VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

SYSTÉM SLOVENSKÉ MORFOLOGIE ZALOŽENÝ NA VZORECH

SLOVAK PATTERN-BASED MORPHOLOGY

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **ANDREJ KLOCOK**
AUTHOR

VEDOUCÍ PRÁCE **Doc. RNDr. PAVEL SMRŽ, Ph.D.**
SUPERVISOR

BRNO 2017

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2016/2017

Zadání bakalářské práce

Řešitel: **Klocok Andrej**
Obor: Informační technologie
Téma: **Systém slovenské morfologie založený na vzorech**
Slovak Pattern-based Morphology
Kategorie: Umělá inteligence

Pokyny:

1. Seznamte se s metodami morfologické analýzy a reprezentace zdrojových i zpracovaných dat morfologického slovníku v češtině a příbuzných jazycích.
2. Shromážděte slovníková a korpusová data, pokrývající co nejlépe současnou slovenštinu.
3. Navrhněte a implementujte systém, který na základě dat vytvoří systém technických vzorů pro flektivní morfologii slovenštiny, z něj odvozený morfologický analyzátor, případně další odvozené nástroje.
4. Vyhodnoťte úspěšnost vytvořeného systému a diskutujte možnosti dalšího zkvalitňování slovníkové báze i morfologického analyzátoru.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 0-262-13360-1.

Pro udělení zápočtu za první semestr je požadováno:

1. Funkční prototyp

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Smrž Pavel, doc. RNDr., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2016
Datum odevzdání: 17. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Cieľom tejto práce je zoznámenie sa s metódami morfologickej analýzy, reprezentáciou dát morfologických slovníkov, vytvorením systému technických vzorov pre flektívnu morfológiu slovenčiny. Z tohto systému je odvodený morfologický analyzátor, ktorý vstupné slová lematizuje, určí ich vzor a morfologickú značku, nástroj pre porovnávanie a vyhodnocovanie stemerov, ktorý hodnotí stemery na základe derivačného slovníka, nástroj na rekonštrukciu diakritiky, ktorý vznikol ako pomocný nástroj. V posledných kapitolách práce sú jednotlivé nástroje zhodnotené, analyzátor je porovnaný s dostupnou alternatívou, pomocou nástroja na hodnotenie stemerov sú porovnané dve implementácie slovenských stemerov a je naznačený ďalší vývoj jednotlivých nástrojov.

Abstract

The aim of this thesis is to get acquainted with methods of morphological analysis, representation of data of morphological dictionaries, creation of system based on technical patterns for flective morphology of Slovak language. From this system is derived a morphological analyzer, which lemmatizes input words, determines their pattern and a morphological tag, a tool for comparison and evaluation of stemmers, which evaluates stemmers based on a derivative dictionary, a tool for reconstruction of diacritics, which was created as an auxiliary tool. In the last chapters of thesis, individual tools are assessed, morphological analyzer is compared with available alternative, two implementations of Slovak stemmers are evaluated by the tool for stemmer evaluation and the further development of tools is indicated.

Kľúčové slová

morfologická analýza, morfologický analyzátor, lematizácia, lema, morfologický slovník, trie, nástroj pre porovnanie a vyhodnocovanie stemerov, nástroj na rekonštrukciu diakritiky, slovenský korpus

Keywords

morphological analysis, morphological analyzer, lemmatization, lemma, morphological dictionary, trie, tool for comparison and evaluation of stemmers, tool for reconstruction of diacritics, Slovak corpora

Citácia

KLOCOK, Andrej. *Systém slovenské morfologie založený na vzorech.* Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Doc. RNDr. Pavel Smrž, Ph.D.

Systém slovenské morfologie založený na vzorech

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána doc. RNDr. Pavla Smrža, Ph.D. Ďalšie informácie mi poskytol pán Vladimír Benko a vedenie Jazykovedného Ústavu Ľudovíta Štúra Slovenskej akadémie vied v Bratislave. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Andrej Klocok
16. mája 2017

Poďakovanie

Veľmi rád by som sa chcel poďakovať vedúcemu práce pánovi doc. RNDr. Pavlovi Smržovi, Ph.D za jeho podporu pri písaní tejto bakalárskej práce a poskytnutie dát na vyhotovenie morfologického slovníka. Tiež by som chcel poďakovať pánovi Vladimírovi Benkovi za poskytnutie slovenskej časti korpusu rodiny Aranea a vedeniu Jazykovedného Ústavu Ľudovíta Štúra Slovenskej akadémie vied v Bratislave za poskytnutie časti paralelného Slovensko-českého korpusu.

Obsah

Kapitola 1

Úvod

Komunikácia je dôležitá v živote človeka. Každý človek vyrastá so svojím materinským jazykom, v ktorom sa počas života zdokonaľuje, komunikuje pomocou neho a vďaka nemu naberá vedomosti ohľadom iných kultúr a ich používaných jazykov v komunikácii. Jazyk teda môžeme považovať za kľúčovú vlastnosť každého z nás a na každom z nás záleží, ako moc si túto vlastnosť vyvinie.

Každý jazyk sa neustále vyvíja. Napríklad taká slovenčina sa vyvíja už od piateho storočia, kedy slovenské nárečia začínajú vytvárať praslovanský základ slovenčiny. V ďalších storočiach sa slovenčina postupne vyvíja, no zostáva po dlhé storočia len jazykom ľudu a ľudovej slovesnosti. Dnes je slovenčina oficiálne úradným jazykom Slovenskej republiky. Slovenský jazyk patrí medzi západoslovanské jazyky. Slovenčina je úzko spätá s češtinou, oba jazyky sa používali v rámci jednej Československej republiky.

Slovenčina je charakteristická hlavne svojim flektívnym jazykovým typom, ktorý sa prejavuje hlavne v oblasti skloňovania, kde jednou gramatickou príponou môžeme vyjadriť viac gramatických kategórii. V slovenčine sa tiež vyskytuje aj mnoho výnimiek.

Pri spracovaní textu je potrebné z neho vyfiltrovať podstatné informácie, samotné slová. Tieto slová sa nachádzajú vo viacerých slovných tvaroch. V rámci získania informácii musíme tieto slová upraviť na ich základný tvar. Tento proces sa nazýva lematizácia. Pri lematizácii vznikajú komplikácie, napríklad morfologická disabiguancia (viď sekcia 2.7).

Tento proces je veľmi dôležitý, napríklad pri vyhľadávaní termínov v dokumentoch. Bez použitia tohto procesu sa vyhľadávaný termín, slovo nachádzajúce sa v istom slovnom tvare, nemusí nahádzať v žiadnom dokumente. Vďaka lematizácii môžeme previesť vyhľadávací termín na jeho základný tvar a následne ho vyhľadávať medzi jednotlivými základnými tvarmi v dokumente, čo zvýši šancu výskytu.

Bakalárska práca je priamo zameraná na systém slovenskej morfológie. V nasledujúcej kapitole sú popísané morfologické typy jazykov, slovenská morfológia, slovná zásoba a tvorenie slov, morfologická analýza, morfologické značky, formáty uloženia slovníkov a pojem korpus (viď kapitola 2). V kapitole 3 sa nachádza podrobný popis návrhu a implementácie jednotlivých nástrojov ako morfologický analyzátor, ktorý pomocou vzorov dokáže popísať jednotlivé vstupné slová, určiť ich základný tvar, morfologickú značku, slovný vzor, nástroj na vyhodnocovanie a porovnávanie stemerov a nástroj na rekonštrukciu diakritiky, ktorý vznikol ako pomocný nástroj. V kapitole 4 je popísané zhodnotenie vytvorených systémov, porovnanie analyzátora s obdobným nástrojom, porovnanie slovenských stemerov na základe nášho systému na porovnávanie a vyhodnocovanie stemerov, popísané úskalia systémov. V záverečnej kapitole zhŕňame dosiahnuté výsledky a zhodnocujeme ďalší vývoj nástrojov (viď kapitola 5).

Kapitola 2

Analýza témy

V tejto kapitole postupne vysvetľujeme pojmy ako morfológia, morfologické typy jazykov, morfológiu slovenčiny, tvorenie slov. Ďalej naviažeme na morfologickú analýzu, proces lematizácie slov, uvedieme typy používaných morfologických značiek, morfologický slovník a jeho najpoužívanejší formát uloženia a nakoniec vysvetlíme pojem korpus.

2.1 Úvod do morfológie

Morfológia, nazývaná aj tvaroslovie, je jazykovedná náuka o gramatických tvaroch slov a o slovách, ktoré majú funkciu tvarov. Morfológiu teda považujeme za náuku o tvarovej rovine v jazykovom systéme.

Jazykový systém sa skladá z niekoľkých rovín, ktoré sa nazývajú aj plány. Jednotlivé roviny sa vzájomne dopĺňajú. Každá rovina jazyka obsahuje prvky rozličnej stavby. Rozlišuje následné jazykové roviny: fonická, lexikálna, morfologická, syntaktická a štýlová rovina. Všetky roviny jazykového systému sú samostatné a ohraničené. Prechodné javy, ktoré existujú medzi jednotlivými jazykovými rovinami, je nutné vykladať práve z hľadiska obidvoch susedných jazykových rovín [8].

Najužšie sa morfologická rovina primkýna práve k syntaktickej jazykovej rovine a spolu tvoria gramatickú stavbu jazyka. Morfologická rovina obsahuje jednotky bilaterálnej povahy. Každá morfologická jednotka, čiže každý tvar alebo gramatické slovo má dve stránky: formu a obsah. Dialektické spätie formy a obsahu tvorí z tvaru jednotku schopnú vývinu a zapojenú do tvarových a významových systémov. Forma a obsah sú paralelné, ale nie sú symetrické. Forma aj obsah tvaru, gramatického slova i morfémy, majú síce svoju vlastnú, vnútornú zákonitosť, ale súčasne aj rešpektujú zákonitosť svojho pendanta. Štruktúra morfologickej roviny jazyka je komplexná, vybudovaná na slovných druhoch, na morfologických kategóriách a na rozličných tvarotvorných postupoch [8].

2.2 Morfologické kategórie

Všeobecný významový prvok slov nesúcich vlastný význam vyjadrený ustálenými tvarovými prostriedkami nazývame morfologická kategória [8]. Najčastejšie sa jedná o vyjadrenie vzťahov tvarmi slova.

Obsahom morfologickej kategórie je všeobecný význam týkajúci sa celého radu slov alebo celého plnovýznamového slovného druhu. Morfologické kategórie majú len plnovýznamové slovné druhy. Obsah morfologickej kategórie je vyššou abstrakciou, než ich lexikálny vý-

znam. Bez lexikálneho významu sa morfologický význam totiž ani nemôže uplatniť, keďže obsahom morfologickej kategórie je vzťah a obsahom lexikálnej jednotky je súhrn druhových vlastností vecí samých.

2.3 Gramatický tvar

Vyjadrenie gramatického významu, t. j. gramatickej kategórie vonkajšími jazykovými prostriedkami, nazývame gramatický tvar slova [8]. Gramatické tvary majú len plnovýznamové slová. Napríklad citoslovcia nemajú gramatické tvary, lebo sú tvarovo amorfné.

Gramatický tvar, podobne ako slovo, sa skladá z morfém. Morféma je najmenšia časť tvaru (alebo slova) vydelená významom alebo funkciou. Tvar i slovo môže mať rozličnú morfematickú štruktúru (stavbu). Napríklad pre slová *voda, vody, vode, vodár, vodáreň* existuje spoločná morféma *„vod“*.

2.4 Morfologický typ

V každom jazyku môžeme sledovať takzvané črty rozličných typov. Ani jeden jazyk nie je typologicky jednoliaty. Typy sa v ňom kombinujú svojským spôsobom. Na druhej strane, vo väčšine prípadov práve jeden typ jazyka prevláda a ostatné typy sú zastúpené len nepatrnou mierou [8]. Typy jazyka môžeme rozdeliť do piatich typov:

- **flektívny typ**
- **aglutinačný typ**
- **analytický typ**
- **introflektívny typ**
- **polysyntetický typ**

2.4.1 Morfologický typ slovenčiny

Slovenčina si zachovala základné črty morfologickej stavby slovanských jazykov. Morfologická stavba je výrazne flektívna. Podstata flektívneho typu spočíva vo vyjadrovaní niekoľkých gramatických významov jediným spoločným formálnym prvkom v rámci jedného slova. Napríklad gramatické významy pádu, čísla a rodu sa vo tvaroch podstatných mien vyjadrujú jednou príponou. Inštrumentál singuláru podstatných mien ženského rodu sa vyjadruje príponou *-ou*: *žen-ou, ulic-ou, gazdin-ou atď.* Flektívny typ je najvýznamnejší v oblasti skloňovania a výrazne sa prejavuje veľkým rozdielom v sústave tvarov jednotlivých ohybných slovných druhov [8].

2.4.2 Morfologický typ cudzích jazykov

V krátkosti uvedieme aj ostatné typy jazykov ako aglutinačný, analytický, introflektívny a polysyntetický typ jazyka, pričom ku každému typu uvedieme aj konkrétny jazyk, v ktorom dominuje. Vychádzame z morfologickej príručky [8].

V turečtine ale aj v maďarčine prevláda aglutinačný typ jazyka. Tento typ jazyka sa zakladá na korešpondencii gramatického a formálneho vyjadrenia príponou v rámci jedného slova. Gramatický tvar môže mať niekoľko prípon. V dôsledku tejto korešpondencie medzi

gramatickým významom a príponou je v aglutinačnom type jazyka obmedzený počet vzorov, tým pádom sa jazyk stáva málo ohybným. Často existuje práve jeden vzor v každom ohybnom slovnom druhu.

Vo francúzštine a angličtine prevláda takzvaný analytický jazykový typ. Vyznačuje sa hlavne tým, že gramatické významy sa v ňom vyjadrujú osobitnými pomocnými slovami – samostatnými gramatickými morfémami. Plnovýznamové slová sú neohybné.

Arabčina sa zakladá na introflektívnom type, tzn. že gramatické významy vyjadrujú variáciu lexikálnej morfémy slova. Napríklad arabské slovo *„kalbun“* (pes) a *„kilabún“* (psy). Tento typ sa dopĺňa flektívnym typom.

Čínština sa zakladá na polysyntetickom type. Polysyntetický typ je založený na skladaní základov slov pri tvorení slov. Vyznačuje sa nedostatkom v oblasti ohýbania plnovýznamových slov.

2.5 Slovné druhy

Slovné druhy sú komplexné lexikálno-gramatické triedy slov. Vyznačujú sa lexikálnymi, morfologickými a syntaktickými vlastnosťami. Slovné druhy sa rozčleňujú pomocou lexikálnych vlastností slov na desať slovných druhov:

- **substantíva** – podstatné mená
- **adjektíva** – prídavné mená
- **prononymá** – zámená
- **numerálie** – číslovky
- **verbá** – slovesá
- **adverbá** – príslovky
- **prepozície** – predložky
- **konjunkcie** – spojky
- **partikuly** – častice
- **interjeckcie** – citoslovcia

Nestoja na jednej rovine, lebo sa nevyčleňujú zo slovnej zásoby podľa jediného kritéria. Základom slovných druhov je ich lexikálna stránka. Následne si priblížime jednotlivé slovné druhy [8].

2.5.1 Podstatné mená (Substantíva)

Podľa monografie [28] sú podstatné mená plnovýznamové ohybné slovné druhy, ktoré pomenúvajú samostatne existujúce substancie. Ich pojmovosť, relatívna presnosť a významová vyhranenosť znamená, že sa využívajú ako termíny v odbornom, administratívnom a publicistickom štýle. Substantíva sa delia na:

- **vlastné mená** – nazývané aj propriá. Delíme ich na:

  - pravé popriá – *Prešov, Trhovište, Novák...*
  - nepravé popriá – skupinové vlastné mená, prechodná skupina medzi všeobecnými menami a vlastnými menami (napr. *Slovák*) a obyvateľské mená (napr. *Prešovčan*)

- **všeobecné mená** – nazývané aj apelatíva. Delíme ich na:
  - konkrétne – väčšinou spočítateľné: *oko, mesto...*
  - abstraktné – väčšinou nespočítateľné: *zrak, let, mladosť...*
  - nepravé abstraktné – pomenovanie vymedzeného javu, plurál implikuje počet, napríklad *deň*

- **z aspektu spočítateľnosti** – všetky podstatné mená sa delia na:
  - spočítateľné – nazývané aj singulatívne: *jeden stôl, päť stolov...*, ale aj pomnožné podstatné mená: *nožnice, dvere...*
  - nespočítateľné – takzvané nesingulatívne podstatné mená: *zrak, Košice, Bratislava...*

Majú tri základné gramatické kategórie (rod, číslo a pád). Menný rod podstatného mena môže byť mužský, ženský, stredný. V rámci mužského rodu rozoznávame životnosť. Podstatné mená sa vyskytujú v jednotnom čísle (singulár) alebo v množnom čísle (plurál). Na základe čísla podstatného mena určujeme pády podstatných mien. V každom čísle existuje šesť pádov (nominatív, genitív, datív, akuzatív, lokál, inštrumentál). V dnešnej slovenčine sa vokatív z morfologického hľadiska takmer nevyskytuje, zachovalo sa len zopár tvarov, napríklad *„otče“*.

U podstatných mien určujeme aj vzor. V slovenčine sa vyskytuje až 132 skloňovacích vzorov podstatných mien, medzi ktoré patrí aj sedemnásť hlavných skloňovacích vzorov ako: chlap, hrdina, dub, stroj, kuli, žena, ulica, dlaň, kosť, gazdiná, pani, idea, mať, mesto, srdce, vysvedčenie, dievča [9].

2.5.2 Prídavné mená (Adjektíva)

V morfologickej príručke [8] sa uvádza, že prídavné mená sú ohybné plnovýznamové slová. Pomenúvajú statické príznaky vecí, osôb, zvierat, predmetov a javov pomenovaných podstatným menom. Gramatické kategórie sa zhodujú v rode, čísle a páde s nadradeným podstatným menom. Statickým príznakom sa rozumejú:

- priame vlastnosti vecí, napr. *silný muž, krásna ulica...*
- vlastnosti vyplývajúce zo vzťahu k iným veciam, dejom alebo okolnostiam, napr. *tehlový byt, pľúcna infekcia, nočný svit...*
- charakteristika označujúca vlastnícky vzťah, napr. *bratov kľúč, štátny sviatok...*

Zhoda prídavného mena a podstatného mena sa opiera o skutočnosť vzťahu príznaku a jeho podstaty, že príznak neexistuje samostatne na veci. Preto sa tvar prídavného mena obmieňa podľa rodu, čísla a pádu nadradeného podstatného mena, napr. *prísny riaditeľ, dlhá cesta...*

Stupňovanie, ako vyjadrovanie miery vlastnosti tvarovou obmenou adjektív, je dané istými vlastnosťami, ktoré sa vyskytujú na veciach v rozličnej miere, porovnávanie vlastností

podľa ich kvantity. Poznáme stupne: pozitív (*úzke šaty*), komparatív (*užšie šaty*), superlatív (*najužšie šaty*).

Tvarmi skloňovania sa vyjadruje zhoda. Prídavné mená majú sedem skloňovacích vzorov a to pekný, krásny, cudzí, rýdzi, otcov, papagáji, môj [9].

2.5.3 Zámená (Prononimá)

Zámená patria k plnovýznamovým ohybným slovným druhom, ktoré zastupujú podstatné mená, prídavné mená, číslovky a príslovky. Neoznačujú skutočnosť priamo, ale ukazujú alebo odkazujú na ňu. Vzťahujú sa na osobu, zviera, vec, jav. Ich význam sa dá pochopiť iba v kontexte. Nadobúdajú gramatické kategórie slov, ktoré zastupujú [15]. Zámená môžeme deliť podľa:

- **vecného významu** – delíme zámená do šiestich kategórii a to na zámená osobné, zvratné, ukazovacie, opytovacie, neurčité a vymedzovacie
- **gramatického významu** – v závislosti na slovnom druhu, ktorý zastupujú: substantívne, adjektívne, číslovkové, príslovkové

2.5.4 Číslovky (Numerálie)

Číslovky sú komplexný plnovýznamový slovný druh. Sú ohybné i neohybné slová, ktorými sa pomenúvajú pojmy počtu, číselné príznaky vecí, dejov. U čísloviek sa uplatňujú s istými obmedzeniami gramatické kategórie podstatných mien, prídavných mien a prísloviek.

Číslovky delíme podľa špecifikácie ich číselného významu na:

- **základné (kardináliá)**
- **skupinové (kolektíva)**
- **násobné (multiplikatíva)**
- **radové (ordináliá)**
- **druhové (speciáliá)**

Podľa ich vzťahu k číselným pojmom ich môžeme rozdeliť na určité a neurčité. Určité číslovky presne vyjadrujú počet alebo číselné určením veci, kým neurčité číslovky vyjadrujú počet a číselné určenie nepresne, na základe odhadu alebo subjektívneho postoja [8].

2.5.5 Slovesá (Verbá)

Slovesá sú ohybné slová, ktorými sa pomenúvajú dynamické príznaky vecí. Slovesá majú gramatické kategórie osoby, času, spôsobu, slovesného rodu, vidu. Príznakom vecí sú javy, ktoré sú nesamostatné a existujúce na niečom. V tom sa slovesá a prídavné mená zhodujú. Na druhej strane prídavné mená vyjadrujú statické príznaky vecí. Slovesá ako slovný druh predpokladajú nadradenú vec pomenovanú podstatným menom alebo zámenom. [8].

Slovesá obsahujú 24 slovesných vzorov, napríklad *piť, kresliť, vládnuť atď.* Slovesá rozdeľujeme na:

- **plnovýznamové**
- **pomocné**

  - modálne (*chcieť, smieť...*)
  - fázové (*začať, zostávať...*)
  - limitné (*ísť, mať...*)
  - sponové (*byť, stať sa...*)

Pomocné slovesá nemajú úplný význam a spájajú sa s iným plnovýznamovým slovesom [27].

2.5.6 Príslovky (Adverbá)

Príslovky sú neohybné plnovýznamové slová, vyjadrujúce okolnosť alebo vlastnosť vzťahujúcu sa na slovesný dej (*počínal si dobre*), na vlastnosť vyjadrenú prídavným menom alebo príslovkou (*celkom dobre*), alebo na okolnosť vyjadrenú príslovkou (*včera ráno*), a to v podstate bez pomoci gramatických kategórií. Príslovky delíme na príslovky:

- **miesta** (*hore, vonku...*)
- **času** (*zajtra, dnes...*)
- **spôsobu** (*rýchlo, zle...*)
- **príčiny** (*úmyselne, bezdôvodne...*)

Príslovky môžeme stupňovať, a to napríklad: skoro (*pozitív*), skoršie (*komparatív*), najskoršie (*superlatív*) [24].

2.5.7 Predložky (Prepozície)

Predložky čiže propozície sú neohybné slová. Spolu vo spojení s nepriamymi pádmi podstatných mien, zámen a čísloviek vyjadrujú vzťahy okolnostné, predmetové a prívlastkové. Vzťahy vyjadrujú nesamostatne, t. j. nevyskytujú sa samostatne, ale v spojení v predložkovej konštrukcii (napr. *po domácky*) [8]. Delíme ich na:

- **jednoduché** (*v, nad...*)
- **zložené** (*popod, popred...*)

2.5.8 Spojky (Konjunkcie)

Spojky sú neohybné slová vyjadrujúce syntagmatické vzťahy medzi jazykovými jednotkami samostatne, t. j. ich hlavná funkcia je spájanie slov [8]. Spojky delíme na:

- **priraďovacie**
  - zlučovacie (*a, i...*)
  - stupňovacie (*ba, ba aj...*)
  - odporovacie (*ale, však...*)
  - vylučovacie (*alebo, buď...*)
- **podraďovacie** – používajú sa v podraďovacích súvetiach (*keď, lebo...*) .

2.5.9 Častice (Partikuly)

Sú to neohybné slovné druhy. Najsilnejší je súvis častíc a spojok. Táto súvislosť so spojkami je daná tým, že častice zahŕňajú dva významy, nadväznosť na známu situáciu alebo kontext, pripojenie nového výrazu alebo výpovede a vyjadrenie hľadiska podávateľa, t. j. subjektívne modifikovanie nejakého výrazu alebo celej výpovede [8].

2.5.10 Citoslovce (Interjeckcia)

Citoslovce sú neohybné slová, ktoré delíme na dve skupiny: na vlastné citoslovce a na zvukomalebné slová. K vlastným citoslovciam patria slová z oblasti citu a vôle (napr. *ach, jaj, haló*). K zvukomalebným slovám patrí skupina slov napodobňujúcich prírodné zvuky (napr. *kikirikí, mú*) [8].

2.6 Slovná zásoba

Slovnú zásobu nazývame lexika. Lexika každého jazyka reaguje na spoločenský vývin. Obohacuje a rozširuje sa v súvislosti s potrebami v rôznych komunikačných sférach a situáciách. Hlavne sa jedná o pomenovanie nových javov, pojmov alebo o nové pomenovanie, tzv. inováciu existujúcich vyjadrovacích prostriedkov [21]. Rozširovanie slovnej zásoby zahŕňa:

- tvorenie slov (lexikálnych jednotiek)
- modifikovanie lexikálnych jednotiek
- deriváciu slov, založenú na sémantickej motivácii
- preberanie cudzích slov
- transflexiu slov
- konverziu – slovnodruhové prechody bez zmeny tvaroslovnej formy.

2.6.1 Tvorenie slov

Náuka o tvorení slov sa nazýva derivatológia [21] alebo slovotvorba. Jej hlavnou podstatou je slovotvorná motivácia, ktorá predstavuje najzávažnejší systémotvorný činiteľ. Je to formálno-sémantický vzťah medzi minimálne doma lexikálnymi jednotkami, z ktorých sa jedna chápe ako východisková a druhá sa na nej formálne i sémanticky zakladá, napríklad *„uči-ť“* a *„uči-teľ“*.

Slovotvorný základ je nositeľom onomaziologického príznaku, reprezentuje význam základového slova. Slovotvorným základom býva koreň slova. Tvarový základ je lexikálna zložka slova bez gramatickej morfémy a je rovnaká pre všetky gramatické formy daného slova. Nositeľ onomaziologickej bázy je slovotvorný formant [21], resp. tá časť odvodeného slova, ktorá sa pripojí k slovotvornému základu a významovo ho modifikuje. Za slovotvorné formanty považujeme:

- **súbory gramatických morfém** (*loviť – lov-0*)
- **zvratný komponent** (*stať sa*)
- **affixy** – slovotvorné prípony a predpony

- sufix – slovotvorná prípona(*chlap-ec*)
- prefix – slovotvorná predpona (*pre-písať*)
- prefixu a sufixu (*ná-byt-ok*)
- prefix a súbor gramatických morfém (*pred-mest-ie*)
- sufix a zvratný komponent (*vy-piť si*)
- prefix, sufix, zvratný komponent (*do-pros-ovať sa*)

2.6.2 Slovotvorné postupy

Odvodzovanie (derivácia) a skladanie (kompozícia) slov patrí medzi základné slovotvorné postupy [21], pri ktorých sa realizuje onomaziologická štruktúra myšlienkového obsahu.

Derivácia je základný spôsob tvorenia slov, proces vytvárania nového slova na základe iného slova, zmenou morfologickej stavby slova. Najčastejšie sa používa práve afixácia, t. j. použitie slovotvorných predpôn a prípon. Ďalej sa používajú aj postfixy, transflexia, t. j. súbor gramatických morfém, reflexivizácia, teda využitie voľnej, zvratnej alebo derivačnej morfémy. Týmto spôsobom tvorenia slov vznikajú odvodené slová.

Transflexia je postup, pri ktorom sa ako formant považuje gramatická morféma, t. j. odvodené slovo tvoria gramatické morfémy, ich zmenou sa mení aj paradigma a slovotvorná charakteristika odvodeného slova.

Reflexivizácia je slovotvorný postup využívajúci sa pri tvorení slovies. Slovesá sa tvoria pomocou derivačnej morfémy sa/si. (napr. *stratiť – stratiť sa*). Obmenu afixov nazývame aj výmenná derivácia (napr.*vy-kloniť, pred-kloniť*).

Kompozícia je slovotvorný postup, pri ktorom sa spájajú lexikálne morfémy do jednoslovného útvaru. Morfémy sú spojené pomocou tzv. spájacej morfémy. Slovotvorné postupy pre vytváranie slov skladaním rozdeľujeme na:

- **vlastná kompozícia** – tvorenie slov pomocou spájacej morfémy (*-o- , -e-*)
- **nevlastná kompozícia** – tvorenie slov bez spájacej morfémy, na jej mieste stojí gramatická morféma prvého člena (napr. *naničhodný*)

Okrem základných postupov tvorenia slov poznáme aj:

- **akronymizácia** – tvorenie slov skracovaním tak, že sa použijú len iniciály (napr. *USA*)
- **abrevizácia** – skrátenie pomenovania (napr. *prof.*)
- **univerbizácia** – proces, pri ktorom z viacslovných pomenovaní vznikajú jednoslovné pomenovania (napr. *hlavný čašník – hlavný*)
- **multiverbizácia** – proces opačný univerbizácii, t. j. z jednoslovných pomenovaní sa tvoria viacslovné pomenovania (napr. *zapísať – urobiť zápis*)

2.7 Morfologická analýza

Morfologická analýza je základným prostriedkom skúmania prirodzeného jazyka. Zaoberá sa rozlišovaním a generovaním správnych gramatických tvarov slovných výrazov, ktoré vzniknú ohýbaním slov. Výsledkom je sada značiek, ktoré popisujú gramatické kategórie daného tvaru, hlavne základný tvar (lema) a slovný vzor [17].

Samotné automatické rozlíšenie slovného tvaru sa môže využiť ako pomôcka pri značkovaní korpusov alebo pri poloautomatickom vytváraní slovníkov. Morfologická analýza je tiež východzí bod spracovania textu pre syntaktickú a sémantickú analýzu. Najväčší problém v tejto oblasti je morfologická desambiguancia. Jedná sa o zjednotenie gramatického značkovania. Napríklad slovo *„hlavný“* sa môže označovať ako substantívum alebo aj adjektívum [14]. Rozlišujeme dva typy morfologickej analýzy:

- úplná morfologická analýza
- lematizácia

Ako cieľ úplnej morfologickej analýzy je získať úplné morfologické informácie o základných gramatických kategóriách ako rod, číslo, pád apod. [31].

2.8 Lematizácia

Jednotlivé slová sa vyskytujú v rôznych morfologických tvaroch, t. j. *pádoch, osobách, číslach...* Existuje nutnosť ich prevádzať na základné tvary, ktoré sa nazývajú lemy. Lema je základným *„slovníkovým“* tvarom slova, lexikálnou jednotkou. Napríklad slovo *„krajší“* ma základný tvar *„pekný“*, slovo *„chlapom“* má základný tvar *„chlap“*. Pri podstatných a prídavných menách je to prvý pád jednotného čísla, pri slovesách neurčitok.

Lematizáciu považujeme za proces, ktorý určí základný tvar slova, tzv. lemu, najčastejšie odstránením slovotvorných, pádových a iných prípon a predpôn. Izolácia koreňa slova, z anglického slova *stemming*, je proces podobný lematizácii, pri ktorom sa z daného slova odstránia všetky jeho morfologické prípony a predpony tak, že ostáva len jeho koreň. Vďaka lematizácii a stemmingu dochádza k významnej redukcii počtu termov. Redukciu môžeme vykonávať pomocou nasledovných troch spôsobov [1]:

- **Slovník koreňov** – za hlavnú výhodu tejto metódy sa považuje jej minimálna chybovosť, no na druhej strane treba počítať s rozsiahlosťou slovníka a jeho obmedzením na špecifický odbor a skutočnosťou, že slovník je prakticky vždy neúplný. Vlastné mená a názvy sa v ňom nenachádzajú, no pritom nesú dôležité informácie o obsahu textu.
- **Odstránenie afixov** – metóda, pri ktorej algoritmus odstraňuje afixy, t. j. prípony a predpony, na základe známeho a vopred definovaného zoznamu prípon a predpôn alebo pomocou sady pravidiel, podľa ktorých sa generujú jednotlivé prípony a predpony.
- **Štatistická metóda** – na základe variety po sebe nasledujúcich znakov sa stanovuje pomocou frekvencie jednotlivých zhlukov znakov v slovách, či sa jedná o prefix, t. j. predponu alebo o sufix, t. j. príponu. Výhoda tejto metódy spočíva v tom, že je nezávislá na konkrétnom použitom jazyku textu.

Vo flektívnych jazykoch, ako je napríklad slovenčina, je situácia komplikovaná. Nastávajú problémy, ktoré spôsobujú časté nejednoznačnosti a rôzne typy nepravidelností jazyka, s ktorými je potrebné sa pri lematizácii vysporiadať. Napríklad slovo *„mier“* má nejednoznačnú lemu, môže označovať sloveso *„mieriť“*, podstatné meno *„mier“* alebo *„miera“*. Potrebné je zistiť kontext slova v texte, odvodiť slovný druh a morfologické kategórie slova a potom sa dá lema určiť presnejšie [1].

2.9 Morfologické značky

V rámci morfologickej analýzy sa slovám priradí ich príslušný základný tvar, t. j. lema. Súčasne sa slová ohodnocujú morfologickými kategóriami, na základe ktorých sa slovám priradia gramatické značky [1], nazývané aj tagy. Morfologická značka určuje slovný druh a v závislosti od slovného druhu aj ďalšie gramatické kategórie ako rod, číslo, pád atď. Samotná reprezentácia morfologických značiek je daná ich súborom a spôsobom zápisu. Uvedieme pozičný a kódovací typ značiek.

2.9.1 Pozičný typ

Každý z typov gramatických značiek má definovanú vlastnú pevnú pozíciu v reťazci znakov. Napríklad pražský systém je pozičný a obsahuje v reťazci šestnásť znakov. Značka je konštruovaná tak, aby každá pozícia odpovedala jednej morfologickej kategórii. V značke sa vyskytujú prevažne veľké písmená abecedy, po prípadne aj iné znaky ako malé písmená, čísla, interpunkčné znamienka apod. Viac informácii o pozičnom type sa môžete dozvedieť z dokumentu [14], v ktorom sa nachádza detailný popis jednotlivých pozícií ako:

1. **slovný druh** – označuje hlavný slovný druh, napr. substantívum (N), adjektívum (A) apod.
2. **detailné určenie slovného druhu** – slúži k zachyteniu ďalších relevantných morfologických kategórii, napr. číslovka „koľko“ (?), skratka ako substantívum (;) apod.
3. **slovný rod** – ženský rod (F), stredný rod (N) apod.
4. **číslo** – plurál (P), singulár (S) apod.
5. **pád** – nominatív (1), genitív (2) apod.
6. **privlastňovací rod** – rod subjektu/objektu, ktorému sa privlastňuje adjektívum, napr. ženský rod (F), mužský rod životný (M) apod.
7. **privlastňovacie číslo** – uplatňuje sa u zámen, napr. plurál (P), singulár (S) apod.
8. **osoba** – prvá osoba (1), druhá osoba (2) apod.
9. **čas** – budúci čas (F), prítomný čas (P) apod.
10. **stupeň** – 1. stupeň (1), 2. stupeň (2) apod.
11. **negácia** – afrimitív (A), negatív (N) apod.
12. **aktívum/pasívum** – aktívum (A), pasívum (P) apod.
13. **nepoužitá pozícia**
14. **nepoužitá pozícia**
15. **variant** – štýlový príznak slova, napr. rovnocenný variant (1), riedky variant (2) apod.
16. **vid slovesa** – dokonavé sloveso (P), nedokonavé sloveso (I) apod.

2.9.2 Kódovací typ

Pri kódovacom type značiek sú značky vyjadrené postupnosťou dohodnutých kódov atribútov, ktoré pozostávajú vždy z dvoch znakov. Prvý znak reprezentuje kódovací atribút, označuje sa malým písmenom a druhý znak predstavujúci hodnotu atribútu, označený veľkým písmenom alebo číslicou. Tento spôsob kódovania sa využíva napríklad pri brnenskom systéme kódovania [1]. Tento typ je viac popísaný v sekcii 3.1.3.

2.10 Morfologické analyzátory

Pomocou morfologických analyzátorov sa vykonáva proces lematizácie a tagovania. Môžu byť buď automatické alebo pracujú vo forme konečného automatu. Súbory značiek a pravidlá bývajú uložené v morfologickom slovníku [1].

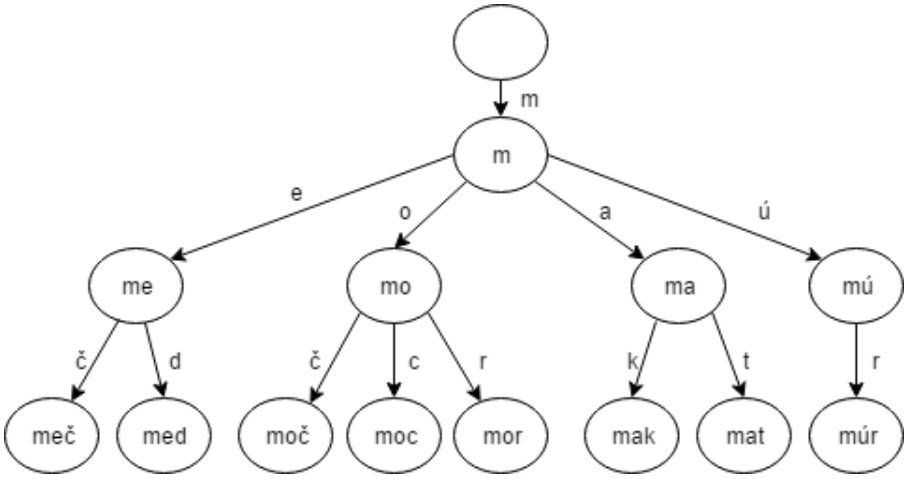
Pri analyzovaní slovného tvaru si analyzátor v prvom behu zistí, či sa daný tvar nenachádza v slovníku. V prípade úspechu analyzátor priradí lemu k slovnému tvaru. V prípade neúspechu, v tzv. pravidlových systémoch, analyzátor obsahuje pravidlá, pomocou ktorých dokáže slovo spracovať a následne vytvoriť lemu. Tieto pravidlá môžu pozostávať napríklad z odstránenie afixov zo slova [31].

2.11 Morfologický slovník

Súčasťou morfologického analyzátora je morfologický slovník [31]. V morfologickom slovníku sú uložené tvary slov, ich lemy a značky v stanovenom formáte. Dôležitá je jednak kvalita slovníka a rýchlosť komunikácie morfologického analyzátora so slovníkom. S týmto súvisí aj spôsob uloženia slovníka. Najčastejším formátom uloženia slovníka sú prefixové stromy.

2.11.1 Trie

Trie, nazývaná aj prefixový strom [32], je dátová stromová štruktúra uchovávajúca asociatívne pole, t. j. abstraktný dátový typ, zložený z kolekcie dvojíc: kľúč a hodnota. Kľúčom zvyčajne býva textový reťazec. Koreň trie tvorí prázdny reťazec. Všetci potomkovia daného uzlu majú spoločný prefix jeho reťazca. Strom v každom uzle obsahuje všetky podreťazce, ktorými môže pokračovať reťazec v prehľadávanej ceste. Častou aplikáciou trií je uloženie slovníkov.



Obr. 2.1: *Príklad trie*

Na obrázku 2.1 môžeme vidieť dátovú štruktúru trie, obsahujúcu kľúče med, mak, mat, meč, moc, mor, moč, múr. Jednotlivé kľúče zdieľajú prefixy, napríklad kľúče mak a mat zdieľajú prefix *„ma“*.

Existujú rôzne implementácie trií. Tieto implementácie môžeme hodnotiť podľa viacerých kritérií ako veľkosť pamäťového priestoru, ktorý slovník zaberá, možnosť aktualizácie slovníka za behu, potrebná pamäť a čas pri zostavení a spracovaní slovníka, rýchlosť vyhľadávania kľúča v slovníku, čas potrebný na uloženie kľúča atď.

Vybrané implementácie:

- **Fsa** – konečný automat, spočiatku implementovaný ako prefixová dátová štruktúra s Daciukovou inkrementačnou DAFSA konštrukciou. DAFSA značí deterministicko-cyklický konečný automat. V súčasnosti obsahuje port generického FSA z GATE balíku. Táto implementácia je optimalizovaná pre veľkosť slovníka a rýchlosti vyhľadávania v ňom, ale nepodporuje modifikáciu FSA [12].

- **Cedar** – knižnica implementovaná v jazyku C++, poskytuje aktualizovateľnú dvojúrovňovú prefixovú dátovú štruktúru, ktorá ponúka rýchle vyhľadávanie a aktualizovanie dotazov v reálnych dátach [34].

- **MARISA** – vyhľadávajúci algoritmus s rekurzívne implementovaným uchovávacím priestorom, knižnica implementovaná v jazyku C++. Využíva statickú dátovú štruktúru trie. Slúži ako slovník, podporujúci vyhľadávanie pomocou presnej zhody, reverzné vyhľadávanie, prefixové a prediktívne vyhľadávanie. Dátová štruktúra zaberá menej miesta ako binárny strom alebo hashovacia tabuľka [33].

- **Pytries** – dátové štruktúry, ktoré využívajú prefixové stromy pre efektívnejšie uloženie slovníka a následne prácu s ním. Jednotlivé implementácie pytries [16]:

  - **Marisa-trie** – statická pamäťovo efektívna trie štruktúra, založená na C++ knižnici MARISA. Reťazec v tejto štruktúre zaberá až stokrát menej pamäte než v slovníku jazyka Python.
  - **DAWG** – alebo aj orientovaný acyklický slovný graf, je dátová štruktúra odpovedajúca slovníku v jazyku Python, vychádza z *dawgdic* C++ knižnice. Reťazce

uložené v tejto štruktúre zaberajú až dvestokrát menej pamäte než štandardne v slovníku jazyka Python.

- **Datrie** – dátová štruktúra, ktorá zaberá štyrikrát menej pamäte než slovník jazyka Python, ale je dva až šesťkrát pomalšia než slovník jazyka Python. Modul je založený na knižnici *libdatrie*, implementovanej v jazyku C.
- **Hat-trie** – modul, ktorý založený na knižnici *hat-trie*, implementovanej v jazyku C. Je 1,5krát rýchlejšií ako *datatrie*, ale zaberá viac pamäte.

Ako najlepšie *pytries* štruktúry sa javia implementácie *DAWG* a *marisa-trie*. Na porovnanie bol vykonaný test [5], pri ktorom bolo potrebné uchovať dáta troch miliónov ruských slov v dátovej štruktúre. Najhoršie si viedol predvolený slovník jazyka Python so šesťsto megabajtami zabranej pamäte. *Marisa-trie* zaberala len sedem megabajtov pamäte a *DAWG* len dva megabajty pamäte.

2.12 Korpus

Podľa dokumentu [7], slovo korpus pochádza z latinského slova *„corpus“*, ktoré značí telo, teleso. Korpus je rozsiahly súbor autentických textov, ktorý je prevedený do elektronickej podoby v jednotnom formáte tak, aby sa v ňom mohli jednoducho vyhľadávať slová a slovné spojenia. Zobrazuje slová a slovné spojenia v ich prirodzenom kontexte, pričom umožňuje vytvárať lingvistický výskum na reálnych dátach.

Korpus slúži ako čo najobjektívnejší model jazykovej empírie. Dnes je bežné, že korpusy presahujú hranicu sto miliónov slov. V dôsledku tohto rozsahu je nutné použitie špeciálnych nástrojov na vyhľadávanie a filtrovanie slov. Rozsah korpusu nie je jediným kritériom kvality korpusu, je len vzorkou jazyka.

Často sú samotné texty, ktoré sa nachádzajú v korpuse, anotované. Obsahujú metainformácie o textoch ako ich pôvod, autorstvo atď. a aj doplnkové informácie o slovách. Lematizácia je príkladom takejto anotácie, kedy každému slovnému tvaru sa priradí ich základný slovníkový tvar, morfologická značka.

Špeciálnym typom korpusu je paralelný korpus, ktorý je hlavným predstaviteľom viacjazyčných prekladových textov, ktorý umožňuje porovnávanie jazykov. V dôsledku nedostatočného počtu takýchto korpusov sa využívajú aj takzvané zarovnané korpusy, ktoré sú založené na textoch (napr. noviny, manuály) s takým istým tematickým oborom.

Zoznam a popis korpusov, ktoré sme využili v práci, sa nachádza v sekcii 4.2.1.

Kapitola 3

Návrh a implementácia

V tejto kapitole vysvetlíme postup a kroky potrebné k vytvoreniu a hodnoteniu morfologického analyzátora (viď podkapitola 3.1), ktorý lematizuje vstupné jednoslovné pomenovania a na základe určenia ich vzoru generuje ich všetky slovné tvary, morfologického slovníka, s ktorým analyzátor pracuje, nástroj pre hodnotenia a porovnávania stemerov (viď podkapitola 3.2), ktorý hodnotí stemery na základe testovacieho slovníka, nástroj na zrekonštruovanie diakritiky (viď podkapitola 3.3), ktorý je založený na základe unigramového vyhľadávania.

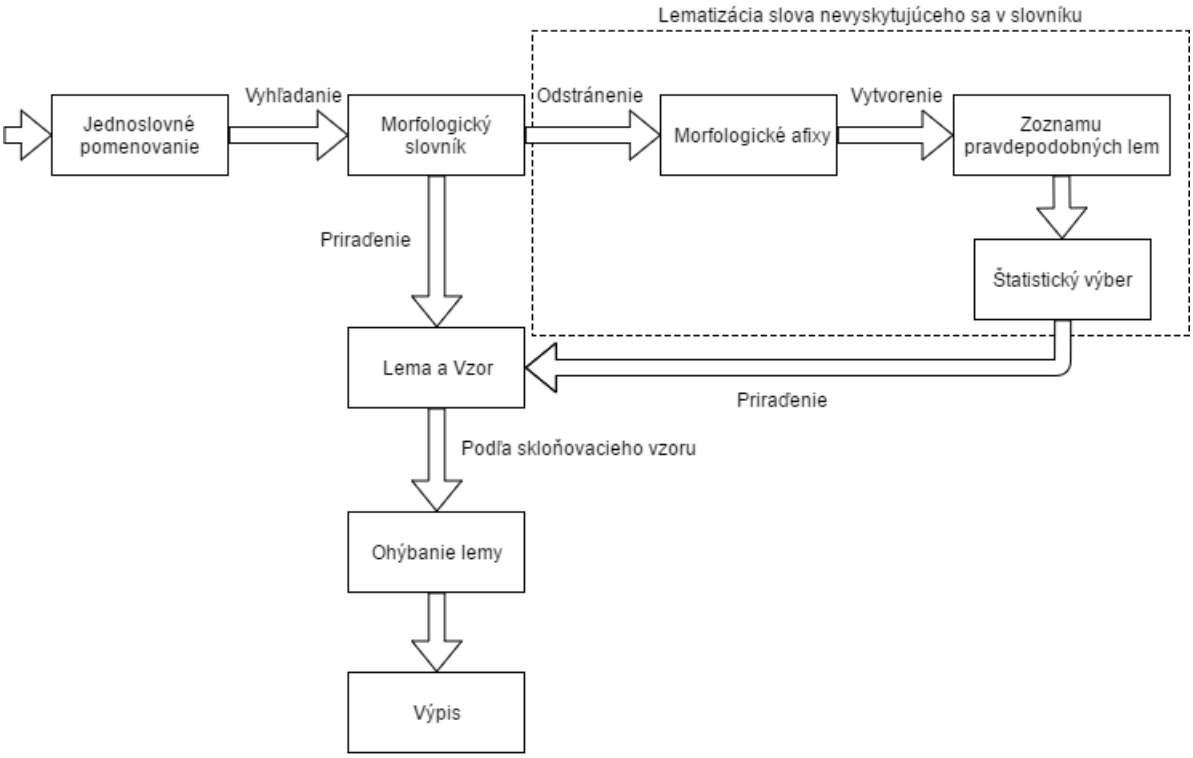
3.1 Morfologický analyzátor

Morfologický analyzátor lematizuje vstupné slová na základe morfologického slovníka, pričom tieto slová sa nemusia nachádzať v slovníku. Hlavná požiadavka analyzátora sa viaže na správnosť lematizácie, resp. priradenia správneho slovotvorného základu a skloňovacieho vzoru, vstupného slova. Táto požiadavka závisí na kvalite slovníka, ktorý analyzátor používa.

S týmto procesom sa spája aj požiadavka na jeho rýchlosť. Analyzátor by mal lematizovať slová, ktoré sa nachádzajú v slovníku, okamžite a slová, ktoré sa v ňom nenachádzajú, v nie príliš dlhom časovom rozmedzí.

Za požiadavku pre nástroj môžeme považovať aj veľkosť pamäťového priestoru, ktorý slovník zaberá. Táto požiadavka sa viaže aj na kvalitu slovníka. Nie je pravidlom, že čím obsiahlejší je slovník, tým je kvalitnejší.

Slovník, ktorý používa analyzátor, je vhodné aktualizovať. Z tohto tvrdenia vyplýva požiadavka na schopnosť analyzátora aktualizovať svoj morfologický slovník. Slovo, ktoré sa nenachádza v slovníku, analyzátor analyzuje, t. j. priradí mu jeho slovotvorný základ a skloňovací vzor, uloží všetky jeho tvary do slovníka a pri nasledujúcej analýze sa čas potrebný na analýzu skráti.



Obr. 3.1: *Algoritmus analýzy vstupného slova*

Obrázok 3.1 zobrazuje algoritmus analýzy vstupného jednoslovného pomenovania morfologickým analyzátorom. Jednoslovné pomenovanie najskôr vyhľadáme v morfologickom slovníku, určíme jeho lemu a vzor. V prípade ak sa nevyskytuje v slovníku, tak nastáva proces lematizácie slova nevyskytujúceho sa v slovníku, na základe ktorého určíme zo slovného tvaru lemu a jej príslušný vzor. Ďalej nastáva fáza skloňovania a výpisu. Následne uvedieme popis jednotlivých fáz algoritmu.

3.1.1 Vstupné dáta

Pri lematizácii sa ako vstupné dáta považujú jednotlivé slová slovenského jazyka. Ako obmedzenie považujeme fakt, že morfologický analyzátor pracuje len s jednoslovnými pomenovaniami, teda pomenovaniami vecí, javov, dejov jedným slovom. Tieto pomenovania môžu byť ktoréhokoľvek slovného druhu či už podstatné meno, citoslovce.

3.1.2 Morfologický slovník

Morfologický slovník je dôležitou časťou morfologického analyzátora. Dá sa považovať za základnú vedomosť analyzátora, resp. za jeho jadro, s ktorým analyzátor pracuje.

Vytvorenie slovníka

Cieľom je vytvorenie systému technických vzorov, na základe ktorých môžeme zostaviť výslednú podobu slovníka.

Najskôr sme z poskytnutých dát museli pomocou skriptov extrahovať slovník paradigiem, ktorý obsahuje jednotlivé vytypované vzory a ich pravidlá ohýbania a slovník lem, ku ktorým je priradený ich skloňovací vzor.

Pri vytváraní slovníka paradigiem sme museli najskôr previesť vstupné dáta do formátu: `tvar slova # lema # morfologická značka`

Zo vzniknutého súboru sme vytvorili slovník pravidiel, ktorý obsahuje záznamy vo formáte:

- lema
- poznámka
- morfologická značka : reťazec na odobratie : reťazec na pridanie

Z tohto slovníka sme postupne vytvorili z každej lemy vzor jej skloňovania a pre lemy, ktorých pravidlá sa zhodovali, sme určili jednu ako zástupcu vzoru a referenciu uložili do slovníka lem. Takto nám vznikli oba slovníky.

Tieto vzory sme roztriedili podľa slovných druhov, ktoré reprezentujú, do samostatných súborov. V rámci každého súboru sme zjednotili vzory podľa podobných pravidiel ohýbania, tiež pomocou skriptov, pričom sa zredukoval ich pôvodný počet. V konečnej fáze sme zostavili zo všetkých súborov výsledný slovník paradigiem a slovník lem, priradených k zjednoteným vzorom.

Slovník paradigiem (súbor *slovak.paradigms*) obsahuje vzory, ktoré sledujú chovanie vzorov vypísaných v morfologických príručkách. Slovenčina obsahuje veľa výnimiek, vďaka ktorým sú niektoré vzory vzájomne podobné, pričom sa líšia len v pár pravidlách, hlavne podstatné mená v ženskom rode. Slovník obsahuje aj vzory, ktoré sú nespisovné, resp. nesledujú chovanie spisovných vzorov. Tieto vzory sú označené pomocou štylistického príznaku *w* v ich morfologickej značke a v rámci lematizácie slov, ktoré sa nevyskytujú v slovníku, ich ignorujeme.

Slovník paradigiem obsahuje záznamy vo formáte:

- názov vzoru – kľúč slovníka
- morfologická značka základného tvaru
- pravidlá ohýbania vzoru

Pravidlá ohýbania predstavujú zoznam pravidiel, ktorého každé pravidlo obsahuje položky, oddelené znakom dvojbodky, ktorých formát má tvar:

- morfologická značka charakterizujúca tvar
- reťazec, ktorý sa zo slova odoberie
- reťazec, ktorý sa k slovu pridá

Keďže spočiatku boli zástupcovia názvu vzorov vytypovaný, museli sme z množiny lem patriacich k danému vzoru vybrať najfrekventovanejšiu lemu, ktorej názov bude zastupovať názov vzoru. Podstatou tohto procesu bol fakt, že nie všetky vzory mali vhodné názvy, ktoré by ich mali vo výslednom systéme reprezentovať. Na základe dostupných korpusov, o ktorých sa môžete dozvedieť viac v sekcii 4.2.1, sme vytvorili frekvenčný slovník, ktorý obsahoval frekvencie jednotlivých slovných tvarov v korpuse. Pomocou tohto slovníka sme

určili najfrekventovanejšie lemy z množiny lem patriacich pod daný vzor, ktoré zastupujú názvy vzorov.

stránka	k1gFnSc1	k1gFnSc1:: k1gFnSc2:a:y k1gFnSc3:a:e k1gFnSc4:a:u k1gFnSc6:a:e k1gFnSc7:a:ou k1gFnPc1:a:y k1gFnPc2:ka:ok k1gFnPc3::m k1gFnPc4:a:y k1gFnPc6::ch k1gFnPc7::mi

Obr. 3.2: *Príklad záznamu slovníka paradigiem*

Ako môžeme vidieť na obrázku 3.2, na ktorom je zobrazený jedno paradigmum (vzor) zo slovníka. Jeho názov odpovedá leme *„stránka“*, morfologická značka základného tvaru *„k1gFnSc1“*, ktorá značí, že sa jedná o podstatné meno ženského rodu. V poslednom stĺpci sú uvedené jednotlivé pravidlá ohýbania.

Slovník lem (*slovak.lpn*) obsahuje záznamy dvojíc základný tvar slova, t. j. lema, skloňovací vzor, podľa ktorého sa lema ohýba. Lema je kľúčom v slovníku.

Formát morfologického slovníka

Pri morfologickej analýze analyzátor *„komunikuje“* so slovníkom. Je potrebné minimalizovať čas, za ktorý sa tento dotaz vykoná a pritom nie je moc výhodné samotný slovník načítať do operačnej pamäte.

Ako formát uloženia morfologického slovníka sme zvolili databázu, konkrétnejšie SQLite3[1]. SQLite priamo číta a zapisuje do bežných súborov, celá databáza, ktorá obsahuje viacero tabuliek, indexov, triggerov atď., je uložená v jednom súbore. Alternatívou bol shelve[2] modul jazyka Python, ktorý slúži na uloženie slovníku podobného objektu, textový súbor, prípadne zvolenie dátovej štruktúry trie. Popis jednotlivých implementácii trií je uvedený v sekcii 2.11.1.

Dôležitá vlastnosť slovníka je jednak jeho veľkosť a rýchlosť komunikácie s ním. Tieto vlastnosti spĺňajú hlavne implementácie využívajúce dátové štruktúry prefixového stromu, trie. Pre jednoduchosť implementácie sme však zostali pri databáze SQLite3. Pri databáze sme zostali aj vďaka možnosti rýchleho aktualizovania slovníka. Nakoniec sme však otestovali aj priame uloženie morfologického slovníka do štruktúry *marisa-trie*, viac informácii sa môžete dozvedieť v sekcii 4.2.6.

Morfologický slovník reprezentuje tabuľka obsahujúca päť stĺpcov:

- tvar slova
- slovníkový tvar
- značka – morfologický reťazec
- anotácia – poznámka, napr. pN – od takéhoto slova sa netvoria negácie
- vzor – podľa ktorého sa riadi skloňovanie slovníkového tvaru

Dôležitý je výber správneho primárneho kľúča tabuľky. Samotný tvar slova nemôžeme zvoliť ako primárny kľúč, pretože daný tvar slova sa môže v jednotlivých pádoch opakovať.

[1] Podrobná dokumentácia k SQLite sa nachádza na https://www.sqlite.org/docs.html

[2] Viac informácii o shelve module môžete nájsť v dokumentácii jazyka Python dostupnej z https://docs.python.org/3.4/library/shelve.html

Napríklad v rámci toho istého slovného druhu tvar slova *„chlapa“* označuje pády genitív (*„od chlapa“*) a akuzatív (*„chlapa“*). Čiže je potrebná kombinácia hodnôt.

Kombinácia tvaru slova a lemy odpadá tiež kvôli vyššie určenému dôvodu, ale aj vďaka morfologickej desambiguancii. Kombinácia tvaru slova a morfologickej značky je ideálna, lebo práve morfologická značka obsahuje potrebné informácie o tvare slova a ich kombinácia je unikátna v slovníku.

Slovník sme rozširovali na základe iterácii cez dostupné slovenské korpusy. O iteráciách sa môžete dozvedieť viac v sekecii 4.2.1. Po skončení iterovania, náš slovník obsahoval cez 16 miliónov tvarov slov. Je to spôsobené tým, že každý tvar je identifikovaný svojou značkou. Jednému tvaru môže pripadať viac morfologických značiek. Situácia je najhoršia u prídavných mien, kedy vzniká takýchto tvarov najviac, kedy rozoznávame dve gramatické čísla, štyri rody, šesť pádov, tri stupne, kladný a záporný tvar.

Tvar slova	Slovníkový tvar	Značka	Anotácia	Vzor
schopnejší	schopný	k2eAgMnPc1d2		slovenský

Obr. 3.3: *Príklad záznamu morfologického slovníka*

Ako môžeme vidieť na obrázku 3.3, na ktorom je zobrazený záznam tabuľky, rozdelený do piatich stĺpcov. Prvý stĺpec označuje tvar slova, v tomto prípade *„schopnejší“*. Tomuto slovnému tvaru je priradený základný slovníkový tvar, resp. lema *„schopný“* a súčasne slovný tvar je popísaný morfologickou značkou. Z morfologickej značky sa dozvieme, že sa jedná o kladné prídavné meno mužského životného rodu v nominatíve plurálu druhého stupňa, t. j. komparatív. Ďalej slovnému tvaru, resp. leme, je priradený aj vzor, podľa ktorého sa daná lema ohýba, v tomto prípade sa jedná o vzor *„slovenský“*.

3.1.3 Značkovanie slovníka

Morfologické značky ohodnocujú jednotlivé tvary slov morfologickými kategóriami. Slúžia ako detailný popis jednotlivých tvarov slov, na základe ktorého môžeme slová rozlíšiť v morfologickom slovníku. Slovník používa kódovací typ značiek. Tento typ značenia pozostáva z kódovacieho atribútu a jeho hodnoty. Detailný popis značenia je uvedený v dokumente *Ajka tagset* [23]. Tento dokument je priložený na pamäťovom médiu, v adresári *morphological_analyzer* pod názvom *tags.pdf*. Tento tagset je využitý aj u morfologických analyzátorov *Ajka, Majka,* morfologického analyzátora *ma* výskumného strediska KNOT.

3.1.4 Lematizácia

Morfologický analyzátor pracuje s morfologickým slovníkom. Vstupné slovo, resp. jednoslovné pomenovanie, sa v tomto slovníku môže nachádzať, ale nemusí. Na základe tohto faktoru rozdeľujeme proces lematizácie na dve časti:

- vyhľadanie základného tvaru v morfologickom slovníku
- lematizáciu slovného tvaru

3.1.5 Vyhľadanie základného tvaru v morfologickom slovníku

Každé jednoslovné pomenovanie sa najskôr vyhľadá v morfologickom slovníku. Pri tejto akcii sa pošle požiadavka do slovníka na vyhľadanie základného slovníkového tvaru vstupného jednoslovného pomenovania a jeho vzoru.

V slovníku sa môžu vyskytovať rôzne základné slovníkové tvary, t. j. lemy tých istých slovných tvarov, preto sa vyhľadajú všetky lemy a spolu s lemami aj ich vzory. Vznikne nám pole dvojíc: lema, vzor. Pri výpise sa jednotlivé lemy vyskloňujú podľa ich príslušných vzorov. Vzory sú uložené v osobitnom súbore, ktorý analyzátor pri svojom spustení načíta.

Na druhú stranu by bolo možné poslať ďalší dotaz do databáze alebo skombinovať ho s predchádzajúcim dotazom na vrátenie všetkých potrebných informácii o leme a jej vyskloňovaných tvaroch. Hlavným dôvodom nezvolenia tohto prístupu je čas, za ktorý databáza vráti dáta, ktorý je v tomto prípade vysoký a to hlavne ak sa k slovnému tvaru nájde viac lem.

3.1.6 Lematizácia slovného tvaru

Do tejto fáze sa dostávajú jednoslovné pomenovania v prípade, že sa k ich tvaru nenašla lema v morfologickom slovníku. Základná informácia, s ktorou pracujeme, je práve vstupné slovo. Potrebné je získanie základného tvaru slova. Využijeme metódu osamostatnenia koreňa slova. Na začiatku zo slova odstránime možné predpony a generatívne časti slov ako: *pod-, čierno-, kanadsko-* atď.

Ďalej slovo prechádzame zľava doprava a vytvárame podreťazec, ktorý porovnávame so vzorovými príponami, v prípade zhody si uložíme vzniknutý stem, príponu a množinu vzorov, ktorej odpovedá. Ak sa ani jeden vzniknutý podreťazec nezhoduje so vzorovou príponou, tak sa môže jednať o slovo v základnom tvare, nezmyselné slovo, skratku. V tomto prípade analyzátor nie je schopný určiť vzor vstupného slova.

Po vytvorení zoznamu, ktorý pozostáva zo stemu a príslušnej množiny vzorov, je potrebné vytvoriť príslušné lemy zo stemu práve pomocou množiny vzorov, t. j. vytvorenie zoznamu odhadovaných lem slovného tvaru.

3.1.7 Zoznam odhadovaných lem

Pomocou metódy osamostatnenia koreňa sme vytvorili zo vstupného slova zoznam, pozostávajúci zo stemov a ich množiny vzorov. Existuje otázka týkajúca sa podoby lemy: *„Ktorý vzor sa použije pre vytvorenie lemy?“*. Na základe vstupného slova sa to nedá presne určiť, preto je potrebné vytvoriť zoznam odhadovaných lem.

V tomto zozname sa nachádzajú všetky lemy, ktoré sú vytvorené z ich príslušného stemu a jeho množiny vzorov. Každá lema z tohto zoznamu spĺňa pravidlá ohýbania svojich vzorov.

3.1.8 Štatistický výber

V tejto fáze nastáva štatistický výber najpravdepodobnejšej lemy. To znamená zistenie výskytu koncoviek lem zo zoznamu.

Intuitívne by sme hľadali výskyt jednotlivých koncoviek pomocou požiadaviek do morfologického slovníka. Tento spôsob nie je výhodný, čo sa týka času, za ktorý sa vyhľadá výskyt každej koncovky. Pri každom spustení si analyzátor vytvorí špeciálny slovník, ktorý pozostáva z lemy, ktorá je kľúčom, morfologickej značky, ktorá je hodnotou. Tento slovník, ktorý reprezentuje dátová štruktúra slovníka *dict* jazyka Python, sa vytvorí pomocou dvoch

súborov. Jedným je súbor, ktorý obsahuje slovník lem a ich vzorov (*slovak.lpn*) a druhým je slovník paradigiem (*slovak.paradigms*).

Následne vyhľadáme v tomto slovníku výskyt koncovky každej lemy, zo zoznamu lem a na základe značky určíme najpravdepodobnejší (najfrekventovanejší) vzor z množiny vzorov pre danú lemu. Ku koncu určíme najčastejšie vyskytovanú kombináciu lema a vzor.

Koniec koncov jedná sa o prostý štatistický výber a tento spôsob určenia lemy nie je dokonalý. Nastávajú prípady kedy ku danému slovu sa vyberie štatisticky *„lepšia“* lema, pričom v porovnaní s ručnou lematizáciou sa jedná o nezmysel.

3.1.9 Ohýbanie lemy

Ohýbanie alebo aj skloňovanie lemy je konečná fáza, kedy na základe priradeného vzoru a jeho pravidiel ohýbania lemy sa vygenerujú príslušné slovné tvary. Tieto pravidlá sú založené na ohýbaní podľa vzoru lemy, kedy sa vytvárajú slovné tvary pomocou stemu a prípon vzoru. Súčasne sa ku každému tvaru pridá slovný vzor a jeho príslušný morfologický reťazec, resp. morfologická značka, ktorá presne určuje gramatické kategórie daného tvaru slova ako sú rod, číslo, pád, osoba, čas... Tieto informácie sa následne vypíšu.

3.1.10 Uloženie lemy do slovníka

V prípade zvolenia možnosti uloženia novovzniknutej lemy do slovníka, sa jednotlivé tvary lemy spolu s jej morfologickými značkami a vzorom uložia do morfologického slovníka, pričom pri ďalšom hľadaní budú dostupné ihneď, bez zbytočného čakania procesu lematizácie neznámeho slovného tvaru.

3.1.11 Hodnotenie systému

Po vytvorení systému je potrebné systém ohodnotiť s ostatnými morfologickými analyzátormi. Otestovanie systému môže slúžiť ako jeho hodnotenie, zistenie či systém pracuje správne alebo je potrebné odstránenie istých nedostatkov systému. Hodnotenie morfologického analyzátora sme vykonali na základe testovacích sád. O jednotlivých výsledkoch sa dozviete viac v kapitole 4.

3.1.12 Použiteľnosť dát inými systémami

Slovník nasleduje tvar hlavného slovníka (súbor *.wltne*), využívaného morfologickým analyzátorom pre češtinu, vyvinutého v rámci výskumnej skupiny KNOT na Fakulte informačných technológii v Brne. Počas vývoja analyzátora bol slovník pomocou pomocných skriptov prevedený do formátu, ktorý môže načítať morfologický analyzátor *ma*. Bohužiaľ sa jednalo len o starý systém, ktorý načítal dáta do pamäte. Na základe tohto faktu by mohol byť slovník prevedený aj do formátu *fsa*, ktorý analyzátor *ma* dokáže načítať tiež a zredukoval by veľkosť slovníka.

3.1.13 Zhrnutie informácií

Pri vytváraní morfologického analyzátora sme sa držali postupu, ktorý sme uviedli vyššie.

Morfologický analyzátor je napísaný v jazyku Python. Pri vytváraní tohto systému bolo vytvorených niekoľko skriptov pre zostavenie morfologického slovníka, v rámci vytvorenia systému technických vzorov, jeho modifikáciu.

Morfologický slovník predstavuje databáza, ktorá využíva SQLite3 ako svoj *„database engine“*. Slovník sme rozširovali v iteráciách cez dostupné korpusy, o ktorých sa môžete dozvedieť viac v sekcii 4.2.1.

Na základe ručne vytvorených testovacích sád sa jednotlivé nedostatky analyzátora ladili. Na koniec bol analyzátor otestovaný na základe testovacích sád a bola vyhodnotená jeho úspešnosť v porovnaní s dostupnou alternatívou. Viacej o hodnotení analyzátora sa môžete dozvedieť v kapitole 4.

Zvolenie databázy ako formátu pre uloženie slovníka so sebou nesie isté nevýhody. Očakávame, že pri hodnotení analyzátora bude rýchlosť vyhľadávania v slovníku podstatne pomalšia ako u analyzátorov, ktoré využívajú slovník, uložený v štruktúre trie alebo v podobe konečného automatu (*fsa*). Túto nevýhodu sa snažíme kompenzovať vytvorením špeciálneho slovníka, ktorý sa využíva pri analyzovaní slov, ktoré sa nevyskytujú v slovníku.

3.2 Nástroj pre porovnávanie a vyhodnocovanie stemerov

Slovenčina ako flektívny jazyk je charakteristická svojou bohatou slovnou zásobou. Za základný slovotvorný postup tvorenia slov považujeme deriváciu, kde sa nové slová tvoria na základe zmeny morfologickej štruktúry slova pomocou afixov.

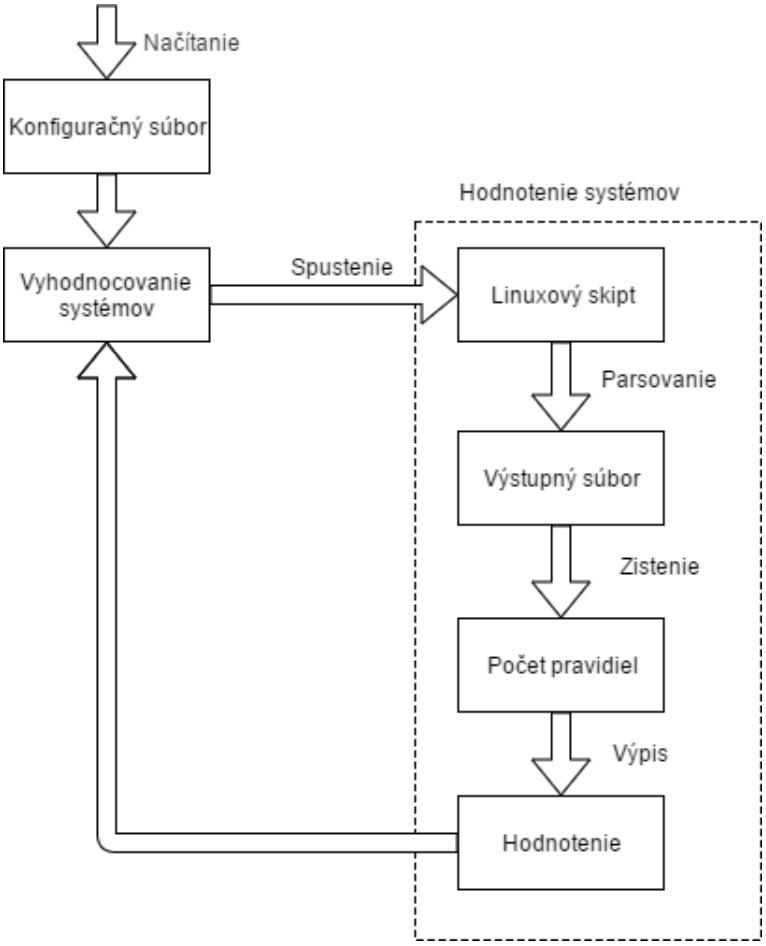
Nástroje, ktoré sa používajú na odstránenie afixov zo slova, sa nazývajú stemery, ktoré tieto afixy odstránia zo slova pomocou rôznych pravidiel, pričom ich výstupom je stem, t. j. koreň daného slova. Tieto nástroje je vhodné porovnať a vyhodnotiť.

Hlavná požiadavka na nástroj sa viaže na kvalitu slovníka, na základe ktorého hodnotí stemery a na jeho rozsah. Slovník by mal obsahovať záznamy, ktorých kľúč predstavuje koreň a hodnotou by mali byť všetky slovné tvary, ktoré daný koreň zdieľajú.

Vybratie správnych kritérií, na základe ktorých bude vyhodnocovanie prebiehať, tiež môžeme považovať ako požiadavku na cieľový systém.

Za požiadavku môžeme považovať aj podporu rôznych implementácii stemerov, t. j. systém by mal byť schopný vyhodnocovať rôzne implementácie stemerov.

Čas, za ktorý prebehne vyhodnotenie systému, je tiež dôležitý faktor a môžeme ho považovať aj za požiadavku. Vhodné by bolo tento čas minimalizovať.



Obr. 3.4: *Algoritmus hodnotenia stemerov*

Obrázok 3.4 zobrazuje algoritmus, ktorým sa riadi nástroj na porovnávanie a hodnotenie stemerov. Nástroj najskôr načíta konfiguračný súbor a na základe neho začne postupne vyhodnocovať systémy. Následne uvedieme popis jednotlivých fáz algoritmu.

3.2.1 Konfiguračný súbor

Konfiguračný súbor predstavuje základné vstupné informácie systému na vyhodnocovanie a porovnávanie stemerov. Každý správne zadaný záznam v tomto súbore reprezentuje stemer, ktorý nástroj vyhodnocuje. Podrobný popis formátu konfiguračného súboru je uvedený v manuály B.2, ktorý je priložený ako príloha.

3.2.2 Vyhodnocovanie systémov

Potrebné informácie pre vyhodnocovanie a porovnávanie systémov sú uložené v konfiguračnom súbore, ktorý následne spracujeme. Nad každým systémom nástroj spustí externý linuxový skript. V tomto skripte sa v cykloch privádzajú na vstup stemeru jednotlivé slová, uložené v testovacom slovníku, ktorý je odvodený od derivačného slovníku.

3.2.3 Derivačný slovník

Derivačný slovník je vytvorený z morfologického slovníka, ktorý používa morfologický analyzátor. Kľúčom v slovníku je stem slova. Ku každému stemu sú pridružené základné slovníkové tvary spolu s ich vzormi. Týmto spôsobom dostaneme pre každý tvar lemy ten istý stem. Zjednodušenú variantu derivačného slovníka použijeme ako testovaciu sadu pri porovnávaní a vyhodnocovaní systémov pre stematizáciu slov.

Vytvorenie slovníka

Pri vytváraní derivačného slovníka si musíme uvedomiť, že v morfologickom slovníku sa nachádzajú slová plnovýznamové aj neplnovýznamové. Primárny účel slovníka je jeho použitie pri porovnávaní a vyhodnocovaní stemerov.

Každý slovníkový tvar je nutné stematizovať a združiť tie lemy, ktorých stem sa zhoduje. K lemám je pridružený vzor, podľa ktorého je možné generovať všetky možné tvary lemy, ktoré zároveň zdieľajú ten istý stem.

V rámci procesu stematizácie slov sa najskôr slovo vyhľadá v zozname *stop slov*[3], ak sa nájde tak sa odstráni. Ďalej je potrebné slovo rozdeliť na dve časti podľa jeho prvej slabiky tak, že jeho druhá časť bude nasledovať spoluhláskou po prvej slabike slova. Napríklad pre slovo *„chlapec“* sa vytvorí *„chlap“* a *„ec“*. Pritom ak je slovo jednoslabičné tak sa považuje ako stem.

V druhej časti slova sa hľadajú jednotlivé morfologické afixy. V prípade zhody afixu sa afix odstráni a v prípade, že po odstránení sa na poslednej pozícii prehľadávanej časti slova nachádza samohláska, tak sa odstráni tiež. Týmto spôsobom sa odstránia všetky morfologické afixy a vzniká koreň slova.

abeced	abecedný#slovenský abecedovaný#nový abecedovať#potrebovať abecedovanie#obdobie abecedne#sociálny abecedno#sociálny abeceda#streda abecedník#jazyk abecedár#mesiac

Obr. 3.5: *Príklad riadku derivačného slovníka*

Ako môžeme vidieť na obrázku 3.5, ktorý zobrazuje riadok derivačného slovníka, koreň *„abeced“* zdieľajú lemy: *abecedný* so vzorom *slovenský*, *abecedovať* so vzorom *potrebovať* atď.

Morfologické afixy

Morfologické afixy sú uložené v osobitných súboroch obsahujúce sufixy a prefixy slovných druhov. Pri vytváraní slovníka ich skript načítal a v prípade zhody odstránil. Jednotlivé sufixy a prefixy sme vypísali z kníh Onomaziologická štruktúra slovenčiny [10] a Základy slovenskej lexikológie [21]. Postačujú morfémy základných tvarov slov, pričom ku stemu slova sa ukladá aj príslušná lema, ku ktorej je pridružený aj vzor.

3.2.4 Hodnotenie systémov

Jednotlivé systémy, ktoré stematizujú slová je nutné ohodnotiť. Systémy hodnotíme podľa troch kritérií ako:

[3] V tomto zozname sa nachádzajú slová, ktoré sa v danom jazyku vyskytujú s vysokou frekvenciou, pričom nenesú žiadnu významovú informáciu.

- počet pravidiel, s ktorými stemer pracuje
- počet prípadov, kedy tvary, patriace k sebe, budú mať rôzny kmeň
- počet prípadov, kedy tvary, nepatriace k sebe, budú mať taký istý kmeň

3.2.5 Hodnotenie na základe počtu pravidiel

Táto kategória je cielená na systémy, ktoré sú založené na pravidlách. Systémy na základe pravidiel odstraňujú afixy, t. j. predpony a prípony, zo vstupného slova. V rámci tohto hodnotenia je potrebné zistiť počet pravidiel, na základe ktorých daný systém rozhoduje či zo slova odtrhne daný afix.

Najjednoduchší spôsob zistenia počtu pravidiel je hľadanie reťazcov (sufixov alebo prefixov) v rámci podmienkových blokov, t. j. zistenie počtu pravidiel, ktoré vyhľadávajú sufixy v slove. Preto toto vyhľadávanie je špecializované pre moduly, ktoré obsahujú práve podobný automat.

3.2.6 Hodnotenie na základe tvarov

Cieľom tejto kategórie je vyhodnotenie funkčnosti systémov, ktoré slúžia na stematizáciu slov. Tieto systémy porovnávame na základe slovníka, obsahujúceho stemy a príslušné lemy, ktoré ich tvoria. Pre jednoduchosť hodnotenia privádzame na vstup systémov len slovotvorné základy, čím zamedzíme prípadom nepravidelného skloňovania slovných tvarov. Napríklad slovnému tvaru *„kôz“* patrí koreň *„koz“*.

Tieto slová sú privedené ako vstup pre jednotlivé systémy, pričom sa očakáva, že stemer vyhodnotí pre všetky slová ten istý stem. Týmto spôsobom určíme počet prípadov, kedy stemer generuje stemy, ktoré práve nie sú zhodné pre všetky vstupné slová. Následne si uložíme výpis stemeru do textového súboru, ktorý ďalej spracujeme.

Tento textový súbor spracujeme a hľadáme tie prípady, kedy stemer vygeneroval tie isté stemy pre slová, ktorých stem je v skutočnosti rôzny. Výstupný súbor môže slúžiť aj ako log z testu. Pri opätovnom spustení testovania sa *„starý“* súbor vymaže a nahradí sa novým.

3.2.7 Porovnávanie systémov

Po skončení hodnotenia jedného systému nasleduje výpis výsledkov na štandardný výstup a v prípade výskytu viacerých systémov, popísaných v konfiguračnom súbore, nasleduje hodnotenie ďalšieho systému.

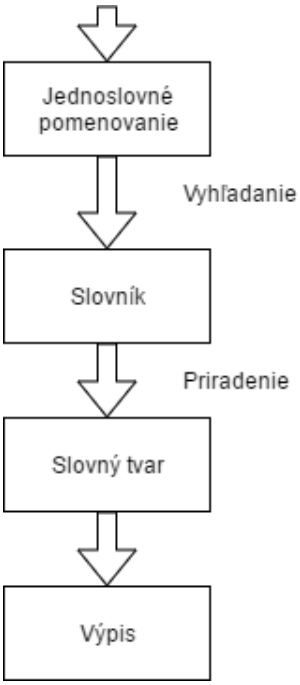
3.2.8 Zhrnutie

Nástroj pre porovnávanie a vyhodnocovanie stemerov je napísaný v jazyku Python3. Systém spúšťa pri testovaní jednotlivých stemerov linuxový skript, ktorý privádza na vstup stemerov slová a vyhodnocuje jednotlivé stemy. Jednotlivé slová pochádzajú z jednoduchšej verzie derivačného slovníka. Derivačný slovník bol vytvorený z morfologického slovníka pomocou skriptov, ktoré jednotlivé slová stematizovali. Systém sme použili na vyhodnotenie dvoch slovenských pravidlových stemerov. O ich úspešnosti sa dozviete viac v kapitole 4.

Očakávame, že čas potrebný na vyhodnotenie stemeru, bude závisieť na konkrétnej implementácii stemera a veľkosti testovacieho slovníka. Stemery generujú koreň slova na základe istých pravidiel a očakávame, že stemery pomocou týchto pravidiel nebudú generovať presné korene slov.

3.3 Nástroj na rekonštrukciu diakritiky

Nástroj na rekonštrukciu diakritiky vznikol ako vedľajší nástroj, pri testovaní pokrytia morfologického analyzátora.



Obr. 3.6: *Algoritmus rekonštruovania diakritiky*

Obrázok 3.6 zobrazuje algoritmus rekonštrukcie diakritiky, ktorý nástroj využíva. Pozostáva z načítania jednoslovného pomenovania, jeho vyhľadanie v slovníku, priradenia správneho tvaru a výpisu. Následne uvedieme popis jednotlivých fáz algoritmu.

3.3.1 Diakritické znamienka

Diakritické znamienka sú rozlišovacie znamienka, ktoré menia výslovnosť slova, vyznačujú tón reči, rozlišujú homonymá. V slovenčine rozlišujeme štyri typy diakritických znamienok:

- **dĺžeň**
- **mäkčeň**
- **dve bodky**
- **vokáň**

3.3.2 Slovník

Nástroj pracuje so slovníkom, ktorý vychádza opäť z morfologického slovníka.

Vytvorenie slovníka

Slovník je vytvorený z morfologického slovníka. Z každého slovného tvaru sme postupne odstránili diakritiku, tým sme dostali všetky tvary bez diakritiky z morfologického slovníka.

Tieto slová sme v ďalšom kroku zoradili podľa ich frekvencie z frekvenčného slovníka, ktorý sme vytvorili z dostupných korpusov. Na základe týchto údajov sme nechali len najčastejšiu variantu prepisu slov bez diakritiky. To znamená, že nástroj vždy zmení slovo *„byt“* na slovo *„byť“*.

Formát slovníka

Slovník, podobne ako morfologický slovník, reprezentuje databáza SQLite3, ktorej tabuľka obsahuje tri stĺpce:

- tvar bez diakritiky, ktorý je primárnym kľúčom
- najfrekventovanejší gramaticky správny tvar
- výskyt najfrekventovanejšieho prepisu vo frekvenčnom slovníku

Tvar bez diakritiky	Gramaticky správny tvar	Výskyt varianty prepisu
byt	byť	18 889 780

Obr. 3.7: *Príklad záznamu slovníka*

Na obrázku 3.7 môžeme vidieť, že najčastejšou formou prepisu slova *„byt“* je tvar *„byť“*, ktorý sa vyskytoval viac ako 18,8 miliónkrát.

3.3.3 Algoritmus

Nástroj rekonštruuje diakritiku vstupného slova, t. j. jednoslovného pomenovania. Algoritmus rekonštrukcie je založený na jednoduchej požiadavke na výskyt vstupného tvaru slova. V prípade, že tvar sa vyskytuje v slovníku, nástroj zistí jeho gramaticky správny tvar a vypíše ho na výstup v takom formáte jednotlivých znakov, aký dostal na vstupe. Môže nastať prípad, kedy sa daný vstupný tvar slova nenachádza v slovníku. V tedy sa slovo považuje za diakriticky správne, pretože v slovníku sa vyskytujú len *„známe“* tvary slov bez diakritiky.

Túto metódu rekonštrukcie diakritiky môžeme považovať za unigramovú metódu. Vyhľadávanie slova je založené len na vyhľadaní jedného slova, t. j. nevychádzame z kontextu slovného spojenia, prípadnej vety. Napríklad slovné spojenie *„novy byt“* nám nástroj zrekonštruuje na spojenie *„nový byť“*.

3.3.4 Zhrnutie

Nástroj vznikol ako vedľajší produkt vytvorenia morfologického analyzátora. Je napísaný v jazyku Python3. Slovník, ktorý používa, predstavuje databáza, ktorá využíva SQLite3 ako svoj *„database engine“*.

Kapitola 4

Hodnotenie vytvorených systémov

Táto kapitola popisuje vyhodnotenie vytvorených nástrojov. Podkapitola 4.1 zhŕňa stanovené ciele práce.

V podkapitole 4.2 uvádzame celkové zhodnotenie morfologického analyzátora, proces iterovania cez dostupné korpusy a ich krátky popis, testovacie sady, alternatívne systémy, výsledky iterovania, porovnanie analyzátora s dostupnou alternatívou, prevod slovníka do štruktúry *marisa-trie* a ku koncu zhodnocujeme problémy analyzátora a navrhujeme ich možné riešenia.

V podkapitole 4.3 zhodnocujeme nástroj pre porovnávanie a vyhodnocovanie stemerov, testovací slovník, popisujeme implementácie testovacích stemerov a ich následné vyhodnotenie. Ku koncu uvádzame zhodnotenie problémov nástroja a ich možných riešení.

4.1 Cieľ práce

Ako cieľ práce sme si stanovili vytvorenie systému vzorov pre flektívnu morfológiu slovenčiny, z neho odvodený morfologický analyzátor, schopný určiť morfologické kategórie slov za pomoci ich lematizácie, pričom tieto slová sa nemusia nachádzať len v slovníku. Nástroj by mal vygenerovať všetky tvary zlematizovaného vstupného slova, ku ktorým priradí ich morfologický reťazec, slovný vzor a lemu. Tento systém je závislý na slovníku, s ktorým komunikuje. Rozsiahlosť, takéhoto slovníka, je dôležitý faktor. Všeobecne platí čím kvalitnejší, resp. rozsiahlejší slovník, tým je väčšia pravdepodobnosť, že slovník bude obsahovať hľadané slovo a tým sa zvýši aj jeho úspešnosť. Slovenčina neustále priberá nové slová a dôležitosť existencie takéhoto nástroja, ktorý dokáže slovám určiť morfologické kategórie, resp. ich lematizovať, stúpa. Morfologický analyzátor je schopný svoj slovník aktualizovať, na základe použitia vhodných argumentov pri jeho spúšťaní.

Ako ďalší cieľ práce sme si stanovili vytvorenie nástroja, ktorý dokáže porovnávať a vyhodnocovať systémy stematizujúce slová na základe pravidiel. Tieto systémy hodnotíme na základe jednoduchšej verzie derivačného slovníka, ktorý vznikol z morfologického slovníka. U týchto systémov sme sa rozhodli sledovať jednak počet pravidiel, s ktorými systémy pracujú, ako aj sledovať ich výstup. Presnejšie sledujeme jednotlivé prípady, kedy systém určí rôzny kmeň slova skupine slov, ktoré podľa testovacieho slovníka majú práve ten istý kmeň a prípady, kedy systém určí ten istý kmeň slovám, ktoré nemajú podľa slovníka ten istý kmeň.

Ako vedľajší cieľ práce vznikol nástroj na rekonštrukciu diakritiky, ktorý sme využili pri práci s korpusmi. Za pomoci tohto nástroja sme pri iteráciách rozširovali slovník o slová, ktorých diakriticky správny tvar sa nenachádza v slovníku.

4.2 Morfologický analyzátor

Morfologický analyzátor spracúva jednotlivé vstupné jednoslovné pomenovania, ktoré prevedie na ich základný slovníkový tvar (lemu), ktorý vypíše, prípadne generuje všetky slovné tvary vychádzajúce zo vzoru, ktorým priradí základnému tvaru.

4.2.1 Iterácie

Slovník morfologického analyzátora sme trénovali na slovenských korpusoch a vybrali sme si jeden, na ktorom sme porovnávali pokrytie slov pred iteráciou a úspešnosť prevádzania slovných tvarov na ich základný slovníkový tvar v porovnaní s iným systémom. V jednotlivých iteráciách cez dostupné korpusy sme rozširovali slovník o slovné tvary, ktoré sa v ňom nevyskytovali.

Toto rozširovanie slovníka sa konalo poloautomaticky, kedy analyzátor určil zo slovných tvarov ich slovotvorný základ, ku ktorému pridružil aj jeho vzor. Následne sme dáta uložili do slovníka. V súčasnom stave, analyzátor dokáže sám rozširovať slovník v prípade, keď vstupné jednoslovné pomenovanie sa nenachádza v slovníku.

Zo všetkých korpusov sme museli vytvoriť ich príslušný frekvenčný slovník, na základe ktorého prebiehali samotné iterácie.

V krátkosti uvedieme všetky použité korpusy.

Slovensko-český paralelný korpus

Slovensko-český paralelný korpus, presnejšie jeho slovenskú časť sme si vybrali za *„vzorový“* korpus, pretože obsahoval jednotlivé slovné tvary pridružené s ich lemou a morfologickou značkou. Tento korpus sa nachádza na Jazykovednom Ústave Ľudovíta Štúra Slovenskej akadémie vied v Bratislave. Celý korpus obsahuje 418,5 miliónov tokenov, pričom každá časť obsahuje 209,3 miliónov tokenov. Texty sú automatický zarovnané vo vetách. Informácie o korpuse sú dostupné na internetovej stránke slovenského korpusu [26]. Časť korpusu nám bola poskytnutá na základe dohovoru s pracovníkmi ústavu. Táto časť obsahuje cez 653 tisíc viet.

Slovensko-anglický paralelný korpus

Slovensko-anglický paralelný korpus sme využili v procese iterácií. Korpus je len časťou väčšieho systému korpusov nazývaného *Europarl: A Parallel Corpus for Statistical Machine Translation*. Celý systém obsahuje jednotlivé paralelné korpusy prevažne jazykov využívajúcich sa v Európskej únii. O tomto systéme korpusov sa môžete dozvedieť viac v práci Philippa Kehna *Conference Proceedings: the tenth Machine Translation Summit* [13]. V rámci práce sme využili len slovenskú časť paralelného korpusu dostupnú z internetovej stránky[1]. Korpus dokopy obsahuje vyše 640 tisíc viet.

[1] http://www.statmt.org/europarl/ [Online; navštívené 15.12.2016]

SkTenTen

Korpus je súčasťou systému TenTen korpusov. Korpusy sú vytvorené na základe metódy *„web crawling“*, t. j. proces prechádzania webových stránok s cieľom získania informácii ako obsah webovej stránky, metadáta, odkazy na ďalšie stránky atď. Bližšie informácie ohľadom TenTen korpusov sú uvedené v práci *The TenTen Corpus Family* [11]. V rámci práce sme využili práve slovenskú variantu skTenTen dostupnú z portálu sketchengine [25]. Tento korpus obsahuje 876 miliónov tokenov.

Aranea

Aranea je rodinou korpusov, podobne ako TenTen korpusy. V rámci zhotovenia korpusov bol opäť použitý proces *„web crawlingu“* nástrojom *SpiderLing*[2]. Táto rodina korpusov je pripravená za pomoci Vladimíra Benka ako súčasť projektu inštitúcii Department of Plurilingual and Intercultural Communication a Ústavu Ľudovíta Štúra Slovenskej akadémie vied. Bližšie informácie môžete nájsť v práci Vladimíra Benka *Aranea: Yet Another Family of (Comparable) Web Corpora* [4]. V práci sme využili slovenský korpus, ktorý nám poskytol Vladimír Benko.

4.2.2 Testovacie sady

V konečnej fáze, po skončení iterácií, je potrebné nástroj ohodnotiť. Na základe testovania zistíme celkovú úspešnosť morfologického analyzátora. Testovanie musí zostať objektívne, resp. musíme zabezpečiť to, aby testovacie sady neobsahovali len slová nachádzajúce sa vo slovníku. V rámci zachovania tohto faktora sme pripravili niekoľko testovacích sád, ktoré obsahovali náhodné články, publikácie z novín, odborných prác, kníh.

Odborný článok

Odborný článok obsahuje inú kategóriu slov, ktorá sa nevyužíva v bežnej komunikácii. Pri vytvorení tejto testovacej sady sme vybrali náhodný článok *Sofiológia ako príklad integrálnej vedy a vzdelávania v tradícii slovanov* napísaný Emilom Pálešom [19]. Testovacia sada obsahuje 963 slov.

Článok časopisu Pravda

Hlavným dôvodom vytvorenia tejto testovacej sady je fakt, že tieto články obsahujú slová používajúce sa v bežnej komunikácii a to, že sledujú aktuálne dianie na Slovensku. Tento článok bol náhodne vybraný z časopisu *Pravda* [30].Táto testovacia sada obsahuje spolu 1085 slov.

Knižný úryvok

Ako poslednú testovaciu sadu sme zvolili prvú kapitolu z knihy *Harry Potter a Dary smrti* [22]. Jedná sa o poslednú knihu série príbehov kúzelníka Harryho Pottera. Táto kniha završuje príbeh Harryho Pottera a jeho kamarátov v boji proti lordovi Voldemortovi. V tejto testovacej sade sa nachádza 1578 slov.

[2]Podrobné informácie o nástroji sú dostupné na stránke https://is.muni.cz/publication/1095720/cs

4.2.3 Existujúce systémy

Existuje hneď niekoľko systémov, ktoré sa využívajú na lematizáciu slov, prípadne stemming. Jednými z nich sú nástroje Majka, Morphodita, Morče, prípadne Snowball.

Majka

Majka je rýchly morfologický analyzátor, ktorý dokáže spracovať približne milión slov za sekundu. Jej predchádzajúci systém pre morfologickú analýzu sa nazýva Ajka. Majka má kompletne inú implementáciu založenú na konečných automatoch, preto je aj mnohonásobne rýchlejšia. Systém k vstupnému tvaru slova priradí základný tvar slova a gramatickú značku. Oba systémy sú napísané v jazyku C. V súčasnosti Majka obsahuje slovníky pre češtinu ale aj slovenčinu, poľštinu, nemčinu, taliančinu atď. Viacej informácii o analyzátore sa dozviete z práce Pavla Šmrka a Pavla Rýchlého *Majka – rýchly morfologický analyzátor* [35].

Morphodita

Morfologický značkovač je *„open-source“* nástroj pre morfologickú analýzu, tagovanie, tokenizovanie a je distribuovaný ako samostatný nástroj alebo knižnica spolu s natrénovaným lingvistickým modelom, českým alebo anglickým. Viac informácii o morfologickom značkovači Morphodita sa môžete dozvedieť v publikácii autorov nástroja Jany Strakovej a Milana Straku *Open–Source Tools for Morphology, Lemmantization, POS Tagging and Named Entity Recognition* [29].

Morče

Projekt Morče, ktorého názov je skratka od pojmu „morfológia češtiny“, je nástroj na morfologické značkovanie českých textov. Algoritmus nástroja je založený na skrytom Markovom modele s priemerným perceptronom[3]. Nástroj je napísaný v jazyku C, poskytovaný pod GPL[4] licenciou bez registrácie. Viac informácii o tomto nástroji sa môžete dozvedieť z internetovej stránky [20].

Snowball

Snowball je jazyk pre spracovanie malých reťazcov, navrhnutý pre vytváranie takzvaných *„stemming“* algoritmov, ktoré sa využívajú pri získavaní informácii. Podľa týchto algoritmov sa implementujú jednotlivé nástroje, nazývané stemery. Algoritmy sú založené na odtrhávaní sufixov zo slova, ktoré sa nachádzajú v príslušnom jazyku. Viacej informácii o Snowball sa môžete dozvedieť z práce Martina Portera *Snowball: A language for stemming algorithms* [18].

4.2.4 Hodnotenie úspešnosti s iným systémom

Náš nástroj sme sa rozhodli porovnať s morfologickým analyzátorom Majka, ktorý rovnako ako náš analyzátor využíva metódu lematizácie. Na začiatku sme v iteráciách cez dostupné korpusy rozširovali slovník a porovnávali správnosť určenia lem nášho analyzátora a Majky pomocou frekvenčného slovníka, vytvoreného zo slovensko-českého paralelného korpusu. Pre

[3] Tento algoritmus je viac popísaný v práci Michaela Collinsa [6]

[4] Licencia pre slobodný softvér

zjednodušenie testovania sa vo frekvenčnom slovníku nachádzali slová, ktoré sa vyskytovali v korpuse aspoň stokrát, čím testovací prípad obsahoval dokopy 7042 slov.

Jednotlivé testovacie sady sme lematizovali ručne, aby sme predišli chybám oboch systémov a následne sme ich spustili na oboch systémoch. Dostali sme výsledky, na základe ktorých sme určili úspešnosť oboch systémov. Tieto výsledky sme zobrazili do tabuliek a ku koncu vypočítali priemernú úspešnosť oboch systémov.

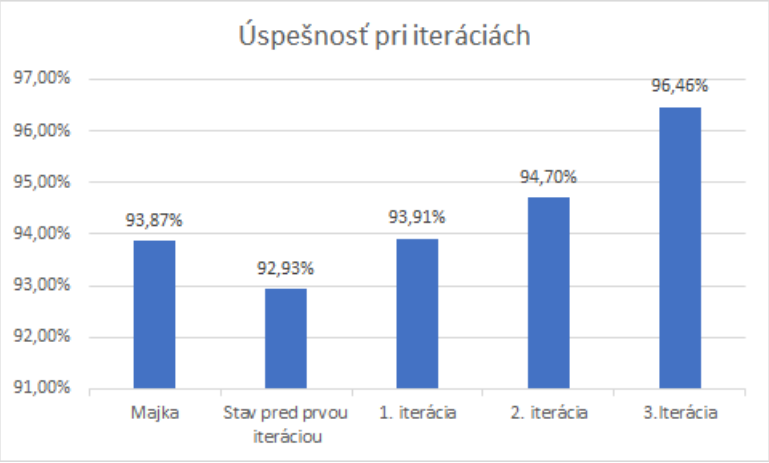
Iterovanie

Analyzátor určoval slovotvorný základ, ku ktorému pridružil aj vzor u slov s istou frekvenciou výskytu.

Prvá iterácia prebiehala cez slovensko-anglický paralelný korpus. V tomto korpuse sa vyskytuje cez dvanásť miliónov slov. Analyzátor lematizoval slová, ktoré sa vyskytovali aspoň stokrát. Vzniklo 203 nových slovotvorných základov, ktoré sa nevyskytovali v slovníku. U 176 slovotvorných základov analyzátor dokázal určiť ich vzor, pričom nedokázal určiť vzor 27 slovných tvarov. Prevažne anglicizmy a cudzie vlastné mená.

Druhá iterácia prebiehala cez korpus skTenTen. V tomto korpuse sa vyskytuje cez 737 miliónov slov. Najskôr sme otestovali pokrytie slov, vyskytujúcich sa aspoň tisíckrát, pričom nám vznikol zoznam slov, z ktorého sme odstránili tie slovná, ktorých diakriticky správny tvar sa vyskytoval v slovníku. Analyzátor určil 220 najčastejšie vyskytovaných slovotvorných základov zo zoznamu slov, ktoré sa nevyskytovali v slovníku. Analyzátor dokázal určiť vzor u 190 nových lem, nedokázal určiť vzor u 30 lem. Dôvodom tohto počtu je fakt, že analyzátor začal určovať nesprávne slovotvorné základy slov, v dôsledku výskytu nespisovných slov, anglicizmov a cudzích vlastných mien.

Posledná iterácia prebiehala cez slovenskú časť korpusu z rodiny Aranea. V tejto časti korpusu sa vyskytuje 853 miliónov slov. Pri poslednej iterácii sme postupovali podobne ako pri iterácii cez skTenTen. Otestovali sme pokrytie slov, ktoré sa vyskytujú aspoň tisíckrát, pričom sme získali zoznam slov, z ktorého sme odstránili slová, ktorých diakriticky správny tvar sa vyskytoval v slovníku. Analyzátor určil 402 najčastejšie vyskytovaných slovných základov nevyskytujúcich sa v slovníku. Analyzátor dokázal určiť vzor u 350 slovotvorných základov, pričom u 52 lem vzor nedokázal určiť.



Obr. 4.1: *Graf zobrazujúci úspešnosť pri testovaní nad slovensko-českým paralelným korpusom*

Ako môžeme vidieť na grafe 4.1, slovník nášho analyzátora bol horší v porovnaní s Majkou pred prvou iteráciou o 0,94%. Následne po prvej iterácii sa k Majke priblížil na rozdiel 0,04% a od druhej iterácie mal náš analyzátor lepšiu úspešnosť nad morfologickým analyzátorom Majka, po druhej iterácii o 0,83% a po tretej iterácii o 2,59%.

4.2.5 Hodnotenie

Prvá testovacia sada

Náš analyzátor	Majka
98,44%	95,63%

Tabuľka 4.1: *Úspešnosť systémov*

Prvá testovacia sada obsahovala odborný článok. Ako môžete vidieť v tabuľke 4.1, náš analyzátor určil o 2,81% viac správnych lem ako morfologický analyzátor Majka.

Druhá testovacia sada

Náš analyzátor	Majka
97,95%	95,78%

Tabuľka 4.2: *Úspešnosť systémov*

Testovacia sada obsahovala článok z časopisu *Pravda*. Výsledky testovania sú zobrazené v tabuľke 4.2, kde náš analyzátor určil o 2,17% viac správnych lem ako Majka.

Tretia testovacia sada

Náš analyzátor	Majka
92,10%	90,56%

Tabuľka 4.3: *Úspešnosť systémov*

Posledná testovacia sada obsahovala celú prvú kapitolu knihy *Harry Potter a Dary smrti*. Výsledky sú zobrazené v tabuľke 4.3, kde náš analyzátor určil o 1,54% viac správnych lem ako morfologický analyzátor Majka.

Priemerná úspešnosť

Po vykonaní všetkých testov a iterácii sme vypočítali priemernú úspešnosť nášho morfologického analyzátoru a Majky osobitne, podľa vzorca uvedeného v rovnici 4.1.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.1}$$

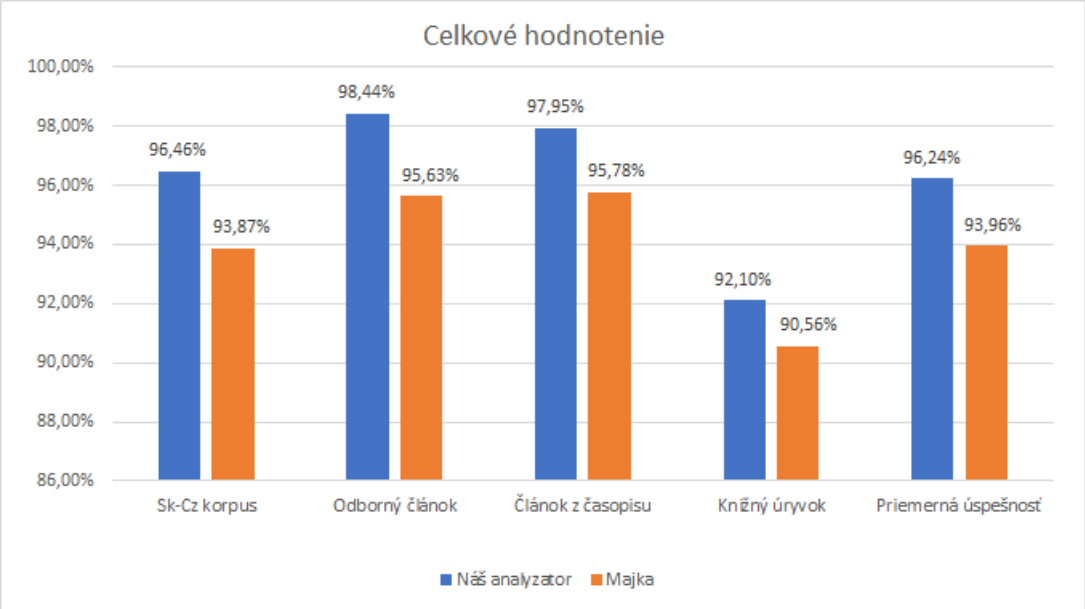
Kde:

μ – priemerná úspešnosť

x_i – úspešnosť testu

n – počet testov

Priemerná úspešnosť nášho morfologického analyzátora je 96,24% a morfologického analyzátora Majka 93,96%.



Obr. 4.2: *Graf zobrazujúci poslednú iteráciu, jednotlivé testovacie sady a priemernú úspešnosť*

Graf 4.2 zobrazuje zhrnutie úspešnosti pri hodnotení nášho analyzátora a Majky po tretej iterácii na Slovensko-českom paralelnom korpuse, jednotlivých testovacích sadách a priemernú úspešnosť nástrojov pre lepšiu prehľadnosť.

4.2.6 Marisa-trie

V rámci testovania zaberanej veľkosti slovníka sme pôvodný slovník previedli do formátu dátovej štruktúry trie, konkrétnejšie *marisa-trie*, ktorá je popísaná v sekcii 2.11.1. Morfologický slovník v podobe databázy *SQLite3* zaberá približne 1,9 GB pamäte. Keď tento morfologický slovník prevedieme do do dátovej štruktúry *marisa-trie*, jeho veľkosť klesne až na približne 94 MB, čo je približne 5% veľkosti priestoru, ktorý zaberá databáza. Nevýhodou tejto implementácie je fakt, že sa jedná o statickú trie štruktúru, t. j. pri aktualizovaní slovníka by sa musela celá trie štruktúra zostaviť nanovo, čo nie je výhodné.

4.2.7 Zhodnotenie problémov a ich možné riešenia

Morfologický analyzátor pracuje so svojím morfologickým slovníkom a na jeho základe sa snaží odhadovať základný tvar, vzor vstupného slova aj pre slová, ktoré sa nenachádzajú v slovníku.

Systém nie je dokonalý, nevie určiť všetky slová, ktoré sa nachádzajú v slovenčine, pretože slová určuje pomocou morfologického slovníka a ak sa slovo nenachádza v ňom, tak sa ho snaží zlematizovať a určiť jeho vzor.

Úskalia analyzátora sú slová neplnovýznamové ako citoslovce, spojky, skratky, predložky, častice, slová, ktoré nenesú nejakú informáciu. Ďalej slová nespisovné, resp. slová bez diakritiky, priezviská, hlavne ženského rodu a vlastné mená, slová cudzieho pôvodu. Analyzátor tieto slová môže určiť nesprávne. S týmto súvisí aj určenie vzorov pri slovách, ktoré sa nachádzajú v ich základnom slovníkovom tvare, pričom ich základný slovníkový tvar neobsahuje morfologické sufixy, jedná sa o koreň slova. V takomto prípade analyzátor nenájde vzor, podľa ktorého by sa vstupná lema skloňovala.

Tento problém by sa mohol čiastočne riešiť zlepšením metódy štatistického výberu koncoviek slov, prípadne rozšírením slovníka paradigiem a následným znovu zostavením morfologického slovníka. Ďalším riešením by bolo použite lepšie označovaných dát na zostavenie slovníka alebo použitie iného, lepšie spracovaného slovníka. Vylepšením by bolo aj zavedenie slova do kontextu prípadnej vety, vďaka ktorému by sa príslušné informácie o slove dali určiť presnejšie.

Za riešenie môžeme považovať aj využitie anotovaných korpusov, ktoré obsahujú v rámci slovného tvaru aj jeho lemu, gramatickú značku. Na základe takýchto korpusov by sme mohli slovník očistiť od chybne určených tvarov a pridať nové slovné tvary s ich lemami, ktoré sa nevyskytujú v slovníku.

Podstatou iterácii cez dostupné korpusy bolo rozširovanie morfologického slovníka, čo sa ukázalo ako vhodná metóda. Po skončení iterovania, morfologický slovník obsahoval frekventované slová, ktoré sa vyskytovali v nami dostupných korpusoch. Korpusy skTenTen, slovenská časť rodiny Aranea sú výťahom *„slovenského internetu“*, teda obsahujú aj veľa frekventovaných, diakriticky nesprávnych slov, nespisovných slov, anglicizmov apod. Ako lepší kandidáti na trénovanie slovníka sa skôr javia korpusy, ktoré sú založené na literárnych dielach ako beletria, populárno-vedecká literatúra apod. V ktorých sa vyskytujú prevažne spisovné slová.

Ďalším problémom analyzátora je veľkosť jeho slovníka. Databáza zaberá veľa miesta na diskovom médiu. Na to nadväzuje aj rýchlosť komunikácie so slovníkom. V rámci lematizovania neznámeho slova tento problém čiastočne riešime tým, že vytvárame nový slovník, uložený v dátovej štruktúre slovníka (*dict*) jazyka Python, pri spúšťaní analyzátora, čo znižuje oneskorenie pri lematizácii slova, ktoré sa nevyskytuje v morfologickom slovníku.

Dôležitým faktorom je typ média, na ktorom sa morfologický slovník nachádza. Čas, potrebný na analýzu slova, je priamoúmerný čítacej rýchlosti použitého média. Napríklad na klasickom pevnom disku je toto zdržanie oveľa väčšie ako na SSD disku.

Tieto dva problémy sa dajú vyriešiť lepším návrhom slovníka. Databáza nie je moc vhodným formátom uloženia slovníka. Vhodnou zmenou by bolo využitie dátovej štruktúry trie, ktorá nie je statická, ale ponúka aj aktualizovanie svojej štruktúry ako napríklad *cedar* (viď 2.11.1).

4.3 Nástroj pre porovnávanie a vyhodnocovanie stemerov

Nástroj vyhodnocuje prácu systémov určených na stematizáciu slov, t. j. stemerov. Jednotlivé systémy hodnotí pomocou testovacieho slovníka, ktorý je odvodený od derivačného slovníka. Podstata derivačného slovníka spočíva v združení lem a ich vzorov, ktoré zdieľajú ten istý koreň.

Na základe testovacieho slovníka systém pomocou linuxového skriptu privádza na vstup jednotlivých stemerov slová, ktoré zdieľajú ten istý koreň a na základe ich výstupu ich hodnotí. Hodnotenie spočíva v tom, že stemer nemusí generovať presný koreň, ale pre všetky slová, ktoré koreň zdieľajú podľa derivačného slovníka, musí byť generovaný koreň zhodný.

4.3.1 Testovací slovník

Ako sme sa už niekoľko ráz zmienili, stemery hodnotíme na základe testovacieho slovníka, ktorý je odvodený od derivačného slovníka. Derivačný slovník má až 84 tisíc záznamov, pričom pri testovaní využívame jeho jednoduchšiu verziu, kedy riadok testovacieho slovníka obsahuje minimálne dve lemy, ktoré zdieľajú ten istý koreň, pričom informáciu o vzore neuvádzame. Týmto spôsobom nám vznikol testovací slovník obsahujúci cez 31 tisíc záznamov a dokopy 123 385 slov.

4.3.2 Testovacie systémy

Pomocou nášho nástroja pre porovnávanie a vyhodnocovanie stemerov sme ohodnotili dva slovenské stemery. V krátkosti si ich uvedieme.

Stemm-sk

Stemer pre Slovenský jazyk, napísaný v jazyku Python, ktorého autor je pán Marek Šuppa. Jedná sa o adaptáciu českého stemeru vytvoreného Luísom Gomesom, tiež v jazyku Python. Český stemer je portom algoritmu implementovaného v jazyku Java, ktorý je vytvorený Ljiljanom Dolamicom. Tento stemer funguje vo dvoch módoch nazvaných *„light“* a *„aggressive“*, pričom v móde *„aggressive“* má stemer lepšie výsledky. Viacej o tomto stemery sa môžete dozvedieť z github repozitára [36], z ktorého sme aj nástroj prevzali.

Stemmer-sk

Stemmer-sk je nástroj na stematizáciu slov, napísaný v jazyku Java. Jedná sa o časť zo skupiny nástrojov, ktoré boli vyhotovené v rámci diplomovej práce pána Filipa Bednárika [2], v ktorej sa môžete dozvedieť viac informácii ohľadom tohto nástroja. Nástroj bol prevzatý z github repozitára [3].

4.3.3 Hodnotenie a porovnávanie systémov

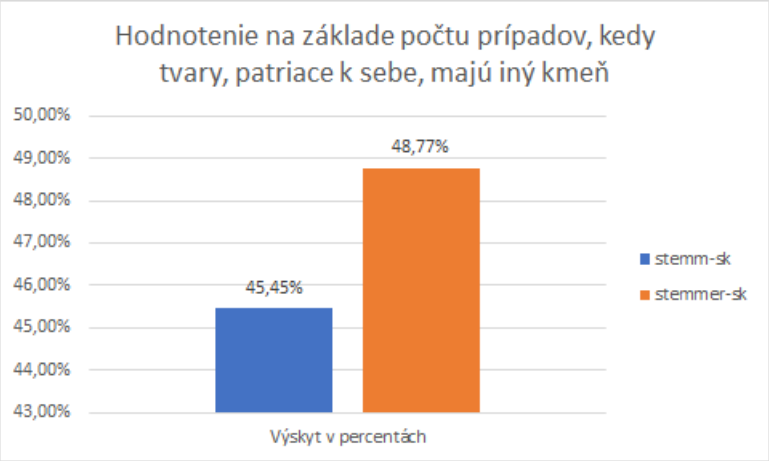
Nad oboma systémami sme spustili náš nástroj pre hodnotenia a porovnávania stemerov s nasledujúcimi výsledkami.

Prípady, kedy tvary, patriace k sebe, budú mať iný kmeň

V tejto hodnotiacej kategórii sa zameriavame na schopnosť stemerov, generovať ten istý kmeň pre všetky vstupné slová, ktoré podľa testovacieho slovníka zdieľajú ten istý kmeň.

Napríklad slová: *administrovaný, administrátorstvo, administrátorský, administrácia, administrovanie, administrátor, administrovať, administračný, administrátorka* zdieľajú ten istý kmeň *„administr“*. Stemer by mal pre tieto slová vygenerovať ten istý kmeň.

Chybné určenie kmeňa predstavujú prípady, kedy stemer slovu *„administrátorka“* určí kmeň *„administrátor“*, slovu *„administrovanie“* určí kmeň *„administrovani“* atď.



Obr. 4.3: *Graf zobrazujúci percená prípadov, kedy tvary, patriace k sebe, majú iný kmeň*

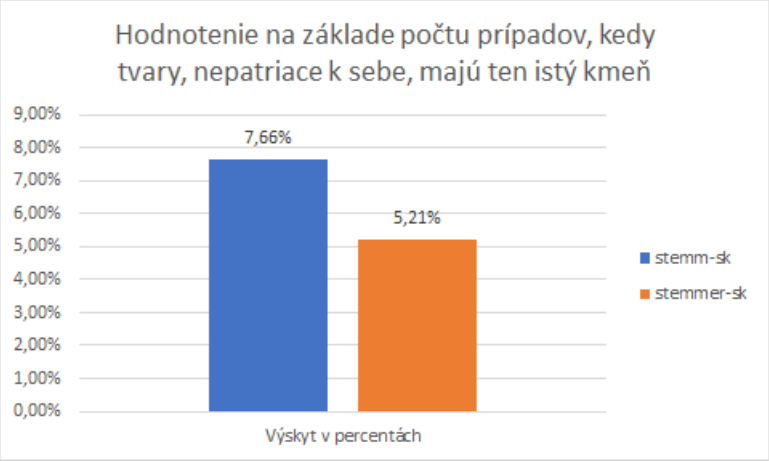
Ako môžeme vidieť na grafe 4.3, kde systém *stemm-sk*, ktorý je implementovaný v jazyku Python, generuje o 3,32% menej prípadov, kedy vstupné tvary nezdieľajú ten istý koreň v rámci vstupu z testovacieho slovníka, ako systém *stemmer-sk* implementovaný v jazyku Java.

Prípady, kedy tvary, nepatriace k sebe, budú mať ten istý kmeň

V tejto kategórii prechádzame výstup stemera a hľadáme prípady, kedy stemer určil ten istý kmeň pre slová, ktoré ho nezdieľajú podľa testovacieho slovníka.

Napríklad slová: *schváliť, schválne, schválený, schválenie, schváleno* zdieľajú kmeň *„schvál“* a slová: *schválny, schválnosť* zdieľajú kmeň *„schváln“*.

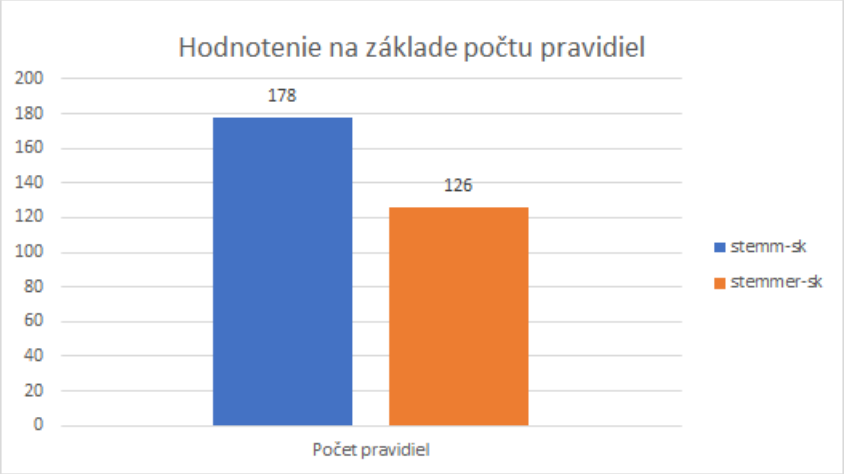
Chybné určenia kmeňa predstavujú prípady, kedy stemer slovu *„schválne“* určí kmeň *„schváln“*, slovu *„schválny“* určí kmeň *„schvál“* atď.



Obr. 4.4: *Graf zobrazujúci percená prípadov, kedy tvary, nepatriace k sebe, majú ten istý kmeň*

Výsledky tohto hodnotenia sú zobrazené na grafe 4.4, kde naopak systém *stemmer-sk* generuje o 2,45% menej prípadov, kedy tvary, ktoré k sebe nepatria, zdieľajú ten istý kmeň.

Hodnotenie na základe zistených pravidiel



Obr. 4.5: *Graf zobrazujúci hodnotenie na základe zisteného počtu pravidiel*

Toto hodnotenie reprezentuje graf 4.5, v ktorom vidíme, že systém *stemm-sk*, implementovaný v jazyku Python, obsahuje o 52 pravidiel viac ako systém *stemmer-sk*.

4.3.4 Vyhodnotenie

Na základe týchto troch hodnotení, systém *stemm-sk* zvíťazil v dvoch z troch hodnotení nad systémom *stemmer-sk*. Nutné je zmieniť, že systém *stemmer-sk* pracuje s počtom pravidiel o 52 menším ako systém *stemm-sk*, pritom dosahuje skoro podobné výsledky v prvej testovacej kategórii a v druhej kategórii je dokonca lepší.

4.3.5 Zhodnotenie problémov a ich riešení

Systém je založený na derivačnom slovníku, ktorého jednoduchšia verzia sa používa pri samotnom hodnotení stemerov. V prípade zlého vyhodnotenia stemu a jeho príslušnej skupiny lem, ktoré ho zdieľajú, vzniká chyba, ktorá ovplyvní výsledok hodnotenia. Výskyt týchto chýb sme sa snažili obmedziť.

Samotné hodnotenie jedného systému môže trvať až niekoľko desiatok minút, v rámci ktorých linuxový skript privádza na vstup jednotlivých systémov slová, ktoré zdieľajú ten istý koreň. Najdlhšie trvá vyhodnotenie stemeru *stemmer-sk*, ktorý je implementovaný v jazyku Java.

Tento problém by sa dal vyriešiť zmenšením a skvalitnením testovacieho slovníka, ktorý by obsahoval menší počet kmeňov slov, ale ku jednotlivým kmeňom by patrilo viac lem, ktoré ho zdieľajú, prípadne aj vyskloňované tvary lem. Iným riešením by bola zmena algoritmu, na základe ktorého komunikuje linuxový skript so stemermi.

Ako ďalší problém sa javí neizolovanosť *„automatu“*, na základe ktorého sa odtrhávajú jednotlivé prefixy zo slova. V prípade, že sa vyskytuje celý kód v jednom súbore, dochádza k chybnému určeniu počtu pravidiel systému, keďže systém považuje za pravidlo suffix ako reťazec v podmienkovom bloku. Tento problém sa môže riešiť komplexnejším vyhľadávaním pravidiel.

Súčasná verzia systému na vyhodnocovanie a porovnávanie stemerov podporuje priamo implementácie v jazykoch Python a Java. V prípade testovania stemerov napísaných v iných jazykoch je potrebné pridať podporu týchto jazykov, čo ale nie je zložité. Stačí len pridať spôsob, ako bude linuxový skript komunikovať s daným systémom.

Kapitola 5

Záver

Zadanie práce predstavuje vytvorenie systému technických vzorov pre flektívnu morfológiu slovenčiny, z neho odvodený morfologický analyzátor, nástroj pre porovnávanie a vyhodnocovanie stemerov, nástroj na zrekonštruovanie diakritiky.

Nami vytvorený morfologický analyzátor vychádza z morfologického slovníka, s ktorým neustále komunikuje. Pri zostavovaní slovníkov sme vytvorili niekoľko skriptov, ktoré slúžili na vytvorenie slovníka paradigiem (vzorov), slovníka lem a ich príslušných vzorov. Na základe týchto dvoch slovníkov sme vytvorili morfologický slovník. Morfologický analyzátor je implementovaný v jazyku Python3. Algoritmus analyzátora pozostáva zo zistenia vstupného jednoslovného pomenovania, jeho vyhľadania v slovníku. Ak sa vstupné jednoslovné pomenovanie nenachádza v slovníku, tak sa jednoslovné pomenovanie lemnatizuje pomocou štatistickej metódy na základe zoznamu pravdepodobných lem. Ďalej nasleduje fáza určenia slovotvorného základu a vzoru a následný výpis. Ako formát uloženia slovníka sme zvolili databázu SQLite3.

Morfologický analyzátor sme porovnali s existujúcim morfologickým analyzátorom Majka, ktorý tiež analyzuje vstupné slová, určí ich slovotvorný základ, morfologickú značku. Na základe iterácii sme rozširovali morfologický slovník analyzátora, pričom pri každej iterácii sme ohodnotili slovník pomocou testovacieho korpusu. Súčasne sme ohodnotili aj samotný slovník analyzátoru Majka a z výsledkov sme zistili, že po každej iterácii sa úspešnosť zvyšovala, pričom už po prvej iterácii bola úspešnosť oboch slovníkov približne rovnaká. Priemerná úspešnosť nášho analyzátora v testovacích sadách bola lepšia o 2,27% než u analyzátora Majka. Dôvodom je to, že Majka sa nesnaží lematizovať nové slová, resp. lematizuje len tie, ktoré sa nachádzajú v jej slovníku.

Na druhej strane Majka pracuje so slovníkom, ktorý zaberá neporovnateľne menej pamäťového priestoru, zatiaľ čo náš slovník zaberá priestoru dosť a aj samotné spracovanie slov je pomalšie ako u Majky. Je to preto, lebo Majka využíva slovník, ktorý využíva formát *fsa*. Tento formát je popísaný v sekcii 2.11.1, čo je v porovnaní s databázou vhodnejší formát uloženia dát slovníka.

V ďalšom vývoji analyzátora by sme určite zmenili formát uloženia dát v slovníku, čo by zvýšilo rýchlosť analyzovania slov a zmenšila by veľkosť miesta, ktoré slovník zaberá. Ďalej by sme použili anotovaný korpus, pomocou ktorého by sme vyčistili morfologický slovník od chybne určených tvarov, prípadne by sme ho rozšírili o nové tvary a ich lemy. Vhodné by bolo aj zavedenie vstupného slova do kontextu prípadnej vety, na základe ktorého by analýza slova prebehla presnejšie.

Ďalším nástrojom, vytvoreným v rámci bakalárskej práce, je nástroj na porovnávanie a vyhodnocovanie stemerov. Systém využíva testovací slovník pri hodnotení, ktorý pred-

stavuje jednoduchšiu verziu derivačného slovníka, ktorý sme vytvorili na základe morfologického slovníka. Systém je implementovaný v jazyku Python3, pričom pri komunikácii so stemermi využívame linuxový skript, ktorý privádza na vstup stemerov jednotlivé slová. Stemery hodnotíme na základe troch kritérií: počtu pravidiel, počtu prípadov kedy tvary, patriace k sebe, budú mať rôzny kmeň a počtu prípadov, kedy tvary, nepatriace k sebe, budú mať taký istý kmeň.

Potrebné informácie o stemeroch sú uložené v konfiguračnom súbore, na základe ktorého systém pracuje. Pre každý správne zadaný záznam, ktorý reprezentuje stemer, systém vykoná hodnotenie. Porovnávali sme dva slovenské stemery, *stemm-sk* implementovaný v jazyku Python a *stemmer-sk* implementovaný v jazyku Java. V hodnotení zvíťazil *stemm-sk*, ktorý pracuje s viacerými pravidlami, a generuje o 3,32% menej prípadov, kedy tvary, patriace k sebe, majú iný kmeň. Na druhej strane *stemmer-sk* generuje o 2,45% menej prípadov, kedy tvary, ktoré k sebe nepatria, zdieľajú ten istý kmeň.

V rámci ďalšieho vývoja systému je nutné priamo podporovať viacej implementačných jazykov než len Python a Javu, prípadne testovací slovník rozšíriť o viac tvarov. Tento systém je možné využiť aj v rámci iných jazykov ako len slovenčina, len musí byť dostupný testovací slovník v danom jazyku.

Posledným nástrojom, vytvoreným v rámci bakalárskej práce, je nástroj na rekonštrukciu diakritiky, ktorý používa slovník tvarov bez diakritiky a ich najčastejších prepisov z frekvenčného slovníka, pričom tento slovník je tiež odvodený od morfologického slovníka. Nástroj je implementovaný v jazyku Python3 a slovník je uložený v databáze SQLite3. Tento nástroj sme využili pri testovaní pokrytia slov morfologickým analyzátorom.

Nástroj sme neporovnávali s ostatnými obdobnými nástrojmi na rekonštrukciu diakritiky, pretože je jednoduchý, dotazuje sa slovníka a nevychádza z kontextu slov, ktorý je pri rekonštrukcii dôležitý. V rámci ďalšieho vývoja by bolo vhodné rozšíriť algoritmus zostavenia diakritiky o zistenie kontextu slova vo vete, na základe ktorého sa zostaví diakritika slova.

Literatúra

[1] Babič, F.,Bednár, P., Butka, P., Furdík, K., Paralič, J., Sarnovský, M., Tutoky, G.: *Dolovanie znalostí z textov*. Equilibria, 2010, ISBN 978-80-89284-62-7.

[2] Bednárik, F.: *Extrakcia informácií z textu*. Diplomová práce, Slovenská Technická Univerzita v Bratislave, Fakulta informatiky a informačných technológií, Bratislava, 5 2016, vedúci práce Ing. Marián Šimko, PhD.

[3] Bednárik, F.: Stemmer so slovenskou podporou pre Lucene/SOLR. 2016, [Online; navštívené 5.04.2017].
URL https://github.com/essential-data/stemmer-sk

[4] Benko, V.: Aranea: Yet Another Family of (Comparable) Web Corpora. In *TSD*, 2014, s. 247–254, doi:http://dx.doi.org/10.1007/978-3-319-10816-2_31.

[5] Blanchard, D., Hickford, M., Korobov, M., Moiseenko, A., Wilk, J., Yamada, I.: *DAWG documentation*. 2015, [Online; navštívené 21.04.2017].
URL http://dawg.readthedocs.io/en/latest/

[6] Collins, M.: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, s. 1–8.

[7] Cvrček, V.: *Korpus*. 2014, [Online; navštívené 23.04.2017].
URL http://wiki.korpus.cz/doku.php/pojmy:korpus

[8] Dvonč, L. a Ružička, J. a Ústav slovenského jazyka (Slovenská akadémia vied): *Morfológia slovenského jazyka*. Bratislava: Vydavateľstvo Slovenskej akadémie vied, 1966, ISBN 71-024-66.

[9] Emiš, P.: *Parafrázovač slovenčiny*. Bratislava: Vydavateľstvo Slovenskej akadémie vied, 1994, ISBN 80-224-0109-9.

[10] Horecký, J.: *Onomaziologická štruktúra slovenčiny*. Spisy Slovenskej jazykovednej spoločnosti pri SAV. Suplement, Slovenská jazykovedná spoločnost' pri SAV, 2003, ISBN 9788089037100.

[11] Jakubíček, M.; Kilgarriff, A.; Kovář, V.; aj.: The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*, Lancaster, 2013, s. 125–127, [Online; navštívené 2.04.2017].
URL http://ucrel.lancs.ac.uk/cl2013/

[12] Kazennikov, A.: Deterministic Acyclic Finite State Automaton implementation for morphological analysis. 2017, [Online; navštívené 15.04.2017].
URL https://github.com/kzn/fsa

[13] Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, AAMT, AAMT, 2005, s. 79–86, [Online; navštívené 1.04.2017].
URL http://mt-archive.info/MTS-2005-Koehn.pdf

[14] Křen, M.: *Morfologické značky (tagy)*. [Online; navštívené 13.04.2017].
URL https://wiki.korpus.cz/doku.php/seznamy:tagy

[15] Lingea s.r.o.: *Gramatika súčasnej slovenčiny*. Lingea s.r.o., 2013, ISBN 9788081450686.

[16] Mikhail Korobov, S. L.: *Pytries Data Structures for Python*. 2017, [Online; navštívené 20.04.2017].
URL https://github.com/pytries

[17] *Morfologie*. [Online; navštívené 13.04.2017].
URL https://nlp.fi.muni.cz/cs/Morfologie

[18] Porter, M. F.: Snowball: A language for stemming algorithms. October 2001, [Online; navštívené 14.04.2017].
URL http://snowball.tartarus.org/texts/introduction.html

[19] Páleš, E.: *Sofiológia ako príklad integrálnej vedy a vzdelávania v tradícii slovanov*. [Online; navštívene 3.04.2017].
URL http://www.sophia.sk/sites/default/files/Sofiologia_ako_priklad_integralnej_vedy.pdf

[20] Raab, J.: *Morče - Český morfologický značkovač*. 2017, [Online; navštívené 28.04.2017].
URL http://ufal.mff.cuni.cz/morce/index.php

[21] Ripka, I.; Imrichová, M.: *Základy slovenskej lexikológie*. Vysokoškolské učebné texty, Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied, 2003, ISBN 9788080682071.

[22] Rowling, J.: *Harry Potter and the Deathly Hallows*. Harry Potter, Pottermore, 2015, ISBN 9781781102435.

[23] Sedláček, R.: *Ajka tagset*. 2006, [Online; navštívené 2.03.2017].
URL https://nlp.fi.muni.cz/projekty/ajka/tags.pdf

[24] Šikra, J.: *Sémantika slovenských prísloviek*. Jazykovedné štúdie, Veda - vydavatel'stvo Slovenskej akadémie vied, 1991, ISBN 9788022403221.

[25] *SkTenTen – webový korpus slovenčiny*. 2011, [Online; navštívené 20.12.2016].
URL https://the.sketchengine.co.uk

[26] *Slovensko-český paralelný korpus*. 2016, [Online; navštívené 24.04.2017].
URL http://korpus.juls.savba.sk/skcs.html

[27] Sokolová, M.: *Sémantika slovesa a slovesný rod*. VEDA, Vyd. Slovenskej Akadémie Vied, 1993, ISBN 9788022403436.

[28] Sokolová, M.: *Nový deklinačný systém slovenských substantív*. Filozofická fakulta Prešovskej univerzity v Prešove, 2007, ISBN 80-8068-550-9.

[29] Straková, J.; Straka, M.; Hajič, J.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland: Association for Computational Linguistics, June 2014, s. 13–18, [Online; navštívené 1.04.2017].
URL http://www.aclweb.org/anthology/P/P14/P14-5003.pdf

[30] Stupňan, I.: *Žitný: Páchatelia únosu Kováča ml. do väzenia nepôjdu*. Pravda, [Online; navštívene 3.04.2017].
URL https://spravy.pravda.sk/domace/clanok/424934-zitny-pachatelia-unosu-kovaca-ml-do-vazenia-nepojdu/

[31] Weisheitelová, J.: *K některým problémům automatické morfologické analýzy a lemmatizace*. [Online; navštívené 14.04.2017].
URL http://sas.ujc.cas.cz/archiv.php?art=2413

[32] Wikipedia: Trie — Wikipedia, The Free Encyclopedia. 2017, [Online; navštívené 21.04.2017].
URL http://en.wikipedia.org/w/index.php?title=Trie&oldid=773327685

[33] Yata, S.: *MARISA: Matching Algorithm with Recursively Implemented StorAge*. 2017, [Online; navštívené 20.04.2017].
URL https://github.com/s-yata/marisa-trie

[34] Yoshinaga, N.: *Cedar - C++ implementation of efficiently-updatable double-array trie*. 2014, [Online; navštívené 20.04.2017].
URL http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/cedar/

[35] Šmerk, P.; Rychlý, P.: Majka – rychlý morfologický analyzátor. Technická zpráva, Masarykova univerzita, 2009, [Online; navštívené 20.03.2017].
URL http://nlp.fi.muni.cz/ma/

[36] Šuppa, M.: A stemmer for Slovak language. 2015, [Online; navštívené 5.04.2017].
URL https://github.com/mrshu/stemm-sk

Príloha A

Obsah priloženého pamäťového média

Priložené pamäťové médium obsahuje:

- adresár *technical_patterns* – obsahujúci skripty, použité pri vytváraní systému technických vzorov
- adresár *morphological_analyzer* – obsahujúci zdrojové súbory morfologického analyzátora
- adresár *stem_system* – obsahujúci zdrojové súbory systému pre hodnotenie a porovnávanie stemerov
- adresár *sk_accent* – obsahujúci zdrojové súbory nástroja na rekonštrukciu diakritiky
- adresár *thesis*
  - adresár src – obsahujúci zdrojové súbory pre technickú správu
  - súbor xkloco00-wis.pdf – technická správa vo formáte PDF, ktorá bola odovzdaná do informačného systému
  - súbor xkloco00-tisk.pdf – technická správa vo formáte PDF, ktorá bola vytlačená a zviazaná
- súbor *poster.pdf* – plagát reprezentujúci prácu
- súbor *readme* – obsah pamäťového média

Príloha B

Manuál

B.1 Morfologický analyzátor

Nástroj sa nachádza v priečinku *morphological_analyzer*. Adresár obsahuje:

- adresár *slovak_kx* – obsahujúci slovníky systému technických vzorov
- súbor *analyzer.py* – spúšťací skript
- súbor *slovak.db* – morfologický slovník
- súbor *statistic.py* – modul, ktorý obsahuje algoritmus lematizácie
- súbory *loadings.py* – modul, ktorý obsahuje algoritmy pre načítanie súborov do rôznych formátov
- súbor *UnknownLemantize.py* – modul, ktorý obsahuje triedu, na základe ktorej prebieha lematizácia neznámeho slova
- súbor *tags.pdf* – tagset, ktorý využívame v slovníku
- súbor *manual.txt* – manuál k nástroju

Morfologický analyzátor komunikuje so slovníkom *slovak.db*. V adresári *slovak_kx* sa nachádzajú aj výsledné slovníky systému technických vzorov. Jednotlivé súbory:

- súbor *slovak.paradigms* – slovník paradigiem (vzorov) a pravidiel ich ohýbania
- súbor *slovak.lpn* – slovník lem a ich vzorov
- súbor *slovak.generative* – zoznam generatív
- súbory *.prip .pred* – zoznamy afixov, ku ktorým sú priradené jednotlivé vzory

Analyzátor k svojmu bezchybnému behu potrebuje všetky vyššie zmienené súbory, t. j. moduly, morfologický slovník a všetky súbory v adresári *slovak_kx*.

Ovládanie nástroja

Analyzátor číta zo štandardného vstupu slová, ktoré analyzuje. Spúšťa sa pomocou príkazu v terminály: `python3 analyzer.py` s voliteľnými argumentami:

-h – výpis nápovedy k nástroju

-add – prepínač na uloženie vyskloňovaného tvaru lemy do slovníka

-d – prepínač zapne výpis vyskloňovania lemy vstupného slova

-m – prepínač obmedzí výpis vyskloňovania lemy na 20 riadkov

-abr – kombinácia prepínačov -add a -abr pridá skratky zo stdin do slovníka

-gen – kombinácia prepínačov -add a -gen pridá generatíva zo stdin do systému

Po spustení nástroja musíme počkať, dokým sa nezobrazí hlavička *„Morphological analyzer“* a až potom môžeme analyzátoru privádzať slová na jeho vstup. Analyzátor ukončíme pomocou zadania znaku '**.**' alebo klávesovej skratky *CTRL + D*. Príklad výpisu nástroja po spustení s argumentom *-d*:

```
chlapcom
Tvar> chlapcom
Lema: chlapec
chlapec:k1gMnSc1:zamestnanec
chlapca:k1gMnSc2:zamestnanec
chlapcovi:k1gMnSc3:zamestnanec
chlapca:k1gMnSc4:zamestnanec
chlapcovi:k1gMnSc6:zamestnanec
chlapcom:k1gMnSc7:zamestnanec
chlapci:k1gMnPc1:zamestnanec
chlapcov:k1gMnPc2:zamestnanec
chlapcom:k1gMnPc3:zamestnanec
chlapcov:k1gMnPc4:zamestnanec
chlapcoch:k1gMnPc6:zamestnanec
chlapcami:k1gMnPc7:zamestnanec
```

B.2 Nástroj pre porovnávanie a vyhodnocovanie stemerov

Nástroj sa nachádza v priečinku *stem_system*. Adresár obsahuje:

- adresár *scripts* – obsahujúci zdrojové súbory (skripty), použité pri vytváraní tohto systému
- adresár *stemmer-sk* – obsahujúci implementáciu stemeru v jazyku Java, ktorú sme použili v hodnotení, tento stemer je dielom pána Filipa Bednárika
- adresár *stemm-sk* – obsahujúci implementáciu stemeru v jazyku Python, ktorú sme použili v hodnotení, tento stemer je dielom pána Mareka Šuppu
- súbor *stemming.conf* – konfiguračný súbor
- súbor *morphTestGood.deriv* – testovací slovník

- súbor *stemTesting.py* – spúšťací skript
- súbor *test.sh* – linuxový skript
- súbor *morphPatterns.deriv* – derivačný slovník
- súbor *manual.txt* – manuál k nástroju

Na komunikáciu so stemerom *stemmer-sk* bola vytvorená a skompilovaná trieda *Main.class*, ktorá mu predáva na štandardný vstup jednotlivé slová, pričom ju využíva pri hodnotení linuxový skript *test.sh*.

Nástroj na základe konfiguračného súboru *stemming.conf*, slovníka *morphTestGood.deriv* vyhodnocuje stemery. Stemerom privádza testované slová na vstup linuxový skript *test.sh*.

Konfiguračný súbor

Každý jednotlivý riadok konfiguračného súboru musí obsahovať päť položiek oddelených bodkočiarkou:

Language – implementačný jazyk stemera. Napríklad python3, java atď.

Directory – adresár, ktorý obsahuje potrebné súbory, jeho názov by mal charakterizovať názov stemeru, musí sa vyskytovať v rámci adresárovej štruktúry nástroja.

File – spúšťací súbor.

Arguments – argumenty príkazového riadku, ktoré sú potrebné na spustenie systému.

Rules – modul, ktorý obsahuje pravidlá, na základe ktorých systém vykonáva proces stematizácie.

```
Language=python3;Directory=stemm-sk;File=stemmsk.py;Arguments=aggressive;Rules=stemmsk.py
```

Obr. B.1: *Príklad riadku konfiguračného súbora*

Príklad formátu riadku konfiguračného súboru môžme vidieť na obrázku B.1, na ktorom vidíme potrebné položky: *Language, Directory, File, Arguments, Rules*, ktoré sú oddelené bodkočiarkou. Z príkladu vidíme, že testovací stemer je napísaný v jazyku Python3, názov spúšťacieho súboru je *„stemmsk.py“*, ktorý je spúšťaný s jedným argumentom *„aggressive“*. Pravidlový systém sa nachádza v tom istom súbore, t. j. *„stemmsk.py“*. Stemer sa nachádza v adresári *„stemm-sk“*.

Ovládanie nástroja

Nástroj sa spúšťa pomocou príkazu: `python3 stemTesting.py stemming.conf` v terminály. Jediným povinným argumentom nástroja je jeho konfiguračný súbor (*stemming.conf*). Pri spustení nástroja s argumentom *-h* sa zobrazí nápoveda.

Príklad výpisu, kedy konfiguračný súbor obsahuje riadok, ktorý je zobrazený na obrázku B.1, vyzerá nasledovne:

```
Testing stemm-sk
Prípady, kedy tvary patriace k sebe, majú
Ten istý kmeň 67312
Iný kmeň 56073
Všetky prípady 123385
Tvary nepatriace k sebe, majú spoločný kmeň v: 9452 prípadoch
Stemer pracuje s: 178 zistenými pravidlami
```

B.3 Nástroj na rekonštrukciu diakritiky

Nástroj sa nachádza v adresári *sk_accent*. Adresár obsahuje:

- adresár *scripts* – obsahujúci zdrojové súbory (skripty), použité pri vytváraní tohto systému
- súbor *sk_accent.py* – spúšťací skript
- súbor *diacritic.db* – slovník
- súbor *manual.txt* – manuál k nástroju

Nástroj číta zo štandardného vstupu slová, u ktorých rekonštruuje diakritiku na základe slovníka *diacritic.db*.

Ovládanie nástroja

Nástroj sa spúšťa pomocou príkazu v terminály: `python3 sk_accent.py`. Nástroj ukončíme pomocou zadania znaku '**.**' alebo klávesovej skratky *CTRL + D*. Pri spustení nástroja s argumentom *-h* sa zobrazí nápoveda.
Príklad výstupu na základe príkazu:

```
echo najvyssí clovek moze | python3 sk_accent.py
najvyšší človek môže
```