J. Yaghob (Ed.): ITAT 2015 pp. 81–87
Charles University in Prague, Prague, 2015



# Spracovanie prirodzeného jazyka pre interaktívne rečové rozhrania v slovenčine

Ján Staš, Daniel Hládek, Stanislav Ondáš, Daniel Zlacký a Jozef Juhár

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Park Komenského 13, 042 10 Košice, Slovenská republika
{jan.stas, daniel.hladek, stanislav.ondas, daniel.zlacky, jozef.juhar}@tuke.sk

*Abstrakt:* V príspevku sú zhrnuté priebežné výsledky aplikovaného výskumu v oblasti spracovania prirodzeného jazyka v úlohách orientovaných na výskum a vývoj modulov rečových rozhraní medzi človekom a strojom, ktorý prebieha v Laboratóriu rečových a mobilných technológií na KEMT FEI TU v Košiciach. Zahrnutie hovorenej reči, ako najprirodzenejšieho komunikačného nástroja medzi ľuďmi, má svoje nezastupiteľné miesto aj pri návrhu a vývoji interaktívnych rečových rozhraní. Pri prechode od rozpoznávania ľudskej reči k jej porozumeniu strojom je potom nevyhnutné vykonať aj dodatočnú analýzu textu po automatickom prepise. To zahŕňa aj proces transformácie textu po rozpoznaní na reprezentáciu určitého typu znalostí, ktorému dokáže stroj porozumieť. Tento zložitý proces všeobecne pozostáva z tokenizácie, automatickej korekcie a dodatočnej morfologickej, syntaktickej a sémantickej analýzy textu. Nami navrhnuté moduly a výsledky automatického spracovania textu v slovenskom jazyku budú postupne predstavené v tomto príspevku.

## 1 Úvod

S príchodom výpočtovej techniky sa stala potreba počítačového spracovania prirodzeného jazyka aktuálnou celosvetovou témou. Vedci sa po celom svete snažia podchytiť charakter takmer každého jazyka s cieľom zjednodušiť interakciu medzi ľuďmi a strojmi a komunikáciu medzi ľudmi samotnými. Oblasť spracovania prirodzeného jazyka zahŕňa širokú škálu disciplín, ako napr. vyhľadávanie informácií, štatistické modelovanie jazyka, strojový preklad, automatické rozpoznávanie a porozumenie reči a pod. Jednotlivé disciplíny však vo väčšine prípadov úzko súvisia, dopĺňajú sa, a pomocou nich je možné ľuďom uľahčiť prácu, štúdium, komunikáciu, či zábavu. Jednou z najaktuálnejších úloh v oblasti spracovania prirodzeného jazyka je aj automatické rozpoznávanie reči (ARR), ktorému sa v našom laboratóriu intenzívne venujeme. Vďaka viacerým zlepšeniam v oblasti automatického rozpoznávania reči v slovenčine sme schopní rozpoznať ľudskú reč s dostatočnou presnosťou v mnohých aplikačných úlohách, avšak komplexné porozumenie významu je v súčasnosti jednou z najnáročnejších úloh pri návrhu rôznych interaktívnych rečových rozhraní a to nielen v slovenčine. V tomto článku budú predstavené nami navrhnuté prístupy na spracovanie prirodzeného jazyka pre interaktívne rečové rozhrania v slovenčine.

## 2 Zdrojové dáta

### 2.1 Dolovanie textu

Rozsiahly korpus písaných textov použitý vo viacerých oblastiach spracovania prirodzeného jazyka v Laboratóriu rečových a mobilných technológií bol vytvorený pomocou nami navrhnutého systému *na dolovanie a spracovanie textových dokumentov z webových stránok* dostupných na sieti Internet s názvom webAgent [1, 2]. Systém doluje textové dáta z rôznych domén a elektronických zdrojov písaných v slovenčine a pomocou preddefinovaných pravidiel na prepis čísloviek, symbolov a skratiek ich spracúva do ich vyslovovanej podoby. Systém je navyše rozšírený o metódy tokenizácie, segmentácie na vety, metódy na kontrolu duplicity na úrovni adresy zdroja textu a obsahu dokumentu, a tiež o metódy filtrácie viet obsahujúcich veľké množstvo gramaticky nespisovných slov, číslic, akronymov, symbolov, skratiek, a iných cudzojazyčných a mimoslovníkových slov. Spracovaný text je následne rozdelený do menších celkov, t.j. podkorpusov, pomocou účinných metód na kategorizáciu textových dokumentov. Súčasný korpus písaných textov v slovenčine obsahuje približne 2,25 mld. tokenov obsiahnutých v 125 mil. vetách.

### 2.2 Kategorizácia textu

S narastajúcim množstvom textových dokumentov stiahnutých zo siete Internet a potrebou vytvárať čoraz presnejšie doménovo orientované interaktívne rečovo-založené systémy a rozhrania, sa vynorila otázka kategorizovať textové dáta nielen podľa adresy (URL) zdroja textu, odkiaľ daný textový dokument pochádza, ale aj na úrovni jeho obsahu. Navyše webová adresa zdroja textu nemusí byť hneď jednoznačným identifikátorom obsahu dokumentu, vychádzajúc tiež z predpokladu, že jeden dokument môže pojednávať o viacerých témach. Kategorizácia textu má preto veľký význam pri návrhu a tvorbe robustných doménovo orientovaných systémov na automatické rozpoznávanie reči, ale aj v iných úlohách využívajúcich textové dáta ako zdroj informácií, napr. pri návrhu a vývoji interaktívnych rečovo-založených rozhraní.

Narozdiel od *metód zhlukovania*, kde dokumenty s využitím štatistických prístupov spájame do určitého počtu zhlukov, v ktorých tému vopred nepoznáme, pri *kategorizácii dokumentov* sa snažíme zadeliť dokumenty do dvoch

alebo viacerých tried na základe ich minimálnej vzdialenosti, resp. sémantickej podobnosti, udávajúcej prienik slov alebo celých fráz medzi dokumentami. V oboch prípadoch je nutné identifikovať tému v danom zhluku, resp. triede, a to buď pomocou prístupov založených na extrakcii kľúčových slov, pravdepodobnostných prístupov založených na výpočte podobnosti dokumentov pomocou dištančných metrík, alebo ich kombináciou.

Počiatočný výskum v oblasti kategorizácie textových dokumentov bol venovaný metódam zhlukovania pomocou iteračných algoritmov založených na *k-means* a *k-medoid zhlukovaní* a na *hierarchickom zhlukovaní* využívajúcom *aglomeračné* a *divizívne kritérium* [3]. Ako najvhodnejším prístupom sa ukázalo hierarchické zhlukovanie textu s využitím aglomeračného kritéria pri zhlukovaní článkov zo slovenskej Wikipédie. Tento spôsob zhlukovania dokumentov sme porovnali s nami navrhnutou metódou založenou na *klasifikácii textových dokumentov pomocou kľúčových fráz* využívajúcou F-skóre ako hodnotiace kritérium [4]. Nami navrhnutý spôsob klasifikácie sa výsledkami ukázal porovnateľný k hierarchickému zhlukovaniu, avšak hlavnou nevýhodou navrhnutej metódy je nutnosť mať k dispozícii zoznamy kľúčových slov, resp. fráz pre jednotlivé domény a v procese klasifikácie textu je nutné správne (v ideálnom prípade automaticky) nastaviť vhodný prah delenia. Tento typ klasifikácie sme v nasledujúcom výskume zameranom na kategorizáciu textu v úlohe robustného doménovo orientovaného modelovania jazyka rozšírili o ďalšie tri metriky určujúce vzdialenosť, resp. podobnosť medzi dokumentami, konkrétne o Bhattacharyyaov koeficient, Jaccardov index a Jensenovu-Shannonovu divergenciu. Ako najvhodnejšou mierou v úlohe klasifikácie textu sa javí použitie Jaccardovho indexu pri výpočte podobnosti dokumentov [5].

Iným spôsobom je použitie metód nekontrolovaného učenia v úlohe kategorizácie textových dokumentov. Vo viacerých súčasných výskumoch zaoberajúcich sa modelovaním jazyka sme pri kategorizácii dokumentov siahli po *latentnej Dirichletovej alokácii* (z angl. „*latent Dirichlet allocation*“, skr. LDA). LDA je charakterizovaná ako generatívny pravdepodobnostný model, ktorý vychádza z multinomického a Dirichletovho rozdelenia pravdepodobnosti [6]. Zavedenie LDA v úlohe modelovania slovenského jazyka pri automatickom prepise diktovaných súdnych rozhodnutí prinieslo tiež výrazné zníženie perplexity modelov a miery chybovosti systému ARR [7].

## 3 Identifikácia tokenov a vetných hraníc

Prvým krokom v spracovaní textu je jeho príprava a prepis číslic, symbolov a skratiek do ich vyslovovanej podoby. Úlohou je pomocou sústavy pravidiel identifikovať také textové jednotky, ktoré sú zaujímavé z hľadiska ďalšieho spracovania, t.j. úprava na jednotný spôsob zápisu a eliminácia nepodstatných častí. Predspracovanie je tak nevyhnutným krokom pre akékoľvek ďalšie štatistické spracovanie, zvlášť v prípade textov stiahnutých z Internetu.

Najdôležitejšou časťou predspracovania textu je jeho *tokenizácia*. Jej cieľom je identifikácia jednotlivých slov a vetných hraníc, ktoré môžu slúžiť ako vstup do ďalšieho spracovania. V tomto kroku sa tiež snažíme zjednotiť spôsob zápisu číslic, diakritiky, interpunkcie, akronymov, symbolov, skratiek a iných významových jednotiek.

Tokenizácia sa zvyčajne vykonáva postupnou aplikáciou vhodne zapísaných regulárnych výrazov, ktoré obsahujú pravidlá pre identifikáciu textových jednotiek, dôležitých pre ďalšie spracovanie. Nepodstatné časti textu, ktoré nie sú pokryté pravidlami, sú z textu vynechané. Nami navrhnutý tokenizátor identifikuje tieto časti textu: diaktritika, slová, akronymy, symboly, skratky, zoznamy, odseky, čísla, e-mailové adresy a adresy URL. Identifkácia vetných hraníc je ďalej vykonávaná pomocou rozlíšenia významu bodky, jej desambiguáciou. V slovenských textoch môže byť bodka súčasťou označenia číselného poradia, skratky alebo e-mailovej alebo webovej adresy.

Na začiatku procesu identifikácie významových častí je vstupný reťazec porovnaný so všetkými pravidlami v databáze. Pravidlo, ktoré vyhovuje najdlhšiemu textu, je vybraté a jeho zodpovedajúci text je prepísaný podľa požiadaviek. Tento text je potom odstránený zo vstupného reťazca. Ak nevyhovuje žiadne pravidlo, vstupný reťazec je skrátený o jeden znak a prehľadávanie bázy pravidiel pokračuje. Výsledkom tokenizácie je text, kde sú textové jednotky oddelené medzerou a vety novým riadkom.

Proces identifikácie tokenov je zvyčajne výpočtovo náročný. Pre urýchlenie sme všetky pravidlá zapísali v špeciálnom jazyku Ragel [8] a spojili do jediného stavového automatu v programovacom jazyku C, z ktorého je zvyčajným spôsobom vytvorený spustiteľný súbor resp. knižnica [9]. Podrobnejšie informácie možno nájsť v [1].

## 4 Anotácia textu

Tam kde to je možné, využívame pre anotáciu textu prístupy založené na štatistickom modelovaní. V trénovacej databáze sú zvyčajne tokenom priradené určité triedy alebo morfologické značky. Štatistický klasifikátor analyzuje trénovací korpus a je schopný priradiť najpravdepodobnejšiu značku aj takým kontextom, ktoré sa v trénovacej databáze nevyskytujú. Slovenčina sa vyznačuje relatívne voľným poradím slov vo vetách, vysokým počtom morfologických tvarov slov a gramatických výnimiek. Počet možných kontextov tak môže byť veľmi vysoký, a to sťažuje úlohu natrénovania čo možno najpresnejšieho štatistického klasifikátora.

### 4.1 Rozpoznávanie pomenovaných entít

Z dôvodu nedostatku trénovacích dát pre rozpoznávanie pomenovaných entít v súčasnosti využívame systém založený na pravidlách. Systém využíva sadu slovníkov, regulárnych výrazov a viacslovných pomenovaní, ktoré sú spojené do unifikovaného systému na automatický prepis

| korpus | anotácia | TreeTagger | Dagger |
|---|---|---|---|
| Národný korpus jazyka poľského | manuálna | 15,63 | 11,83 |
| Český akademický korpus (v2.0) | manuálna | 13,10 | 9,46 |
| Slovenská časť korpusu W2C | automatická | 10,30 | 4,47 |
| Maďarský webový korpus | automatická | 2,55 | 1,97 |

Tabuľka 1: Miera chybnej klasifikácie [v %] morfologických analyzátorov TreeTagger a Dagger

pomenovaných entít. Tento na pravidlách založený systém pracuje podobne ako tokenizátor, pomocou stavového automatu. Rozpoznané pomenované entity je možné využiť v rôznych úlohách spracovania prirodzeného jazyka v slovenčine. Vzhľadom na to, že autorom systému na rozpoznávanie pomenovaných entít nie je v súčasnosti známy žiaden iný porovnateľný nástroj vytvorený pre slovenčinu, ani databáza vhodná na testovanie, nie je preto možné tento nástroj správne ohodnotiť a vyčísliť jeho úspešnosť.

### 4.2 Morfologická analýza textu

Morfologické značky sú jedným z najdôležitejších príznakov v spracovaní prirodzeného jazyka. Z toho dôvodu sme *morfologický klasifikátor* Dagger [10] navrhli tak, aby bral do úvahy špecifické vlastnosti flektívnych jazykov. Klasifikátor je založený na skrytom Markovovom modeli (z angl. „*hidden Markov model*“, skr. HMM) druhého rádu a najpravdepodobnejšia postuponosť morfologických značiek je vyhľadávaná Viterbiho algoritmom.

Nami navrhnutý HMM klasifikátor Dagger pre morfologickú analýzu flektívnych jazykov sa skladá z nasledujúcich štyroch častí:

1. lexikón, ktorý navrhuje množinu možných značiek na základe slova alebo jeho koncovky;
2. model prechodov, ktorý vyjadruje pravdepodobnosť nasledujúcej značky na základe dvoch predchádzajúcich,
3. model pozorovaní, ktorý vyjadruje pravdepodobnosť slova na základe možnej značky;
4. a v prípade, že skúmané slovo sa nenachádza v trénovacej databáze, využije sa dodatočný model pozorovaní, ktorý vyjadruje pravdepodobnosť stavu na základe koncovky daného slova.

Je vhodné poznamenať, že algoritmus obsiahnutý v morfologickom analyzátore Dagger využíva vlastný algoritmus na automatickú identifikáciu koncoviek slov založený na minimálnej opisnej dĺžke.

Na natrénovanie klasifikátora pre slovenský jazyk sme využili početnosti trigramov slov z *ručne morgologicky anotovaného korpusu* r-mak-2.0 [1] a množinu *morfologických značiek* [2], získaných zo Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra na Slovenskej akadémii vied v Bratislave. Algoritmus sme uplatnili najmä pri morfologickej analýze korpusu písaných textov v slovenčine a pri trénovaní štatistických modelov jazyka založených na triedach slov v systémoch na automatické rozpoznávanie plynulej reči v slovenčine [11, 12].

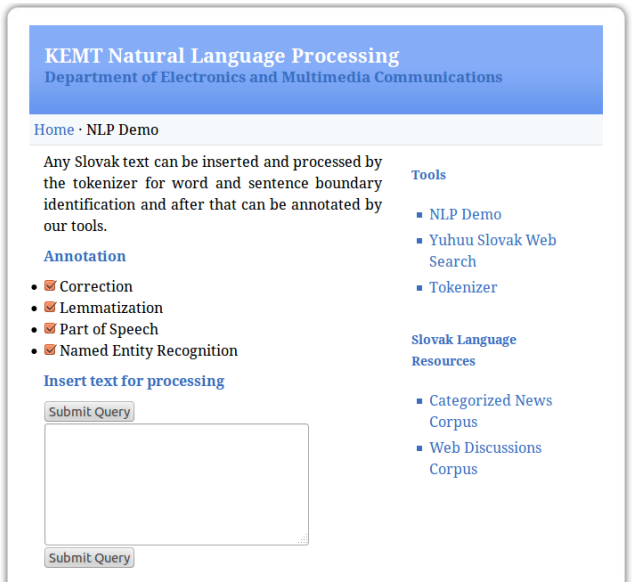
Presnosť klasifikácie nami navrhnutého morfologického analyzátora Dagger sme porovnali s dobre známym slovnodruhovým (z angl. „*part-of-speech*“, skr. PoS) značkovačom TreeTagger [13], ktorého algoritmus je založený na rozhodovacích stromoch. Pre porovnanie presnosti klasifikácie sme morfologické analyzátory vyhodnotili na štyroch rôznych manuálne resp. automaticky anotovaných textových korpusoch, a to na Národnom korpuse jazyka poľského [14], Českom akademickom korpuse, vo verzii 2.0 [15], Maďarskom webovom korpuse [16], a na slovenskej časti textového korpusu Web to Corpora [17][3]. V Tab. 1 sú znázornené výsledky miery chybnej klasifikácie, ktoré ukazujú, že nami navrhnutý algoritmus morfologickej analýzy dosahuje porovnateľnú presnosť s klasifikáciou obsiahnutou v nástroji TreeTagger.

### 4.3 Doplňovanie diakritiky

Častým javom pri komunikácii medzi ľudmi prebiehajúcej na sieti Internet je vysoký výskyt preklepov a chýbajúca diakritika. Hoci človeku to väčšinou pri porozumení správy nerobí problém, pri počítačovom spracovaní prirodzeného jazyka je potrebné nájsť vhodný spôsob pre rozlíšenie významu nejednoznačných zápisov na základe okolitého kontextu. Z toho dôvodu sme sa venovali aj problému automatického doplňovania diakritiky slov [18].

Podobne ako pri návrhu morfologického analyzátora sme pri rekonštrukcii diakritiky využili algoritmus využívajúci skrytý Markovov model. V tomto prípade je však matica prechodov tvorená trigramovým jazykovým modelom a matica pozorovaní je trénovaná pomocou algoritmu pre generovanie nesprávnych zápisov na texte, ktorý je pokladaný za správny. Úspešnosť navrhnutého systému na automatické doplňovanie diakritiky v korpusoch textov z blogov písaných v slovenskom jazyku dosahuje úroveň až 85%. Podobný nástroj na rekonštrukciu diakritiky pre slovenčinu, využívajúci štatistické modely jazyka vysokého rádu, bol vytvorený tiež tímom pracovníkov v Slovenskom národnom korpuse Jazykovedného ústavu Ľudovíta Štúra na Slovenskej akadémii vied v Bratislave[4].

[1] http://korpus.sk/ver_r(2d)mak.html
[2] http://korpus.juls.savba.sk/attachments/morpho/tagset-www.pdf
[3] https://lindat.mff.cuni.cz/repository/xmlui/
[4] http://korpus.juls.savba.sk/diakritik.html




Obr. 1: Webové rozhranie k systémom na spracovanie prirodzeného jazyka na KEMT FEI TU v Košiciach

### 4.4 On-line webové rozhranie

Pre účely demonštrácie a testovania presnosti tokenizácie, morfologickej analýzy a automatického doplňovania diakritiky slov sme k nami navrhnutým nástrojom na počítačové spracovanie prirodzeného jazyka v slovenčine vytvorili aj jednoduché on-line webové rozhranie[5], ktoré je znázornené na Obr. 1 a opísané v článku [11].

## 5 Aktivity v oblasti aplikovaného výskumu

### 5.1 Štatistické modelovanie jazyka

Konštrukcia štatistického modelu jazyka pre slovenčinu, ktorá patrí do skupiny vysoko flektívnych jazykov, je oveľa obtiažnejšia, než vytvorenie štatistického modelu pre jazyk anglický. Prvým dôvodom je *neusporiadanosť* slovenského jazyka, čo vedie k voľnejším pravidlám reťazenia slov do viet. Druhým je samotná *flektívnosť* jazyka, ktorá vytvára predpoklad pre mnohonásobne väčší slovník, než je to v prípade jazyka anglického.

Súčasný stav v oblasti štatistického modelovania jazyka v slovenčine v doposiaľ navrhnutých systémoch na interakciu človeka so strojom hovorenou rečou a automatické rozpoznávanie a prepis plynulej reči to textu sa opiera o poznatky z oblasti modelovania príbuzných jazykov, najmä jazyka českého, poľského, slovinského, srbochorvátskeho, či ruského. Čo sa týka samotného modelovania pomocou štatistických metód, hlavným predpokladom pri tvorbe kvalitného jazykového modelu je dôsledné predspracovanie textového korpusu, ktorý vstupuje do procesu trénovania. Zvýšenú pozornosť je vhodné preto venovať najmä prepisu čísloviek a skratiek, a v neposlednom rade aj generovaniu ohybných tvarov vlastných podstatných mien, ktoré tvoria kritickú časť pri tvorbe akéhokoľvek štatistického modelu jazyka. Taktiež slovník, ktorý vstupuje do procesu trénovania, ale aj samotného rozpoznávania reči, musí byť v podmienkach reálneho nasadenia systému ARR obmedzený čo do počtu slov. Ukázalo sa, že dobré výsledky modelovania slovenského jazyka sa dosahujú už pri veľkosti 100–150 tisíc slov pri doménovo orientovanom automatickom prepise diktovanej reči do textu a 300–400 tisíc slov v úlohách všeobecného rozpoznávania spontánnej reči [19].

V oblasti *adaptácie modelov jazyka* na vybranú tému alebo prehovor rečníka sa ukázalo, že metódy, ktoré vykazovali pozoruhodné výsledky pre štatisticky viac závislé jazyky (ako napr. angličitna), v prípade slovenčiny nebolo pozorované výrazné zlepšenie. Z toho dôvodu je použitie metódy *lineárnej interpolácie* tematicky zameraných modelov jazyka viac než postačujúce, pričom výpočet interpolačných váh by mal byť určený *minimalizáciou perplexity modelov na množine odložených dát*. Ako adaptačné dáta je vhodné použiť buď textové dáta získané z prepisov rečových nahrávok obsiahnutých v rečových databázach, alebo písané texty čo možno najviac príbuzné doméne, v ktorej rozpoznávanie reči prebieha [20].

Kvalitu jazykového modelu, ako aj úspešnosť samotného rozpoznávania reči je možné zlepšovať množstvom optimalizačných techník. Jednou z možností je modelovať vysoko frekventované javy pomocou *viacslovných výrazov*. Takéto výrazy pokrývajú zväčša kontext dvoch–troch slov a zvyčajne sú tvorené odbornými termínmi, resp. spojením predložky s podstatným, či prídavným menom. Na základe experimentov opísaných v [19] konštatujeme, že viacslovné výrazy, aj keď len v malej miere, dokážu prispieť k zlepšeniu presnosti rozpoznávania plynulej reči, a to najmä na začiatku rečového prejavu pri rozpoznávaní krátkych jednoslabičných slov a nadobúdajú na význame aj v čiastkových úlohách pri reprezentácii viacslovných pomenovaní v jazykovom modeli, a tým prispievajú aj k postspracovaniu dát po rozpoznaní systémom ARR.

Ďalšou možnosťou je modelovať málopočetné javy v jazyku pomocou *morfémových modelov*. Delením singletónov a javov s veľmi malým výskytom vo vybranom jazyku na subslovné jednotky (koreň a koncovku), je možné štatisticky pokryť aj také javy, ktoré sa priamo v jazykovom modeli nevyskytujú. Výsledky modelovania slovenského jazyka pomocou morfémových modelov ukazujú výraznú redukciu počtu mimoslovníkových tvarov a perplexity modelov približne o jednu tretinu [21].

Naopak javy, ktoré sa v danom jazyku menia dynamicky a počet všetkých možných tvarov ohybných slov nie je v jazyku limitovaný, je vhodné modelovať pomocou *modelov založených na triedach slov*. Medzi takéto javy možno zahrnúť najmä vlastné podstatné mená, ako sú krstné mená, priezviská, geografické názvy alebo číslovky. Experimentálne výsledky modelovania slovenského jazyka pomocou slovných tried odvodených od koncoviek slov ukazujú mierne zlepšenie presnosti rozpoznávania

[5] http://nlp.web.tuke.sk/

reči oproti štandardným modelom približne o 5% v relatívnej miere a sú určitým kompromisom medzi modelmi založenými na slovných druhoch a štandardnými modelmi, čo do počtu slovných tried, percenta mimoslovníkových slov, či perplexity jazykového modelu [22].

Pri tvorbe jazykových modelov použiteľných v systémoch na automatický prepis spontánnej reči je často nevyhnutné sa vysporiadať aj s rôznymi mimorečovými prejavmi, ktoré pochádzajú priamo do rečníka. Tie sú spôsobené zlou výslovnosťou, nevhodnou artikuláciou a nedokonalosťou rečového prejavu. *Modely vložených páuz a dysfluentných javov* sa preto snažia podchytiť a zahrnúť do jazykového modelu rôzne suprasegmentálne javy obsiahnuté v rečovom prejave, ako napr. zaváhanie, prolongovanie a opakovanie slov resp. fráz, skomolenie slov, či časté dýchanie. Tieto javy vo vysokej miere vplývajú aj na celkovú chybovosť systému ARR. Bolo dokázané, že vhodným výberom a správnou reprezentáciou vybraných typov vložených páuz a dysfluentných javov v slovníku výslovnosti a v modeli jazyka je možné dosiahnuť zlepšenie presnosti rozpoznávania reči relatívne až do 10%.

V neposlednom rade, kvalitu jazykového modelu je možné zvýšiť aj na úrovni *rozširovania štatistík* pomocou internetových vyhľadávačov [7], prekladom slov alebo slovných párov z príbuzných jazykov a pod.

Oblasť štatistického modelovania slovenského jazyka má za sebou krátku minulosť a donedávna jej nebola venovaná taká pozornosť ako napr. v susednej Českej republike. Z toho dôvodu bolo nevyhnutné pri tvorbe štatistických modelov slovenského jazyka, ako aj pri samotnom spracovaní textových dát, ktoré sa používajú najmä v procese ich trénovania a adaptácie, podrobne naštudovať aj oblasť komputačnej lingvistiky, a vytvoriť tak rad programových nástrojov na počítačové spracovanie slovenského jazyka. Absencia dostupnosti niektorých kľúčových nástrojov slúžiacich najmä k morfologickej či syntaktickej analýze tiež obmedzili použitie jazykových modelov založených na triedach slov v plnom rozsahu, aj keď prvé kroky v tejto oblasti už boli uskutočnené. Napriek týmto obmedzeniam, modely slovenského jazyka dosahujú vysokú úspešnosť na úrovni 84–95% v reálnych aplikáciách systému ARR, ktorých výsledky sú zhrnuté v Tab. 2, od jednoduchých hlasových rozhraní slúžiacich na ovládanie robotických systémov, cez jednoduché rečové dialógové manažéry poskytujúce hlasové, interaktívne, či multimodálne služby, až ku komplexným diktačným a transkripčným systémom, ktoré pracujú s veľmi veľkými slovníkmi, nezávisle od rečníka, dokážu sa adaptovať na vybranú tému, či konkretného rečníka, sú robustné a prebiehajú v reálnom čase [7, 12, 20, 23].

Vďaka narastajúcemu záujmu o interaktívne rečové technológie v slovenskom jazyku sa ďalšie smerovanie v tejto oblasti uberá cestou využitia doménovo orientovaných modelov jazyka pri tvorbe diktačných systémov aj pre takú oblasť ako je medicína, ďalej systémov na automatický prepis akademických prednášok, spravodajských relácií, športových prenosov, televíznych, rozhlasových, či parlamentných debát, resp. obchodných rokovaní v malých konferenčných miestnostiach s viacerými účastníkmi, a to s využitím robustných algoritmov na adaptáciu jazykových modelov na vybranú tému alebo rečníka [20], ale aj ďalších aplikácií určených pre robotické systémy.

| úloha | Acc* [%] |
|---|---|
| hlasové ovládanie ramena robota (SNR=10dB) | 90,19 |
| automatický prepis diktátu z oblasti súdnictva | 95,11 |
| elektronické služby Generálnej prokuratúry | 94,65 |
| prepis reálnych parlamentných debát | 90,93 |
| prepis spravodajských relácií | 84,33 |

* Presnosť (z angl. „*accuracy*“, Acc) automatického prepisu vychádza z miery chybovosti slov (z angl. „*word error rate*“), definovanej ako minimálna vzdialenosť medzi referenčnou sekvenciou slov a automatickým prepisom po rozpoznaní.

Tabuľka 2: Celková presnosť jazykových modelov v doposiaľ navrhnutých systémoch na automatické rozpoznávanie a prepis reči do textu v slovenskom jazyku

### 5.2 Porozumenie prirodzeného jazyka

K tomu, aby bolo možné obohatiť počítačové systémy o schopnosť skutočného *porozumenia prirodzenej reči a jazyka*, je potrebné realizovať proces *sémantickej analýzy*, a získaný význam tak vhodným spôsobom zachytiť a zosúladiť s databázou znalostí. Je možné konštatovať, že ide o neľahkú úlohu, vzhľadom na komplexnosť slovenského jazyka, jej sémantiky a ostatných faktorov, ktoré súvisia s porozumením. Aj keď sémantická analýza nie je hlavným zameraním nášho laboratória, pre využitie systému ARR v aplikáciách interakcie človeka so strojom hovorenou rečou, je nevyhnutné sa vysporiadať s interpretáciou vyjadrení tiež v prirodzenom jazyku.

V systéme IRKR [24] sme na tento účel implementovali v jednotke riadenia dialógu podporu jazyka pre sémantickú analýzu – W3C SISR, ktorý umožňuje vložiť interpretačné inštrukcie priamo do *deterministických gramatík*, napísaných podľa W3C odporúčania SRGS. V danom prípade sa jednalo iba o limitované porozumenie, ktoré bolo zamerané skôr na naplnenie doménovo špecifických sémantických slotov hodnotami získanými z rečoveho prejavu používateľa. Takéto riešenie bolo v tom čase pomerne komfortné a postačujúce pre celý rad rečových aplikácií a rozhraní, ktoré poskytujú tzv. rečovo-založené dialógové systémy.

Pri riešení projektu zameraného na implementáciu hlasového ovládania do robotickej platformy, kde kvôli kompaktnosti a rýchlosti systému na limitovanom hardvéri nebolo možné použiť rečové gramatiky, keďže komplexnosť riadiacich povelov bola omnoho väčšia, implementovali sme pre tento účel tzv. „*keyword-spotting*“ techniku sémantickej analýzy. Vytvorili sme viacero doménovo-špecifických sémantických slotov, ktoré zachytávali predefinované slová z rečového dopytu používateľa [23].

Spolu s návrhom a vývojom systémov ARR, nielen pre rozpoznávanie jednoduchých povelov a fráz, ale aj pre diktovanú a spontánnu reč, vzrastali nároky na ich interpretáciu a predchádzajúce prístupy pre ne neboli použiteľné. Pre potreby rozpoznávania plynulej reči sa namiesto deterministických gramatík začali vo veľkom využívať štatistické modely jazyka, nakoľko plynulá reč poskytuje podstatne väčšiu výrazovú variabilitu.

Zvlášť dôležitou sa sémantická analýza a interpretácia ukázala pri experimentovaní s *virtuálnym konverzačným agentom*, ktorý má ľudský zjav [25]. Pri takomto druhu komunikácie má človek tendenciu očakávať od virtuálneho konverzačného agenta podobné výrazové prostriedky ako majú ľudia, predovšetkým v oblasti komunikačných schopností a porozumenia, ktoré spolu úzko súvisia. Ďalším špecifikom je, že systémy s hlasovým rozhraním sa stávajú viac doménovo nezávislými, teda umožňujú dialógovú interakciu v rámci množstva tém (ako napr. Apple SIRI[6] a pod.), čo posúva interpretáciu významu ďalej, od relatívne „bezpečných“ doménových sémantických slotov k viac všeobecným sémantickým roliam.

T. E. Payne v [26] definuje sémantické roly nasledovne: „*Sémantická rola predstavuje základný vzťah, ktorý daná entita má k hlavnému slovesu vo vete.*“ Ďalej vysvetľuje, že: „*Sémantická rola je aktuálna rola, ktorú participant hrá v nejakej reálnej alebo imaginárnej situácii, bez ohľadu na lingvistickú realizáciu danej situácie.*“

Aj keď je teória sémantických rolí a s nimi súvisiacich valenčných rámcov slovies pomerne dobre rozpracovaná pre rôzne jazyky, neexistujú však žiadne systémy na *automatické určovanie sémantických rolí* (z angl. „*automatic semantic roles labeling*“, skr. ASRL) v slovenčine. Za veľmi dôležitú prácu v oblasti automatického určovania sémantických rolí pre slovenčinu možno považovať prácu E. Paleša, ktorý detailne opísal proces porozumenia prirodzeného jazyka na jednotlivých vrstvách a vyvinul prvý systém SAPFO – Parafrázovač slovenčiny, ktorého súčasťou bol aj modul pre určovanie sémantických rolí [27]. Tento systém však nie je podľa našich vedomostí voľne dostupný. Navyše sa jedná o deterministický systém, ktorý ako konštatuje M. Laclavík [28], nie je možné uspokojivo skonštruovať pre analýzu slovenského jazyka, z dôvodu veľkého množstva výnimiek. Z tohto pohľadu sú štatistické metódy jednoznačne lepšou voľbou.

Štatistické metódy pre systémy ASRL využívajú tzv. štatistické modelovanie typické pre rôzne úlohy v oblasti spracovania prirodzeného jazyka. Pre natrénovanie štatistických modelov je potrebná textová databáza anotovaná na úrovni sémantických rolí, ktorá v prípade slovenčiny doposiaľ prakticky neexistovala. Označenie vetných participantov pomocou sémantických rolí je náročná úloha a vyžaduje tiež dobré lingvistické znalosti.

Vzhľadom na neexistenciu databázy pre slovenčinu anotovanej na úrovni sémantických rolí, sme sa rozhodli vytvoriť aj takýto druh korpusu. Korpus SEMIENKO [29] aktuálne obsahuje tristo viet v slovenskom jazyku anotovaných podľa nami *modifikovanej dvojúrovňovej schémy na označovanie sémantických rolí*, prevzatej z anotačnej schémy podľa E. Paleša [27] a upravenej pre potreby automatickej sémantickej analýzy. Sémantická anotácia korpusu SEMIENKO je ilustrovaná na nasledujúcom príklade:

$$\text{AGS|KOG}[\textit{Ján}]\ \text{VRB}[\textit{spoznal}]\ \text{PAC|FEN}[\textit{Máriu}]\ .$$

Úloha automatického určovania sémantických rolí všeobecne pozostáva z dvoch základných častí, a to z rozdelenia viet na vetné participanty a následného priradenia sémantických rolí daným participantom. Pre klasifikáciu vetných participantov sme experimentálne vyskúšali dve techniky. Prvá metóda modeluje jednotlivé pravdepodobnosti nepriamo pomocou $n$-gramových modelov [30], pričom účinnosť klasifikácie na danom korpuse dosahuje úspešnosť na úrovni 48%. Druhá metóda využíva modifikovaný HMM klasifikátor, obsiahnutý v nástroji Dagger [10], ktorý v procese prehľadávania výstupnej sekvencie implementuje Viterbiho dekódovanie. Úspešnosť tohto typu klasifikácie v súčasnosti dosahuje úroveň až 56%, čo je vzhľadom na veľkosť trénovacej množiny adekvátne. Na základe predbežných výsledkov sémantickej analýzy v slovenskom jazyku môžeme konštatovať, že pre ďalšie zlepšenie je nevyhnutné významne rozšíriť manuálne anotovaný korpus, čo je však neľahká a veľmi pracná úloha.

## 6 Záver

V tomto príspevku boli predstavené úlohy z oblasti spracovania a modelovania slovenského jazyka, ktorým sa v Laboratóriu rečových a mobilných technológií na KEMT FEI TU v Košiciach v súčasnosti intenzívne venujeme. Je možné konštatovať, že úspešnosť nami navrhnutých algoritmov stále dobieha úroveň svetových výskumov, avšak súčasné výsledky je možné už teraz aplikovať v rôznych systémoch na rozpoznávanie a porozumenie reči, ale aj v iných systémoch intrakcie človeka so strojom hovorenou rečou, ktoré na našom pracovisku vyvíjame.

## Poďakovanie

Táto práca vznikla realizáciou projektu Univerzitný vedecký park TECHNICOM pre inovačné aplikácie s podporou znalostných technológií (kód ITMS: 26220220182) vďaka podpore operačného programu Výskum a vývoj spolufinancovaného zo zdrojov Európskeho fondu regionálneho rozvoja (25%) a výskumných projektov: Výskum a vývoj modulov pre jazykovo-adaptívne multimodálne rozhrania na základe Zmluvy č. SK-HU-2013-0015 podporujúcej spoluprácu medzi organizáciami v Slovenskej republike a v Maďarsku (50%), a Slovník viacslovných pomenovaní (lexikografický, lexikologický a komparatívny výskum) v rámci projektu APVV-0342-11 (25%), realizovaných vďaka podpore Agentúry na podporu výskumu a vývoja financovanej z prostriedkov Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky.

[6] https://www.apple.com/ios/siri/

## Literatúra


[1] Hládek, D., Staš, J.: Text mining and processing for corpora creation in Slovak language. Journal of Computer Science and Control Systems. **3, 1** (2010) 65–68

[2] Hládek, D., Staš, J., Juhár, J.: Building organized text corpora for speech technologies in the Slovak language. Jazykovedné štúdie XXXI: Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu). **31** (2014) 173–181

[3] Zlacký, D., Staš, J., Juhár, J., Čižmár, A.: Slovak text document clustering. Acta Electrotechnica et Informatica. **13, 2** (2013) 3–7

[4] Zlacký, D., Staš, J., Čižmár, A.: Supervised text document clustering algorithm with keywords in Slovak. In Proc. of the $7^{th}$ Int. Workshop on Multimedia and Signal Processing, Redžúr 2013. Smolenice, Slovakia (2013) 31–34

[5] Staš, J., Juhár, J., Hládek, D.: Classification of heterogeneous text data for robust domain-specific language modeling. EURASIP Journal on Audio, Speech and Music Processing. **2014, 14** (2014) 1–12

[6] Zlacký, D., Staš, J., Juhár, J., Čižmár, A.: Text categorization with latent Dirichlet allocation. Journal of Electrical and Electronics Engineering. **7, 1** (2014) 161–164

[7] Staš, J., Hládek, D., Juhár, J.: Recent advances in the statistical modeling of the Slovak language. In Proc. of the $56^{th}$ Int. Symp. ELMAR 2014. Zadar, Croatia (2014) 39–42

[8] Thurson, A.: Parsing computer languages with an automaton compiled from a single regular expression. In Implementation and Application of Automata: Proc. of the $18^{th}$ Intl. Conf. CIAA 2013, Halifax, NS, Canada. Ibarra, O. H., Yen, H. Ch. (Eds.). LNCS **4094**. Springer Berlin Heidelberg (2006) 285–286

[9] Ćavar, D., Jazbec, I.-P., Stojanov, T.: CroMo - Morphological analysis for standard Croatian and its synchronic and diachronic dialects and variants. In Proc. of the $8^{th}$ Int. Conf. on Finite-State Methods and Natural Language Processing, FSMNLP 2009. Pretoria, South Africa (2009) 183–190

[10] Hládek, D., Staš, J., Juhár, J.: Dagger: The Slovak morphological classifier. In Proc. of the $54^{th}$ Int. Symp. ELMAR 2012. Zadar, Croatia (2012) 195–198

[11] Hládek, D., Ondáš, S., Staš, J.: Online natural language processing of the Slovak language. In Proc. of the $5^{th}$ IEEE Int. Conf. on Cognitive Infocommunications, CogInfoCom 2014. Vietri sul Mare, Italy (2014) 315–316

[12] Rusko, M., Juhár, J., Trnka, M., Staš, J., Darjaa, S., Hládek, D., Sabo, R., Pleva, M., Ritomský, M., Ondáš, S.: Recent advances in the Slovak dictation system for judicial domain. In Proc. of the $6^{th}$ Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, LTC 2013, Poznań, Poland (2013) 555–560

[13] Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In Proc. of Int. Conf. on New Methods in Language Processing. Manchester, UK (1994) 44–49

[14] Przepiórkowski, A., Górski, R. L., Łaziński, Pęzik, P.: Recent developments in the National corpus of Polish. In Proc. of the $7^{th}$ Int. Conf. on Language Resources and Evaluation, LREC 2010. Valletta, Malta (2010) 994–997

[15] Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., Raab, J.: The Czech academic corpus 2.0 guide. The Prague Bulletin of Mathematical Linguistics. **89** (2008) 41–96

[16] Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In Proc. of the $4^{th}$ Int. Conf. on Language Resources and Evaluation, LREC 2004. Lisbon, Portugal (2004)

[17] Majliš, M.: W2C – Web to Corpus – Corpora. LINDAT/CLARIN Digital Library at Institute of Formal and Applied Linguistics, UFAL, Charles University in Prague, Czech Republic (2011)

[18] Hládek, D., Staš, J., Juhár, J.: Unsupervised spelling correction for the Slovak text. Advances in Electrical and Electronics Engineering. **11, 5** (2013) 392–397

[19] Juhár, J., Staš, J., Hládek, D.: Recent progress in development of language model for Slovak large vocabulary continuous speech recognition. In New Technologies - Trends, Innovations and Research, Volosencu, C. (Ed.). InTech Open Access, Rijeka, Croatia (2012) 261–276

[20] Staš, J., Hládek, D., Juhár, J.: Language model speaker adaptation for transcription of Slovak parliament proceedings. In Proc. of the $17^{th}$ Int. Conf. on Speech and Computer, SPECOM 2015, Athens, Greece (2015) to be published

[21] Staš, J., Hládek, D., Juhár, J., Zlacký, D.: Analysis of morph-based language modeling and speech recognition in Slovak. Advances in Electrical and Electronic Engineering. **10, 4** (2012) 291–296

[22] Staš, J., Hládek, D., Juhár, J.: Morphologically motivated language modeling for Slovak continuous speech recognition. Journal of Electrical and Electronics Engineering. **5, 1** (2012) 233–236

[23] Ondáš, S., Juhár, J., Holcer, R.: Methodology for training small domain-specific language models and its application in service robot speech interface. Journal of Electrical and Electronics Engineering. **7, 1** (2014) 107–110

[24] Ondáš, S., Juhár, J.: Development and evaluation of the spoken dialogue system based on the W3C recommendations. In Product and Services; from R&D to Final Solutions, Fuerstner, I. (Ed.). Scyio, Rijeka, Croatia (2010) 315–330

[25] Ondáš, S., Juhár, J., Trnka, M.: SIMONA – The Slovak embodied conversational agent. Intelligent Decision Technologies. **8, 4** (2014) 277–288

[26] Payne, T. E.: Describing morphosyntax: A guide for field linguists. Cambridge University Press, Cambridge (1997)

[27] Páleš, E.: SAPFO - Parafrázovač slovenčiny. Veda. Bratislava, Slovenská republika (1994)

[28] Laclavík, M., Ciglan, M., Krajči, S., Hluchý, L., Furdík, K.: Dostupné zdroje a výzvy pre počítačové spracovanie informačných zdrojov v slovenskom jazyku. In Proc. of the $1^{st}$ Workshop on Intelligent and Knowledge Oriented Technologies, WIKT 2006. Bratislava, Slovakia (2006) 92–98

[29] Staš, J., Hládek, D., Ondáš, S., Juhár, J.: On building the Slovak example-based meaning corpus. In Proc. of the $8^{th}$ Int. Conf. on NLP, Corpus Linguistics, Lexicography, Slovko 2015. Bratislava, Slovakia (2015) to be published

[30] Ondáš, S., Hládek, D., Juhár, J.: Semantic roles labeling system for Slovak sentences. In Proc. of the $5^{th}$ IEEE Int. Conf. on Cognitive Infocommunications, CogInfoCom 2014. Vietri sul Mare, Italy (2014) 161–166