Univerzita Pavla Jozefa Šafárika v Košiciach

Prírodovedecká fakulta

SPRACOVANIE PRIRODZENÉHO JAZYKA

Študijný odbor: Informatika
Školiace pracovisko: Ústav informatiky
Vedúci práce: doc. RNDr. Stanislav Krajči, PhD.

Košice 2015

Bc. Július Mareš

Poďakovanie

Rád by som poďakoval vedúcemu diplomovej práce doc. RNDr. Stanislavovi Krajčimu, PhD. za cenné pripomienky počas tvorby tejto práce.

Abstrakt

V tejto práci sa zaoberáme vyvinutím modelu, ktorý určuje vetné členy vo vybraných typoch jednoduchých dvojčlenných viet a súvetí. Využívame pri tom databázu tvarov slovenského jazyka Tvaroslovník, z ktorej získavame slovný druh a gramatické kategórie každého slova vo vete. Na základe týchto informácií určujeme vetné členy a vetné sklady.

Abstract

In this thesis we create a model that focuses on finding constituents of some types of simple sentences with both subject and predicate and in complex and compound sentences. For this purpose the database of word forms in Slovak language Tvaroslovník is used, which we retrieve part of speech and grammatical categories of each word in a sentence from. With use of this information we find constituents of a sentence and their hierarchy.

Obsah

Úvod

Spracovávanie textu je oblasť informatiky, v ktorej sa skúmajú texty, modifikujú sa a určujú ich vlastnosti. Niektoré texty, ktoré sú predmetom skúmania tejto disciplíny, sú napísané prirodzenými jazykmi – jazykmi, ktorými sa dorozumievajú ľudia. V tejto práci spracovávame prirodzený jazyk – slovenčinu.

Oblasť, ktorou sa v tejto práci zaoberáme, je vetný rozbor. Vetný rozbor je disciplínou syntaxe – náuky o tvorbe viet,ktorá je jednou z častí gramatiky. Vetný rozbor sa zaoberá určovaním, akými vetnými členmi sú slová vo vete a aké vetné sklady sú medzi týmito slovami.

V práci sa zaoberáme automatizáciou vetného rozboru. Na určovanie vetných členov a vzťahov medzi slovami vo vete využívame Tvaroslovník – databázu tvarov slov slovenského jazyka. Cieľom práce je navrhnúť a implementovať model, ktorý zisťuje funkcie a vzťahy jednotlivých slov vo vete.

Implementácia navrhnutého modelu po vykonaní vetného rozboru má znázorňovať analyzovanú vetu vo forme grafu. V tomto grafe vrcholy zodpovedajú vetným členom a hrany spájajúce vrcholy zodpovedajú vetným skladom.

V rámci práce analyzujeme jednoduché vety a aj priraďovacie a podraďovacie súvetia. Pri súvetiach určujeme vzťahy medzi jednotlivými vetami a aj vetné členy pre každú vetu zo súvetia zvlášť.

V prvej kapitole je úvod do problematiky a popis systému Tvaroslovník. Pre každý slovný druh je v tejto kapitole uvedené, aké charakteristiky má uvedené v databáze a ktoré z nich v práci využívame.

V druhej kapitole je návrh modelu na vykonávanie vetného rozboru. V tejto kapitole opisujeme spracovávanie slov vety na základe slovného druhu, gramatických kategórií a pozície slov vo vete. Pri niektorých tvaroch slov výsledky dopytov v databáze nie sú jednoznačné – výsledok má viac ako jeden riadok. Návrh modelu sa zaoberá aj týmto problémom.

V tretej kapitole je popis implementácie systému. Zaoberá sa aj spôsobom, akým

sa vizualizujú vetné členy a vetné sklady.

V štvrtej kapitole opisujeme dosiahnuté výsledky a pri akých typoch slovenských viet dokáže systém správne urobiť vetný rozbor.

Kapitola 1

Úvod do problematiky

Veta sa skladá z vetných členov. Tie sú spojené vetnými skladmi.

Hlavnými vetnými členmi dvojčlennej vety sú podmet a prísudok.

Podmetom je vykonávateľ činnosti, alebo nositeľ stavu. Podmet býva vyjadrený podstatným menom alebo zámenom v nominatíve, prípadne iným plnovýznamovým slovným druhom.

Prísudok vyjadruje činnosť, stav alebo vlastnosť podmetu. Prísudky sa delia na slovesné a menné. Slovesný prísudok je vyjadrený slovesom, menný prísudok je vyjadrený spojením sponového slovesa a plnovýznamového slovného druhu.

Okrem základných vetných členov sa v rozvitých vetách nachádzajú pomocné vetné členy. Pomocnými vetnými členmi sú predmet, príslovkové určenie a prívlastok.

Predmet je zvyčajne vyjadrený podstatným menom alebo zámenom. Viaže sa s pádmi rôznymi od nominatívu. Rozvíja prísudok.

Príslovkové určenie je vetný člen, ktorý rozvíja sloveso, prídavné meno alebo príslovku. Pýtame sa na ňu otázkami kde?, kedy?, ako?, prečo?. Môže ju tvoriť príslovka alebo podstatné meno.

Prívlastok je vedľajší vetný člen, ktorý rozvíja podstatné meno. Prívlastky sa delia do dvoch kategórií: na zhodné a nezhodné prívlastky. Zhodné prívlastky majú rovnaké gramatické kategórie ako slová, ktoré rozvíjajú. Môžu byť vyjadrené prídavným menom, príslovkou. Nezhodné prívlastky majú iné gramatické kategórie ako ich nadradený vetný člen. Môžu byť vyjadrené napríklad podstatným menom.

1.1 Tvaroslovník

Na určovanie, akým vetným členom dané slovo vo vete je, potrebujeme poznať jeho vlastnosti. Tie zisťujeme pomocou prístupu k elektronickému slovníku. Využívame systém Tvaroslovník.

Tvaroslovník [4] je databáza tvarov slov slovenského jazyka, ktorá je dostupná na [6]. Obsahuje približne 320 000 základných tvarov slov. Spolu s prechýlenými tvarmi je v databáze približne 30 000 000 riadkov. Okrem tvaru slova každý záznam obsahuje aj informáciu o tom, akého je slovného druhu a príslušné gramatické kategórie daného slova.

Údaje v Tvaroslovníku boli získané zo Slovníka slovenského jazyka a Veľkého slovníka cudzích slov. Tieto slovníky obsahujú iba základné tvary slov, preto po zelektrizovaní slov ich prechýlené tvary museli byť vygenerované. Slová boli uložené do textových súborov, z ktorých boli následne načítané do databázy. Použitý bol databázový systém MySQL.

Všetky údaje sú v databáze uložené v jednej tabuľke. Každý jej riadok obsahuje jeden z tvarov slova spolu so všetkými informáciami o ňom.

Zoznam stĺpcov v tabuľke je nasledujúci:

- idSlovo – jedinečný celočíselný identifikátor slova,
- idTvar – jedinečný celočíselný identifikátor tvaru slova (základný tvar má vždy identifikátor 0),
- tvar – textový tvar slova,
- slovnýDruh – textová informácia o slovnom druhu daného slova. V prípade prídavných mien, ktoré sú privlastňovacie, je okrem slovného druhu v tomto atribúte uvedená aj táto skutočnosť – záznam má tvar „prídavné meno privlastňovacie“,
- charakteristika – textový zoznam hodnôt gramatických kategórií slova. Zoznam kategórií závisí od slovného druhu.

Primárny kľúč v tabuľke je tvorený kombináciou prvých dvoch atribútov idSlovo a idTvar. Pomocou týchto dvoch atribútov vieme jednoznačne určiť slovo a jeho konkrétny tvar. Atribút tvar nemôže byť primárnym kľúčom, keďže v slovenčine existujú slova s rovnakým tvarom, ale rôznymi vlastnosťami.

Jednotlivé slovné druhy majú svoje vlastné množiny charakteristík. Podstatné mená majú charakteristiky rod, číslo, pád. Ak ide o životné podstatné mená, v stĺpci charakteristika je uvedená aj kategória podrod, ktorá môže nadobúdať hodnoty životné alebo neživotné.

Prídavné mená majú uvedené rod, číslo, pád, ak ide o prídavné mená mužského rodu, aj podrod. Akostné a vzťahové prídavné mená majú aj uvedené, v akom sú stupni. Ak ide o prídavné meno privlastňovacie, tento fakt je uvedený v stĺpci slovnýDruh.

Zámená, ktoré sú osobné, majú v charakteristike uvedené tieto kategórie: osoba, číslo, pád, rod. Osobné privlastňovacie zámená v základnom tvare majú uvedenú poznámku, že sú privlastňovacie od nejakého osobného zámena. Opytovacie zámená majú uvedený pád a ak sa to dá z tvaru slova zistiť, aj rod a číslo, pri mužskom rode aj podrod. Ukazovacie, zvratné zámená sa a si, neurčité a vymedzovacie zámená, ktoré sú nesklonné, nemajú v charakteristike uvedené žiadne vlastnosti. Sklonné ukazovacie, tvary zvratných zámen seba a sebe, neurčité a vymedzovacie zámená majú uvedený pád, rod, číslo, v prípade mužského rodu aj podrod.

Základné, radové, druhové číslovky majú uvedený pád, rod (v mužskom rode aj podrod), číslo. Skupinové číslovky majú uvedený pád a číslo. Násobné a nesklonné neurčité číslovky nemajú uvedené žiadne charakteristiky.

Slovesá majú v charakteristike uvedené číslo, čas, spôsob. Slovesá, ktoré môžu byť aj zvratné, majú uvedenú aj zvratnosť spolu so zvratným zámenom, s ktorým sa viažu. Slovesá, ktoré sú prechodníky, trpné alebo činné príčastia, túto skutočnosť majú uvedené v položke charakteristiky forma. Neurčitky majú v stĺpci charakteristika uvedené iba forma: neurčitok, okrem tejto položky tam nie sú uvedené žiadne ďalšie položky.

Príslovky môžu mať charakteristiku stupeň. Príslovky, ktoré nemožno stupňovať, nemajú v stĺpci charakteristika uvedenú žiadnu položku.

Predložky majú v charakteristike uvedenú položku väzba, v ktorej sú uvedené pády, s ktorými sa viažu.

Existujú slová, ktoré môžu byť v závislosti od kontextu predložkami alebo príslovkami. Takéto slová majú v atribúte slovnýDruh hodnotu predložka, príslovka. V charakteristike majú uvedené , s akými pádmi sa viažu, ak sú predložkou.

Spojky, častice a citoslovcia nemajú v atribúte charakteristika uvedené žiadne položky.

V tejto práci pri určovaní vetných členov pracujeme s charakteristikami podstat-

ných, prídavných mien, čísloviek, slovies a zámen. Ostatné slovné druhy (príslovky, spojky, častice, citoslovcia, predložky) buď nemajú v stĺpci charakteristika uvedené žiadne gramatické kategórie, alebo ich kategórie nepotrebujeme pre účely tejto práce využívať.

Pri podstatných menách z ich charakteristík využívame pád, rod. Pri práci s prídavnými menami prídavnými menami využívame gramatickú kategóriu pád.

Kapitola 2

Návrh riešenia

Pri analýze vety postupne prechádzame jej všetky slová. Pri každom slove vo vete sa vykoná dopyt v Tvaroslovníku pýtajúci sa na slovný druh, charakteristiku, tvar slova a id slova. Id slova je potrebné preto, lebo v niektorých prípadoch je potrebné nájsť základný tvar slova.

Dopyt môže vrátiť viacero riadkov, keďže rôzne slová v slovenčine môžu mať ten istý tvar. Ak dopyt v databáze nevráti žiadny výsledok (vráti 0 riadkov), predpokladá sa, že používateľ spravil typografickú chybu. V takomto prípade program používateľa vyzve, aby toto slovo z vety vymazal alebo ho opravil.

Interpunkčné znamienka sú písané hneď za slovami. V tejto práci sa zaoberáme vetami, v ktorých sa vyskytujú interpunkčné znamienka bodka a čiarka. Pri každom spracovávanom slove sa kontroluje, či na jeho konci nie je nejaké z týchto interpunkčných znamienok. Ak tam takéto znamienko je, tak sa zo slova odstráni a v databáze sa dopytuje takto novovzniknuté slovo. Obe slová aj s vlastnosťami z databázy uložíme do zoznamu.

Keďže databáza pri dopytoch nerozlišuje diakritiku, program po každom z nich kontroluje všetky riadky výsledku, či je v nich tvar slova zhodný so slovom z dopytu. Ak tvar slova zhodný nie je – líši sa v diakritike, tento riadok vylúčime z výsledku dopytu. Táto situácia nastane napríklad pri dopyte slova má, keď výsledok obsahuje aj riadok, v ktorom je v stĺpci tvar hodnota ma a preto tento riadok z výsledku dopytu vyradíme.

V slovenčine sa vyskytujú vymedzovacie zámena, ktoré pozostávajú z dvoch slov, napríklad ten istý, taký istý, tak isto [1]. Ak program pri spracúvaní slov vo vete narazí na základný alebo na jeden z prechýlených tvarov slov ten, taký, tak, skontroluje, či sa bezprostredne za ním nachádza príslušný tvar slova istý, resp. isto. V takomto prípade

program tieto slová spojí do jedného tokenu. V ďalšom priebehu vetného rozboru ich bude považovať iba za jedno slovo.

2.1 Spracovávanie jednotlivých slovných druhov

Pri spracovávaní slova postupne pre slovné druhy, ktoré môžu byť súčasťou vetného členu, zisťujeme, či nimi dané slovo môže byť. Ak zistíme, že slovo môže byť daným slovným druhom, spracujeme ho a ak určíme, že je nejakým vetným členom, v rámci iterácie neskúmame, či môže byť ďalším vetným členom, ale prechádzame na ďalšie slovo vety.

Udržujeme zoznam, v ktorom je pre každé slovo uvedený index vetného člena, ktorého je slovo súčasťou. Umožňuje nám to získať charakteristiky slov, ktoré sú časťou už spracovaného vetného člena, ktorého vlastnosti zisťujeme.

Najprv zisťujeme, či slovo nie je častica by. Toto slovo je súčasťou slovesa v podmieňovacom spôsobe a teda aj súčasťou prísudku. Preto ho pridáme k prísudku, ak už v zozname vetných členov príslušnej vety je. Ináč do nej pridáme by ako slovesný prísudok.

Ak slovo je častica nie a nachádza sa pred tvarom slovesa byť, program ho označí ako slovesný prísudok. Uvažujeme prípady, v ktorých sa nachádza pred tvarom slovesa byť, ktoré program pridá do prísudku, keď naň v rámci načítania slov vo vete narazí. Častica nie sa môže vyskytovať aj v samostatne ako jediné slovo vety alebo vety súvetia, v takom prípade ju nepovažujeme za vetný člen. Rovnako sa môže vyskytovať aj pred inými slovami, napríklad vo vete: Nie vždy jazdil autom. Prípadmi, v ktorých sa vyskytuje samostatne, sa nezaoberáme.

Zisťujeme, či slovo môže byť slovesom. Slovesá sú súčasťou slovesného aj menného prísudku. V slovenčine sa môže stať, že sa slovesný tvar skladá z viacerých slov, napríklad „išiel som“. Pri spracovávaní slova, ktoré môže byť slovesom, sa teda v zozname vetných členov kontroluje, či sa v príslušnej vete už spracovával prísudok. Ak áno, práve spracovávané sloveso k nemu pripojíme.

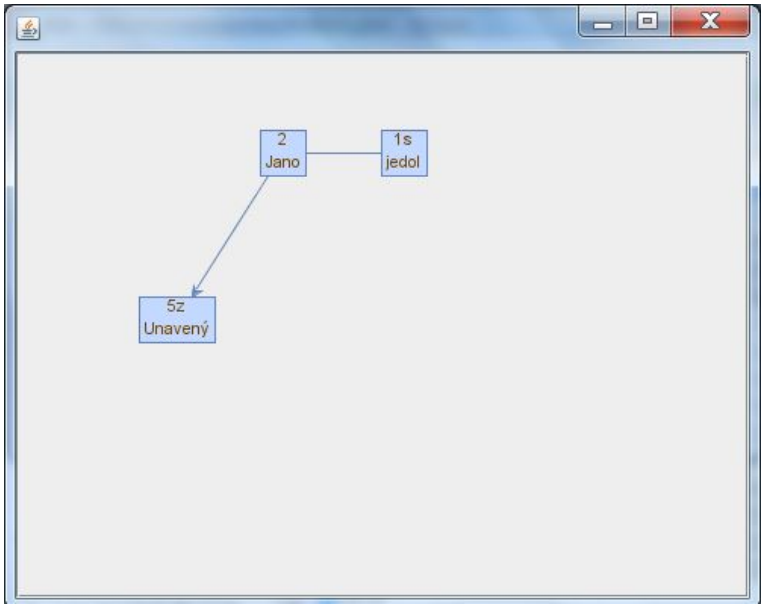
Ak je sloveso v jednom z tvarov slova byť alebo stať, predbežne ho považujeme za súčasť menného prísudku, keďže slová stať sa a byť môžu byť sponovými slovesami. Ak sa vo vete už vyskytuje prísudok, pridaním slovesa vznikne menný prísudok len vtedy, ak je toto sloveso neurčitok byť, prípadne ak je toto sloveso v prechýlenom tvare slova byť a zároveň predchádzajúce slová prísudku neobsahujú žiadne sloveso alebo prísudok už obsahuje jednu z foriem slovesa byť. V opačnom prípade ide o slovesný

prísudok.

Pri pridaní prísudku sa kontroluje, či medzi spracovanými vetnými členmi príslušnej vety nie je aj podmet. V prípade, že medzi nimi podmet je, skontroluje sa, či rod a číslo slova, ktoré tvorí podmet, zodpovedá osobe a číslu slovesa. Ak nie, označenie tohto vetného členu sa zmení na predmet.

Osoba a číslo slovesa, sa vyhodnotia na základe prípony slovesa a prítomnosti slov charakteristických pre niektorú z osôb v niektorom čísle. Kontrolujeme komplexne celý tvar slovesa, keďže jednotlivé časti slovesa môžu byť v inej osobe a čísle ako celé sloveso.

Trpné aj činné príčastia mávajú rovnakú funkciu ako prídavné mená [5]. Preto ich v rámci rozboru nepovažujeme za slovesá a ak sú súčasťou prísudku, môže to byť len menný prísudok, nie slovesný. V Obr. 2.1 je ukázané spracovanie slova unavený, ktoré môže byť prídavné meno aj trpné príčastie. V tejto vete má funkciu zhodného prívlastku.



Obr. 2.1: Rozbor vety: Unavený Jano jedol.

Následne kontrolujeme, či slovo môže byť spojkou alebo sa vo vete končí čiarkou. Spojky a čiarky môžu spájať viacnásobné vetné členy alebo vety súvetia. Spojky nie sú súčasťou žiadneho vetného členu.

Pri spracovávaní spojky alebo čiarky (spoločne ich označujeme separátor) berieme do úvahy niekoľko syntaktických javov. Ak posledné slovo pred výskytom čiarky alebo spojky je prísudok, veta je súvetím, ktoré rozdeľuje separátor. Takisto ide o súvetie,

ak prvé slovo za separátorom je slovesom.

Ak posledné slovo pred separátorom je príslovka, táto príslovka bude príslovkovým určením. Ak sa za separátorom nachádza vzťažné alebo neurčité zámeno alebo spojenie predložky a vzťažného alebo neurčitého zámena, veta je podraďovacie súvetie, ktoré sa delí na dve vety v mieste spracovávaného slova.

Vzťažné zámeno je opytovacie zámeno vo funkcii spojovateľa dvoch viet. V databáze nie je informácia o tom, či zámeno je opytovacie ani neurčité, preto identifikáciu niektorých vzťažných a neurčitých zámen vykonávame na základe ich začiatočných a koncových písmen.Používame bežne používané predpony a prípony neurčitých zámen, niektoré začiatočné písmená opytovacích zámen „ke“, „ko“, „kd“, „kt“, „ak“ a opytovacie zámeno „prečo“.

Nasleduje kontrola, či slovo môže byť predložkou. Niektoré predložky môžu mať rovnaký tvar ako príslovka. Pri každej z nich preto kontrolujeme, či sa za ňou (v tej istej vete, ak program spracováva súvetie) nevyskytuje podstatné meno alebo osobné zámeno v tom istom páde, pričom sa medzi nimi nenachádza žiadne sloveso. Ak vo vete je podstatné meno alebo zámeno, ktoré spĺňa tieto podmienky, slovo, ktoré sa spracováva, bude programom považované za predložku.

V prípade, že vo vete podstatné meno, ktoré má rovnaký pád ako potenciálna predložka, nie je, bude slovo vyhodnotené ako príslovka a v rámci vetného rozboru bude považované za príslovkové určenie rozvíjajúce prísudok. Ak sa za predložkou alebo príslovkou nachádza ďalšie slovo, ktoré môže byť predložkou, rekurzívne ho spracovávame rovnakým spôsobom. Po tom, čo sa toto ďalšie slovo spracuje, sa vrátime k spracovávaniu pôvodného slova, pričom platí, že ak druhé slovo je predložkou, tak slovo, s ktorým sa spája, nebude môcť byť spojené s prvou spracovávanou potenciálnou predložkou.

Predložku spojíme s príslušným slovom a spolu budú tvoriť jeden vetný člen. Na základe vlastností podstatného mena alebo zámena určíme, o aký vetný člen ide. Do úvahy tu pripadajú 3 vetné členy: nezhodný prívlastok, predmet alebo príslovkové určenie.

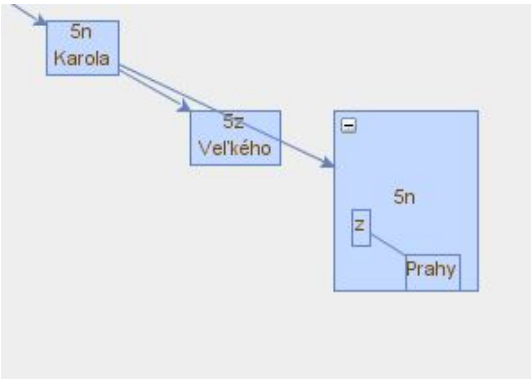
Pri spojeniach predložiek a podstatných mien však nerozlišujeme medzi predmetom, keďže sa gramaticky nelíšia, rozdielny je iba spôsob, akým sa na ne pýtame. Na predmet sa pýtame pádovými otázkami a na príslovkové určenia otázkami kde?, kedy?, ako?, prečo?.Taktiež nevieme určiť, akou z otázok sa pýtame na príslovkové určenie, a tak neurčujeme, či ide o príslovkové určenie miesta, času, spôsobu alebo príčiny.

Ak sa v tej istej klauzule pred tokenom nachádza iné podstatné meno v inom páde a medzi nimi nie je sloveso, token je nezhodným prívlastkom. V opačnom prípade určíme, že ide o predmet.

Ak slovo môže byť príslovkou alebo zámenom, ktorého charakteristika v Tvaroslovníku neobsahuje pád, pre potreby tejto práce toto slovo považujeme za príslovkové určenia, aj keď niektoré z týchto zámen takúto funkciu v spracovávanej vete nemusia mať.

Ak je slovo zámenom a nemá určený pád, ide o nesklonné zámeno ukazovacie (napríklad tu, tam), opytovacie (napríklad prečo), vymedzovacie alebo neurčité (napríklad niekedy). Na základe charakteristiky v Tvaroslovníku sa nedá určiť, o aký druh zámena ide a medzi zámenami tej istej kategórie existujú rozdiely v ich funkcii vo vete.

Ak slovo môže byť prídavným menom, kontrolujeme, či pred alebo za ním sa v príslušnej vete nachádza podstatné meno v rovnakom páde, pričom medzi nimi nie je sloveso. Ak vo vete také podstatné meno je, prídavné meno program prehlási za zhodný prívlastok. Prídavné meno sa môže viazať na podstatné meno, ktoré sa nachádza za ním len vtedy, ak medzi nimi nie je predložka, ktorá sa tiež viaže na toto podstatné meno.



Obr. 2.2: Príklad predložky medzi prídavným menom a podstatným menom

Na obrázku je znázornený príklad, v ktorom je prídavné meno Veľkého pred podstatným menom Prahy a za prechýleným tvarom podstatného mena Karol. Keďže hneď za slovom Veľkého je predložka z, nebude sa viazať na slovo Prahy, hoci obe slová môžu byť v genitíve a teda sa môžu na seba v inom kontexte viazať. Namiesto toho bude slovo Veľkého zhodným prívlastkom, ktorý bude rozvíjať podstatné meno Karola.

V prípade, že vo vete v okolí prídavného mena podstatné meno v rovnakom páde nie je, program skontroluje, či sa v zozname spracovaných vetných členov nachádza menný prísudok. Ak áno a prídavné meno je v nominatíve alebo inštrumentáli, tak prídavné meno sa pripojí k mennému prísudku. Ak nie, prídavné meno bude podmetom, ak jeho pád bude nominatív. Ak je prídavné meno v inom páde ako nominatív, bude predmetom.

Ak sa v programe vyskytuje za sebou skupina slov, ktoré podľa databázy môžu byť prídavné alebo podstatné mená a všetky z nich majú rovnaký pád, posledné slovo z nich program považuje za podstatné meno, predošlé slová budú považované za prídavné mená.

Nasleduje kontrola, či slovo môže byť osobné privlastňovacie zámeno. Osobné privlastňovacie zámená majú vo vete rovnakú funkciu ako prídavné mená, preto slovo, ktoré môže byť takýmto zámenom, je spracovávané rovnakým spôsobom ako slová, ktoré môžu byť prídavnými menami.

Nasleduje spracovávanie slov, ktoré môžu byť podstatnými menami alebo osobnými základnými zámenami. Pri týchto slovách skúmame, či môžu byť v nominatíve. Ak áno, skontrolujeme, či sa pred nimi nenachádza iné podstatné meno, pričom medzi nimi nie je prísudok. V takom prípade bude podstatné meno označené ako nezhodný prívlastok.

V navrhovanom modeli má pri určovaní podmetu prioritu to slovo, ktoré sa vo vete vyskytuje skôr, keďže preferujeme objektívny slovosled. Ak teda spracovávané slovo môže byť podmetom, ale jeden podmet sa už vo vete vyskytuje, spracovávané slovo bude iným vetným členom.

Ak podstatné meno, ktoré môže byť v nominatíve, nie je nezhodným prívlastkom a nastane aspoň jedna z týchto možností: podstatné meno nemôže byť v akuzatíve alebo sa v zozname spracovaných vetných členov nenachádza podmet, pridáme slovo ako menný prísudok, ak sa taký vyskytuje v zozname vetných členov, je pred ním v jeho blízkosti a jeho posledné slovo je sloveso.

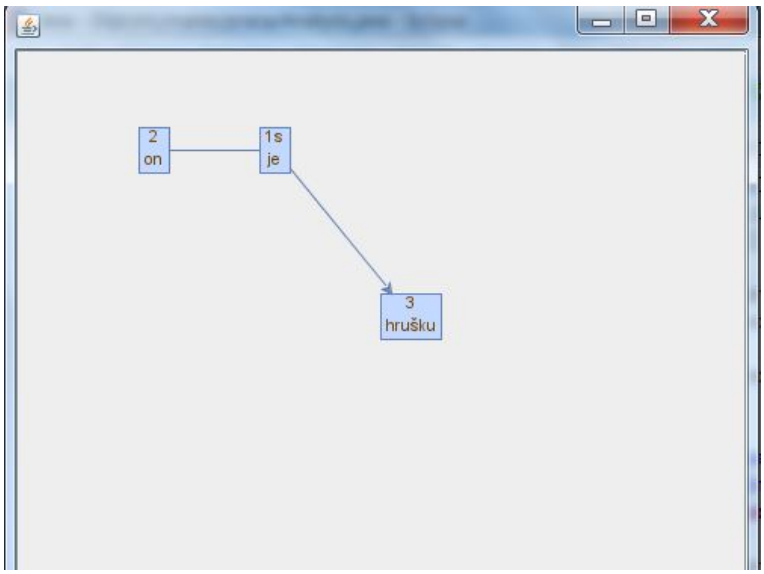
Ak podstatné meno, ktoré môže byť v nominatíve, nie jemenným prísudkom, môže to byť podmet alebo predmet. Ak jeho rod a číslo korešponduje s prísudkom, označíme ho ako podmet. Ak spracovávané slovo nemôže byť podmetom ani menným prísudkom, bude označené ako predmet.

Ak slovo môže byť predmetom, najprv kontrolujeme, či sa pred ním nenachádza iné podstatné meno. V prípade, že áno, slovo bude nezhodným prívlastkom, ktoré rozvíja tento predošlý člen. Inač sa zisťuje, či je v zozname vetných členov menný

prísudok, a ak sa v ňom ešte nenachádza menná časť a spracovávané slovo môže byť v nominatíve alebo inštrumentáli, pridáme ho k mennému prísudku. Ináč ho označíme ako predmet.

Po spracovaní všetkých slov vo vete kontrolujeme každý menný prísudok, či sa končí podstatným menom, prídavným menom alebo príslovkou. Ak sa žiadnym z týchto slovných druhov potenciálny menný prísudok nekončí, zmeníme jeho označenie na slovesný prísudok.

Program dokáže správne určiť menný prísudok, ak sponové sloveso sa vo vete nachádza pred mennou časťou. V prípade, že to tak nie je, program vyhlasi prísudok za slovesný.



Obr. 2.3: Príklad vety, v ktorej podstatné meno nie je v nominatíve ani v inštrumentáli

Na Obr. 2.3 je znázornený príklad spracovania prechýleného tvaru slova hruška, ktoré môže byť predmetom, pričom pred ním sa nachádza sloveso je. Tu je vo význame jesť ale rovnaký tvar má aj 3. osoba jednotného čísla slovesa byť. Toto sloveso v prípade, že je sponové, je súčasťou menného prísudku. Keďže však slovo hrušku nemôže byť v nominatíve ani v inštrumentáli, nemôže byť súčasťou menného prísudku, a teda bude označené ako predmet. Sloveso bude označené programom ako slovesný prísudok.

2.2 Určovanie zamlčaného podmetu

Každá jednoduchá veta alebo veta v priraďovacom alebo podraďovacom súvetí obsahuje neslovesný vetný základ alebo sloveso, pričom môže mať vyjadrený podmet, zamlčaný podmet alebo podmet nemusí vôbec mať (napríklad veta: Snežilo.). Vetným rozborom jednočlenných viet, ktoré sloveso neobsahujú alebo nemajú podmet, sa tento program nezaoberá a vždy predpokladá, že na vstupe je dvojčlenná veta, resp. súvetie obsahujúce iba dvojčlenné vety.

Po spracovaní všetkých slov vo vete program skontroluje, či medzi vetnými členmi je podmet. Ak sa spracováva súvetie, prítomnosť podmetu sa kontroluje vo všetkých vetách súvetia. Ak sa v niektorej z viet podmet nevyskytuje, pridá sa do nej osobné základné zámeno v nominatíve, keďže v modeli predpokladáme, že spracovávaná veta má zamlčaný podmet.

Tvar zamlčaného podmetu je určený podľa času a osoby prísudku, ku ktorému sa viaže. Ak je sloveso v príslušnej vete v množnom čísle, zamlčaný podmet môže mať tieto tvary: vy (v prípade, že sloveso obsahuje slovo ste), my (v prípade, že sloveso obsahuje slovo sme).

V prípade, že sloveso v množnom čísle neobsahuje žiadne zo slov sme a ste, tvar zamlčaného podmetu bude oni/ony. Pri tejto možnosti sa nedá určiť rod slova, ktoré je podmetom (mužský, ženský alebo stredný rod), a teda ani presný tvar zamlčaného podmetu, keďže pre všetky tri rody v množnom čísle 3. osoby sú prípony slovesa vo všetkých časoch rovnaké.

Ak sloveso v príslušnej vete je v jednotnom čísle, tiež môže nastať niekoľko možností. Ak sloveso obsahuje slovo som, zamlčaný podmet má tvar ja. Ak dané sloveso obsahuje slovo si, zamlčaný podmet má tvar ty. V prípade, že nenastala žiadna z týchto možností, zamlčaný podmet je v 3. osobe a môže byť v jednom z týchto tvarov: on, ona alebo ono.

V prípade že sloveso je vo vete v 3. osobe jednotného čísla v prítomnom čase alebo v budúcom čase, nie je možné určiť rod zamlčaného podmetu z toho dôvodu, že pre všetky tri rody sú prípony takéhoto slovesa rovnaké. Zamlčaný podmet bude teda mať vo výstupe vetného rozboru iba nešpecifický tvar osobného základného zámena on/ona/ono.

V prípade, že sloveso je v 3. osobe jednotného čísla v minulom čase, tvar zamlčaného podmetu sa určuje na základe prípony slovesa. Zamlčaným podmetom bude slovo on, ak sa sloveso končí príponou -l. V prípade, že sa dané sloveso končí príponou

-la, zamlčaný podmet má tvar ona. V prípade, že sloveso je zakončené príponu -lo, zamlčaným podmetom bude slovo ono.

Kapitola 3

Implementácia

Program vykonávajúci vetný rozbor je implementovaný v programovacom jazyku Java. Na vizualizáciu vetných členov je použitá knižnica JGraphX dostupná z [3]. K pripojeniu programu k databáze je použitá knižnica MySQL Connector/J dostupná z [2].

Knižnica JGraphX slúži na zobrazovanie grafov. Umožňuje vykreslenie vrcholov s textom vo vnútri, ich spojenie hranami. Je možné aj umiestniť vrchol vo vnútri iného vrchola, čo využívame, ak sa vetný člen skladá z viacerých slovných druhov (napríklad z predložky a podstatného mena).

Knižnica MySQL Connector/J slúži na vytvorenie spojenia Java aplikácie s MySQL serverom. V rámci tejto práce ju používame na spojenie s Tvaroslovníkom a dopytovanie tabuľky slov.

3.1 Implementované triedy

V programe sa nachádzajú tieto triedy:Main, Dao, SentenceFrame, Analysis, PartOfSentence, Visualisation, Constants.

Program je spúšťaný pomocou triedy Main. Jej spúšťacia metóda vytvorí objekt triedy SentenceFrame, ktorá plní funkciu užívateľského rozhrania.

Pripojenie k Tvaroslovníku zabezpečuje trieda Dao. Pomocou tejto triedy sa spájame s databázou a získavame z nej informácie o každom slove vety. Trieda zároveň zabezpečuje dopyt, pomocou ktorého sa získajú z databázy všetky vyskytujúce sa základné tvary slova.

Trieda SentenceFrame dedí od triedy JFrame. Obsahuje textové pole, do ktorého používateľ vpíše vetu. Po tom, čo používateľ stlačí tlačidlo, trieda zavolá metódu

partsOfSentence() triedy Analysis, pomocou ktorej program urobí vetný rozbor vety na vstupe.

Vetný rozbor zabezpečuje trieda Analysis. Vstupom jej metódy partsOfSentence() je veta, ktorej rozbor sa prevedie. Výstupom je zoznam obsahujúci vetné členy reprezentované triedou PartOfSentence. Vetné členy sa v zozname nachádzajú v rovnakom poradí, v akom sa vo vete vyskytovali.

Informácie o slove spolu s výsledkom dopytu v databáze sú uložené v objekte triedy Word. Táto trieda má inštančné premenné, v ktorých sú uložené tvar slova, ktorý má slovo vo vete, tvar slova po odstránení interpunkčných znamienok, ktoré sa môžu vyskytovať na jeho konci, a zoznam všetkých riadkov tabuľky, ktoré dopyt v databáze vrátil.

Vetný člen je reprezentovaný triedou PartOfSentence. Táto trieda má v inštančných premenných uvedené akým vetným členom je, slová, z ktorých vetný člen pozostáva, zoznam rodičov jednotlivých slov vetného členu hovoriaci o tom,aké sú hierarchické vzťahy medzi týmito slovami. Ďalšími uloženými informáciami v inštančných premenných sú index vetného členu v rámci jeho vety. Ak bola analyzovaná zložená veta, tak obsahuje celočíselnú premennú obsahujúcu index vety súvetia, v ktorej sa nachádza. Ak robíme vetný rozbor jednoduchej vety, index vety súvetia má u všetkých vetných členov hodnotu 0.

Trieda Visualisation slúži na vizualizáciu vetných členov a skladov. Spracováva zoznam vetných členov, ktorý bol vytvorený triedou Analysis. Na základe tohto zoznamu vytvorí a priestorovo usporiada vrcholy grafu, ktorý je reprezentovaný objektom triedy mxGraph z knižnice JGraphX. Následne nastaví graf tak, aby pri vizualizácii užívateľ nemohol jeho vrcholy a hrany modifikovať. Potom vytvorí nové okno, pričom tento graf v ňom nechá zobraziť.

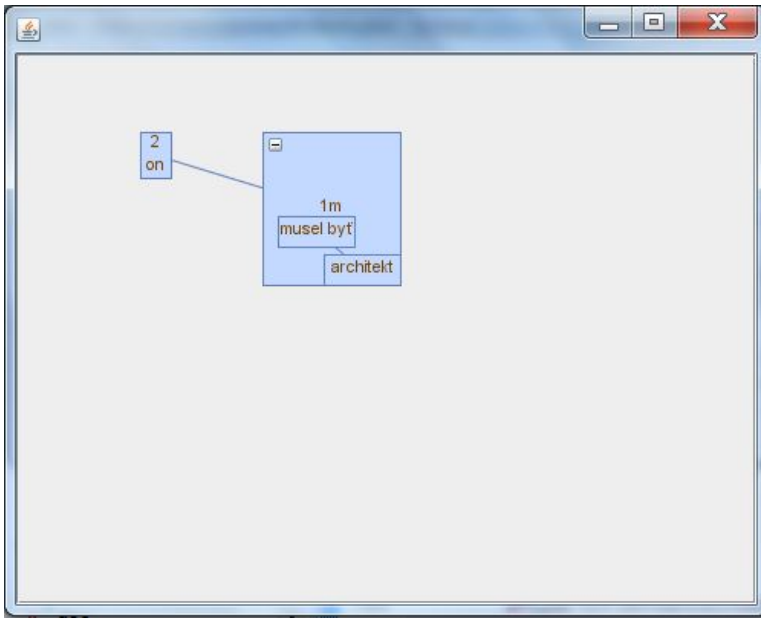
Trieda Constants obsahuje konštanty – textové reťazce, ktoré sa frekventovane používajú v ostatných triedach. Sú to názvy všetkých vetných členov, slovných druhov a pády.

3.2 Vizualizácia

V každom vrchole grafu je uvedený tvar vetného člena a jeho zaužívaná značka. Prísudok je označovaný číslovkou 1, pričom ak ide o slovesný prísudok, tak je označovaný 1s, menný prísudok je označovaný značkou 1m. Podmet má označenie 2, predmet 3, príslovkové určenie 4. Prívlastky sú označované pomocou číslovky 5 a to nasledujúcim

spôsobom: zhodný prívlastok má označenie 5z a nezhodný prívlastok má označenie 5n.

Niektoré vetné členy sa skladajú z viac ako jedného slova. Sú to napríklad menné prísudky, predmety a príslovkové určenia skladajúce sa z predložky a podstatného mena. Za viacslovný vetný člen nepovažujeme viacslovné slovesné tvary. Takéto vetné členy sú zobrazované pomocou vnútorných vrcholov, ktoré sa nachádzajú vo vrchole reprezentujúcom vetný člen.



Obr. 3.4: Vnorenie vrcholov pri viacslovných vetných členoch

Pred vizualizáciou sa vytvorí ďalší dvojrozmerný zoznam, ktorého každá položka obsahuje zoznam indexov všetkých detí vetného člena, ktorý je asociovaný s indexom položky. Vetné členy sú zobrazované v rôznych výškových úrovniach v rámci okna, rozvíjajúce členy sú o úroveň nižšie ako ich nadradené vetné členy, ktoré sú nimi rozvíjané.

V každej úrovni sú vetné členy umiestnené v takej vzájomnej vzdialenosti, aby tie členy, ktoré sú vo vete pred ich nadradeným vetným členom, boli naľavo a tie, ktoré sú za ním, boli umiestnené napravo. V najvyššej vrstve sa nachádzajú podmety a prísudky.

V prípade súvetia je každý podmet spojený s prísudkom z rovnakej vety. V priraďovacom súvetí sú všetky hlavné vetné členy každej vety zobrazené v rovnakej úrovni.

Program umiestňuje pri vizualizácii vetné členy vo vzdialenosti od ľavého okraja okna v závislosti od toho, ktoré v poradí sú vo vete. Vzájomné horizontálne umiestne-

nie medzi všetkými dvojicami vetných členov v rámci grafu je určené ich vzájomnou polohou vo vete – grafová reprezentácia vetného člena, ktorý sa vo vete vyskytol skôr, bude umiestnená bližšie k ľavému okraju okna ako reprezentácia druhého, neskoršie sa vyskytujúceho vetného člena.

Kapitola 4

Výsledky

Na základe navrhnutého modelu sme vyvinuli systém zaoberajúci sa vetným rozborom slovenských viet. Tento systém reflektuje niektoré typy vetných väzieb v slovenskom jazyku.

Systém sa zaoberá iba dvojčlennými vetami – takými, ktoré obsahujú podmet aj prísudok. Systém nerieši vetný rozbor vetných základov. Predpokladá, že vety, ktoré sú analyzované, sú oznamovacie, jedinými interpunkčnými znamienkami, ktoré sa v nich vyskytujú, sú bodka a čiarka. Vety neobsahujú citoslovcia a častice s výnimkou slov by, nie.

Ak sa vo vete vyskytne slovo, ktoré sa nenachádza v Tvaroslovníku, predpokladá sa, že používateľ spravil vo vstupe typografickú chybu. Vetný rozbor neprebehne a používateľ bude upozornený, aby dané slovo upravil. Deje sa tak aj v prípadoch, v ktorých bolo slovo zadané správne, ale nenachádza sa v Tvaroslovníku, lebo ide o neologizmus alebo názov.

Niektoré tvary slov v slovenčine môžu byť viacerými slovami a slovnými druhmi. Riešime to uprednostňovaním niektorých vetných členov pred inými v rámci spracovávania slova.

Ak je vo vete zamlčaný podmet, program ho doplní do grafu. Tvar zamlčaného podmetu zodpovedá prísudku.

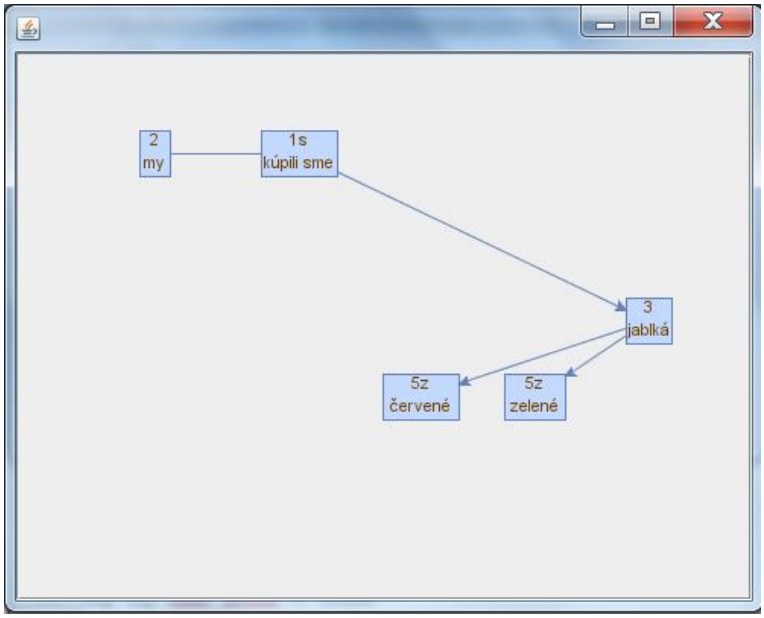
Program rozlišuje slovesný a menný prísudok. Za menný prísudok považuje spojenie sponového slovesa byť alebo stať sa a podstatného alebo prídavného mena v nominatíve alebo v inštrumentáli.Keďže niektoré neživotné podstatné mená mužského rodu majú rovnaký tvar v nominatíve aj v akuzatíve, program pri takýchto slovách nedokáže spoľahlivo určiť ich pád, keďže sa nezaoberá sémantikou slov. Môžu teda nastať prípady, v ktorých slovo v akuzatíve bude súčasťou menného prísudku, hoci vo

vete mal byť slovesný prísudok.

Zaoberáme sa aj prípadmi, v ktorých je sponové sloveso v mennom prísudku spojené s iným pomocným slovesom. Ak už nejaké sloveso bolo spracované, pri spracovávaní sponových slovies považujeme prísudok za menný len vtedy, ak je sponové sloveso v neurčitku (z toho predpokladáme, že sa sa spája s pomocným slovesom) alebo predchádzajúce sloveso obsahuje tvar slova byť (predpokladáme, že ide o podmieňovací spôsob v minulom čase). Pri menných prísudkoch sa nezaoberáme slovosledom, v ktorom nie je objektívne poradie (objektívne poradie – také, v ktorom sa podmet nachádza pred prísudkom). Program napríklad neoznačí v spojení policajtom bol prísudok za menný, ale za slovesný. Zároveň slovo policajtom bude označené ako predmet.

Nevieme rozlíšiť predmet od príslovkového určenia v prípadoch, že sú kandidátmi na vetný člen pri podstatných menách. V takom prípade bude slovo označené ako predmet.

Program rieši aj prípady, v ktorých sa vo vete vyskytujú viacnásobné prívlastky. Medzi takýmito prívlastkami je čiarka alebo spojka.



Obr. 4.5: Vetný rozbor vety: Kúpili sme červené a zelené jablká.

Na Obr. 4.4 je príklad vetného rozboru, ktorý vykonal naimplementovaný program. Na začiatok pridal vrchol obsahujúci zamlčaný podmet my zistený podľa tvaru slovesa. Hoci spracovával sloveso sme, prísudok bude slovesný, keďže predchádzajúce sloveso neobsahovalo tvar slova byť. Pri slovách červené a zelené vzniká v Tvaro-

slovníku nejednoznačnosť. Tieto slová môžu byť podľa Tvaroslovníka aj podstatnými menami. Keďže však program považuje prídavné mená za prioritné, sú označené ako zhodné prívlastky.

Záver

V tejto práci sme sa zaoberali vetným rozborom vybraných typov viet v slovenskom jazyku. Využili sme na to Tvaroslovník – databázu tvarov slovenského jazyka, z ktorej sme získavali informácie o slovách vo vetách. V rámci vetného rozboru sme určovali podmet, slovesný prísudok, menný prísudok, predmet, príslovkové určenie, zhodný a nezhodný prívlastok.

Na základe rozličných vzťahov, ktoré v slovenčine existujú medzi slovami vo vete, sme vyvinuli systém, ktorý skúmal akými vetnými členmi jednotlivé slová sú. Vzťahy a vetné členy sme následne vizualizovali vo forme grafu. Pri viacslovných vetných členoch sme túto skutočnosť znázorňovali pomocou vnorených vrcholov.

V tejto práci sme skúmali nie všetky typy viet v slovenčine. Zaoberali sme iba vetami, ktoré sú dvojčlenné. Skúmali sme vybrané možnosti spájania slov vo vetách vzhľadom na veľké množstvo existujúcich kombinácií. Ostáva teda možnosť tento model vylepšiť.

Zoznam použitej literatúry

[1] Dvonč, L. a kol.: Morfológia slovenského jazyka [online]. Bratislava: Vydavateľstvo Slovenskej akadémie vied, 1966, s. 302 [cit. 2015-01-28], dostupný na http://www.juls.savba.sk/ediela/msj/msj-hq.pdf

[2] http://dev.mysql.com/downloads/connector/j/

[3] https://github.com/jgraph/jgraphx

[4] Krajči, S., Novotný, R.: Projekt Tvaroslovník – slovník všetkých tvarov všetkých slovenských slov, Znalosti 2012, zborník príspevkov 11. ročníka konferencie: 14. - 16. október 2012, Mikulov (Česko), Praha, MATFYZPRESS, Vydavatelství MFF UK v Praze, 2012, ISBN 9788073782207, s. 109–112

[5] Mistrík, J.: Gramatika slovenčiny. Bratislava: Slovenské pedagogické nakladateľstvo 1994

[6] projekt Tvaroslovník, http://tvaroslovnik.ics.upjs.sk/