Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

FIIT-5208-5657

Bc. Jozef Harinek

Anotácia rozsiahlych textov za využitia sily davu

Diplomová práca

Študijný program : Informačné systémy

Študijný odbor : 9.2.6 Informačné systémy

Miesto vypracovania : Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava

Vedúci práce : Ing. Marián Šimko, PhD.

máj 2015

ANOTÁCIA

Slovenská Technická Univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Študijný program: Informačné systémy

Autor: Jozef Harinek

Diplomový projekt: Anotácia rozsiahlych textov za využitia sily davu

Vedúci diplomového projektu: Ing. Marián Šimko, PhD.

máj, 2015

V našej práci sa venujeme získavaniu syntaktických anotácií pre slovenčinu, za využitia sily davu. Prostredníctvom prirodzeného jazyka sa v textoch uchováva obrovské množstvo informácií, ktoré je potrebné pre ďalšie strojové využitie spracovať. Jeho spracovanie je proces ktorý začína rozpoznávaním zvukov a ide až po určovanie sémantiky textu. Jedným z krokov je rozpoznanie syntaxe.

Zameriavame sa na syntaktickú anotáciu rozsiahlych textov v prirodzenom jazyku, pri ktorej využívame silu davu. Doménou sú texty v slovenskom jazyku v rôznych žánroch. Navrhnutú metódu na anotáciu rozsiahlych textov overujeme v softvérovom prototype na vykonávanie úloh spätých so spracovaním jazyka.

Navrhli sme metódu, ktorá využíva dav žiakov základných stredných škôl. Títo študenti majú počas vyučovania Slovenského jazyka zahrnutú aj časť o syntaktickej analýze. Toto učivo musia precvičovať a teda syntaktické anotácie musia vykonávať aj tak. Navrhujeme využiť túto skutočnosť a poskytnúť žiakom nástroj, v ktorom anotácie môžu vykonávať a využívať ich na anotovanie textov.

Zistili sme, že nami navrhovanou metódou vieme získavať syntaktické anotácie pre jednotlivé tokeny s presnosťou 0,80 pri dosiahnutí úplnosti 0,69. Správne určiť vzťahy medzi vetnými členmi sme dokázali s úspešnosťou 66,67 %. Tieto syntaktické anotácie síce nemajú takú kvalitu a granularitu ako anotácie vytvorené expertmi, avšak na podporu niektorých úloh spracovania prirodzeného jazyka sú postačujúce.

ANNOTATION

Slovak University of Technology Bratislava

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Degree Course: Information systems

Author: Bc. Jozef Harinek

Diploma Thesis: Crowdsourcing for large scale texts annotation

Supervisor: Ing. Marián Šimko, PhD.

2015, May

In our project, we focus on syntactic analysis of Slovak language by employing crowdsourcing techniques. There is a huge amount of information stored in natural language. In order to be able to process it by computer, we need to preprocess the text into machine understandable form. It is a process that starts from sounds recognition and goes up to text semantics recognition. One of the processing steps is syntactic analysis.

We focus on large scale texts syntactic annotation by employing crowdsourcing principles. We work in domain of Slovak language and use texts from different genres. Proposed method was verified in a software prototype for language-related problems solving.

We proposed a method that harnesses crows of elementary and high school students. These students have in their curriculum included a part about syntactic analysis. In order to gain this knowledge they have to complete various assignments and therefore perform syntactic analysis. We propose to make use of this fact and give the students a tool in which they can complete the assignments and at the same time produce syntactic annotations.

We showed that proposed method can be used to produce syntactic annotations for tokens with precision 0.80 when recall is 0.69. The relations between tokens were correctly assigned in 66.67 % of cases. Syntactic annotations produced by our method do not have the quality of annotations created by experts but still can be used to support other natural language processing tasks.

Čestne prehlasujem, že som prácu vypracoval samostatne a len za použitia citovaných zdrojov.

Jozef Harinek

Ďakujem doktorovi Mariánovi Šimkovi za odborné vedenie a množstvo cenných rád a podnetov počas riešenia projektu a takisto členom skupiny PeWe za množstvo cenných podnetov z ich strany.

Obsah

1 Úvod

V súčasnej dobe zaznamenávame obrovský nárast informácií na webe. Je to spôsobené nástupom Webu 2.0, vďaka ktorému sa bežný používateľ stal nielen pasívnym príjemcom obsahu, ale aj jeho aktívnym tvorcom (Šimko a Bieliková 2014). Tieto informácie sú často uchovávané vo forme prirodzeného jazyka a sú teda strojovo náročne spracovateľné v ich pôvodnej forme. Je potrebné dostať ich do takej formy, ktorú dokáže stroj jednoduchšie spracovať a následne ich možno použiť na ďalšie úlohy (Paralič et al. 2010). Tejto úlohe sa venuje oblasť spracovania prirodzeného jazyka.

V našej práci sa zaoberáme dvoma celkami. Prvým je získavanie syntaktických anotácií ako podkladový korpus pre ďalšie úlohy spojené s prirodzeným spracovaním jazyka (angl. *Natural Language Processing – NLP*). Druhým celkom sú metódy čerpania z davu, ktoré chceme pre účely získavania syntaktických anotácií navrhnúť a použiť. V práci sa zaoberáme syntaktickými anotáciami pre slovenský jazyk. Identifikovali sme potenciálne veľkú skupinu žiakov druhého stupňa základných škôl a študentov stredných škôl (gymnázií), ktorí sa počas hodín Slovenského jazyka učia aj syntaktickú analýzu viet. Tento dav by sme preto chceli využiť a ponúknuť im vzdelávací nástroj, pomocou ktorého bude možné zároveň zbierať syntaktické anotácie. Našim cieľom je preskúmať silu davu pre plnenie úlohy získavania syntaktických anotácií, analyzovať granularitu a kvalitu anotácií, ktoré týmto spôsobom dokážeme získať.

Druhou oblasťou, ktorou sa v práci zaoberáme, je čerpanie z davu (angl. *crowdsourcing*). Čerpanie z davu je populárne a atraktívne riešenie problému rýchleho a lacného spracovania veľkého množstva dát. Prístupy založené na čerpaní z davu využívajú namiesto ľudí, ktorí by úlohu plnili tradične, širokú verejnosť, ľudí – laikov, ktorí nie sú v danej oblasti expertmi (Howe 2006a; Jeff 2009; Howe 2006b). Pri využití takéhoto prístupu samozrejme vzniknú výstupné dáta, v ktorých sa nachádza aj veľa zle spracovaných vzoriek: niektorí účastníci experimentu sa môžu pokúsiť o podvádzanie, alebo aj s dobrým úmyslom pomôcť sa dopustia chyby, pretože nepochopia inštrukcie. Takisto, nie všetci účastníci dokážu vyriešiť zadanú úlohu aj po pochopení inštrukcií bezchybne (Quinn a Bederson 2011). Preto je potrebné mať dobre navrhnutú metódu získavania informácií pomocou čerpania z davu. K dobrej špecifikácií metódy pomôže zaradenie do jednotlivých dimenzií čerpania z davu, ako sú opísané v (Quinn a Bederson 2011). Tieto dimenzie rozlišujú jednotlivé prístupy a rozdeľujú ich na základe motivácie, kontroly kvality výstupných dát, agregácie výstupných dát, schopností, ktoré musia účastníci mať, postupnosti vykonávania úlohy, kardinality rozdelenia úloh medzi účastníkov (Quinn a Bederson 2011).

Oblasť strojového spracovania prirodzeného jazyka je sama osebe veľmi zaujímavá. V práci sú opísané rôzne úlohy, pri ktorých je možné využiť syntaktické anotácie na zlepšenie výsledkov existujúcich prístupov. Niektoré prístupy samotné sú založené na využití syntakticky anotovaných dát v danom jazyku. Táto oblasť ponúka množstvo výziev, hlavne čo sa týka spracovania slovenského jazyka, ktorému sa venujeme. Syntaktická analýza slovenčiny je jednou zo zaujímavých oblastí, v ktorej je v súčasnom stave stále priestor pre zlepšenie. Úskalia sú spôsobené zložitosťou slovenského jazyka ako takého (synonymia, homonymia), množstvom variácií pri vetnej skladbe a pod. (Ondáš et al. 2011). Je potrebné sa s nimi vysporiadať a automatizované parsery to nie vždy dokážu. V oblasti spracovania prirodzeného jazyka sme identifikovali viacero možností použitia anotovaného textu, resp. viacero spôsobov anotovania. Pri spracovaní prirodzeného jazyka je viacero úrovní anotácie textu. Prvým stupňom je rozpoznanie zvuku, pokiaľ začíname pri zvukovej analýze jazyka. Ďalej je to rozpoznanie jednotlivých hlások, slabík, slov, viet, vetných členov a správnych pádov až po sémantiku jazyka. Tieto úrovne spracovania sa dajú s rôznou úspešnosťou strojovo vykonať, avšak čím vyššie v úrovni spracovania ideme, tým je menšia úspešnosť a komplexnejšie spracovanie. Existujúce poloautomatické prístupy potrebujú dohľad experta a sú veľmi časovo (a teda aj finančne) náročné.

Našim zámerom je využiť potenciál davu pri spracovaní úloh z oblasti anotovania rozsiahlych textov. Ak sa tento problém rozdelí na veľa menších častí, jednotliví anotátori z davu nie sú príliš časovo zaneprázdnení a spoločne sa anotuje veľký korpus.

V našej práci sa chceme zamerať na dav tvorený študentami, ktorí prichádzajú do styku so slovenčinou. Títo študenti pri svojom štúdiu plnia rôzne úlohy späté so slovenským jazykom a preto pri použití vhodného nástroja a metódy môžu nielen oni profitovať z naučenej oblasti, ale aj my môžeme získať cenné dáta. Pri správnom rozdelení jednotlivých úloh vieme rôznymi metódami overenia správnosti výstupu zaručiť pomerne kvalitné výstupné dáta. V našom prípade je to rozdelenie jednotlivých úloh medzi žiakov tak, aby jednu úlohu riešili viacerí žiaci (študenti). Čo sa týka motivácie, máme veľkú výhodu vďaka skutočnosti, že žiaci v škole preberajú učivo ohľadne syntaktických anotácií a pre potreby precvičenia, resp. štúdia musia tieto anotácie vykonávať tak či tak. Túto skutočnosť vieme veľmi dobre využiť, ak im poskytneme systém, v ktorom bude možné precvičovať učivo zo školy, alebo systém, ktorý bude možné zo strany učiteľov využiť na manažovanie domácich úloh.

Cieľom našej práce je preskúmať silu davu žiakov pri tvorbe syntaktických anotácií, identifikovať kvalitu takto získaných anotácií. Zamýšľame sa aj nad možnosťami použitia získaných anotácií pri podpore úloh spracovania prirodzeného jazyka. Pre podporu

dosiahnutia našich cieľov sme implementovali softvérový nástroj, ktorý bol použitý žiakmi základných a stredných škôl pri precvičovaní syntaktickej analýzy vety.

V oblasti spracovania prirodzeného jazyka (NLP) je ešte mnoho problémov, ktoré je potrebné riešiť. Syntaktická anotácia, ktorej sa venujeme, je jedným z krokov k zložitejším úlohám v rámci NLP. Takýmito problémami sú napríklad automatická sumarizácia, strojový preklad textu, rozpoznávanie pomenovaných entít a mnoho ďalších.

Práca je členená nasledovne. V kapitole 2 opisujeme spracovanie prirodzeného jazyka a syntaktickú analýzu jazyka. V závere sú priblížené otvorené problémy NLP.

V kapitole 3 sa venujeme metódam čerpania z davu a zamýšľame sa nad ich možným využitím v oblasti spracovania prirodzeného jazyka.

V kapitole 4 sú následne opísané ciele práce, v ktorej sa snažíme prepojiť oblasti syntaktickej anotácie slovenského jazyka a metód čerpania z davu. V kapitole 5 navrhujeme metódu využívajúcu potenciál davu pri anotácii. Kapitola 6 opisuje návrh a implementáciu softvérového nástroja, ktorý bol navrhnutý na podporu realizácie tejto úlohy.

V kapitole 7 prinášame správu o experimentálnom overení našej metódy a kapitola 8 predstavuje zhrnutie celej práce a možnosti ďalšieho smerovania.

V prílohách sa nachádza technická dokumentácia, používateľská a inštalačná príručka, článok publikovaný na študentskej vedeckej konferencii IIT.SRC.

2 Spracovanie prirodzeného jazyka

V súčasnosti je množstvo informácií na webe uchovávaných vo forme prirodzeného jazyka. Prirodzený jazyk je primárnym prostriedkom komunikácie ľudí medzi sebou, vyjadrujú ním svoje názory, presvedčenia, túžby, otázky. Nakoľko obsah webu tvoria primárne ľudia pre ľudí, je prirodzené, že väčšina týchto informácií je na webe vo forme, ktorá je pre strojové spracovanie neštruktúrovaná. Ak chceme lepšie využiť informáciu zachytenú prirodzeným jazykom, musíme ho vedieť strojovo spracovávať. Avšak získavanie významu z prirodzeného jazyka je komplexný proces.

Spracovaniu prirodzeného jazyka sa venuje mnoho výskumných tímov. Niektoré viac, iné menej úspešne. Strojová analýza jazyka je však vždy závislá od konkrétneho jazyka, alebo minimálne skupiny jazykov s podobnou skladbou a pravidlami (Cimiano 2006). V našej práci sa zaoberáme spracovávaním slovenského jazyka. Slovenský jazyk má podobnú štruktúru ako český jazyk, je teda možné čerpať z prác venujúcim sa výskumu českého jazyka.

Spracovanie prirodzeného jazyka sa delí do viacero úrovní (Ondáš et al. 2011). Tieto úrovne spracovania jazyka na seba nadväzujú a každá ďalšia stavia na tej predošlej. :

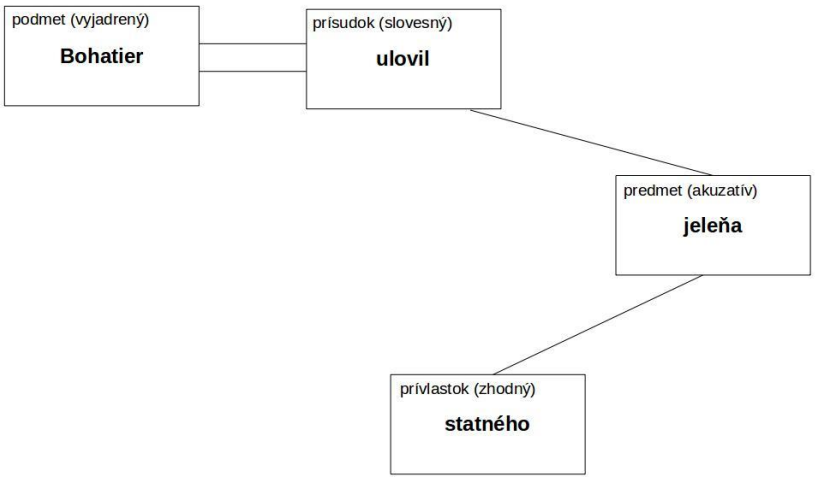
- *Morfologická vrstva* – analýza slov a extrakcia ich gramatických kategórií.
- *Syntaktická vrstva* – rozpoznanie syntaxe. Syntax je reprezentovaná vetnými členmi a vzťahmi medzi nimi.
- *Morfematická vrstva* – k predošlým dvom vrstvám pridáva informáciu o morfematickej štruktúre slov.
- *Sémantická vrstva* – analyzujú sa dôležité časti viet a identifikuje sa v nich sémantická štruktúra.
- *Kontextová vrstva* – dopĺňa sémantickú vrstvu tým, že ju zasadzuje do kontextu daných slov.

Pomocou vykonávania analýzy v týchto úrovniach sa vieme postupne dopracovať až k významu viet. V našej práci sa venujeme syntaktickej analýze textu, ktorá je priblížená v nasledujúcej kapitole.

2.1 Syntaktická analýza textu

Syntaktická analýza má za cieľ odhaliť štrukturálne vzťahy medzi slovami a vetnými konštrukciami. Identifikuje vetné členy, ich skladbu do viet a priraďuje im syntaktické roly – podmet, prísudok, predmet, prívlastok, atď. Výsledkom takejto analýzy je tzv.

syntaktický strom, ktorý znázorňuje vzťahy medzi jednotlivými vetnými členmi (obrázok 1).



Obrázok 1: Príklad syntaktického stromu

Syntaktická analýza má dve úrovne (Cimiano 2006). Prvou je identifikácia menných fráz a druhá je parsovanie textu.

Identifikácia menných fráz vo vete (angl. chunking)

Výstupom tejto analýzy sú čiastočne anotované vety, kde nie sú určené všetky vetné členy, ale iba niektoré, dôležité pre danú frázu. Tento spôsob spracovania sa nazýva aj plytká analýza.

Syntaktické jednotky pri tomto spracovaní textu sa všeobecne nazývajú kusy textu (angl. *chunk*). Takéto kusy textu majú dve základné vlastnosti a to:

- Sú nerekurzívne – jeden kus textu nemôže byť časťou iného.
- Sú nevyčerpávajúce – veta môže obsahovať slová, ktoré sa medzi kusy textu nedostali.

Plytké analyzátory teda vytvárajú zhluky slov, ktoré tvoria syntaktickú jednotku. Zvyčajne používajú techniky z konečných stavových automatov v takzvaných kaskádach, kedy spracovávajú vstup iteratívnym spôsobom, kde výstup z jednej fázy tvorí vstup ďalšej fázy. Plytké analyzátory vo všeobecnosti nerozoznávajú gramatické vzťahy ako podmet,

prísudok, atď. V snahe zachovať jednoduchosť a vyvarovať sa chybám väčšiny plytkých parserov sa nesnažia rozlíšiť sémantické, alebo syntaktické nejednoznačnosti.

Parsovanie textu (angl. parsing)

V tejto úrovni sa v kontraste k plytkému parsovaniu určuje celá syntaktická štruktúra vety. Avšak kvôli zložitosti kompletnej syntaktickej štruktúry je tento prístup oveľa náchylnejší na chyby, keďže sa snaží zachytiť celú syntaktickú štruktúru, nie iba jej časť.

Nástroje ktoré vykonávajú automatickú syntaktickú analýzu sa nazývajú parsery. Delíme ich na dve skupiny:

- *Parsery založené na pravidlách:* Parsery založené na pravidlách fungujú ako konečné automaty s definovanými pravidlami pre anotáciu textov. Príkladom takéhoto parsera je SET, určený pre český jazyk, alebo parser vytvorený Čižmárom a kol. (2011).
- *Štatistické parsery:* Štatistické spravidla využívajú metódy strojového učenia. Je ich preto potrebné najprv natrénovať pomocou dostatočne veľkej vzorky už anotovaných dát. Príkladom takéhoto parsera je The Stanford Parser[1]. Tento parser bol natrénovaný na korpusoch ručne anotovaných dát.

2.2 Syntaktická analýza Slovenského jazyka

V súčasnosti sa slovenskému jazyku v oblasti syntaktickej analýzy nevenuje veľa prác. Problémov je viacero. Jeden je skutočnosť, že v našom jazyku nemáme pevne daný slovosled vo vete a teda aj táto voľnosť v skladbe vety sťažuje úlohu automatickej syntaktickej analýzy – nevieme s určitosťou povedať v akej pozícií majú jednotlivé vetné členy vo vete voči sebe byť. Ďalej je tu homonymia (jedno slovo má viacero významov), nejednoznačnosť hovoreného jazyka. Problémom je teda zložitosť jazyka, ktorú je potrebné pri automatickom spracovaní prekonať (Ondáš et al. 2011).

Pri syntaktickej analýze slovenského jazyka sa chceme zamerať najmä na identifikovanie základných vetných členov, ktoré sú:

- Podmet – odpovedá na otázku „*Kto vykonáva tú činnosť?“* Môže byť vyjadrený, alebo nevyjadrený (napr. na rozdiel od angličtiny, kde sa musí vo vete vždy nachádzať). Napríklad *„(Ona) Varí.“* tu predstavuje zámeno „O*na*“ nevyjadrený podmet.

[1] http://nlp.stanford.edu/software/lex-parser.shtml

- Prísudok – odpovedá na otázku *„Čo sa stalo?/Čo sa deje?/Čo robí?“* Predstavuje teda sloveso vo vete.
- Predmet – odpovedá na otázku *„Čo robí? S kým niečo robí?“*
- Prívlastok – špecifikuje a dopĺňa vlastnosti podmetu a predmetu. Napríklad v slovnom spojení „milý človek“ slovo *milý* bližšie určuje vlastnosti *človeka.*
- Príslovkové určenie – odpovedá na otázku *„Ako/Kedy/Prečo/Kde on/ona niečo robí?“* V závislosti od otázky ktorou sa pýtame rozlišujeme 4 druhy príslovkového určenia: spôsobu (ako?), času (kedy?), príčiny (prečo?) a miesta (kde?).

Vzťahy medzi vetnými členmi vyjadrujú vetné sklady. Tu rozlišujeme tri typy skladov:

- Prisudzovací – vyjadruje vzťah medzi podmetom a prísudkom.
- Priraďovací – vyjadruje vzťah medzi dvoma rovnocennými vetnými členmi (napr. dva predmety).
- Určovací – vyjadruje vzťah nadradeného a podradeného vetného člena (napr. podmet a prívlastok).

Výskumu syntaktickej analýzy sa na Slovensku nevenuje veľa prác, avšak (Ondáš et al. 2011) mali snahu vytvoriť nástroj na porozumenie prirodzeného jazyka. Ich riešenie zahŕňa morfologickú a syntaktickú vrstvu analýzy jazyka. Tieto sa vykonávajú v moduloch na to určených, nás zaujímal hlavne syntaktický analyzátor (parser).

Tento parser je založený na pravidlách, sformulovaných podľa Pravidiel slovenského pravopisu[2]. Experiment s týmto parserom bol vykonaný na vzorke 127 viet z novinových článkov. Na základe rôzneho spôsobu morfologickej analýzy a kvality výstupu z tejto fázy, pracoval parser nasledovne: dokázal správne rozoznať až 98 % prísudkov a ostatné vetné členy rozoznal s úspešnosťou rozpätí 72 – 85 %. Experimentálne bolo preukázané, že úspešnosť tohto parsera úzko súvisí s kvalitou morfologickej analýzy (Ondáš et al. 2011). Táto práca je jednou z mála prác zaoberajúcich sa syntaktickou analýzou slovenčiny za posledné roky.

Slovenský závislostný korpus

Korpus textov je špecifický súbor jazykových dát, ktorý je dostupný v elektronickej podobe, jeho základom sú texty rôznych žánrov a štýlov, ku ktorým sú priradené lingvistické informácie na rôznych úrovniach (slova, vety, celého textu). Korpusy môžu

[2] http://www.juls.savba.sk/ediela/psp2000/psp.pdf

byť rôzneho typu (obsahujúce morfologické informácie, informácie o syntaxi, atď.)[3]. V našej práci sa zaoberáme korpusmi zo syntaktického hľadiska.

Od roku 2002 existuje v Jazykovednom ústave Ľudovíta Štúra Slovenskej akadémie vied oddelenie Slovenského národného korpusu. V roku 2005 sa začala práca na ručnej syntaktickej anotácií Slovenského závislostného korpusu (SZK). Korpus je momentálne celý zanotovaný a jednotlivé anotácie sa kontrolujú (Gajdošová, 2007). Toto všetko je v súčasnom stave vykonávané manuálne, za účasti expertov, nakoľko nie je k dispozícií vhodný automatický nástroj. Postup anotácie SZK sa inšpiroval Pražskou závislostnou syntaxou a projektom PDT (angl. Prague Dependency Treebank), kde vyvinuli softvérové nástroje na ručnú syntaktickú anotáciu a taktiež navrhli štruktúru ukladania syntaktických informácií[4]. Dôvod, prečo anotátori zvolili ako východisko pražskú závislostnú syntax, je ten, že oba jazyky majú podobnú syntaktickú štruktúru (český aj slovenský jazyk) a kolegovia z Česka už mali bohaté skúsenosti z niekoľkoročnej anotácie závislostného korpusu.

V SZK sa nachádza 34 613 viet z rôznych štýlov. Každá veta bola anotovaná dvoma anotátormi, nezávisle od seba. Štýlovo-žánrová štruktúra SZK je pestrá, jadro tvoria texty nasledovných štýlov:

- Beletria – George Orwel – 1984, Ladislav Ballek – Pomocník, preklady rozprávok, mládežnícky román.
- Odborné texty – historická monografia, lingvisticko-spoločenské štúdie
- Populárno-náučný štýl – texty z portálu *Wikipedia*[5]
- Publicistika – denník SME a internetový časopis InZine

V SZK sú zahrnuté kompletné texty jednotlivých diel z popísaných žánrov. Takto anotované dáta je následne možné použiť napríklad ako trénovaciu množinu pre syntaktické analyzátory využívajúce prístupy založené na strojovom učení. Rôznorodosť žánrov zabezpečuje následné natrénovanie klasifikátora na rôzne druhy textov (žánre) a teda takto natrénovaný klasifikátor by mal dokazovať lepšie výsledky (Buchholz a Marsi 2006).

Pri anotovaní SZK sa autori inšpirovali pražskou závislostnou syntaxou[6] a projektom PDT 1.0[7]. Ako základná príručka bol použitý manuál *Anotace na analytické rovině. Návod pro*

[3] http://korpus.juls.savba.sk/what.html
[4] https://ufal.mff.cuni.cz/pdt3.0
[5] http://sk.wikipedia.org/wiki
[6] Rozbor syntaxe zaoberajúci sa vzťahmi medzi členmi vo vete (súvetí).
[7] https://ufal.mff.cuni.cz/pdt/

anotátory. Rozhodnutie vybrať sa cestou pražskej závislostnej syntaxe bolo podporené viacerými dôvodmi:

- Oba jazyky majú blízku syntaktickú štruktúru.
- Blízkosť teoretických východísk na analytickej rovine.
- V Prahe mali už v tom čase bohaté skúsenosti s anotovaním, mali už vyriešenú technickú stránku spracovania syntaktickej úrovne jazyka.

2.3 Úlohy v oblasti spracovania prirodzeného jazyka

V nasledujúcej kapitole chceme opísať aktuálne úlohy spracovania prirodzeného jazyka a možnosť využitia syntaktických anotácií pri ich riešení.

Automatická sumarizácia (angl. *Automatic text summarization*) – je to proces redukcie textových dokumentov na krátku sumarizáciu dôležitých častí. Táto úloha vznikla spoločne s nárastom informácií na webe a presýtením informáciami, kedy vznikla potreba sumarizovať z dlhých textov ich podstatný obsah do kratších útvarov. Syntaktická štruktúra viet môže pomôcť pri tejto úlohe. So znalosťou syntaktickej štruktúry je možné lepšie rozpoznať dôležité časti vety (Yousfi-Monod a Prince 2005).

Strojový preklad (angl. *Machine translation*) – preklad z jedného jazyka do iného. Aby sa text preložil správne, stroj mu musí úplne rozumieť. Vyžaduje si množstvo rozličných znalostí (znalosť gramatiky, sémantickej štruktúry, znalosť faktov o reálnom svete, atď.). Zaraďuje sa do triedy tzv. UI-kompletných problémov. Znamená to že kompletné riešenie tohoto problému je ekvivalentné problému vytvorenia počítača tak inteligentného ako človeka (Shapiro 1992). Pri využití znalostí syntaxe jazykov medzi ktorými sa prekladá je preukázané zlepšenie výsledkov oproti štatistickému prekladu bez tejto vedomosti (Chiang 2010).

Generovanie prirodzeného jazyka (angl. *Natural Language Generation*)[8] – pri tejto úlohe ide o generovanie prirodzeného jazyka zo znalostí reprezentovaných v databázach, znalostných systémoch a podobne. Túto úlohu je možné pripodobniť prekladaču, ktorý prekladá strojovú reprezentáciu do prirodzeného jazyka. Syntakticky anotovaný korpus pri tejto úlohe slúži na podporu tvorby správnej syntaxe viet pri generovaní prirodzeného jazyka.

Porozumenie prirodzeného jazyka (angl. *Natural language understanding*)[8] – je to proces rozkladania prirodzeného jazyka do štruktúrovanej formy, ktorej rozumie počítač. Do tejto

[8] http://en.wikipedia.org/wiki/Natural_language_processing#Major_tasks_in_NLP

oblasti zapadá široká škála úloh, od rozpoznania jednoduchých povelov robotom až po plné pochopenie umeleckých textov poézie a prózy.

Určovanie slovných druhov (angl. *Part-of-speech tagging*)[8] – Mohlo by sa nazvať gramatické značkovanie, alebo rozlišovanie slovných druhov. V zjednodušenej forme sa táto úloha vyučuje aj v školách na hodinách jazyka – žiaci majú identifikovať slovné druhy jednotlivých slov vo vete. Táto úloha je zložitejšia ako sa môže zdať. Nie je postačujúce mať zoznam slov a ich gramatických kategórií, pretože tie sa môžu meniť v závislosti od kontextu použitia slov. Pri rozlišovaní tohto kontextu pomôže znalosť syntaktickej štruktúry vety, pretože vetné členy prinášajú obmedzenia na to, ktorý slovný druh mohol byť na danom mieste použitý (napr. prísudok – vo väčšine prípadov je to sloveso).

Syntaktická analýza (angl. *Parsing*)[8] – pri automatickom parsovaní textu (syntaktickej analýze) existuje mnoho prístupov používajúcich metódy strojového učenia. Syntakticky anotovaný korpus metódami čerpania z davu, sa môže použiť ako trénovacia množina pre prístup k syntaktickej analýze využívajúci strojové učenie.

Rozpoznanie viet v texte (angl. *Sentence breaking*)[8] – je to úloha pri ktorej je potrebné rozdeliť text na jednotlivé vety. Tie sa väčšinou dajú učiť podľa interpunkčných znamienok, ale nie vždy to musí platiť – napr. ak sa v texte nachádzajú skratky a podobne. Preto aj pri tomto spôsobe spracovania môže pomôcť znalosť vetnej syntaxe – napr. rozpoznanie, či sú všetky potrebné vetné členy prítomné v rozpoznanej vete, alebo nie.

Extrakcia informácií (angl. *Information extraction*)[8] – pri tejto úlohe sa automaticky extrahujú štruktúrované informácie z neštruktúrovaného, alebo čiastočne štruktúrovaného textu z pohľadu strojového spracovania. Vo väčšine prípadov je to spracovanie textov určených pre ľudí metódami spracovania prirodzeného jazyka. Ako extrakcia informácií sa môže považovať automatizovaná anotácia, extrakcia obsahu (sumarizácia) a podobne.

Tvorba zdrojov (angl. *Resource creation*) (Vilnat et al. 2008). Úloha pozostávajúca z viacerých krokov, medzi ktorými je aj získanie syntaktických anotácií.

- Vytvorenie syntaktických anotácií parsovaním
- Syntaktické anotácie vytvárajú, resp. obohacujú lingvistické zdroje (napr. lexikóny, gramatiky, anotované korpusy)
- Lingvistické zdroje vytvorené resp., obohatené syntaktickými anotáciami sú použité na natrénovanie existujúcich parserov
- Tieto parsery sú použité na vytvorenie bohatších (syntakticko-sémantických) anotácií
- Pokračuje sa znova prvým krokom

Identifikácia materinského jazyka (angl. *Native language identification*) – autori v (JojoWong a Dras 2011) použili časti syntaktických stromov na podporu úlohy identifikácie autora, z hľadiska toho, či jazyk ktorým tvoril obsah textu je jeho materinským jazykom, alebo nie. Použitie syntaktických stromov a ich častí zlepšilo výsledky.

2.4 Diskusia

V tejto kapitole sme zhrnuli znalosti z oblasti spracovania prirodzeného jazyka, konkrétne syntaktickej analýzy. Zaoberali sme sa hlavne syntaktickou analýzou slovenského jazyka, kde sme identifikovali nedostatok prác venujúcim sa tomuto problému. V (Ondáš et al. 2011) sa autori zaoberali automatizovanou syntaktickou analýzou slovenského jazyka, kde dosiahli úspešnosť extrakcie jednotlivých vetných členov v rozpätí 72 – 85 % extrahovaných správnych vetných členov. Základné vetné členy (podmet a prísudok – vetný základ) dosahovali najvyššiu úspešnosť, ostatné vetné členy boli v už vyššie spomenutom percentuálnom rozpätí. Úspešnosť takéhoto automatizovaného parsera úzko súvisí s kvalitou predchádzajúcej morfologickej analýzy daného textu. Prístup k získavaniu syntaktických anotácií, ktorý by nevyžadoval predchádzajúcu morfologickú anotáciu, by prinášal v tomto ohľade výhodu – bolo by potrebné vykonať o krok menej.

V kapitole sme tiež zhrnuli aktuálne problémy spracovania prirodzeného jazyka, pri ktorých syntakticky anotovaný korpus je buď nutnou súčasťou riešenia, alebo prináša zlepšenie výsledkov oproti riešeniu daných problémov bez tejto znalosti.

3 Využitie potenciálu davu ľudí na podporu výpočtových úloh

Dav ľudí má obrovský potenciál. Táto sila vedomostí, ktoré ľudia majú, je dobre využiteľná pri riešení rozsiahlych úloh. Preto sa v súčasnosti čoraz viac využíva čerpanie z davu pri riešení rôznych typov úloh. Čerpanie z davu je založené na jednoduchom, ale silnom koncepte – teoreticky hocikto má potenciál vložiť do riešenia hodnotné informácie (Greengard 2011). Na tomto princípe je založených mnoho existujúcich systémov (CQA systémy – StackOverflow, Yahoo answers, rôzne hry s účelom, Wikipédia, atď.). Na rozdiel od prístupov založených na účasti expertov, v metódach založených na dave sa zúčastňujú laici. Výhodou takéhoto prístupu je, že máme k dispozícií oveľa viacej prispievateľov (expertov je menej ako laikov; experti nie sú vždy dosiahnuteľní). Je však potrebné nájsť ten správny spôsob ako úlohy davu prezentovať, aby bol schopný ich riešiť a taktiež zaručiť, aby sa úlohy dostali k tým, ktorí reálne môžu prispieť k riešeniu.

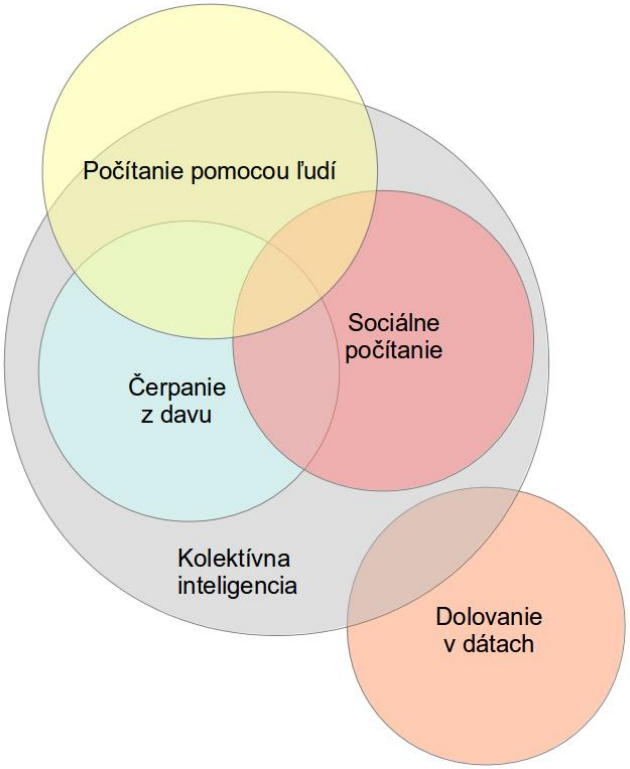
Podľa jednej z definícií, čerpanie z davu znamená vykonať prácu bežne vykonávanú špecializovaným odborníkom, za pomoci bližšie nešpecifikovanej, veľkej skupiny ľudí, formou otvorenej výzvy (Howe 2006a). Keďže ako expertné prístupy, tak aj prístupy čerpania z davu sú založené na ľuďoch, môže sa nám zdať, že tam nie je veľký rozdiel. V čom je ale podstatný rozdiel, je spôsob validácie výstupu. Pri expertných prístupoch je to samotný expert, ktorý v danej oblasti vie najlepšie posúdiť správne riešenie problému. Pri prístupoch založených na dave je pri overovaní výsledku hlavný predpoklad, že ak sa mnoho ľudí nezávisle zhodne na odpovedi, pravdepodobne je to pravda (napr. priradenie rovnakej značky pesničke, obrázku). Aj keď sú výstupy metód založených na čerpaní z davu, na základe vyššie spomenutého princípu dokážu generovať relatívne presné výsledky.

V nasledujúcej kapitole sú analyzované jednotlivé formy využívania potenciálu človeka pri riešení výpočtových úloh. Čerpanie z davu je len jednou z foriem počítania pomocou ľudí (angl. Human Computing).

3.1 Rozdelenie prístupov využívajúcich potenciál ľudí na výpočtové úlohy

Myšlienke využitia potenciálu ľudí na úlohy ktoré počítač nedokáže sám spracovať sa výskum venuje už nejakú dobu. V roku 2005 bola dokončená prvá dizertačná práca na tému *Human Computation*. Od tohoto času sa záujem o toto pole len zvyšuje (Quinn a Bederson 2011).

Úlohy, ktoré sú vykonávané počítačom za asistencie ľudí, sa rozdeľujú do viacerých skupín. Aj keď sa môžu zdať príbuzné až totožné, je medzi nimi rozdiel v spôsobe zahrnutia ľudí do procesu riešenia úlohy (Quinn a Bederson 2011). V nasledujúcej časti chceme načrtnúť ich základné rozdelenie. Na je graficky znázornený presah jednotlivých oblastí riešenia výpočtových úloh za asistencie ľudí.



Obrázok 2: Rozdelenie prístupov využívajúcich ľudský potenciál (Quinn & Bederson, 2011)

- *Počítanie pomocou ľudí (*angl. *Human computation):* Sú to všetky problémy, ktoré vo všeobecnosti pasujú do výpočtovej paradigmy a môžu byť jedného dňa vyriešené len za použitia počítača. V súčasnom stave však potrebujú interakciu človeka. Účasť človeka je riadená daným výpočtovým systémom, resp. procesom (Quinn a Bederson 2011).
- *Čerpanie z davu (*angl. *crowdsourcing):* Čerpanie z davu bolo odvodené Jeffom Howeom zo slova outsourcing. Toto aj vystihuje podstatu tohto slova, keďže je to využitie davu (angl. *crowd*) na vykonanie činnosti, ktorú by mal inak vykonať špecialista (angl. outsourcing). Niektorí si zamieňajú *Human Computing* s čerpaním z davu, ale tieto dva pojmy sú odlišné. Aj keď sa v mnohom prekrývajú, čerpanie z davu zahŕňa úlohy ktoré sa bežne vykonávajú buď iba ľuďmi, alebo iba počítačmi, ale je potreba ich zrýchliť (alebo skvalitniť) použitím princípov čerpania z davu.

Human Computing zahŕňa úlohy, ktoré nie je možné vyriešiť samostatne počítačom.

- *Sociálne počítanie* (angl. *Social Computing):* Táto oblasť zahŕňa prípady v ktorých sú zainteresovaní ľudia v nejakej sociálnej role, kde je komunikácia zabezpečované technológiami. Príkladmi takýchto systémov sú wiki stránky, blogy, online komunity. Rozdiel medzi touto oblasťou a *Human Computing* je v tom, že *Social Computing* pripodobňuje prirodzené ľudské správanie v ktorom je komunikácia zabezpečovaná nejakou technológiou.
- *Dolovanie v dátach* (angl. *Data mining):* Pod dolovaním v dátach sa rozumie aplikovanie špecifických algoritmov na extrakciu vzorov z dát (Fayyad et al. 1996). Pretože tieto algoritmy sú často používané na extrakciu vzorov z dát vytvorených človekom, je aj tento prístup niektorými považovaný za formu počítania pomocou ľudí.

V práci sa venujeme hlavne metódam využívajúcim silu davu, preto v nasledujúcej časti priblížime princípy na ktorých sú tieto metódy založené.

3.2 Čerpanie z davu

Čerpanie z davu (angl. *crowdsourcing*) sa objavilo spolu s vznikom Webu 2.0 (web so sémantikou; web kde nie sú používatelia len pasívnymi prijímateľmi obsahu, ale aj jeho tvorcami). Používatelia tvoria jeho obsah a zároveň aj metadáta k obsahu (Šimko a Bieliková 2014).

Dimenzie čerpania z davu

Jednotlivé metódy čerpania z davu sa navzájom od seba odlišujú zaradením do tzv. *dimenzií čerpania z davu*. Tieto dimenzie slúžia na zaradenie danej metódy do konkrétnej oblasti metód čerpania z davu, nakoľko táto oblasť výskumu je obrovská (Quinn a Bederson 2011). V nasledujúcej časti načrtneme rozdelenie jednotlivých dimenzií spolu s niekoľkými príkladmi ako môžu byť dimenzie nastavené.

Motivácia

Čerpanie z davu je z veľkej časti založené na správnej motivácií účastníkov takejto metódy (riešiteľov problému). Je to jedna z výziev čerpania z davu, nájsť tú najlepšiu cestu motivácie ľudí a od toho závisí aj úspech daného projektu (Quinn a Bederson 2011).

Jedným z najväčších motivačných faktorov sú peniaze a tie sú aj v mnohých prípadoch používané ako primárny motivačný faktor. Tento model odmeňovania funguje na princípe rozdelenia úlohy na veľmi malé čiastkové úlohy za ktoré dostanú riešitelia zaplatenú malú

sumu. Čím viac ich vyriešia, tým väčšiu odmenu dostanú. Takto funguje napríklad *Amazon Mechanical Turk*[9].

Druhým typom motivácie pre ľudí je pocit pomoci dobrej veci. Ak je projekt využívajúci čerpanie z davu niečo, čo dáva ľuďom pocit pomoci všeobecne prospešnej veci, radi sa zapoja.

Ďalším motivačným faktorom je zábava. Ak budú mať ľudia pocit zábavy pri plnení úlohy, je to pre nich dostatočná motivácia. Tento princíp využívajú hry s účelom (angl. *Games with a purpose – GWAP*). Táto oblasť je asi najzaujímavejšia, keďže spája vývoj hier s navrhovaním metód ako hrami pomôcť nejakému výskumnému problému.

Vykonanie práce, za ktorú dostane používateľ uznanie známou organizáciou je tiež motivačným faktorom, ktorý je pre mnohých postačujúci. Uznanie na stránke projektu a možnosť referencovať si účasť na takomto projekte môže byť cenné.

Pre fungovanie princípov čerpania z davu je teda potrebné mať čo najlepšie zakomponovaný niektorý z týchto princípov vo svojej metóde. Ideálne tak, že tí čo sa zúčastňujú na riešení problému, vnímajú hlavne motivačný faktor ako náplň ich činnosti.

Kontrola kvality výstupu prebieha viacerými spôsobmi. Niektoré prístupy sú založené na redundantnosti plnenia úlohy - viacerí pracovníci dostanú tú istú úlohu a akceptovanie výsledku je možné iba po splnení určitej úspešnosti. Ďalej v *Mechanical Turk* funguje model zvýhodňovania, resp. blokovania ľudí s vysokou/nízkou reputáciou. Jedným z dobrých princípov je rozbiť úlohu na tak jednoduché menšie úlohy, že je jednoduchšie splniť úlohu ako oklamať systém.

Kontrola kvality

Dáta, ktoré získame metódami čerpania z davu majú často veľkú mieru „znečistenia“, čiže nesprávnych riešení problému. Je to spôsobené tým, že riešenia problému sa zúčastňujú laici a nie iba experti. Niektorí daný problém pochopia, alebo vedia riešiť do väčšej miery, avšak na druhej strane iní menej. Preto je potrebné mať vyvinutý nejaký mechanizmus na kontrolu dát získaných pomocou čerpania z davu.

- *Zhoda vo výstupe* (Von Ahn a Dabbish 2008)*:* Dvaja, alebo viacerí nezávislí účastníci robia tú istú úlohu a vygenerovaný výstup je akceptovaný len v prípade, ak sa zhodne väčšina. Pre tento predpoklad je dôležitá vzájomná nezávislosť

[9] https://www.mturk.com/mturk/

jednotlivých účastníkov riešenia problému. Napr. hra ESP (Von Ahn a Dabbish 2004).

- *Zhoda vo vstupe* (Von Ahn a Dabbish 2008)*:* Toto je opak zhody vo výstupe. Dvaja nezávislí ľudia dostanú nejaký vstup, ktorý môže, ale aj nemusí byť rovnaký. Ich úlohou je popísať ho jeden druhému a potom sa majú rozhodnúť, či dostali rovnaký vstup alebo nie. Takýto spôsob kontroly kvality bol prvý krát použitý v hre *Tag-a-Tune,* kde hráči popisovali skladby ktoré dostali a mali sa rozhodnúť či s protihráčom dostali rovnakú skladbu na posúdenie, alebo nie (Law a Von Ahn 2009).
- *Defensive task design:* Za týmto spôsobom kontroly kvality výstupu stojí myšlienka vytvorenia tak jednoduchej úlohy, že je ľahšie splniť samotnú úlohu ako podvádzať pri jej plnení a získať tak odmenu za vykonanie práce (Callison-Burch a Dredze 2010). Tento spôsob je relevantný najmä pri systémoch v ktorých účastníci dostanú za vykonanie úlohy finančnú odmenu.
- *Systém reputácie:* V niektorých systémoch je motiváciou vykonať úlohu čo najlepšie systém reputácie. V Mechanical Turk používateľ, ktorý sa často zúčastňuje riešenia úloh a rieši ich úspešne dostane časom prístup k žiadanejším úlohám, avšak naopak, ten čo vždy poskytne zlé riešenie môže byť časom zablokovaný.
- *Redundantnosť:* Pokiaľ je k dispozícii veľa pracovníkov, môže každú úlohu vykonať viac a viac ľudí, čím získame viac vzoriek, z ktorých vieme identifikovať ľudí, ktorí vždy dávajú chybné riešenia a v budúcnosti vieme ich prácu odfiltrovať (Quinn a Bederson 2011).
- *Primiešanie jednoduchých testovacích úloh:* Jeden z bežných prístupov je založený na miešaní úloh s takými úlohami, ktorých riešenie máme zaručene správne. Následne vieme na základe týchto úloh odfiltrovať prispievateľov, ktorí buď zámerne odosielajú zlé riešenia, alebo nepochopili čo je ich úlohou (Quinn a Bederson 2011).
- *Štatistické filtrovanie:* Podstatou je odfiltrovanie takých dát, ktoré sú irelevantné, napríklad nebrať do úvahy dáta, ktoré nepatria do určitej predpokladanej distribúcie dát (Chen et al. 2009).
- *Viacúrovňová kontrola:* Tento spôsob kontroly je založený na dvoch nezávislých skupinách ľudí, z ktorých jedna skupina vykoná úlohu a druhá skontroluje a ohodnotí riešenie. Zložitejšie spôsoby podľa tejto schémy môžu byť podobné, ako napríklad v prípade jedného projektu na *Mechanical Turk*, ktorý je zameraný na zlepšenie schopností autorov pri písaní textov (Bernstein et al. 2010).

- *Kontrola expertom:* Použitie hodnoverného experta, ktorý skontroluje relevantnosť výsledkov práce. Podobnú možnosť poskytuje aj Mechanical Turk, kde zadávateľ úlohy môže skontrolovať riešenia a rozhodnúť sa, či za dané riešenie zaplatí, alebo nie.

Agregácia dát

Čerpanie z davu je založené na rozložení jedného veľkého problému na čo najmenšie podproblémy, preto pri spracovaní dát pomocou sily davu musíme rátať s kombináciou všetkých čiastkových riešení do jedného globálneho riešenia nášho primárneho problému. Na základe spôsobu agregácie dát vieme potom rozlíšiť jednotlivé prístupy k čerpaniu z davu.

Ľudské schopnosti

V závislosti od typu úlohy, čerpanie z davu môže vyžadovať viacero schopností, ktoré musia riešitelia týchto úloh ovládať. Väčšinou sú to schopnosti ktoré má skoro každý človek už „vrodené“ (napr. schopnosť opísať to čo vidí, počuje, atď.). V niektorých prípadoch to však môžu byť aj špeciálne schopnosti, ktoré neovláda každý jedinec (napr. vie písať a čítať po Slovensky). Pri návrhu riešenia je dobré byť čo najšpecifickejší ohľadom schopností, ktoré majú účastníci mať, pretože môžeme zistiť, že niektoré časti problému môžu byť omnoho jednoduchšie vykonané automaticky, strojovo.

Poradie vykonávania procesu

V mnohých systémoch založených na čerpaní z davu sú tri základné role: žiadateľ, pracovník a počítač. Žiadateľ je koncový používateľ, ktorý má úžitok z výsledkov výpočtu (napr. niekto kto si nechá preložiť text do iného jazyka prekladačom spolupracujúcim s davom anotovaným korpusom). Pracovník je pochopiteľne ten, ktorý úlohu vykonáva. Jednotlivé metódy sa odlišujú aj poradím v akom títo jednotliví účastníci používajú a spracúvajú dáta.

Kardinalita mapovania úloh na účastníkov

Tento aspekt určuje, koľko ľudí bude vykonávať úlohy. V závislosti od typu problému môže byť dostatočný relatívne malý počet účastníkov, avšak pri rozsiahlych problémoch potrebujeme veľa pracovníkov, obzvlášť ak je potrebná rozsiahla agregácia výstupných dát. V závislosti od nastavenia tohto faktoru vieme predpokladať koľko času a finančných prostriedkov bude vyriešenie problému stáť.

3.3 Čerpanie z davu v úlohách spracovania prirodzeného jazyka

Ako bolo spomenuté vyššie, metódy čerpania z davu sa používajú pri spracovávaní veľkého množstva dát. Keďže tvorba lingvistických korpusov je práca s veľkými objemami dát, existuje viacero prác, ktoré sa ich využitím v poli NLP zaoberali. V tejto časti by sme chceli priblížiť niektoré z nich.

V (Munro et al. 2010) autori porovnávali rôzne laboratórne experimenty (na tvorbu dát použitých pre psycholingvistiku), pri ktorých lingvistické dáta vytvárali experti, s použitím prístupov založených na čerpaní z davu. Pri niektorých typoch vytváraných dát boli dokonca dosiahnuté výsledky lepšie v prospech čerpania z davu. Autori sa zhodujú v tom, že pri použití čerpania z davu je pri niektorých úlohách dosiahnutá dokonca väčšia rozmanitosť anotácií, ktoré je možné získať. Je možné generovať nové druhy dát, najmä pre rôzne experimentálne paradigmy (napr. pri získavaní dát pre psycholingvistiku[10] a pod.). Pri budovaní modelu viet psycholingvisti skúmajú vplyv kontextu vety na jej spracovanie ľudským mozgom. Pomocou čerpania z davu sa tvorili podklady pre tvorbu štatistických modelov na predikciu ďalšieho slova vety v závislosti od aktuálneho kontextu. Pri porovnaní tohto experimentu s laboratórnym experimentom dosahovala metóda využívajúca dav lepšie výsledky, vďaka tomu, že využitie davu umožňovalo jednoduchšie zapojiť rôznorodejšiu vzorku ako laboratórny experiment. Dáta boli potom pochopiteľne univerzálnejšie.

Čerpanie z davu bolo ďalej použité napríklad pri určovaní podobnosti frázových slovies s ich samostatnými tvarmi v angličtine (Schnoebelen a Kuperman 2010). Frázové slovesá sú také, ktoré v závislosti od použitia v rôznych frázach menia svoj význam, rozdeľujú ho medzi seba a ďalšie časti frázy. V práci zisťovali vzťah medzi samotným slovesom a jeho rôznymi výskytmi vo frázach. Autori zistili, že lepšie výsledky boli dosiahnuté, keď boli úlohy rozdelené v menších dávkach davu oproti výsledkom dosiahnutým keď úlohu vykonávali ľudia vo väčších dávkach úloh.

V (Ambati et al. 2010) sa autori zaberali prekladom textov z jedného jazyka do druhého. Navrhli systém tzv. Aktívneho prekladu davom (angl. *Active crowd translation*), kde pomocou davu získavali preklady textov medzi jazykmi, ktoré nemajú veľkú početnosť zdrojov lingvistických dát. Autori na vytvorenie metódy na strojový preklad spojili jeden z prístupov strojového učenia (aktívne učenie sa – angl. *active learning*) a dav neexpertov získaných pomocou platformy Amazon Mechanical Turk[11] (AMT). Dav postupne prekladá

[10] Psycholingvistika – veda zaoberajúca sa mentálnymi aspektami jazyka. Skúma vzťah medzi jazykom a vedomím človeka.
[11] http://mturk.com

jazykový korpus medzi dvoma jazykmi, kde je heuristickou metódou vždy vybraná veta, ktorú je ďalej potrebné preložiť pre získanie potrebného paralelného dvojjazyčného súboru dát. Každú vetu prekladá viacero ľudí a na zaistenie kvality výstupu sa tieto preklady kombinujú medzi sebou a sú taktiež porovnávané s externými zdrojmi. Tieto dáta sú následne použité na trénovanie modulu využívajúceho aktívne učenie.

V (Snow et al. 2008) autori skúmali možnosti použitia davu neexpertov na riešenie rôznych úloh spojených so spracovaním prirodzeného jazyka. Tieto úlohy boli rozpoznanie afektu (angl. *affect recognition*), rozlíšenie podobnosti slov (angl. *word similarity*), zoradenie udalostí v čase (angl. *event temporal ordering*), zjednoznačňovanie významu slova (angl. *word sense disambiguation*). Na vykonanie týchto úloh za pomoci davu autori použili platformu AMT, kde využili širokú základňu neexpertných anotátorov, ktorí úlohy plnili za finančnú odmenu. Autori zistili, že pre skúmané úlohy stačí iba malý počet neexpertných anotátorov aby sa kvalita anotácií vyrovnala anotáciám získanými expertmi.

Na predchádzajúcich prácach je demonštrované použitie čerpania z davu na získavanie anotácií a taktiež omnoho rozmanitejšie úlohy v lingvistike ako samotné anotovanie. Z výsledkov prezentovaných prác je vidno, že čerpanie z davu má obrovský potenciál využitia v rôznych oblastiach spracovania prirodzeného jazyka.

3.4 Diskusia

V tejto kapitole sme analyzovali prístupy k riešeniu výpočtových úloh, ktoré zapájajú ľudský faktor (angl. *Human computation*). Jedným z nich je aj čerpanie z davu, ktoré chceme použiť pri návrhu našej metódy na získavanie syntaktických anotácií. Metódy využívajúce dav ľudí sa odlišujú nastavením metódy vo viacerých dimenziách, ktoré ich charakterizujú (motivácia, kontrola kvality, kardinalita rozdelenia úloh, potrebná ľudská schopnosť, agregácia dát).

Viacero prác sa už venovalo využitiu čerpania z davu v úlohách spracovania prirodzeného jazyka. Keďže spracovanie jazyka je úloha ktorá je najlepšie vyriešená samotnými ľuďmi, je len prirodzené zamýšľať sa nad využitím čerpania z davu na podporu týchto úloh. V niektorých prípadoch vykazovalo čerpanie z davu lepšie výsledky ako metóda bez použitia davu (viď kapitola 3.3). V našej práci sa aj my zameriavame na prepojenie týchto dvoch oblastí.

4 Ciele práce

V našej práci sa zaoberáme získavaním syntaktických anotácií za využitia sily davu. Ako bolo predostreté v predchádzajúcich kapitolách, tieto dve oblasti (spracovanie prirodzeného jazyka a čerpanie z davu) je možné úspešne prepojiť. S narastaním digitalizácie v školstve je stále väčší priestor na použitie softvérových nástrojov v procese výučby a teda aj priestor pre zapojenie žiakov v procese výučby do úloh, ktoré dáta získané v procese učenia sa žiakov môžu využiť na riešenie niektorého z výpočtových problémov.

Hlavným cieľom práce je navrhnúť a overiť metódu na získavanie syntaktických anotácií pre slovenský jazyk za využitia sily davu tvoreného žiakmi základných a stredných škôl.

Cieľom nie je porovnávať sa s automatizovanými metódami na získavanie syntaktických anotácií, ale skôr preskúmanie anotácií získaných metódami čerpania z davu z viacerých hľadísk.

Sila davu žiakov na základných, prípadne stredných školách

Tento dav je potenciálne veľmi veľký – 285 814 žiakov[12]. Avšak neustále sa meniace sylaby výučby slovenského jazyka ho môžu zmenšovať, dôsledkom toho, že žiaci sa môžu jednotlivé časti syntaxe slovenského jazyka učiť až neskôr. Jedným z cieľov našej práce je teda preskúmať silu tohto davu na tvorbu syntaktických anotácií v ročníkoch, ktoré sa aktuálne učia syntaktický rozbor – 8. a 9. ročník základných škôl a v malom rozsahu aj všetky ročníky stredných škôl. Chceme zistiť do akej miery sú žiaci schopní vykonávať syntaktickú analýzu.

Identifikácia granularity anotácií, ktoré dokážu vytvoriť

Syntaktické anotácie, ktoré sú dostupné v slovenskom národnom korpuse majú vysokú úroveň granularity. Sú určené pre expertov – lingvistov, avšak pre účely spracovania prirodzeného jazyka sú postačujúce aj anotácie s nižšou úrovňou granularity. Sme si vedomí skutočnosti, že syntaktické anotácie, ktoré získame od žiakov základných a stredných škôl, nebudú dosahovať takú úroveň granularity, akú majú anotácie v SNK. Jedným z cieľov práce je teda preskúmať úroveň granularity, ktoré dokážeme získať davom tvoreným žiakmi základných a stredných škôl.

[12] 209 103 základné školy 2. stupeň + 76 711 stredné školy v šk. roku 2013/2014 (http://www.uips.sk/registre/zoznamy-skol-sz-v-exceli)

Vyhodnotiť presnosť, kvalitu získaných anotácií

Ako bolo opísané v kapitole 2.3, na mnohé úlohy spojené so spracovaním prirodzeného jazyka sa dajú využiť syntaktické anotácie. Jedným z našich cieľov je vyhodnotiť kvalitu získaných anotácií a teda aj možnosti ich použitia na podporu niektorých z týchto úloh.

Ďalším z cieľov práce je tiež navrhnúť a implementovať softvérový prototyp, v ktorom je možné realizovať experimenty na overenie navrhnutej metódy. Tento softvérový prototyp môže slúžiť aj ako nástroj na podporu vzdelávania v oblasti výučby (slovenského) jazyka.

5 Metóda na získavanie syntaktických anotácií za využitia sily davu

Čerpanie z davu je jedna z metód na spracovanie veľkého množstva dát. Spracovanie prirodzeného jazyka je výpočtová úloha, ktorá sa väčšinou vykonáva sa nad veľkým množstvom dát. V nasledujúcej časti predstavíme metódu, ktorá spája tieto dve oblasti a slúži na získavanie syntaktických anotácií za využitia sily davu. Pri návrhu metódy sme najprv zmapovali aktuálny stav výučby na školách, čo sa týka osnov slovenského jazyka. Potom sme na základe tejto analýzy navrhli nastavenie metódy z hľadiska čerpania z davu a navrhnuté princípy sme realizovali v softvérovom prototype, ktorý slúžil aj na overenie správnosti navrhnutých konceptov pri konfrontácií návrhu s učiteľmi.

V kapitole sa nachádzajú najprv východiská pre metódu, potom návrh metódy v kontexte princípov využívaných pri čerpaní z davu a nakoniec navrhujeme spôsob určenia správneho riešenia zo zozbieraných dát.

5.1 Východiská pre navrhovanú metódu vo výučbe vetnej syntaxe na školách

Pred realizáciou práce sme uskutočnili prieskum aktuálneho stavu na školách. Zisťovali sme, či je vôbec možné zrealizovať zbieranie syntaktických anotácií v školskom prostredí, aká je forma výučby na základných školách v súčasnosti a aké by boli dobré vlastnosti výučbového systému v ktorom by žiaci mohli precvičovať znalosti z oblasti gramatiky slovenského jazyk, v našom prípade vetnej syntaxe.

Ako prvé sme zisťovali v akom veku sa žiaci učia vetnú syntax. Zistili sme, že s výučbou tejto časti gramatiky sa začína na druhom stupni základnej školy a kompletnú stavbu jednoduchej vety sa žiaci učia v 8. ročníku základných škôl. Preto aj úroveň anotácií, ktoré môžeme získať, je rôzna. Až žiaci 8. a 9. ročníka základnej školy sú schopní vykonať kompletný rozbor základnej vetnej syntaxe v jednoduchých vetách. V súvetiach vedia rozoznať jednotlivé zložky súvetia (vety) a v každej určiť vetné členy, avšak nevedia určiť vzťah medzi jednotlivými vetami súvetia.

Na základných školách bolo v školskom roku 2013/2014 209 133 žiakov v ročníkoch 5. - 9. Na gymnáziách bolo v tom istom školskom roku 76 711 študentov[13]. Všetci títo študenti sa v rámci učiva na hodinách slovenského jazyka venujú v nejakej miere rozboru vetnej syntaxe. To znamená, že všetci sú potenciálni účastníci davu pri realizácií našej metódy. Hodinová dotácia na túto látku zodpovedá približne 1 mesiacu preberania vetnej syntaxe počas jedného školského roka. Učitelia identifikovali ako najvhodnejší spôsob použitia

[13] Údaje o počte žiakov sú zo stránky Ústavu informácií a prognóz školstva: http://www.uips.sk/

systému, ktorý by umožňoval precvičiť učivo syntaktickej analýzy, na riešenie domácich úloh žiakom. Ak počítame, že jedna trieda má v priemere 20 žiakov, ktorí počas jedného mesiaca, kedy sa vyučuje vetná syntax vyriešia týždenne 10 viet ako domáce úlohy, vieme pomocou jednej triedy žiakov získať 800 rôznych anotácií viet (20 žiakov * 10 viet * 4 týždne). Samozrejme, anotácie treba zbierať redundantne pre každú vetu, aby bola zaručená čo najvyššia kvalita výstupu. V akom počte je potrebná redundantnosť anotácií sme zisťovali v experimente. Pri počte žiakov, ktorý je v príslušnom veku na Slovensku, dáva táto skutočnosť obrovský potenciál využitia tohto davu na úlohy spojené so spracovaním prirodzeného jazyka.

Výučbový systém využívajúci vyššie popísané skutočnosti však taktiež potrebuje podporu zo strany pedagógov. Pri návrhu možných scenárov sme identifikovali dva hlavné:

- Použitie priamo v škole počas vyučovania;
    - Učitelia identifikovali ako nereálny scenár, väčšinou nie je na to technické vybavenie v triedach;
- Použitie ako nástroj na plnenie domácich úloh, prípadne na precvičenie učiva z vlastnej iniciatívy;
    - Učitelia túto možnosť vnímali veľmi pozitívne. Automatizovaná oprava domácich úloh a ich jednoduchá definícia cez rozhranie aplikácie by im vedela ušetriť značné množstvo času (čas strávený kontrolou domácich úloh na hodine môže byť až 20 minút, čo je skoro polovica vyučovacej hodiny);

Pri diskusii učitelia načrtli aj možnosť použitia na vysokých školách, ktoré sú zamerané na výučbu jazykov a pedagogiky, kde sa taktiež vyučuje vetná syntax. Avšak na túto oblasť sme sa v práci nezameriavali, nakoľko sme sa chceli sústrediť na zakomponovanie našej metódy do procesu výučby v rámci povinnej školskej dochádzky a teda aj využitiu najväčšieho potenciálneho davu v tejto oblasti.

Za pomoci učiteľov sme sa taktiež snažili zistiť všetky prípustné variácie úloh spojených s vetnou syntaxou v procese výučby. Žiaci 8. ročníka a starší sú schopní vykonať syntaktickú analýzu celej vety. Nižšie ročníky majú čiastkové vedomosti z oblasti syntaxe a vedia len niektoré veci (určenie len niektorých vetných členov, v závislosti od ročníka, v ktorom sú). Ako potenciálne problematické časti vetnej syntaxe učitelia označili:

- Rozlišovanie podmetu od predmetu;
- Rozbor syntaxe v jednočlennej vete;
- Určenie zamlčaného podmetu;
- Nepriamy prívlastok.

Táto vedomosť nám môže pomôcť pri upravovaní váh jednotlivých riešení, ak sa vyskytnú vyššie spomenuté situácie v konkrétnych vetách.

Ako budúce možné rozšírenie platformy učitelia navrhli doplnenie časti, ktorá by komplexne pokrývala výučbu gramatiky na školách a obsahovala by cvičenia na morfológiu (slovné druhy, pády, vzory slov, atď.), cvičenia na dopĺňanie i/y, cvičenia zameraná na lexikológiu (určovanie synoným, slovotvorné postupy a pod.). Pri zameraní sa okrem vetnej syntaxe aj na morfológiu a lexikológiu by mohli byť zaujímavé nové možnosti získavania užitočných anotácií počas procesu výučby. Táto oblasť je však mimo rozsah našej práce.

5.2 Návrh metódy a jej zaradenie do dimenzií čerpania z davu

Na základe zozbieraných informácií a analýzy aktuálneho stavu oblasti syntaktickej analýzy slovenského jazyka sme navrhli metódu na získavanie anotácií za využitia metód čerpania z davu. Identifikovali sme potenciálne veľký dav (žiaci základných a stredných škôl), ktorý musí syntaktickú analýzu vykonávať tak či tak a je možné zapojiť ho do syntaktickej anotácie počas procesu výučby.

Preto navrhujeme softvérové riešenie, ktoré je založené na čerpaní z davu a poskytuje platformu pre riešenie úloh spojených so syntaktickou analýzou vety.

Riešenie je založené na predpoklade, že dav žiakov, ktorí sa učia syntaktickú analýzu, dokáže pri správnej podpore riešenia tejto úlohy spoločne vytvoriť správne anotácie v rozsahu ktorý sa práve v škole učia. Žiaci musia syntaktickú analýzu robiť aj tak, v rámci výučby. Ako podklad pre cvičenia je možné použiť ľubovoľné texty, ktoré sa následne rozdistribuujú medzi žiakov a týmto spôsobom je možné pokryť a vytvoriť rozsiahle korpusy anotovaných dát. Dôležitou rolou je tiež učiteľ, ktorý vystupuje ako jedna z možností kontroly kvality anotácií. Pri identifikácií nejednoznačnosti v anotáciách získaných od žiakov sa učiteľ zapojí do procesu anotácie ako expert.

Naše riešenie teda poskytuje učiteľom a žiakom nástroj, v ktorom učiteľ zadefinuje parametre úloh zameraných na vetnú syntax, tieto úlohy sú vykonávané na ľubovoľných podkladových textoch, pre ktoré chceme získať syntaktické anotácie. Úlohy sa rozdistribuujú medzi žiakov, viacerí riešia rovnakú úlohu. Zo získaných riešení následne extrahujeme správne riešenie. Jednotlivé pravidlá pre metódu sú opísané nižšie. Ak identifikujeme kontroverzné riešenie (žiaci sa nezhodli jednoznačne), posunieme ho na kontrolu učiteľovi.

V nasledujúcej časti je pre lepšie pochopenie metódy jej opis rozdelený do jednotlivých metód čerpania z davu.

Zaradenie navrhovanej metódy do dimenzií čerpania z davu

Ako bolo spomenuté v kapitole 3.2, čerpanie z davu sa zaraďuje do rôznych dimenzií, na základe ktorých je daná metóda opísaná. V ďalšej časti teda chceme opísať našu metódu z pohľadu jednotlivých dimenzií pre jej lepšie zaradenie.

Motivácia

Ako hlavný motivačný faktor pre žiakov bude dobrá známka z predmetu. Metóda je navrhnutá tak, že žiaci vykonávajú syntaktický rozbor v rámci predmetu Slovenský jazyk ako domáce úlohy, resp. doplňujúce cvičenia. Motiváciou v konečnom dôsledku je teda dobrá známka z tejto časti predmetu. Čiastkovou motiváciou je precvičenie si učiva vo vlastnom záujme, aby ho žiak sám lepšie vedel.

Učiteľ je v našom systéme účastný tiež ako člen davu. Vo veľkej miere je prostredník, ktorý definuje aké úlohy majú študenti riešiť, čím vykonáva akési predspracovanie. Preto je potrebné motivovať aj učiteľov, aby systém používali. Motivácia učiteľov je prostredníctvom systému samotného – mali by ho chcieť používať preto, lebo im uľahčí prácu (nemusia vymýšľať vety, automatická oprava a vyhodnotenie úloh, študenti si môžu prejsť viacero úloh, atď.). Našou úlohou je navrhnúť rozhranie pre učiteľa čo najlepšie, aby v konečnom dôsledku nemal so systémom viac problémov ako bez neho.

Kontrola kvality výstupu

Kontrolu správnosti výstupu chceme posudzovať na základe zhody viacerých účastníkov vo výstupe. Táto myšlienka sa zakladá na predpoklade, že ak viacero nezávislých ľudí vyrieši úlohu rovnako, pravdepodobne je toto riešenie správne. Avšak je potrebné predpokladať aj možnosť, že väčšina vyrieši úlohu nesprávne a detegovať takéto prípady. Nakoľko je pri tvorbe syntaktických anotácií potrebná zručnosť vedieť tieto anotácie tvoriť, do davu vyberáme len tých, ktorí túto zručnosť majú. Na základe školského hodnotenia, prípadne vyriešenia niekoľkých kontrolných úloh vieme aj tento dav filtrovať a prispôsobovať váhy anotácií od jednotlivých anotátorov.

Agregácia dát

Na základe stratégie rozdeľovania viet na anotovanie sa bude postupne prechádzať celým korpusom a postupne teda získame anotácie pre celý korpus. Dáta sa teda postupne agregujú a vytvára sa väčší celok. Pre každú vetu zbierame anotácie redundantne, v počte ktorý je možné definovať v nastavení metódy.

Ľudská schopnosť

Pre participáciu v našej metóde je potrebné, aby jednotliví účastníci mali základné znalosti syntaktického rozboru viet. Samozrejme, na inej úrovni sú žiaci základnej školy, ktorí tieto vedomosti ešte len čerstvo získavajú, na inej úrovni sú žiaci strednej školy, ktorí sa učia určovať aj pokročilejšie vlastnosti a na ešte vyššej úrovni sú študenti vysokej školy v lingvistických odboroch. Môžeme teda získať rôzne úrovne syntaktickej anotácie od základov až po kompletnú syntaktickú analýzu vety. Každá z týchto úrovni anotácie je použiteľná na úlohy spojené so spracovaním prirodzeného jazyka, čím však máme podrobnejšie anotácie, tým sú výsledky lepšie.

Poradie vykonávania úloh

Riešenie problému prebieha tak, že učiteľ zadefinuje parametre úlohy, na základe ktorých sa vygenerujú úlohy pre žiakov. Parametre úlohy sú definované v opise konkrétnych cvičení, ale je to napr. zložitosť použitých viet, stratégia rozdeľovania úloh, čas na ich vykonanie a pod. Jedným z parametrov úlohy je aj súbor dát, ktorý sa použije. Je tu možné použiť rôznorodé texty, ktoré chceme anotovať, napr. korpus ktorý sa použije pre ďalšie úlohy spracovania jazyka. Tieto úlohy sú žiakom následne pridelené a musia ich do nejakého časového limitu vyriešiť.

Pracovník (angl. *worker*) sa tu teda skladá z dvoch stupňov. Prvým je učiteľ, ktorý začne riešenie problému tým, že nadefinuje vlastnosti úlohy. Takto vytvorené úlohy následne rieši žiak, ktorý dokončí riešenie problému vykonaním syntaktickej analýzy.

Kardinalita pridelenia úloh účastníkom

Aby bolo možné automaticky vyhodnotiť správnosť anotácií, je potrebné mať čo najviac vzoriek riešenia, nakoľko pri syntaktickej anotácií sa žiaci veľmi ľahko môžu pomýliť a správne riešenie teda nie je jasné hneď z prvej získanej anotácie. Pre každú vetu je teda potrebné získať čo najväčší počet anotácií, z ktorých už vieme vyhodnotiť správnosť riešenia.

5.3 Vyhodnotenie správneho riešenia zo získaných anotácií

Pri vyhodnocovaní správneho riešenia sa opierame hlavne o viacnásobné anotácie jednotlivých viet, kde konečné riešenie určujeme agregáciou všetkých dostupných riešení. Pri zbere anotácií je možné nastaviť želanú násobnosť anotácií pre každú vetu. Pochopiteľne, čím viac riešení je k dispozícií, tým lepšie je možné určiť správne riešenie. Vzťah pre určenie vetného členu (resp. vetného skladu) kombináciou všetkých dostupných riešení:

$$element = \max\{\frac{|opt_1|}{|evaluators|}, \frac{|opt_2|}{|evaluators|}, \dots, \frac{|opt_n|}{|evaluators|}\} \qquad (1)$$

$$max > n$$

element – vetný člen priradený slovu po vyhodnotení riešení

$|opt_i|$ – početnosť možnosti i, kde $\{opt_1, \dots, opt_n\}$ predstavuje množinu všetkých riešení, ktoré boli vygenerované pri anotácií.

|*evaluators*| – početnosť všetkých rôznych anotácií získaných pre dané slovo.

n – koeficient, ktorý určuje minimálnu akceptovateľnú percentuálnu početnosť extrahovaného výsledku.

Všetky zozbierané anotácie sa najprv zosumarizujú a pre každé slovo je viacero možností (*opt*) spolu s ich početnosťou. Z množiny týchto možností sa vyberie tá, ktorá je najpočetnejšie zastúpená, avšak zároveň je táto percentuálna početnosť vyššia ako nastavená minimálna hranica (*n*). Ak taká možnosť nie je (všetky sú pod hranicou *n*), tak pre dané slovo neurčíme žiaden vetný člen.

Do úvahy pri určovaní správneho riešenia je ďalej možné započítať expertnosť jednotlivých používateľov. Tento údaj môže byť získaný buď z oficiálneho školského hodnotenia žiaka, jeho subjektívne sebahodnotenie, alebo pomocou vyhodnotenia kontrolnej vzorky viet a úspešnosti žiaka v ich anotácií. Na základe tohto údaju je možné rozdeliť žiakov do viacerých tried hodnotenia, kde každej triede je priradený koeficient, ktorý určuje váhu daného riešenia v rámci početnosti všetkých vzoriek. Početnosť každého z možných riešení (*opt*) bude vypočítaná ako súčet váh žiakov, ktorí to konkrétne riešenie ponúkli.

Na vylúčenie kontroverzných riešení sme pridali ešte ďalšie kritérium. Ak sú percentuálne ohodnotenia pre dve rôzne triedy príliš podobné (minimálna požadovaná veľkosť ohodnotenia majoritnej triedy je určená parametrom *n*, maximálna možná veľkosť ohodnotenia ostatných tried je určená parametrom *e*), tzn. že viaceré parametre prekročia hranicu určenú parametrom *e*, identifikujeme túto úlohu ako problémovú a je potrebné prekontrolovať ju expertom (učiteľom). Týmto spôsobom zaručíme odfiltrovanie anotácií pri ktorých síce trieda (označenie vetného členu) ktorá bola identifikovaná pre dané slovo presiahla hranicu *n*, ale pravdepodobne nastala kontroverzia v riešeniach ostatných žiakov, nakoľko iná trieda presiahla svojím ohodnotením hranicu *e*.

6 Budzogáň – prostredie pre zber syntaktických anotácií v procese výučby

Jedným z cieľov práce je návrh a implementácia softvérového prototypu, v ktorom je možné overiť navrhované princípy. Pre realizáciu metódy navrhnutej v práci sme vytvorili softvérové riešenie, pomocou ktorého je možné zberať syntaktické anotácie pre texty. Je to webová aplikácia s názvom *Budzogáň*[14], ktorá slúži aj ako platforma pre podporu vzdelávania. Podrobná technická dokumentácia k riešeniu sa nachádza v prílohe A. V tejto kapitole sa nachádzajú najdôležitejšie informácie o implementácií. Podrobnejší opis sa nachádza v prílohe A.

6.1 Špecifikácia požiadaviek

Návrh systému sme realizovali za konzultácie expertov v danej oblasti. Nami navrhnuté vlastnosti sme prezentovali učiteľom slovenského jazyka, od ktorých sme dostali cennú spätnú väzbu. Po vzájomných dohovoroch a úpravách požiadaviek sme dospeli k nasledovným vlastnostiam, ktoré by systém mal mať:

Prostredie na tvorbu syntaktického stromu

V tejto časti sa nachádza hlavný nástroj pre riešenie úlohy. Kľúčovým je čo najjednoduchšie rozhranie. Zvolili sme preto systém ovládania pomocou ťahania objektov myšou. Týmto spôsobom si žiak doplní do stromu potrebné elementy (vetné členy, vetné sklady, slová z vety). Je dostupná tiež možnosť úpravy vlastností jednotlivých vetných členov.

Zoznam úloh

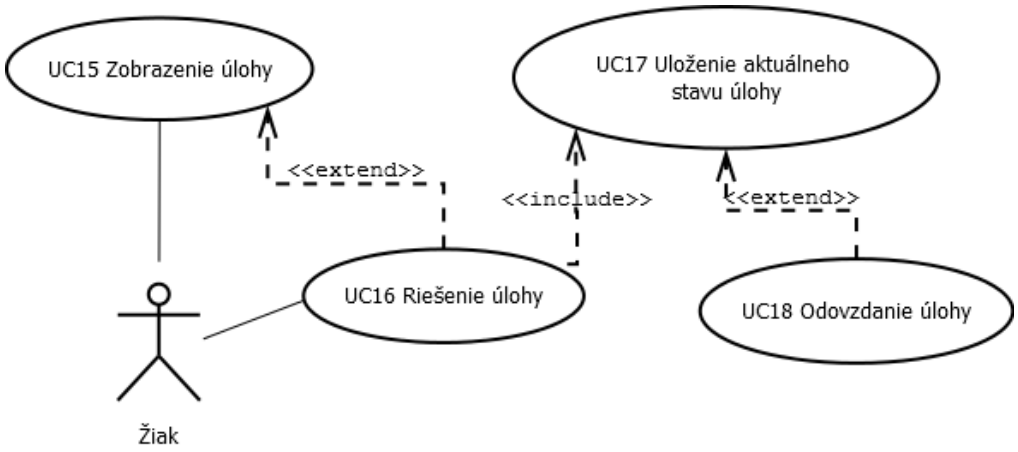
V zozname úloh žiak vidí všetky svoje aktuálne úlohy. Stav úlohy (nová/ rozpracovaná/ odovzdaná) vidí žiak podľa indikátora, ktorý sa nachádza pri každej úlohe.

Prostredie na tvorbu novej úlohy

Prostredie pre tvorbu úlohy sme rozdelili na tri základné časti. V prvej učiteľ zadefinuje vlastnosti úlohy – typ, prvky na ktoré sa má úloha sústrediť, textová špecifikácia zadania. V druhej časti nastaví parametre viet, ktoré sa majú pre danú úlohu použiť. Vety si môže nechať automaticky vygenerovať podľa nastavených pravidiel, prípadne doplniť vlastné. V tretej časti nastaví učiteľ časový plán úlohy a spôsob pridelenia žiakom. Následne sa vygenerujú úlohy pre žiakov.

[14] http://vm36.ucebne.fiit.stuba.sk

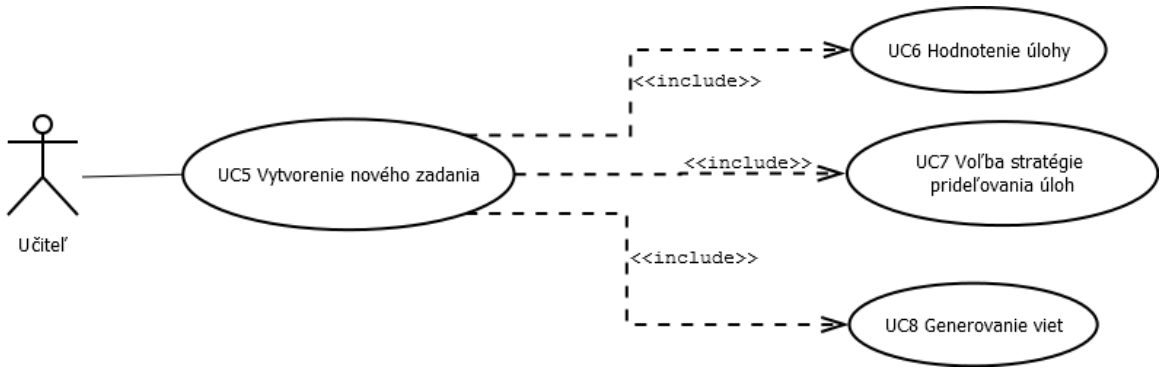
6.1.1 Opis hlavných prípadov použitia



Obrázok 3. Digram prípadov použitia (riešenie úlohy)

Prípad použitia riešenie úlohy (UC16)

Krok	Akcia
1	Otvorenie úlohy.
2	Načítanie aktuálneho stavu úlohy z databázy.
3	Editácia úlohy.
4 (a)	Uloženie rozpracovanej úlohy.
4 (b)	Odovzdanie úlohy.



Obrázok 4. Diagram prípadov použitia (nové zadanie)

Prípad použitia vytvorenie nového zadania (UC5)

Krok	Akcia
1	Zvolenie obrazovky pre definíciu novej úlohy.
2	Výber typu úlohy.
3	Dodefinovanie parametrov zvoleného typu úlohy.

4	Formulácia textu zadania.
5	Pokračovanie na ďalšiu obrazovku.
6	Definovanie vlastností viet, ktoré sa majú použiť (dĺžka, zdroj).
7	Generovanie a editácia vygenerovaných viet.
7 (b)	Doplnenie vlastných viet.
8	Pokračovanie na ďalšiu obrazovku.
9	Výber triedy, ktorá má úlohu dostať.
10	Definovanie obdobia, v ktorom je úloha dostupná.
11	Definovanie stratégie rozdelenia viet.
12	Dokončenie procesu definície novej úlohy.

Prípad použitia vytvorenie pracovnej skupiny

Krok	Akcia
1	Zvolenie obrazovky na vytvorenie novej triedy.
2	Pomenovanie novej triedy/skupiny.
3	Pridanie žiakov do skupiny.
4	Uloženie vytvorenej skupiny.

6.1.2 Identifikované typy úloh na vetnú syntax

V okruhu úloh na precvičenie vetnej syntaxe sme identifikovali niekoľko možných typov cvičení. Na začiatku sa nachádza zoznam akcií, ktoré prostredie má poskytovať, pri každej úlohe sú následne vymenované akcie ktoré je pre jej splnenie potrebné mať k dispozícii.

Každý typ úlohy je popísaný podľa nasledujúcej šablóny:

#ID úlohy	Slovný popis úlohy.
Znenie zadania (príklad):	Príklad ako môže znieť zadanie úlohy
Vstup:	Aké vstupy má žiak k dispozícii pri novej úlohe.
Zoznam akcií:	Zoznam potrebných akcií.

Zoznam akcií

- A1. Potiahnutím myšou a spustením na plátno vytvoriť vetné členy (boxy).
- A2. Potiahnutím myšou a spustením na plátno presunúť slova z vety do boxov.
- A3. Vytvoriť sklady medzi boxami.
- A4. Určiť atribúty vetných členov.
- A5. Určiť atribúty vetných skladov.

U1	Kompletná analýza vety.
Znenie zadania (príklad):	Určte všetky vetné členy a sklady vo vete
Vstup:	Veta.
Zoznam akcií:	A1, A2, A3, A4, A5

U2	Určenie všetkých vetných členov (bez skladov)
Znenie zadania (príklad):	Určte všetky vetné členy vo vete
Vstup:	Veta.
Zoznam akcií:	A1, A2, (A4)

U3	Určenie vybraných vetných členov (bez skladov)
Znenie zadania (príklad):	Nájdite predmet, prísudok a nezhodný prívlastok
Vstup:	Veta, názvy vybraných vetných členov
Zoznam akcií:	A2

U4	Určenie všetkých skladov
Znenie zadania (príklad):	Určte všetky sklady medzi označenými vetnými členmi vo vete
Vstup:	Veta, požadované vetné členy
Zoznam akcií:	A4, A5

U5	Určenie vybraných skladov
Znenie zadania (príklad):	Určte prisudzovací sklad
Vstup:	Veta.
Zoznam akcií:	A1, A2, A3, A4, (A5)

U6	Oprava syntaxe vety
Znenie zadania (príklad):	Nájdi chybu
Vstup:	Veta, určené vetne členy, určené sklady
Zoznam akcií:	A1, A2, A3, A4, A5

6.2 Technické detaily a architektúra systému

Systém Budzogáň je realizovaný ako webová aplikácia implementovaná v rámci na vývoj webových aplikácií *Ruby on Rails*. Tento rámec podporuje tvorbu aplikácie pomocou návrhového vzoru *Model-View-Controller (MVC)*, ktorý zabezpečuje oddelenie aplikačnej

logiky od dát a prezentačnej vrstvy do samostatných logických celkov. Aplikácia Budzogáň je tiež navrhnutá podľa tohto vzoru.

V databázovej vrstve je použitá databáza *PostgreSQL*. Aplikácia využíva javascriptový rámec *JointJS*[15], ktorý slúži na manipuláciu s *SVG*[16] grafikou. To sa využíva pri tvorbe syntaktického stromu vety, v nástroji na editáciu úlohy.

Aplikácia sa skladá z dvoch hlavných častí:

- Prostredie pre žiaka
    - zoznam úloh
    - nástroj na tvorbu syntaktického stromu
- Prostredie pre učiteľa
    - vytvorenie a editovanie tried
    - vytvorenie a priradenie úloh triedam

Vďaka použitiu návrhového vzoru MVC je zdrojový kód aplikácie dobre štruktúrovaný a je jednoduché pridávať nové moduly s funkcionalitou.

6.3 Funkcionalita

Pre systém sme identifikovali tieto základné požiadavky na funkcionalitu:

- Prostredie na tvorbu syntaktickej analýzy vety.
    - Tu sa nachádza nástroj na nakreslenie analyzovanej vety do syntaktického stromu.
    - Môžu tu byť riešené rôzne typy úloh podľa špecifikácie (pozri 6.1.2).
- Administratívne prostredie pre učiteľov na definovanie nových úloh.
    - Učiteľ má možnosť definovať nové úlohy, nastaviť ich parametre a distribuovať ich žiakom.
- Administratívne prostredie pre učiteľov na manažment tried.
    - Učiteľ tu má k dispozícii zoznam žiakov a tried, môže ich vytvárať a editovať.

6.4 Používateľské rozhranie

Používateľské rozhranie sme sa snažili navrhnúť čo najintuitívnejšie, aby bolo jeho používanie čo najjednoduchšie a používateľsky prívetivé. V tejto časti je základný pohľad

[15] http://www.jointjs.com/
[16] Scalable vector graphics - http://www.w3schools.com/svg/

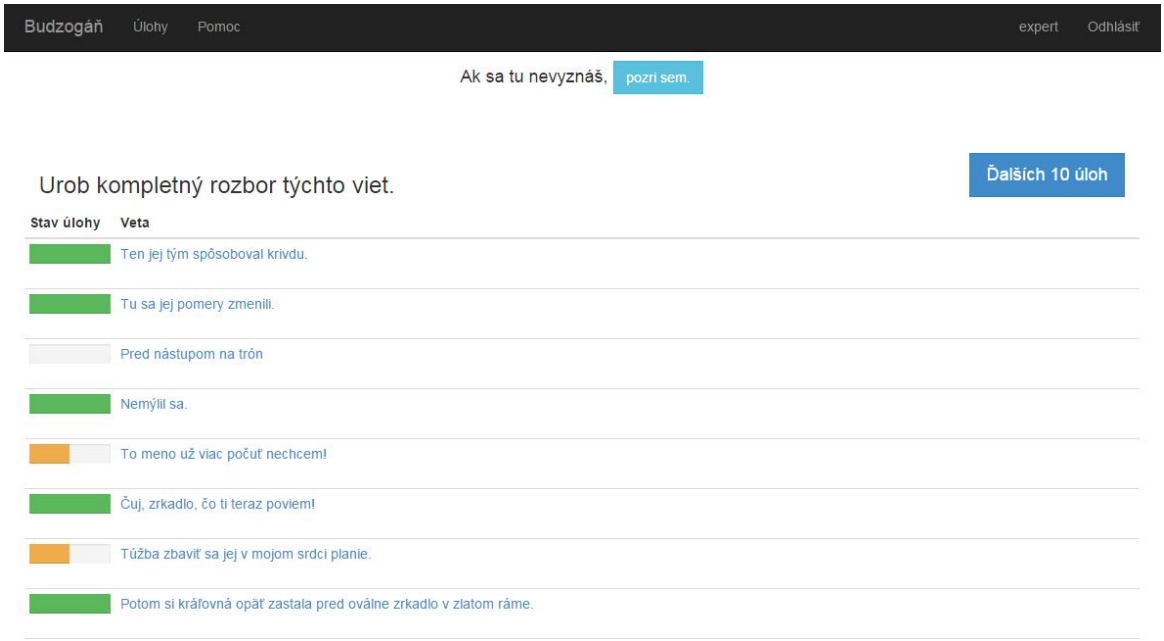
na používateľské rozhranie, všetky funkcie aplikácie sú detailnejšie opísané v prílohe C, kde je používateľská príručka.

Horná lišta – nachádza sa tu navigácia v aplikácií. Položky v lište sú zobrazené v závislosti od role používateľa (žiak/učiteľ)

- Budzogáň – návrat na domovskú stránku (žiak, učiteľ).
- Úlohy – obrazovka so zoznamom úloh pre daného žiaka (žiak)
- Nová úloha – sprievodca na vytvorenie novej úlohy pre triedu (učtiteľ).
- Zoznam tried – zoznam tried, ktoré môže daný učiteľ manažovať (učiteľ).
- Nová trieda – sprievodca na vytvorenie novej triedy (učiteľ).

Obrázok 5. Navigačná lišta

Zoznam úloh (žiak) – žiak tu má zobrazené všetky svoje úlohy. Má ich rozdelené na už vyriešené úlohy a úlohy ktoré ešte nie sú odovzdané, alebo sú nové – ešte neriešené. V úlohách ktoré už odovzdal môže vidieť svoje riešenie aj s hodnotením.



Obrázok 6. Zoznam úloh

Prostredie na tvorbu syntaktického stromu vety (žiak) – tu sa nachádza nástroj na riešenie a editáciu úlohy. Je to prostredie vytvorené prevažne v javascripte, kde je ovládanie myšou, ktorou sa ťahajú potrebné objekty na plátno.

(1) – veta na rozbor

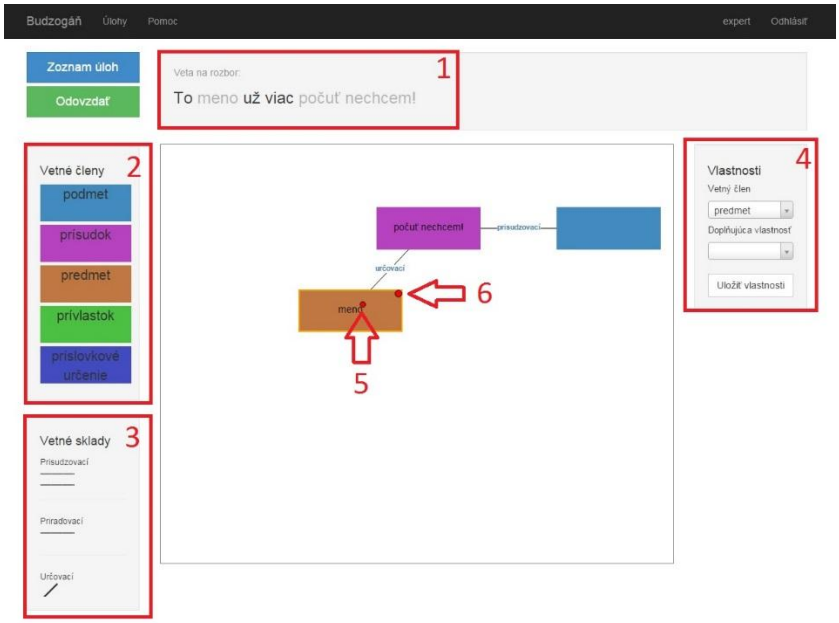
(2) – vetné členy

(3) – vetné sklady

(4) – vlastnosti vetného členu

(5) – zmazanie textu z vetného členu

(6) – zmazanie celého vetného členu



Obrázok 7: Prostredie na tvorbu syntaktického stromu

Prostredie na vytvorenie novej úlohy (učiteľ) – skladá sa z troch obrazoviek (výber typu úlohy – definícia viet pre úlohu – priradenie úlohy žiakom). Podrobnejšie sú opísané v používateľskej príručke.

Prostredie na vytvorenie a editáciu tried žiakov (učiteľ) – zoznam jednotlivých tried pre učiteľa. Triedy môže editovať, alebo vytvárať nové triedy. Pod triedou sa rozumie akákoľvek skupina žiakov, ktorú si učiteľ zadefinuje a ďalej s ňou pracuje pri zadeľovaní úloh.

7 Overenie navrhovanej metódy

Navrhnutú metódu sme overili v dvoch krokoch. Najprv sme realizovali pilotný experiment, v ktorom bolo hlavným cieľom overiť navrhované princípy metódy a získať spätnú väzbu od používateľov na systém. V druhom experimente sme sa zamerali na získanie čo najväčšieho objemu dát, na ktorých sme následne nastavovali parametre pre identifikáciu správneho riešenia a vyhodnocovali úspešnosť metódy.

Pri overovaní navrhnutej metódy sme mali tieto hlavné ciele:

- Overiť navrhovanú metódu na získavanie syntaktických anotácií.
- Overiť jednotlivé aspekty, ktoré sme identifikovali že môžu mať vplyv na správnosť riešenia.
- Analyzovať silu davu žiakov základných a stredných škôl.
- Analyzovať granularitu a kvalitu anotácií, ktoré dokážeme získať.

7.1 Pilotný experiment s kvalitatívnym vyhodnotením

Na overenie konceptu našej metódy bolo zrealizované pilotné testovanie. Na tomto testovaní bolo zúčastnených 20 žiakov základnej školy, ktorí počas jednej vyučovacej hodiny vykonávali syntaktický rozbor viet. Medzi žiakmi sa nachádzali ako výborní, tak i priemerní a podpriemerní žiaci.

V rámci experimentu sme vykonali taktiež kvalitatívne overenie nášho riešenia. Žiaci dostali základné inštrukcie, popis systému, čo je v ňom možné vykonávať. Bolo im povedané, aby prešli cez čo najviac viet a v každej vete vykonali syntaktický rozbor najlepšie, ako sú schopní. Počas experimentu sme sledovali spôsob interakcie žiakov so systémom, všímali si úkony pri vykonávaní ktorých mali problémy. Okrem vykonania syntaktického rozboru žiaci vyplnili aj krátky dotazník. Pri tomto testovaní sme zistili niekoľko nedostatkov používateľského prostredia, ktoré boli pred ďalším testovaním opravené.

Počas experimentu sme mali za cieľ zozbierať dáta – syntaktické anotácie. Pri tomto prvotnom experimente sme získali 186 anotácií, čo v priemere vychádza na 9 anotovaných viet na žiaka.

7.1.1 Použité dáta

Na tento experiment bol použitý súbor viet DATA-TEST.

DATA-TEST boli použité na realizáciu pilotného experimentu. Sú to vety vybrané z učebnice Slovenského jazyka pre ôsmy ročník. V tomto ročníku sa žiaci učia kompletnú

syntaktickú analýzu jednoduchej vety. Vety použité v tomto súbore dát boli ručne anotované učiteľmi slovenského jazyka, pre potreby overenia (zlatý štandard).

7.1.2 Analýza zozbieraných dát

Z týchto dát sme sa snažili odvodiť prvotný vzťah pre identifikáciu správneho, resp. nesprávneho riešenia syntaktickej analýzy vety. Z prvotnej analýzy dát sme zistili, že z tejto vzorky anotácií je možné získať jednotné riešenie problému. V tomto stave sme počítali s pravdou väčšiny, čiže ak je zhoda vo viac ako 50 % anotácií, považujeme toto riešenie za správne.

V rámci pilotného experimentu sme analyzovali úspešnosť pri tvorbe anotácií našou metódou zo štyroch hľadísk – percento správne určených vetných členov (tokenov) všeobecne, percento správne určených vetných skladov všeobecne, percento viet, v ktorých boli všetky vetné členy určené správne, percento viet, v ktorých boli všetky vetné členy aj vetné sklady určené správne (kompletná analýza).

V rámci experimentu sme zozbierali 186 anotácií od 20 žiakov a všetky vety boli anotované aj učiteľom pre potreby vyhodnotenia ako zlatý štandard. Počas experimentu sme zozbierali anotácie pre 11 rôznych viet, z ktorých 8 viet malo počet anotácií viac ako 12. Výsledky experimentu sú uvedené v tabuľke 1.

Tabuľka 1: vyhodnotenie pilotného experimentu

	Percento správnych anotácií
Vetné členy (tokeny) všeobecne	92,68 %
Vetné sklady všeobecne	90,00 %
Vety (iba vetné členy)	85,71 %
Vety (vetné členy + vetné sklady)	71,43 %

Výsledky sme analyzovali zo štyroch hľadísk:

- Vetné členy všeobecne – percento vetných členov určených správne.
- Vetné sklady všeobecne – percento vetných skladov určených správne.
- Vety (iba vetné členy) –percento viet v ktorých sú všetky vetné členy určené správne.
- Vety (kompletne) – percento viet, v ktorých sú všetky vetné členy aj sklady určené správne.

Diskusia

V pilotnom experimente sme zistili, že našou metódou dokážeme získavať korektné anotácie za použitia relatívne malého davu (20 žiakov). Metódy automatickej extrakcie dosahujú úspešnosť 72 – 85 % správne identifikovaných prípadov pri jednotlivých vetných členoch (Ondáš et al. 2011). My sme boli schopní dosiahnuť úspešnosť 92,68 % pri určovaní vetných členov. Aj keď našim cieľom nie je vyrovnať sa automatizovaným metódam, ale skôr určiť kvalitu anotácií ktoré vieme získať za použitia davu. Porovnanie výsledkov však dáva sľubné výsledky a povzbudzuje k ďalšej, podrobnejšej práci.

7.1.3 Kvalitatívny experiment

Pri testovaní sme okrem zbierania dát pre ďalšiu analýzu vykonali aj kvalitatívny experiment. Počas riešenia úloh sme sa rozprávali so žiakmi, pýtali sa ich na prostredie v ktorom majú vykonávať úlohy, aké problémy v ňom vidia. Po ukončení testovania dostali žiaci dotazník, v ktorom sme sa ich pýtali nasledovné otázky:

- Porovnaj si robenie rozboru vety do zošita a robenie rozboru v systéme Budzogáň. Bolo to v Budzogáni rýchlejšie?
- Ak ti rozbor vety trval v Budzogáni dlho, bolo to kvôli zlému prostrediu (napr. pomaly sa s ním pracuje, nevedel/a si sa v ňom zorientovať)?
- Ak ti rozbor vety trval dlho, bolo to kvôli nedostatku tvojich vedomostí (už si to zabudol/zabudla, neučil/a si sa poriadne, ...)?
- Myslíš, že rozbor vety by ti trval kratšie, ak by si mohol/mohla vidieť podobnú vetu aj s riešením ako "žolíka"?
- Motivoval by ťa podobný systém k rýchlejšiemu urobeniu si domácej úlohy?
- Bavilo by ťa viacej robiť úlohy v podobnom systéme viac ako robiť ich na papier?
- Páčil sa ti systém Budzogáň?

Odpovedať na tieto otázky mohli piatimi odpoveďami (podľa vzoru Likertovej škály):

- Určite nie
- Skôr nie
- Neviem
- Skôr áno
- Určite áno

Diskusia

Vyhodnotením dotazníka sme zistili, že žiaci vnímajú myšlienku systému Budzogáň veľmi pozitívne. Otázky zamerané na hodnotenie systému a dojmy žiakov jeho používania

hodnotili systém takmer v 100 % odpovedí pozitívne a riešenie úloh v systéme vnímajú ako jednu z foriem motivácie pre ich rýchlejšie vykonanie, softvérový nástroj je pre nich lákavejší ako manuálne plnenie úloh v zošite.

Na druhej strane žiaci sa vyjadrili neurčito ohľadom rýchlosti plnenia úloh v systéme Budzogáň v porovnaní so zošitom. Časti z nich to šlo rýchlejšie, časti nie. Čas riešenia úloh však identifikovali ako približne rovnaký voči riešeniu v zošite.

Pri návrhu a taktiež aj testovaní systému Budzogáň sme mali tiež neformálne rozhovory s učiteľmi Slovenského jazyka, pri ktorých sme zisťovali ich záujem o podobný systém. Zistili sme, že učitelia sú tejto iniciatíve naklonení. V podstate všetkým sa páčila myšlienka systému, v ktorom môžu žiaci vykonávať úlohy spojené so syntaxou. Motiváciou pre učiteľov je fakt, že týmto spôsobom môžu žiakom zadať viacero úloh a to omnoho jednoduchšie a rýchlejšie ako na papier. Učitelia taktiež vravia, že podobný systém je príťažlivejší aj pre samotné deti, keďže v dnešnej dobe omnoho radšej interagujú s výpočtovými technológiami oproti písaniu do zošita. Takýto systém môže pre nich teda byť lákadlom a motiváciou. Toto bolo v podstate potvrdené dotazníkom po experimente na malej vzorke žiakov, kde 100 % žiakov odpovedalo, že s Budzogáňom by ich viacej bavilo urobiť si úlohu ako ju písať do zošita. Od učiteľov sme teda získali veľmi pozitívnu spätnú väzbu na náš systém. Je však potrebné spomenúť, že pre učiteľov je tiež podstatné, aby bol systém zjednodušením aj pre nich. Tvorba nových úloh by preto mala byť čo najjednoduchšia a pokiaľ možno najrýchlejšia. Pomocou dotazníka sme od žiakov zistili, že systém by privítali vo výučbe a motivoval by ich k rýchlejšiemu splneniu si domácich úloh. Vyskytli sa aj niektoré prípady, kedy povedali, že mali problém s používateľským prostredím. Tieto pripomienky sme zapracovali v ďalšej verzii systému (niekedy chybná definícia hraníc plátna na ktoré sa kreslí graf, problém pri ťahaní elementov).

Pri diskusii o používateľskom rozhraní nám bolo povedané, že by malo byť čo najjednoduchšie na použitie, aby ušetrilo čas v porovnaní s kreslením syntaktického stromu do zošita. Je preto potrebné zamerať sa na čo najprívetivejšie používateľské prostredie ako jeden z motivačných faktorov.

Pri kvalitatívnom experimente sme zistili, že systém Budzogáň je medzi žiakmi aj učiteľmi vnímaný veľmi pozitívne. Vďaka dotazníku sme zistili niektoré chyby v systéme, ktoré bolo potrebné opraviť aby bol dosiahnutý lepší používateľský zážitok.

7.2 Druhý experiment

Po overení konceptu nášho riešenia sme navrhli druhú časť overenia našej práce. Pripravili sme množinu 180 viet, ktoré sme distribuovali žiakom základných a stredných škôl. V tomto experimente sme získavali čo najväčší objem dát, z ktorých sme následne určovali automatizovane správne syntaktické anotácie a analyzovali ich z viacerých hľadísk.

Oslovili sme viacero škôl, ktorých žiaci vykonávali syntaktické anotácie. Väčšina žiakov sa zúčastnila na experimente v školskom prostredí. Ostatní vykonávali anotovanie doma, ako precvičenie si svojich vedomostí. Návrh a nastavenie tohto experimentu sú opísané v nasledovnej časti.

Na začiatku sa každý žiak zaregistroval a vyplnil o sebe základné informácie. Tieto údaje sme následne použili pri vyhodnocovaní výsledkov experimentu na rozdelenie jednotlivých anotátorov do skupín podľa rôznych kritérií. Od žiakov sme chceli nasledovné údaje:

- Ich vek (ročník v škole)
- Sebahodnotenie (známka ako v škole)
- Pohlavie
- Na akej škole študuje (základná, gymnázium)

7.2.1 Použité dáta

Pri experimentoch sme použili dátovú sadu zostavenú z viacerých vzoriek. Dáta sú dostupné na dátovom nosiči v adresári data/:

- DATA-SNK – tieto dáta sa na nosiči nenachádzajú, nakoľko ich licencia nepovoľuje šíriť.
- DATA-NEWS
- DATA-SELECTION

DATA-SNK je syntakticky anotovaný korpus, ktorý máme k dispozícii od Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied (JÚĽŠ-SAV). Vzorka korpusu, s ktorou sme pracovali obsahuje 11 537 viet. Je tu zastúpený beletristický a encyklopedický štýl. Tento korpus bol ručne anotovaný expertami zo Slovenskej akadémie vied. Vety z neho vybrané mali teda zaručene správne syntaktické anotácie, avšak v mierne odlišnej forme ako sme potrebovali pre naše účely. Napr. vzťahy medzi jednotlivými vetnými členmi nie sú explicitne definované, museli sme ich doplniť vo fáze predspracovania na základe toho ktoré vetné členy navzájom spájali.

DATA-NEWS sú vety vybrané z článkov novín denníkN[17]. Tieto vety sme vybrali preto, aby v bol v našich dátach zastúpený aj publicistický štýl. Je tu vzorka rôznych typov viet vyskytujúcich sa v novinárskych článkoch. V tejto dátovej sade je 30 viet. Keďže tento súbor dát bol manuálne zostavený z textov na webe, neexistovali preň syntaktické anotácie, ktoré by bolo možné použiť ako zlatý štandard na overenie výsledkov našej metódy. Vety sme preto nechali anotovať expertom (učitelia slovenského jazyka), aby sme získali zaručene správne syntaktické anotácie pre potreby overenia.

DATA-SELECTION súbor dát obsahujúci 180 viet vybraných z vyššie popísaných súborov dát (DATA-TEST, DATA-SNK, DATA-NEWS). Tieto vety boli vybrané pre potreby overenia navrhovanej metódy.

7.2.2 Metriky použité na vyhodnotenie výsledkov

Pri vyhodnocovaní získaných výsledkov sme použili viaceré metriky. Na ohodnotenie určenia správnosti jednotlivých vzťahov medzi slovami sme použili metriky LAS a UAS (Green 2011).

LAS (*angl. Label Attachment Score*) – percento tokenov, pre ktoré bol správne určený rodič a typ väzby na rodiča.

UAS (*angl. Unlabeled Attachment Score*) – percento tokenov pre ktoré bol správne určený rodič.

Na vyhodnotenie správnosti určenia vetných členov sme použili metriky *Presnosť (angl. Presicion),Úplnosť (angl. Recall)* a *F-score.*

Precision – určuje koľko zo získaných výsledkov bolo relevantných (v našom prípade správne určených vetných členov)

$$Precision = \frac{tp}{tp + fp}$$

Recall – určuje koľko percent vetných členov sme dokázali správne určiť.

$$Recall = \frac{tp}{tp + fn}$$

tp – získané výsledky, ktoré sú relevantné

fp – získané výsledky, ktoré nie sú relevantné

[17] http://www.dennikn.sk

fn – relevantné výsledky, ktoré neboli identifikované ako relevantné

F-score – kombinuje tieto dve metriky (*Precision a Recall*) ako ich harmonický priemer.

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Pri overovaní výsledkov sme použili k-násobnú krížovú validáciu, kde k = 3.

7.2.3 Realizácia experimentu

Počas experimentu sme zozbierali 2 210 anotácií viet od 226 žiakov základných a stredných škôl z rôznych častí Slovenska. Týmto sme sa snažili zaručiť pokrytie čo najrôznorodejšej vzorky žiakov.

Počas experimentu bola nastavená hranica počtu anotácií na jednu vetu na 40 anotácií na vetu. Táto hranica sa využívala pri generovaní viet, kedy sa medzi žiakov rozdeľovali vety podľa nasledovného kľúča:

Každému žiakovi sme generovali vety troch rôznych typov. Na základné rozlíšenie zložitosti viet sme ako metriku použili dĺžku viet. Toto sme vykonali na základe predpokladu, že dĺžka vety súvisí s jej náročnosťou na vetný rozbor. Základné vetné členy, ktoré sa žiaci učia, musia byť najprv zastúpené v každej vete, aby k nim bolo možné pridať zložitejšie vetné členy, ktoré ich rozvíjajú. Takisto, súvetia, ktoré sú náročnejšie na rozbor, sú nutne dlhšie ako jednoduché vety. Na základe tejto metriky sme teda vety rozdelili do troch kategórií:

- Krátke vety: 1 – 5 slov
- Stredne dlhé vety: 6 – 15 slov
- Dlhé vety: viac ako 15 slov

Aby sme si overili schopnosť žiaka vykonať syntaktickú anotáciu základných viet, pri vygenerovaní prvej množiny viet sa žiakom ako prvé 4 vety vygenerujú testovacie vety, ktoré pokrývajú základné časti syntaktickej analýzy. Na týchto vetách sme si overili schopnosť žiaka anotovať vety.

Každému žiakovi sme následne generovali dávky s počtom 10 viet, kde boli zastúpené dlhé - stredne dlhé - krátke vety v pomere 3 - 3 - 4. Žiaci mali za úlohu vykonať čo najviac rozborov viet, nemali určený žiaden limit, ani minimálny, ani maximálny.

7.2.4 Vyhodnotenie experimentu

Zozbierané dáta sme analyzovali s viacerých hľadísk. V nasledujúcej časti sa nachádzajú výsledky podľa jednotlivých kritérií. Vyhodnotenie je realizované za použitia vyššie opísaných metrík, používaných na ohodnocovanie výsledkov syntaktickej analýzy (LAS, UAS, Precision, Recall).

Kritériá:

- Hodnotenie žiaka
- Základná/stredná škola
- DĺžkA viet
- Čas riešenia úlohy

Hodnotenie žiaka

V tomto vyhodnotení sme brali do úvahy sebahodnotenie žiakov. Žiaci pri registrácií ohodnotili svoje vedomosti zo slovenského jazyka na stupnici 1 – 5, kde 1 znamená najlepší, 5 najhorší. Pri vyhodnocovaní sme skúšali podľa zaradenia do jednotlivých tried hodnotenia váhovať riešenia jednotlivých žiakov. Experimentálne sme určili koeficienty pre jednotlivé známky, ktoré sú v tabuľke 2. Koeficienty sme určili tak, že sme najprv vyhodnotili úspešnosť žiakov v jednotlivých triedach a na základe ich maximálnej dosiahnutej presnosti sme odstupňovali rozdiel medzi známkami hodnotenia. Keďže sme chceli zvýšiť váhu anotácií lepších žiakov, pri hodnotení žiaka 1 (najlepší) sme začali s váhou 1,5 a postupne ju znižovali na základe už vyššie spomenutého princípu. K hodnote 1,5 sme prišli tak, že sme skúšali koeficienty v intervale <1, 2> a na základe najlepších výsledkov sme vybrali hodnotu 1,5.

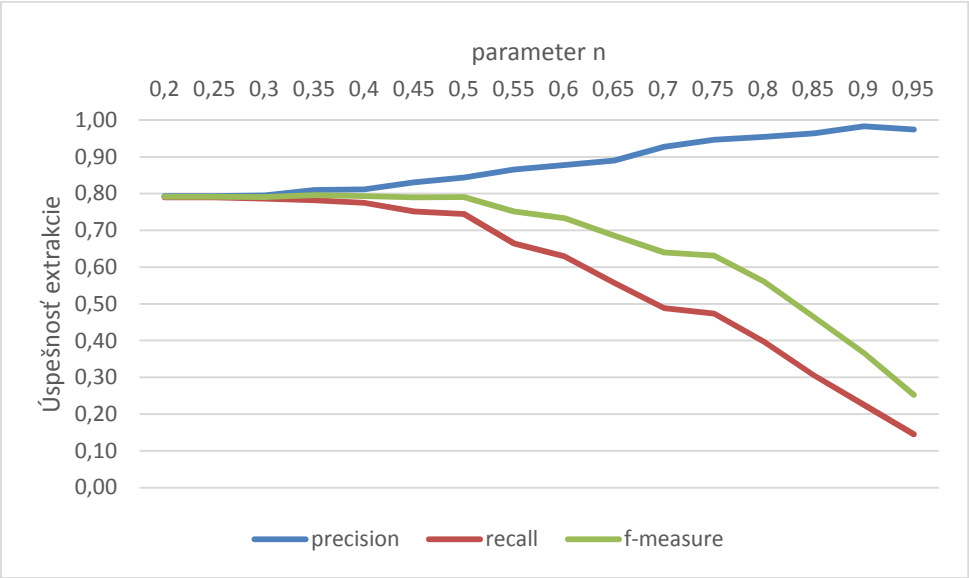
Tabuľka 2. Koeficienty podľa hodnotenia žiaka

Známka	**Koeficient**
1	1,50
2	1,39
3	1,33
4	0,72
5	0,44

Pri následnej agregácií dát sme aplikovali koeficienty na riešenia žiakov. V grafoch nižšie je možné vidieť rozdiel medzi vyhodnotením pred použitím koeficientov a po použití koeficientov.

V tejto časti sú zobrazené výsledky, ktoré sme dosiahli pri rozdelení podľa hodnotenia žiakov. Snažili sme sa reflektovať skutočnosť, že žiaci s lepšími výsledkami majú väčšiu pravdepodobnosť poskytnutia správneho riešenia ako žiaci s horším hodnotením, preto by vplyv riešení jednotlivých skupín žiakov mal byť odlíšený.

Najlepšie výsledky vychádzali pri nastavení parametra n = 0,5 (viď vzťah 1) ak sme nebrali do úvahy váhovanie koeficientom podľa hodnotenia žiaka. Ak sme počítali aj s rôznymi váhami podľa hodnotenia, pri n = 0,75 sme dosiahli najvyššiu hodnotu *f-measure*. V praxi to znamená zvýšenie presnosti (*precision)* s menšou stratou úplnosti (*recall)*. Priebeh úspešnosti so zmenou parametra je vidieť v Graf 1 a 2.



Graf 1: Úspešnosť extrakcie vetných členov bez váhovania študentov

Graf 2. Úspešnosť extrakcie vetných členov s váhovaním študentov podľa hodnotenia

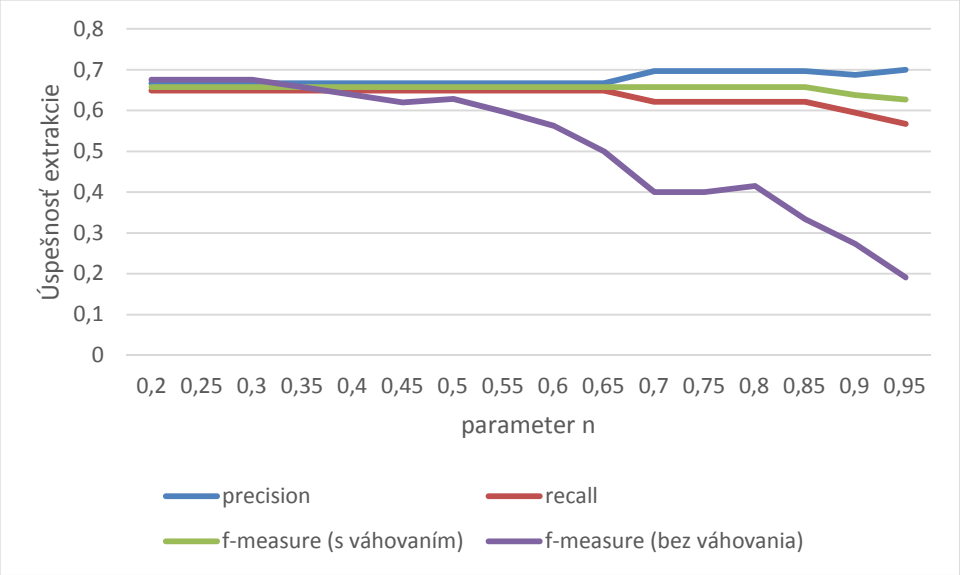
Úspešnosť extrakcie vzťahov medzi vetnými členmi (vetné sklady) je zobrazená v tabuľke nižšie. V tomto nastavení sme analyzovali určovanie vetných skladov nezávisle od dĺžky viet, teda úspešnosť v celom korpuse.

Tabuľka 3. Úspešnosť extrakcie vetných skladov

Parameter n	**LAS - bez váhovania (%)**	**LAS – s váhovaním(%)**
0,2	21,59	38,63
0,25	15,91	27,84
0,3	6,82	20,45
0,35	5,68	15,34
0,4	3,98	9,66
0,45	1,70	5,68
0,5	1,14	5,11

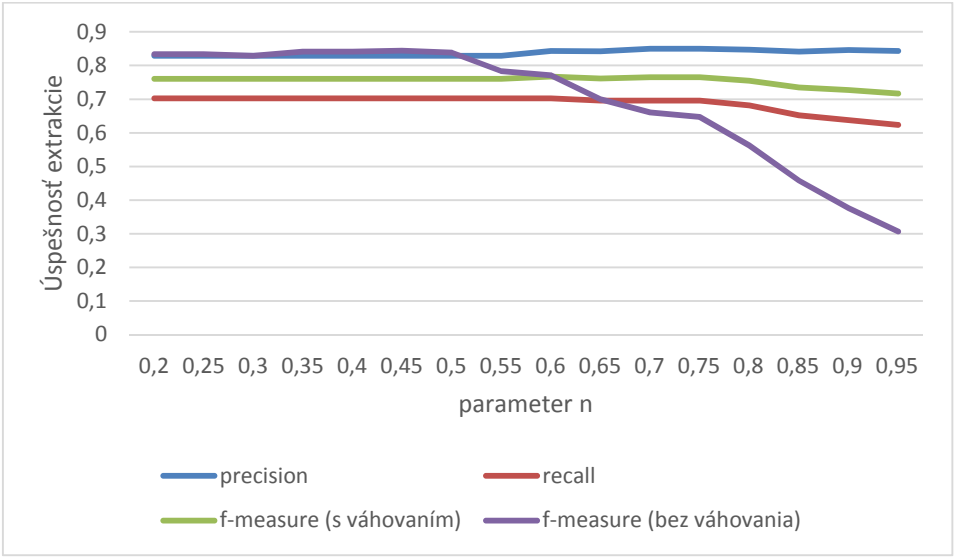
Dĺžka viet

Pri vyhodnocovaní tohto kritéria sme zisťovali úspešnosť s akou dokážu žiaci anotovať vety s rozličnou dĺžkou, resp. náročnosťou. Analyzovali sme úspešnosť určenia správnych anotácií pre tri triedy dĺžky viet, ktoré sme opísali vyššie (krátke, stredne dlhé a dlhé vety).



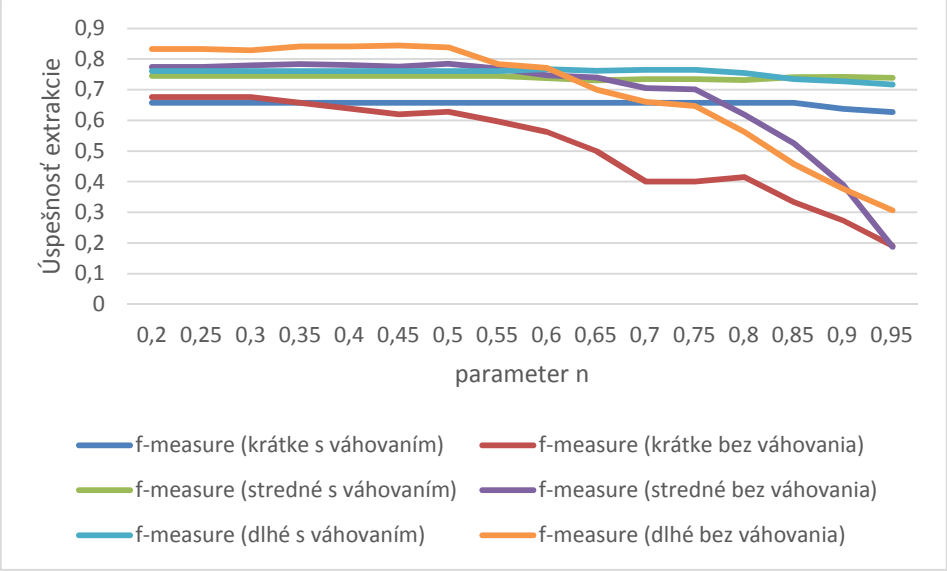
Graf 3: Úspešnosť extrakcie vetných členov - krátke vety

Graf 4: Úspešnosť extrakcie vetných členov - stredne dlhé vety



Graf 5: Úspešnosť extrakcie vetných členov - dlhé vety

Z výsledkov môžeme vidieť, že pri určovaní vetných členov, nemá dĺžka vety signifikantný vplyv na úspešnosť určenia vetných členov. Pripisujeme to skutočnosti, že žiaci dokážu určiť vetné členy nezávisle od typu vety (jednoduchá veta/ súvetie).



Graf 6. Úspešnosť extrakcie vetných členov - porovnanie podľa dĺžky viet

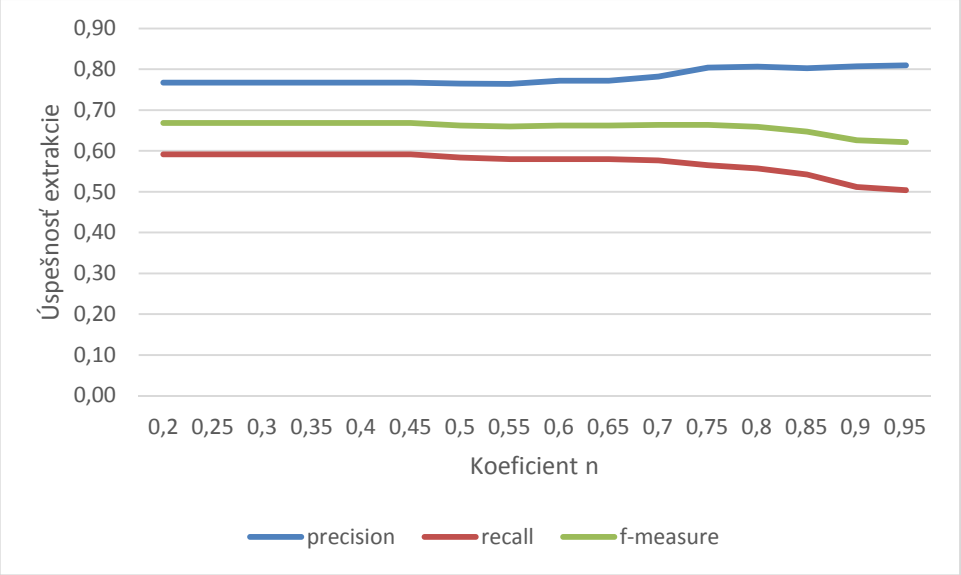
Iná je situácia pri vetných skladoch, kde žiaci dosahovali najvyššiu úspešnosť pri krátkych vetách. Tu je vidno že dĺžka vety naozaj naznačuje jej zložitosť na syntaktickú analýzu z hľadiska vzťahov medzi slovami. Výsledky extrakcie vetných skladov sú v tabuľke nižšie.

Tabuľka 4. Úspešnosť extrakcie vetných členov podľa dĺžky vety

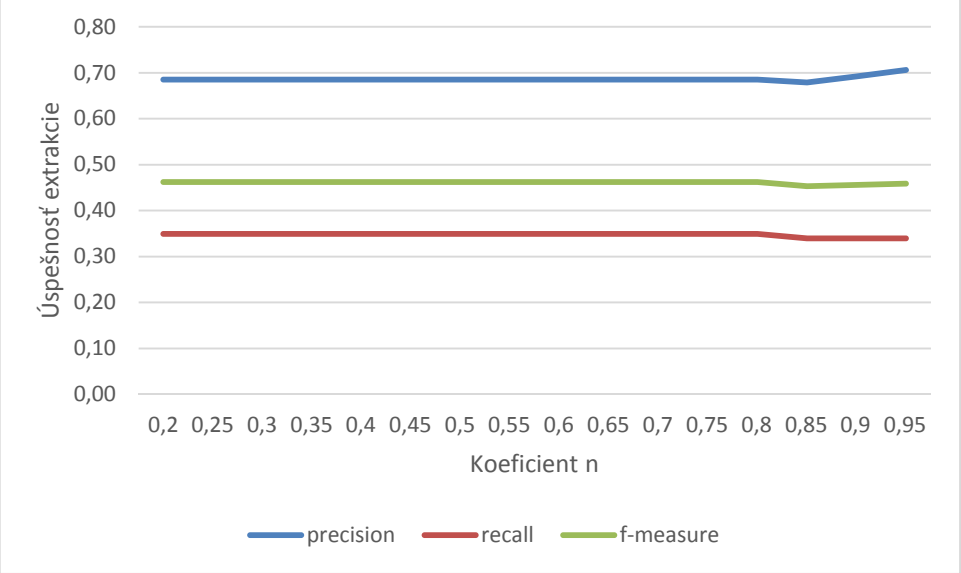
Parameter n	UAS – krátke vety (%)	UAS – stredné vety (%)	UAS – dlhé vety (%)	LAS – krátke vety (%)	LAS – stredné vety (%)	LAS – dlhé vety (%)
0,2	88,89	53,33	40,16	66,67	28,89	36,07
0,25	88,89	35,56	27,87	66,67	22,22	23,77
0,3	77,78	28,89	20,49	66,67	20,00	16,39
0,35	66,67	22,22	14,75	55,56	13,33	12,30
0,4	66,67	8,89	11,47	44,44	6,67	8,20
0,45	66,67	6,67	3,28	44,44	6,67	2,46
0,5	66,67	6,67	2,46	44,44	6,67	1,64

Základná vs. stredná škola

Pri analýze tohto kritéria sme sa snažili porovnať úspešnosť študentov základných a stredných škôl. Na základných školách sa totižto žiaci učia základy vetnej syntaxe, na strednej škole toto učivo najmä opakujú a sčasti prehlbujú o ďalšie poznatky. Týmto pohľadom na dáta sme chceli zistiť, či žiaci stredných škôl dokážu vytvoriť kvalitnejšie anotácie ako žiaci základných škôl.



Graf 7. Úspešnosť získavania vetných členov za využitia žiakov základných škôl



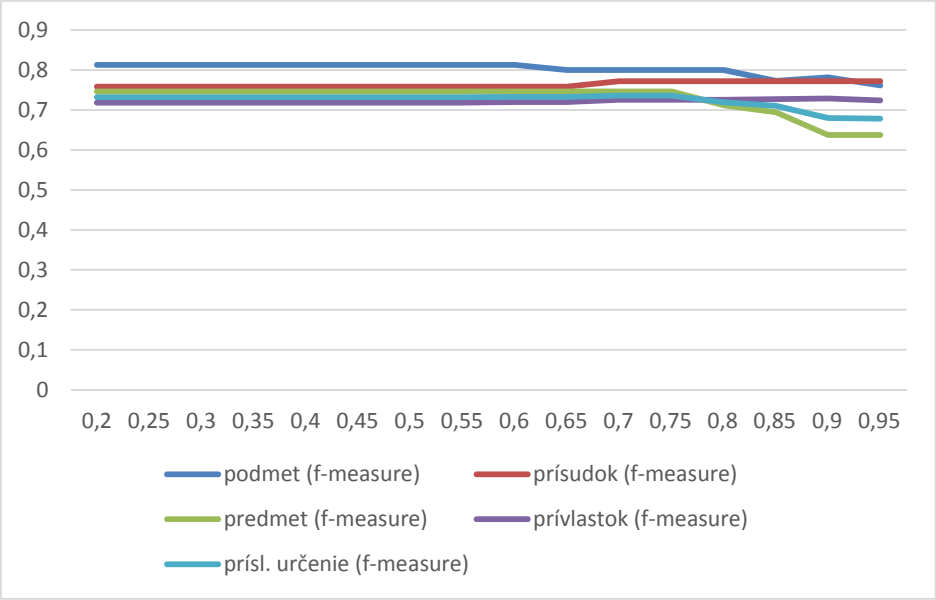
Graf 8 Úspešnosť získavania vetných členov za využitia žiakov stredných škôl

Z výsledkov je vidieť, že náš predpoklad bol chybný. Žiaci základných škôl boli schopní vytvoriť kvalitnejšie dáta na extrakciu vetných členov. Môže to byť spôsobené tým, že žiaci základných škôl preberajú oblasť syntaktickej analýzy intenzívnejšie a sú teda lepšie pripravení na syntaktický rozbor.

Vyhodnotenie podľa jednotlivých vetných členov

Na dáta sme sa pozreli tiež z hľadiska jednotlivých vetných členov. Skúmali sme, s akou úspešnosťou je možné získať správne anotácie pre podmet, prísudok, predmet, prívlastok a príslovkové určenie. Ako najzákladnejšie vnímame určenie vetného základu – podmet a prísudok. Úspešnosť s akou sme dokázali extrahovať jednotlivé vetné členy je vidieť v grafe nižšie.

Ako je vidieť v grafe 8, najlepšie výsledky dosahuje vetný základ – podmet a prísudok, ako sme aj predpokladali. Najspoľahlivejšie teda vieme určiť tieto vetné členy, podobné výsledky dosahujú aj automatizované metódy syntaktickej analýzy.



Graf 9. Úspešnosť extrakcie vetných členov - rozdelenie podľa jednotlivých vetných členov

Čas riešenia úlohy

Na tento atribút úloh sme sa pozreli z hľadiska vyhodnotenia použiteľnosti nášho systému. Je totižto potrebné, aby čas riešenia úlohy v systéme Budzogáň nebol oveľa dlhší ako čas riešenia úlohy na papier. V tabuľke 5 sa nachádzajú priemerné časy riešenia úloh rozdelené podľa jednotlivých skupín žiakov (rozdelenie podľa hodnotenia žiakov). Tento čas vyjadruje, koľko času žiak s daným hodnotením priemerne strávil na jednej úlohe.

Tabuľka 5: Priemerný čas riešenia úlohy podľa hodnotenia

Hodnotenie	Čas riešenia úlohy
1	1 min 56 s
2	2 min 11 s
3	1 min 27 s
4	2 min 4s

Z diskusie s učiteľmi a žiakmi sme zistili, že čas riešenia úlohy na papier je približne rovnaký ako v systéme Budzogáň. Z tohto hľadiska teda systém nepredstavuje prekážku. Identifikovali sme však niektoré vylepšenia, ktoré by mohli ešte viac zrýchliť proces anotácie:

- Určovanie vetného členu spôsobom – kliknem na slovo vo vete a stlačím číslo prislúchajúce vetnému členu – akcia sa vykreslí na plátne.

7.2.5 Diskusia

V grafoch výsledkov experimentov je vidieť, že s menšími hodnotami prahu minimálnej zhody (parameter *n* zo vzťahu 1) je síce menšia presnosť (*precision*), ktorá sa postupne zvyšuje, ale hodnota úspešnosti (*recall)* má klesajúcu tendenciu. Kombináciou týchto dvoch metrík sme dostali hodnotu *f-measure*, ktorá lepšie vyjadruje ich vzájomný priebeh. Pri experimentoch sme zistili, že najlepšie výsledky sa dosahujú, keď je miera zhody okolo 50 % bez použitia váhovania. Avšak ak využijeme informáciu o hodnotení jednotlivých žiakov a váhujeme ich výsledky, vieme dosiahnuť vyššiu presnosť (*precision)* bez straty úplnosti (*recall*). Najlepšie je to pri hodnote n = 0,75. Vtedy je vzájomná kombinácia presnosti a úspešnosti najlepšia. Výsledky každého čiastkového experimentu boli podrobnejšie vysvetlené pri zhrnutí jednotlivých experimentov. Kompletné výsledky experimentov sa nachádzajú na dátovom nosiči v súbore *data/výsledky_experimentov.xlsx*.

Úspešnosť extrakcie vetných členov zo zozbieraných dát od žiakov dosahuje pri použití našej metódy úspešnosť podľa metrík precision a recall hodnoty 0,80 (precision) a 0,69 (recall) pri *f-measure* = 0,74. Znamená to, že 80 % vetných členov, ktoré sme dokázali identifikovať, bolo určených správne a pokryli sme 74 % všetkých vetných členov. Automatizované metódy dosahujú úspešnosť v rozmedzí 72 – 85 % správne určených vetných členov. Vieme sa teda dostať na podobnú úroveň, bez potreby použitia podkladových korpusov na trénovanie.

Extrakcia vzťahov medzi jednotlivými vetnými členmi dosahovala úspešnosť podľa metriky LAS 66,67 % pri krátkych vetách, 28,89 % pri stredne dlhých vetách a 36,06 %

pri dlhých vetách. Štandardné automatizované metódy dosahujú úspešnosť podľa tejto metriky okolo 80 %. Pri krátkych vetách sa teda vieme našou metódou priblížiť výsledkami k iným referenčným metódam. Výsledky metód založených na strojovom učení úzko súvisia s kvalitou korpusu, na ktorom boli natrénované. Výhodou našej metódy je, že nepotrebuje korpus anotovaný na natrénovanie. Avšak relatívne úspešne dokážeme získavať vzťahy medzi slovami iba pre krátke vety.

Celkovo sme zistili, že navrhovanou metódou, ktorá využíva dav základných a stredných škôl je možné úspešne získať syntaktické anotácie pre základné vetné členy (podmet, prísudok) a vzťahy medzi nimi (prisudzovací sklad). Pri ostatných vetných členoch je úspešnosť nižšia. Zistili sme, že žiaci sú schopní vytvoriť anotácie s granularitou zodpovedajúcou tomu, čo sa v škole učia. Je to určenie vetných členov a s menšou úspešnosťou aj vzťahy medzi nimi. Kvalita anotácií je najväčšia pri vetnom základe – podmet, prísudok a pri ostatných vetných členoch klesá.

8 Zhrnutie

V práci sa venujeme spracovaniu prirodzeného jazyka za využitia metód čerpania z davu. Práca sa zameriava na získavanie syntaktických anotácií rozsiahlych textov v slovenskom jazyku. V súčasnosti sa syntaktickej analýze slovenského jazyka nevenuje veľa prác, preto má táto oblasť stále veľký potenciál pre výskum. Jedným z problémov syntaktickej analýzy slovenského jazyka je nepravidelnosť vetnej skladby slovenčiny, ďalej je to homonymia, ktorú nevieme bez kontextu vety rozoznať, atď. Identifikovali sme, že na riešenie týchto problémových faktorov môže byť vhodný ľudský faktor pri riešení problému. Naskytuje sa však otázka, ako zabezpečiť ľuďmi asistovanú syntaktickú anotáciu tak, aby bola efektívna. Tomu sa práve venuje oblasť čerpania z davu, ktorú sme takisto analyzovali. Rozhodli sme sa preto pre spojenie týchto dvoch oblastí – spracovanie prirodzeného jazyka a čerpanie z davu.

Našu metódu sme definovali zasadením do jednotlivých dimenzií čerpania z davu, ktorými sa navzájom odlišujú jednotlivé prístupy. V rámci týchto dimenzií sú opísané kľúčové aspekty našej metódy. Žiaci v školách riešia v rámci vyučovania slovenského jazyka aj úlohy zamerané na vetnú syntax. Tieto úlohy musia splniť aj tak a preto sme sa rozhodli využiť túto skutočnosť a poskytnúť žiakom nástroj v ktorom je možné vykonávať syntaktickú analýzu vety, precvičiť si preberané učivo a zároveň, čo je pre nás najhlavnejšie, týmto spôsobom sa buduje syntakticky anotovaný korpus, ktorý je možné použiť v ďalších úlohách spracovania prirodzeného jazyka.

V softvérovom prototype sme vykonali prvé testovanie, ktorého účelom bolo overiť koncept metódy, získať spätnú väzbu od učiteľov a žiakov a prvotné dáta z ktorých môžeme odvodiť ďalšie smerovanie práce. Nastavenie experimentu bolo nasledovné: 20 žiakov malo k dispozícii 20 viet počas jednej vyučovacej hodiny. Ich úlohou bolo vykonať pokiaľ možno najviac anotácií viet. V priemere za tento čas stihol každý žiak vykonať 9 anotácií, dokopy sme získali 186 vzoriek anotovaných viet. Aby sme mali k dispozícii aj zlatý štandard, učiteľ tiež vykonal anotácie týchto viet. V tomto experimente sme dokázali určiť kompletné syntaktické anotácie viet v 71,43 % viet. Vetné členy sme dokázali určiť s úspešnosťou 92,68 % a vzťahy medzi vetnými členmi sme určili správne v 90,00 % prípadov. Z týchto výsledkov vidíme, že získavanie syntaktických anotácií je možné aj pri nižšom počte žiakov, ktorí vetu anotovali.

Vykonali sme tiež kvalitatívne overenie, kde sme zistili, že naše riešenie je medzi učiteľmi a žiakmi vítané, vnímajú ho ako motiváciu na rýchlejšie splnenie úloh zadaných v škole. Po vykonaní anotácií dostali žiaci za úlohu vyplniť dotazník, v ktorom sme zisťovali ich vnímanie niektorých kľúčových prvkov systému. V zásade sme si overili naše predpoklady

a získali cennú spätnú väzbu k vylepšenia a opravám prostredia na tvorbu syntaktického stromu. Žiaci sa zhodli, že systém pôsobí pozitívne na ich motiváciu, plniť úlohy elektronicky ich láka viac, ako v zošite. Čo sa týka času plnenia úlohy v systéme Budzogáň, zistili sme, že systém nezrýchlil proces vykonávania rozboru vety, avšak systém má potenciál ušetriť čas učiteľovi pri kontrole úloh, kde učiteľ pri kontrole úloh v zošite môže stráviť aj polovicu vyučovacej hodiny, ktorú by mohol využiť inak (viď časť *Východiská pre metódu*).

V druhom experimente sme zozbierané dáta analyzovali z viacerých hľadísk. Vyhodnotili sme úspešnosť získavania anotácií v závislosti od dĺžky analyzovaných viet, zakomponovali sme do ohodnotenia riešení aj hodnotenia žiakov a snažili sa tak váhovať jednotlivé riešenia.

Očakávali sme, že na kvalitu anotácií vplýva hodnotenie žiakov z daného predmetu. Experimentom sme tento predpoklad potvrdili. Síce sme nedosiahli lepšie hodnoty *f-measure,* ale zlepšenie spočívalo v tom, že priebeh hodnôt *f-measure* s narastajúcim minimálnym prahom zhody (parameter *n*) bol iba mierne klesajúci. To znamená že pri zvyšujúcej sa presnosti (*precision*) sa neznižovala úplnosť (recall) tak výrazne ako pred použitím váhovania (porovnaj grafy 1 a 2).

Využitím davu študentov základných a stredných škôl nevieme získať syntaktické anotácie takej kvality, ktorá by bola použiteľná pri zložitejších úlohách spracovania prirodzeného jazyka. Tieto sa kvalitou nevyrovnajú anotáciám od lingvistov a teda nie je možné použiť ich na lingvistické úlohy. Avšak nami vytvorené anotácie je možné použiť na podporu úloh strojového spracovania prirodzeného jazyka, ktoré sú opísané v kapitole 2.3.

Keď porovnáme anotácie ktoré dokážeme vytvoriť našou metódou s anotáciami vytvorenými manuálne odborníkmi, tak naša metóda dokáže pokryť základné vetné členy (podmet, prísudok, predmet, prívlastok, príslovkové určenie). Lingvistický korpus obsahuje ešte aj iné, zložitejšie typy anotácií, avšak tie je nemožné získať použitím davu žiakov základných a stredných škôl. Čo sa týka vzťahov medzi jednotlivými slovami, vieme určiť tie, ktoré spájajú vyššie spomenuté vetné členy – sú to tri základné typy väzieb vyučované na školách (prisudzovací, priraďovací a určovací sklad).

Anotácie vytvorené našou metódou je možné použiť napríklad na natrénovanie automatizovaného syntaktického analyzátora, ktorý využíva strojové učenie. Exitujú princípy, kde sa klasifikátor postupne zdokonaľuje postupným trénovaním na viacerých množinách dát a práve nami vytvorený korpus môže poslúžiť na tento účel.

Ako možné ďalšie smerovanie práce navrhujeme zapojiť do anotovania textov aj študentov vysokých škôl, ktorý študujú lingvistické odbory. Títo študenti sa môžu považovať temer za odborníkov a je teda pravdepodobnosť získania anotácií vyššej kvality.

V rámci rozsahu práce sme nemali príležitosť nasadiť systém Budzogáň v plnom rozsahu ale iba v experimentálnom nastavení, kde sme sa zamerali na zber syntaktických anotácií. Avšak implementovali sme aj prostredie pre učiteľa, v ktorom je možné nadefinovať parametre úlohy pre triedy, prideliť úlohy a manažovať skupiny žiakov. Praktické nasadenie v školách a prípadné rozšírenie platformy o ďalšie úlohy na vetnú syntax môže byť predmetom ďalšieho smerovania projektu.

Zdroje

AMBATI, Vamshi, Stephan VOGEL a Jg CARBONELL, 2010. Active Learning and Crowd-Sourcing for Machine Translation. *Lrec* [online]. 2010, s. 2169–2174. Dostupné na: doi:10.1.1.164.9485

BERNSTEIN, Michael S, Greg LITTLE, Robert C MILLER, Björn HARTMANN, Mark S ACKERMAN, David R KARGER, David CROWELL a Katrina PANOVICH, 2010. Soylent: a word processor with a crowd inside. V: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. s. 313–322.

BUCHHOLZ, Sabine a Erwin MARSI, 2006. CoNLL-X shared task on multilingual dependency parsing. V: *Proceedings of the Tenth Conference on Computational Natural Language Learning*. s. 149–164.

CALLISON-BURCH, Chris a Mark DREDZE, 2010. Creating speech and language data with Amazon's Mechanical Turk. V: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. s. 1–12.

CIMIANO, Philipp, 2006. *Ontology Learning and Population from Text* [online]. ISBN 978-0-387-30632-2. Dostupné na: doi:10.1007/978-0-387-39252-3

FAYYAD, Usama M, Gregory PIATETSKY-SHAPIRO, Padhraic SMYTH a OTHERS, 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. V: *KDD*. s. 82–88.

GAJDOŠOVÁ, Mária Šimková--Katarína, no date. Slovenský závislostný korpus. no date.

GREEN, N, 2011. Dependency Parsing. 2011, s. 137–142.

GREENGARD, Samuel, 2011. Following the crowd. *Communications of the ACM*. 2011, roč. 54, č. 2, s. 20–22.

HOWE, Jeff, 2006a. Crowdsourcing: A definition. *Crowdsourcing: Tracking the rise of the amateur*. 2006.

HOWE, Jeff, 2006b. The rise of crowdsourcing. *Wired magazine*. 2006, roč. 14, č. 6, s. 1–4.

CHEN, Kuan-Ta, Chen-Chi WU, Yu-Chun CHANG a Chin-Laung LEI, 2009. A crowdsourceable QoE evaluation framework for multimedia content. V: *Proceedings of the 17th ACM international conference on Multimedia*. s. 491–500.

CHIANG, David, 2010. Learning to Translate with Source and Target Syntax. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 2010, č. July, s. 1443–1452.

JEFF, Howe, 2009. *Crowdsourcing: Why the power of the crowd is driving the future of business*. 2009. B.m.: Random House Books, New York.

JOJOWONG, Sze-Meng a Mark DRAS, 2011. Exploiting Parse Structures for Native Language Identification. *Emnlp* [online]. 2011, s. 1600–1610. Dostupné na: http://aclweb.org/anthology//D/D11/D11-1148.pdf

LAW, Edith a Luis VON AHN, 2009. Input-agreement: a new mechanism for collecting data using human computation games. V: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. s. 1197–1206.

MUNRO, Robert, Victor KUPERMAN, Tzuyin LAI, Robin MELNICK a Tyler SCHNOEBELEN, 2010. Crowdsourcing and language studies : the new generation of linguistic data. *Linguistics* [online]. 2010, č. June, s. 122–130. Dostupné na: http://portal.acm.org/citation.cfm?id=1866696.1866715

ONDÁŠ, Stanislav, Jozef JUHÁR a Anton ČIŽMÁR, 2011. Extracting sentence elements for the natural language understanding based on slovak national corpus. V: *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*. B.m.: Springer, s. 171–177.

PARALIČ, Ján, Karol FURDÍK, Gabriel TUTOKY, Peter BEDNÁR, Martin SARNOVSK\`Y, Peter BUTKA a František BABIČ, 2010. Dolovanie znalost{í} z textov. *Equilibria, Ko{š}ice*. 2010.

QUINN, Alexander J a Benjamin B BEDERSON, 2011. Human Computation: A Survey and Taxonomy of a Growing Field. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* [online]. 2011, s. 1403–1412. ISSN 1450302289. Dostupné na: doi:10.1145/1978942.1979148

SHAPIRO, Stuart C, 1992. *ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE SECOND EDITION*. B.m.: New Jersey: A Wiley Interscience Publication.

SCHNOEBELEN, Tyler a Victor KUPERMAN, 2010. Using Amazon Mechanical Turk for linguistic research. *Psihologija*. 2010, roč. 43, č. 4, s. 441–464. ISSN 00485705.

SNOW, R., B. O'CONNOR, D. JURAFSKY a a.Y. NG, 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* [online]. 2008, č. October, s. 254–263. Dostupné na: doi:10.1.1.142.8286

ŠIMKO, Jakub a Mária BIELIKOVÁ, 2014. *Semantic Acquisition Games*. B.m.: Springer.

VILNAT, Anne, Gil FRANCOPOULO, Olivier HAMON, Sylvain LOISEAU, Patrick PAROUBEK a Eric DE LA CLERGERIE, 2008. Large Scale Production of Syntactic Annotations to Move Forward. *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation* [online]. 2008, č. August, s. 36–43. Dostupné na: http://www.aclweb.org/anthology/W08-1306

VON AHN, Luis a Laura DABBISH, 2004. Labeling images with a computer game. V: *Proceedings of the SIGCHI conference on Human factors in computing systems*. s. 319–326.

VON AHN, Luis a Laura DABBISH, 2008. Designing games with a purpose. *Communications of the ACM*. 2008, roč. 51, č. 8, s. 58–67.

YOUSFI-MONOD, Mehdi a Violaine PRINCE, 2005. Automatic summarization based on sentence morpho-syntactic structure: narrative sentences compression. V: *NLUCS'05: 2nd International Workshop on Natural Language Understanding and Cognitive Science*. s. 161–167.