UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA, PRÍRODOVEDECKÁ FAKULTA

Stromová reprezentácia slovenskej vety

Autor: Jana Hiľovská

Vedúci práce: RNDr. Ondrej Krídlo, PhD.

Konzultant: doc. RNDr. Stanislav Krajči, PhD.

Úvod

Naša diplomová práca sa bude zaoberať problematikou spracovávania textu písaného v prirodzenom jazyku. Konkrétne sa budeme zaoberať automatizáciou vetného rozboru. Vetný rozbor je disciplínou syntaxe – náuky o tvorbe viet, ktorá je jednou z častí gramatiky. Táto disciplína sa zaoberá určovaním, akými vetnými členmi sú slová vo vete a aké sú vzťahy medzi týmito slovami. Slovenský jazyk má niekoľko slovných druhov (napríklad podmet, prísudok, predmet a ďalšie) medzi ktorými sú priraďovacie, či podraďovacie vzťahy. Našou úlohou v tejto diplomovej práci bude navrhnúť a implementovať algoritmus, ktorý tieto vzťahy medzi slovami určí automaticky. Keďže vzťahy medzi slovami majú hierarchickú štruktúru, výsledkom algoritmu by mala byť stromová štruktúra slovenskej vety.

Asi je prirodzené položiť si otázku, načo je vlastne výskum v takejto oblasti potrebný. Aj keď to možno na prvý pohľad nie je vidieť, automatická analýza textu má využitie v čoraz populárnejšom odbore vedy a techniky – v umelej inteligencii. Ak totiž naučíme počítač (robota) rozoznávať nielen jednotlivé slovné druhy vo vete, ale aj vzťahy medzi týmito slovami, sme o krok bližšie k tomu, aby sme ho naučili rozumieť tomu, čo hovoríme. Ak totiž zvládneme automatizovať syntaktickú rovinu jazyka, ďalším krokom je naučiť počítač sématiku slov a prototyp umelej inteligencie je hotový. Hoci súčasné programy typu Siri, či Cortana od Microsoftu sa umelú inteligenciu snažia simulovať, v skutočnosti významom slov nerozumejú.

Charakteristika problému

Ako sme už spomenuli v predchádzajúcej kapitole, slovenskú vetu tvorí viacero vetných členov. Môžeme ich rozdeliť na hlavné vetné členy a rozvíjacie vetné členy. Medzi týmito entitami sú vo vete vzťahy, ktoré vieme reprezentovať ako strom.

Hlavné vetné členy v slovenskej vete:

- Podmet
    - podľa jazykovedcov vykonávateľ činnosti, alebo nositeľ stavu
    - býva vyjadrený podstatným menom alebo zámenom v nominatíve, prípadne iným plnovýznamovým slovným druhom.
- Prísudok
    - vyjadruje činnosť, stav alebo vlastnosť podmetu.
    - slovesné prísudky sú vyjadrený slovesom
    - menné prísudky sú vyjadrený spojením sponového slovesa a plnovýznamového slovného druhu.

Rozvíjacie vetné členy

- Predmet
    - zvyčajne vyjadrený podstatným menom alebo zámenom
    - viaže sa s pádmi rôznymi od nominatívu. Rozvíja prísudok.
- Príslovkové určenie
    - rozvíja sloveso, prídavné meno alebo príslovku.
    - otázky: kde?, kedy?, ako?, prečo?.
    - tvorené príslovkou alebo podstatným menom.

- Prívlastok
    - rozvíja podstatné meno.
    - 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.

Hlavným problémom pri automatickom spracovávaní viet zo skupiny flektívnych jazykov, do ktorých patrí aj slovenčina je viactvarovosť slov. Na rozdiel od napríklad angličtiny, kde plnovýznamové slová sa nemenia, v slovenčine môže mať slovo s jedným významom niekoľko tvarov, napríklad keď ho vyskloňujeme. Pre implementáciu nášho riešenia to teda bude znamenať pracovať s pomerne veľkým množstvom dát, keďže k jednému slovnému druhu (napríklad podstatnému menu) musíme brať do úvahy aj všetky jeho tvary (napríklad podstatné meno vo všetkých jeho pádoch). Slovenské slová je nutné pred uložením do databázy spracovať, aby sme potom v nich vedeli napríklad vhodne vyhľadávať, či ďalej s týmito slovami pracovať. Ďalším sťažením syntaktickej analýzy je nepravidelný slovosled slovenčiny. Na rozdiel od už spomínanej angličtiny, či napríklad aj španielčiny, ktoré majú pomerne pravidelný slovosled a teda by sme s pomerne vysokou presnosťou vedeli určiť ktorý slovný druh sa vyskytuje na danej pozícii vo vete, pri slovenčine je potrebné urobiť hlbšiu analýzu slov a ich tvarov.

Analýza a návrh riešenia

Tvaroslovník

Na to, aby sme dokázali vetný rozbor zautomatizovať určite potrebujeme vedieť o jednotlivých slovách niekoľko dodatočných informácií. Pri našej diplomovej práci budeme používať databázu tvarov slov slovenského jazyka Tvaroslovník, ktorá bola vytvorená na našej fakulte. V tejto databáze sa nachádza množstvo slovenských slov, ich tvarov a metadát, ktoré budeme používať pri praktickej implementácii nášho riešenia. Tvaroslovník 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.

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, kde slovo v základnom tvare má túto hodnotu nastavenú na 0
- tvar – textový tvar slova
- slovnýDruh – hovorí o slovnom druhu príslušného slova
- charakteristika – textový zoznam hodnôt gramatických kategórií slova. Tieto hodnoty sú oddelené bodkočiarkami a závisia od konkrétneho slovného druhu

Príklad záznamu v databáze Tvaroslovník:
Ak hľadané slovo je „škole“, záznam vyzerá nasledovne:

- škole [škola]
  podstatné meno
  ženský r.; singulár; datív;

- škole [škola]
  podstatné meno
  ženský r.; singulár; lokál;

Je dôležité povedať, že nie všetky záznamy v Tvaroslovníku sú takéto jednoznačné.
V slovenskom jazyku sa v značnej miere vyskytujú slová známe ako homonymá, ktoré majú rovnaký tvar, no rozličný význam.

Preložené do problematiky spracovania prirodzeného jazyka to znamená, že v Tvaroslovníku majú rovnaký záznam v stĺpci *tvar*, no rôzny záznam v stĺpci *slovnyDruh,* čo spôsobuje nejasnosť v syntaktickom rozklade slovenskej vety.

Typickým príkladom je slovo **mám**, pre ktoré nájdeme v Tvaroslovníku nasledujúce záznamy:

- mám [mama]

  podstatné meno

  ženský r.; plurál; genitív;

- mám [mámiť]

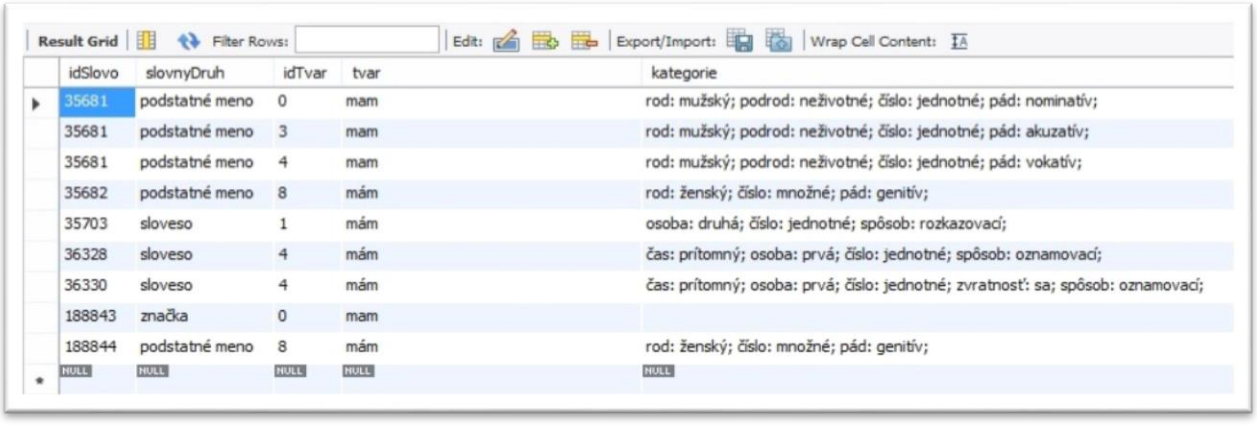
  sloveso

  osoba: druhá; singulár; spôsob: rozkazovací;

- mám [mať]

  sloveso

  čas: prítomný; osoba: prvá; singulár; spôsob: oznamovací;



Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

idSlovo	slovnyDruh	idTvar	tvar	kategorie
35681	podstatné meno	0	mam	rod: mužský; podrod: neživotné; číslo: jednotné; pád: nominatív;
35681	podstatné meno	3	mam	rod: mužský; podrod: neživotné; číslo: jednotné; pád: akuzatív;
35681	podstatné meno	4	mam	rod: mužský; podrod: neživotné; číslo: jednotné; pád: vokatív;
35682	podstatné meno	8	mám	rod: ženský; číslo: množné; pád: genitív;
35703	sloveso	1	mám	osoba: druhá; číslo: jednotné; spôsob: rozkazovací;
36328	sloveso	4	mám	čas: prítomný; osoba: prvá; číslo: jednotné; spôsob: oznamovací;
36330	sloveso	4	mám	čas: prítomný; osoba: prvá; číslo: jednotné; zvratnosť: sa; spôsob: oznamovací;
188843	značka	0	mam	
188844	podstatné meno	8	mám	rod: ženský; číslo: množné; pád: genitív;
NULL	NULL	NULL	NULL	NULL

Obr. 1: Tvaroslovník

Charakteristika slovných druhov

Pri návrhu riešenia bude dôležitým ukazovateľom vzťahu medzi slovami nielen samotný slovný druh, ale špecifické gramatické kategórie, ktoré tomuto slovnému druhu prislúchajú. Nasledujúce údaje sú uvedené v databáze pre každé slovo v stĺpci charakteristika. Pre prehľadnosť rozdelíme v tejto práci slovné druhy na plnovýznamové a neplnovýznamové, pričom hľadíme na to, že len plnovýznamové slovné druhy majú schopnosť stať sa vetným členom, t.j.: majú vetnočlenskú platnosť

- Plnovýznamové slovné druhy
    - podstatné mená majú rod, číslo, pád, vzor a informáciu o tom, či sú životné alebo neživotné
    - Prídavné mená majú rod, číslo, pád, ak ide o prídavné mená mužského rodu, aj podrod. Akostné a vzťahové prídavné mená majú aj uvedené, v akom sú stupni.
    - 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.
    - 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.

- Neplnovýznamové slovné druhy
    - 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 prí- slovkami. 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.

Analýza a návrh riešenia

Aktuálny návrh riešenia pozostáva z postupného prechádzania jednotlivých slov vo vete, medzi ktorými hľadáme vzťahy. K rozpoznaniu vzťahov nám pomáha databáza Tvaroslovník. Na rozdiel od diplomovej práce kolegu Júliusa Mareša, ktorá sa zaoberala týmto odvetvím informatiky sa naša práca primárne nesústredí na určovanie vetných členov. Naopak, zvolili sme opačný prístup, v ktorom najprv určíme vzťahy medzi slovami, z ktorých potom budú viditeľné nielen samotné vetné členy, ale aj celý syntaktický rozbor vety.

Každý vetný člen má svoje špecifické vlastnosti nielen čo sa týka gramatických kategórií, či schopnosti byť vetným členom, ale aj väzbami s inými slovnými druhmi. Typickým príkladom môže byť prídavné meno, ktoré sa v slovenskej vete typicky viaže iba s podstatným menom. Túto skutočnosť sme sa rozhodli využiť aj pri aktuálnom návrhu riešenia. Prechádzame teda vetu

alebo text na vstupe a pre každé slovo postupne hľadáme slovo resp. slová, s ktorými môže byť vo vzťahu. Ak nájdeme k aktuálne spracovávanému slovu dvojicu, ihneď vytvoríme vzťah. Určíme, ktorá z dvojice je nadradená a začneme vytvárať strom tak, že podradený vetný člen umiestnime pod jeho nadradeného partnera. Vznikne teda veta o jedno slovo kratšia, pričom už je známy prvý vetný sklad. Takýmto spôsobom prechádzame všetky slová na vstupe a postupne skracujeme vetu až na dva členy. Pri správnej implementácii by na konci prehľadávania mali zostať dva slovné druhy a to podstatné meno a sloveso, ktoré tvoria prisudzovací sklad.

V predchádzajúcom odstavci sme rozobrali ideálny príklad vstupu – dvojčlennú jednoduchú vetu. V slovenskom jazyku sa však vyskytuje viac typov viet, s ktorými sa implementácia spomenutého návrhu musí vysporiadať. Asi najčastejším sú vety so zamlčaným podmetom, kedy pri prisudzovacom sklade nie je podstatné meno vo vete uvedené. V tomto prípade na konci algoritmu zostane iba jeden slovný druh – sloveso a algoritmus zistí prítomnosť zamlčaného podmetu, ktorý vieme bližšie špecifikovať na základe slovesných kategórií.

Ďalším problémom, s ktorým sa musíme vysporiadať sú homonymá v slovenčine. Keďže nejde vôbec o ojedinelý jav, návrh riešenia aj jeho implementácia sa má vedieť vysporiadať aj s prípadom, že pri hľadaní slova v Tvaroslovníku nepríde jednoznačná odpoveď nielen čo sa týka gramatických kategórií, ale dokonca nemusí byť jasné ani to, o aký slovný druh ide. Typickým prípadom je slovo *mám* spomenuté v predchádzajúcej kapitole. V takomto prípade sa syntaktický rozbor bude vetviť a výstupom z programu bude viacero stromov.

Na tieto výstupy by sa v budúcnosti mohla aplikovať niektorá zo štatistických metód, aby sa určilo, ktorý zo stromov je pravdepodobnejší. Táto štatistika by však mala výpovednú hodnotu iba vtedy, ak by sme mali k dispozícii rozsiahlejšiu zbierku súvislých textov, s čím návrh nášho riešenia zatiaľ nepočíta. Ďalším priestorom na zlepšenie by bolo doplnenie návrhu tak, aby spracovával aj zložené vety, t.j. vety s viacerými prísudkami.

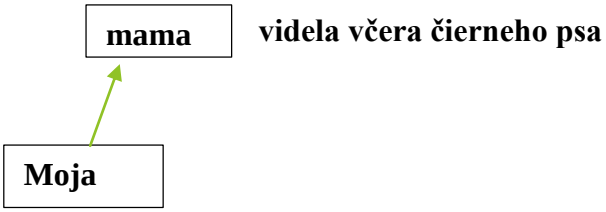
Ilustračný príklad

Návrh riešenia ilustrujeme na príklade. Majme dvojčlennú rozvitú vetu:

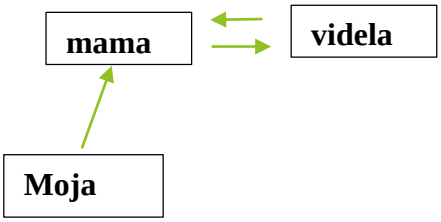
Moja mama videla včera čiernych psov

Priebeh algoritmu:

1. Prvé slovo je slovo **Moja** . Z databázy Tvaroslovník zistíme, že ide výlučne o zámeno, t.j. neexistuje k nemu homonymum. Vieme, že zámeno stojí zvyčajne za podstatným menom. Hľadáme teda najbližšie podstatné meno a utvoríme vzťah

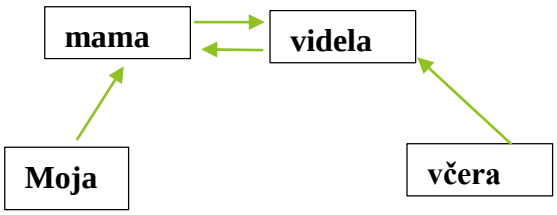


2. Ďalšie slovo je **mama** . Odpoveď z Tvaroslovníka je opäť jednoznačná. Ide o podstatné meno v nominatíve. Znamená to teda, že sa bude viazať s najbližším slovesom v rovnocennom vzťahu.

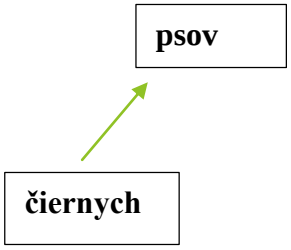


3. Ďalším slovom je **videla** . Toto slovo sme už do vzťahu dali v predchádzajúcim kroku, teda už ho nespracovávame.

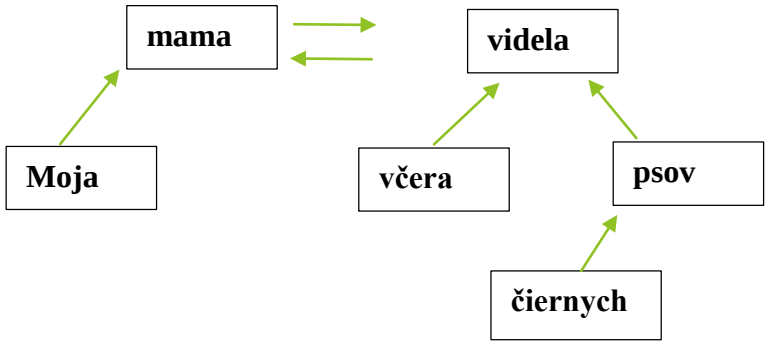
4. Ďalším slovom je **včera** . Tvaroslovník opäť jednoznačne určí, že ide o príslovku. Tá sa viaže so slovesom. Vytvoríme vzťah



5. Ďalším slovom je **čiernych** . V tomto prípade Tvaroslovník vráti dve možnosti. Toto slovo je buď genitív plurálu od slova *čierna* a teda je to podstatné meno alebo je privlastňovacie prídavné meno. V takomto prípade by vznikli dva výstupy, kde v prvom (sématicky nezmyselnom) by slovo **čiernych** bolo ako predmet naviazané na sloveso a v druhom, ktorému sa budeme venovať v ďalších krokoch je slovo **čiernych** správne naviazané na najbližšie prídavné meno. Všimnime si, že sme určili vzťah **čiernych – psov** bez toho, aby sme vedeli, kam sa naviaže slovo *psov*



6. Posledným slovom je **psov** . Tu Tvaroslovník opäť jednoznačne určí, že ide o podstatné meno. Toto podstatné meno nie je v nominatíve, preto ho naviažeme ako podradené pod sloveso. Všimnime si, že slovo *čiernych* putuje do vzťahu so slovesom spolu so slovom *psov* . Keďže ide o posledné slovo, dostávame výsledný strom vety:



V tomto príklade ako výstup figuruje iba jedna (tá správna) stromová reprezentácia vety. Nezabúdajme však na nejednoznačnosť slov, ktorá vznikla v bode 5. skutočným výstupom by teda boli dva stromy.

Implementácia

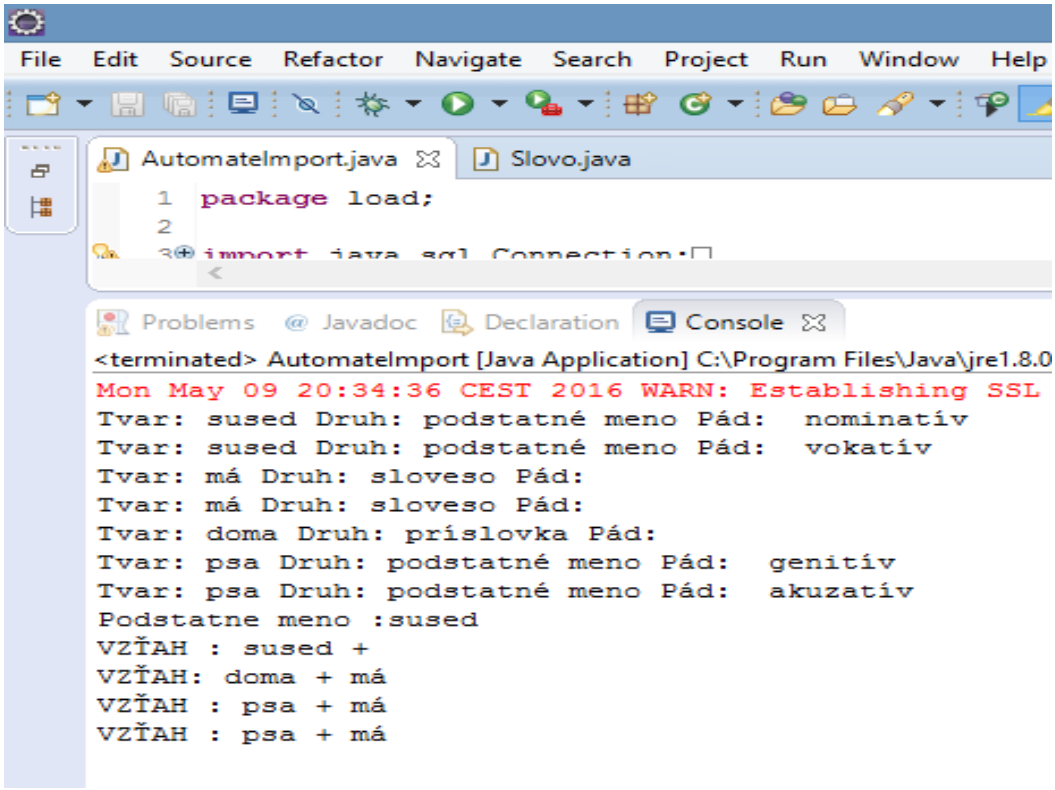
Pred implementáciou samotného algoritmu sme najprv automatizovali import samotnej databázy. Keďže Tvaroslovník je uložený v textových súboroch, ktorých je vyše 200 000, nebolo možné tento import urobiť ručne. Vytvorili sme si teda jednoduchú triedu AutomateImport, ktorá postupne tieto textové súbory prechádzala a ukladala ich do MySQL databázy. Táto trieda, ako aj samotná implementácia algoritmu je v programovacom jazyku Java.

V aktuálnej verzii máme zareprezentované slovo ako základnú jednotku vetného rozboru. Od tejto triedy potom dedia plnovýznamové a neplnovýznamové slová, ktoré túto triedu rozširujú o svoje vlastné gramatické kategórie. Vetu reprezentujeme ako spájaný zoznam, pretože táto dátová štruktúra najlepšie spĺňa požiadavky.

V ďalších verziách však plánujeme vzťah medzi jednotlivými slovnými druhmi nahradiť samostatnou triedou a to najmä z dôvodu toho, že si potrebujeme poznačiť nielen to, ktoré dve slová sú „vedľa seba“ a teda tvoria pár, ale aj to, ktorý slovný druh je nadradený resp. podradený. Samotný strom vety potom budeme vedieť reprezentovať napríklad pomocou knižnice PAZGraph, kde uzly budú práve objekty triedy Slovo a vzťahy sa určia na základe implementovanej reprezentácie.

Pri samotnej implementácii sme sa stretli s niekoľkými problémami. Hoci v ilustračnom príklade sme túto skutočnosť opomenuli, pri implementácii nebolo získanie informácií z Tvaroslovníka takéto jednoznačné. Ako sme uviedli pri charakteristika samotnej databázy, jeden tvar slova môže mať rôzne gramatické kategórie (napr. rod alebo číslo), uvedené v stĺpci *charakteristika* . Hoci teda hľadané slovo nemá homonymá a vyskytuje sa iba ako jediný slovný druh, pri implementácii nám odpoveď z Tvaroslovníka prišla vo viacerých riadkoch.
Asi najzávažnejších problémom sú homonymá, ktoré sme spomínali v predchádzajúcich kapitolách. Tieto nejasnosti pri syntaktickom rozbore súvisia s nutnosťou pridať do rozboru aj sématickú (významovú) zložku - musíme sa vedieť opýtať pádovou otázkou, aby sme vedeli určiť vetný člen.
Práve túto zložku budeme musieť získať na vstupe od používateľa, resp. ju simulovať využitím štatistického prístupu. Aktuálna verzia algoritmu teda je schopná vyhodnotiť iba malé percento špecifických viet a to také, v ktorých sa nevyskytujú homonymá.

Aktuálne implementovaná verzia pracuje v textovom režime, v ktorom vstup zadávame ručne do konzoly a takisto výstup – vzťahy medzi jednotlivými slovami sú používateľovi odprezentované v textovej forme



Obr. 2: Výstup aktuálnej verzie programu

Finálna verzia programu, ktorá by mala byť súčasťou diplomovej práce má byť schopná spracovať dokumenty v textových súboroch. Pri nejednoznačnosti vetného rozboru potom vieme využiť niektorú zo štatistických metód a to v prípade, že bude text dostatočne rozsiahly, prípadne zabezpečíme komunikáciu s používateľom, aby si sám vybral, ktorý z ponúkaných rozborov vety má aj sématický význam. Výstup programu má už byť v grafickej podobe.

Zoznam použitej literatúry

Projekt Tvaroslovník, http://tvaroslovnik.ics.upjs.sk/

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

Konferencia SLovko 2015, http://korpus.sk/~slovko/2015/Proceedings_Slovko_2015.pdf