.

Jeden záznam procedúry  $TESTUJ$  obsahuje 3 konfigurácie a číslo  $i$ , čo spolu zaberá  $\log 5 + 2 \log r + 3 \log w + 2r \log w = O(r \log w) = O(n \log n)$  priestoru.

Počet záznamov na zásobníku závisí od hĺbky rekurzie. Keďže začíname na hodnote  $i = 5rw$  a pri každom volaní je  $i$  zmenšené na polovicu, hĺbka vnorenia bude  $\log(5rw) = O(\log n^2) = O(2 \log n) = O(\log n)$ .

Celkovo  $M$  na zásobníkovej páske zapíše  $O(\log n) \cdot O(n \log n) = O(n \log^2 n)$  priestoru.  $\square$

## 2.2 Popisná zložitosť moderných regulárnych výrazov

V tejto kapitole rozoberieme moderné regulárne výrazy z dvoch hľadísk. Najprv nás bude zaujímať, ako pomohli nové konštrukcie pri popise regulárnych jazykov a potom prejdeme na analýzu dĺžky výrazov pre zložitejšie jazyky.

Lookaround môže výrazne pomôcť pri definovaní konečných jazykov, napríklad le-regex z [Tó13, Poznámka 1.]  $\beta = ((?= (a^m) \$) a^{m+1}) * a \{1, m-1\} \$$ . Aké slová obsahuje?

- $a^{m+1+(m-1)}$
- $a^{2m+2+(m-2)}$
- $\vdots$
- $a^{(m-1)(m+1)+1}$

avšak  $a^{m(m+1)} \notin L(\beta)$ , lebo nám chýba zvyšok 0. Zaujímavé je, že le-regex využíva iteráciu  $(m-1)$ -krát a napriek tomu je konečný.

# Záver

Zhrnutie na záver...

# Literatúra

- [EKSwW13] Keith Ellul, Bryan Krawetz, Jeffrey Shallit, and Ming wei Wang. Regular expressions: New results and open problems. *Journal of Automata, Languages and Combinatorics*  $u(v)w, x-y$ , 2013.
- [EZ] Andrzej Ehrenfeucht and Paul Zeiger. Complexity measures for regular expressions.  
[https://digitool.library.colostate.edu///exlibris/dt1/d3\\_1/apache\\_media/L2V4bGlicmlzL2R0bC9kM18xL2FwYWNoZV9tZWRpYS8xNjYzMDQ=.pdf](https://digitool.library.colostate.edu///exlibris/dt1/d3_1/apache_media/L2V4bGlicmlzL2R0bC9kM18xL2FwYWNoZV9tZWRpYS8xNjYzMDQ=.pdf).
- [Tó13] Tatiana Tóthová. Moderné regulárne výrazy. Bachelor's thesis, 2013.  
<https://github.com/Tatianka/bak/blob/master/tothova-bc.pdf?raw=true> [Online; accessed 22-Nov-2013].