Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky

Automatická identifikácia konceptov v textoch v prirodzenom jazyku

Dizertačná práca

Študijný program:	Umelá inteligencia
Študijný odbor:	9.2.8 Umelá inteligencia
Školiace pracovisko:	Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ:	doc. Ing. Marián Mach, CSc.

Košice 2013 **Ing. Viliam Ročkai**

Analytický list

Autor:	Ing. Viliam Ročkai
Názov práce:	Automatická identifikácia konceptov v textoch v prirodzenom jazyku
Podnázov práce:	
Jazyk práce:	slovenský
Typ práce:	Dizertačná práca
Počet strán:	146
Akademický titul:	PhD.
Univerzita:	Technická univerzita v Košiciach
Fakulta:	Fakulta elektrotechniky a informatiky (FEI)
Katedra:	Katedra kybernetiky a umelej inteligencie (KKUI)
Študijný odbor:	9.2.8 Umelá inteligencia
Študijný program:	Umelá inteligencia
Mesto:	Košice
Vedúci práce:	doc. Ing. Marián Mach, CSc.
Konzultant(i) :	
Dátum odovzdania:	3. 1. 2013
Dátum obhajoby:	31. 8. 2013
Kľúčové slová:	koncept, relácia, AUP, konfabulácia, sémantická podobnosť
Kategória:	Technika, technológie, inžinierstvo a pod.
Citovanie práce:	Ročkai Viliam: Automatická identifikácia konceptov v textoch v prirodzenom jazyku. Dizertačná práca. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky. 2013. 146 s.
Názov práce v AJ:	Automatic discovery of concepts in documents written in natural language
Kľúčové slová v AJ:	concept, relation, AUP, confabulation, semantic similarity

Abstrakt v SJ

Táto dizertačná práca sa venuje problematike dolovania konceptov a relácií z textov v prirodzenom jazyku. Automatizované prístupy sú často závislé na lexikálnych zdrojoch, preto sa predpokladá ich možné rozšírenie pomocou automatického učenia pojmov (AUP). Popis teoretických základov a funkcionality AUP tvorí jadro práce. Nad AUP je vykonaný súbor experimentov v doméne úloh budovania konceptuálnych sietí z textov v prirodzenom jazyku pomocou implementácie Beast. Metódy AUP sú prezentované na rôznych doménových oblastiach a to tak, aby ich bolo možné kvalitatívne a kvantitatívne porovnať. Experimentálne výsledky potvrdzujú spoľahlivosť jeho metód na rozpoznávanie sémantickej podobnosti slov a extrakciu relácií z textov na základe ukážkovej relácie. S ich využitím sa podarilo nájsť plné zhluky konceptov patriacich do jednej sémantickej triedy ako napríklad mesiace v roku, číslovky, mená osôb alebo farby, a to rovnako pre angličtinu ako aj pre slovenčinu a češtinu. Táto práca predstavuje novú univerzálnu metódu dolovania konceptov a relácií z textov v prirodzenom jazyku, ktorá pracuje bez znalostí o jazyku.

Abstrakt v AJ

This thesis deals with a practice of concept and relation mining from natural language texts. Automated approaches are ussually dependent on lexical resources, therefore it would be useful to extend them with associative learning of concepts (ALC) method. Core of the thesis is dedicated to ALC description. With ALC implementation called Beast, several experiments targeted for semantic networks building from natural language texts are executed. Within them, methods of ALC are presented on several domains, so it's possible to compare them in several ways. Experiment results confirm the applicability of ALC methods for word semantic relatedness computation, as well as relation extraction from texts, based on given example relation. ACL succeeded in identification of clusters of words belonging into a single distinct semantic class, like months, numbers, personal names of colors. Similar accuracy was achieved in several languages - English, Czech and Slovak. This thesis presents a new universal method for concept and relation extraction from natural language texts without any knowledge about specific language.

Čestné vyhlásenie

Vyhlasujem, že som dizertačnú prácu vypracoval samostatne s použitím uvedenej odbornej literatúry.

Košice 3. 1. 2013 .....

Vlastnoručný podpis

Poďakovanie

Touto cestou by som sa chcel poďakovať môjmu vedúcemu, pánovi doc. Ing. Mariánovi Machovi, CSc. za cenné rady a pripomienky k mojim školským aktivitám. Za cenné pripomienky a podnety by som sa chcel osobitne poďakovať aj Mgr. Miloslavovi Nepilovi, Phd., Ing. Petrovi Bednárovi, Phd., Ing. Petrovi Kostelníkovi, Phd., Ing. Tomášovi Kasanickému a Ing. Františkovi Šimků.

Predhovor

„Väčšina ľudí si myslí, že intelekt robí veľkého vedca. Nemajú pravdu. Robí to charakter.“ (Albert Einstein)

K výberu tejto dizertačnej práce som sa rozhodol na základe záujmu hlbšie preniknúť do oblasti konceptuálneho modelovania, ktorú som načal už počas inžinierskeho štúdia. Silnou motiváciou mi bola taktiež myšlienka, že môžem sám prispieť do rozvoja oblasti danej problematiky.

Obsah

Zoznam použitých skratiek

ALC Associative learning of concepts - Asociatívne učenie pojmov.

AUP Asociatívne učenie pojmov.

BC Base Concepts - základné koncepty.

BSD Berkeley Software Distribution - softvérová licencia.

DIPRE Dual Iterative Pattern Relation Expansion.

FEI Fakulta elektrotechniky a informatiky.

FI Fakulta informatiky.

GB Gigabytes.

GSD Google similarity distance.

GT Google translate - on-line služba firmy Google.

HTML Hypertext Markup Language.

ILI Interlingual Index - medzijazykový index.

KG Konceptuálny graf.

MB Megabytes.

MI Mutual information - vzájomná informácia.

MRD Machine readable dictionary - Strojovo čitateľný slovník.

MUNI Masarykova univerzita.

NER Named entity recognition - rozpoznávanie pomenovanej entity.

NLP Natural Language Processing - spracovanie prirodzeného jazyka.

OOP Objektovo orientované programovanie.

PMI Pointwise mutual information - bodová vzájomná informácia.

PWN Princeton Wordnet.

SLG Simple language generator.

SN Sémantická sieť.

SNePS Semantic Network Processing System.

TOC Top-Ontology Concepts - vrcholová ontológia výrazov.

TU Technická univerzita.

UI Umelá inteligencia.

UML Unified Modeling Language - zjednotený jazyk na modelovanie.

URL Uniform Resource Locator - jednotný lokátor zdrojov.

WN Wordnet iný ako Princetonský.

WWW World Wide Web - súbor hypertextových dokumentov na internete.

Slovník pojmov

Doménový model je typom konceptuálneho modelu, ktorý sa používa na znázornenie prvkov, ich štruktúry a konceptuálnych ohraničení v rámci nejakej oblasti záujmu.

Hebbovo pravidlo - Keď má axón bunky A excitačný účinok na bunku B a opakovane alebo vytrvalo sa zúčastňuje na jej aktivácii, v jednej alebo v oboch bunkách prebehne nejaký rastový proces alebo metabolická zmena, takže účinnosť bunky A ako jednej z buniek, ktoré aktivujú B, vzrastie [4].

Korpus je v rámci lingvistiky veľká a štrukturovaná množina textov.

Mozgová kôra je sídlom poznávacích (kognitívnych) procesov. Kôra spolupracuje s podkôrovými centrami, ktoré sú umiestnené v strede mozgu a sú evolučne staršie ako kôra [4].

Multiagentové systémy sú zložené z viacerých vzájomne sa ovplyvňujúcich výpočtových prvkov – tzv. agentov. Agenti sú výpočtové systémy s dvoma dôležitými vlastnosťami. Sú schopné autonómnej akcie a sú schopné interagovať s ostatnými agentami a to nielen v podobe prenosu dát, ale aj v podobe analogickej sociálnej aktivity - spoluprácou, koordináciou, vyjednávaním atď.[76]

Petriho sieť je orientovaný graf s dvoma druhmi uzlov, „miestami“ a „prechodmi“, s podmienkou, že žiadna hrana nespája uzly toho istého druhu [15].

Talamus je sústava viacerých neurónových zoskupení v strede mozgu. Je evolučne starší ako mozgová kôra. Hrá úlohu pri spracovaní zmyslových vstupov, budení kôry a chode organizmu [55].

Turingov test bol zostavený Alanom Turingom a slúži na posúdenie inteligencie výpočtového systému. Test je založený na schopnosti stroja v komunikácii presvedčiť svojho ľudského partnera, že sa rozpráva so živým človekom [30].

1 Úvod

„Veľa hovoriť a veľa povedať nie je to isté.“ (Homér)

Spracovanie textov v prirodzenom jazyku (NLP - z „natural language processing“) je jednou z vedných disciplín umelej inteligencie (UI). Za jeho počiatok sa pokladá rok 1950, keď Alan Turing publikoval koncept Turingovho testu [72]. Domáce počítače sú dnes oveľa rýchlejšie ako najrýchlejšie počítače sveta spred pár rokov. Spolu s rozšírením slobodných licencií dochádza k zlepšeniu dostupnosti výpočtových zdrojov a vývoj v rôznych oblastiach NLP sa tak stáva prístupným aj pre bežných ľudí. Výpočty, ktoré by pred desiatimi rokmi zabrali niekoľko dní, je možné vykonať za pár minút. Obrovské korpusy textov v prirodzenom jazyku sú prístupné zadarmo na internete. To ešte viac urýchľuje už teraz veľmi rýchly a neprehľadný pokrok nielen v oblasti NLP, ale vo všetkých oblastiach umelej inteligencie. Je možné, že veľa prístupov z minulosti, ktoré boli zavrhnuté kvôli svojej výpočtovej náročnosti či prehnaným nárokom na veľkosti korpusov, nad ktorými pracovali, sa budú pomaly vracať ako nové a úspešné.

Aj napriek pozitívnemu vývoju je stále v oblasti NLP množstvo nevyriešených problémov. Turingov test [72], automatické rozpoznávanie entít [41] v texte a relácií [39] medzi nimi, extrakcia gramatických pravidiel [44], nejednoznačnosť významu slov [32] a strojový preklad [34], to všetko tvorí iba zlomok oblastí stále otvorených pre nové originálne myšlienky.

Táto práca sa venuje konkrétnej oblasti NLP, a to problematike automatickej extrakcie konceptov a relácií z textov v prirodzenom jazyku. Asociatívne učenie pojmov (AUP) je pôvodný prístup, navrhnutý v tejto práci, ktorý je schopný zvládnuť extrakciu jednoslovných konceptov, využívajúcich jeden typ relácie bez použitia lexikálnych zdrojov a s minimálnym využitím iných znalostí o jazyku, s ktorým pracuje. Pod lexikálnymi zdrojmi sa rozumejú akékoľvek podporné znalosti v podobe jazykových slovníkov. Pod znalosťami o jazyku sa rozumie akákoľvek znalosť o štruktúre

daného jazyka a použitá je jedine znalosť o segmentácii slov a viet. Oblasť extrakcie konceptov a relácií je úzko prepojená s formou reprezentácie znalostí v podobe sémantických sietí, o ktorej táto práca tiež pojednáva. Samotná úloha extrakcie konceptov sa dá transformovať na úlohu vytvárania stromu hierarchicky usporiadaných konceptov, čo je priamo jeden zo základných typov sémantických sietí.

Motivácií pre vznik tejto práce bolo viacero, avšak tou hlavnou bol osobný príspevok do výskumu v oblasti NLP. Závislosť nástrojov extrakcie konceptov a relácií z textov v prirodzenom jazyku na lexikálnych zdrojoch, dáva mnoho priestoru na vylepšenia. V neposlednom rade bolo silným motivačným faktorom aj spoznávanie špecifík výskumnej činnosti a túžba po sebarealizácii v tejto sfére.

1.1 Štruktúra práce

Úvodná kapitola 1 objasňuje motiváciu a načrtáva problematiku, ktorou sa práca zaoberá. Táto kapitola ďalej obsahuje popis štruktúry a je ukončená definíciou cieľov práce.

Kapitola 2 popisuje vybrané druhy sémantických sietí. Špeciálnu pozornosť venuje sieťam, pomocou ktorých je možné modelovať bežné znalosti o svete. Po krátkej a stručnej definícii druhov týchto sietí nasleduje popis vybraných konkrétnych implementácií. Špeciálna pozornosť je venovaná lexikálnej databáze anglického jazyka - projektu WordNet (a jeho jazykovým mutáciám), ktorý je sám reprezentovaný vo forme lexikálnej ontológie (teda konceptuálnej siete) a nachádza široké uplatnenie ako pomocný nástroj pri rôznych metódach vytvárania konceptuálnych sietí.

Kapitola 3 ponúka popis vybraných metód tvorby konceptuálnych sietí, ktoré sú do určitej miery závislé na lexikálnych zdrojoch. Začiatok kapitoly je zameraný na automatické metódy budovania wordnetových (WN) slovníkov, ktoré využívajú znalosť jazyka, nad ktorého textami pracujú. Nasleduje popis metód na extrahovanie relácií z textov v prirodzenom jazyku. V závere kapitoly je predstavený úspešný prístup extrakcie relácií DIPRE spolu s jeho rozšírením Snowball. Tieto automati-

zované prístupy sú závislé na lexikálnych zdrojoch, preto sa predpokladá ich možné rozšírenie pomocou automatického učenia pojmov (AUP).

Kapitola 4 zahŕňa pôvodný prístup učenia konceptuálnych štruktúr - AUP. Obsahuje popis jeho teoretických základov a funkcionality. Základné funkcie sú prezentované na príkladoch. (Táto kapitola) Obsahuje úplný prehľad súčasného stavu AUP.

Stručný opis programu Beast, ktorý je pôvodnou open-source implementáciou AUP v jazyku Java, je obsahom 5. kapitoly.

Kapitola 6 obsahuje experimentálne overenie metódy AUP v doméne úloh budovania konceptuálnych sietí z textov v prirodzenom jazyku pomocou jeho implementácie Beast. Základné funkčné bloky AUP prezentuje na rôznych doménových oblastiach (korpusoch líšiacich sa v jazykoch alebo svojou veľkosťou) a to tak, aby ich bolo možné kvalitatívne a kvantitatívne porovnať.

Dosiahnuté výsledky zhodnocuje kapitola 7. Zároveň rekapituluje najdôležitejšie body tejto práce a navrhuje potrebné smerovanie výskumu do budúcnosti.

1.2 Ciele práce

Táto práca sa snaží prispieť do oblasti objavovania relácií medzi konceptmi v textoch v prirodzenom jazyku. Kladie si viacero cieľov:

- *Identifikácia medzier* v oblasti automatickej extrakcie konceptov a relácií z textov v prirodzenom jazyku.
- *Návrh konkrétnej metódy*, ktorá by na základe vstupu textu v prirodzenom jazyku bola schopná v danom texte zachytávať relevantné koncepty prepojené pomenovanou, jasne určenou reláciou.
- *Experimentálne vyhodnotenie* tejto metódy nad korpusmi s rôznymi charakteristikami.
- *Identifikácia slabých miest* a určenie hraníc použiteľnosti metódy AUP.

Prvý cieľ, *identifikácia medzier*, predstavuje prehľad súčasného stavu techník extrakcie konceptov a relácií z textov v prirodzenom jazyku, ktorému predchádza všeobecný popis sémantických sietí (SN - z ang. „semantic networks“) vrátane ich delenia. Osobitná časť je venovaná ontológiám a ich vybraným implementáciám, ktoré modelujú všeobecné znalosti. Je špeciálne zameraná na súčasný stav projektu WordNet (WN) vrátane jeho jazykových mutácií.

Druhý cieľ, *návrh konkrétnej metódy*, tvorí návrh a popis vlastnej metódy extrakcie konceptov a relácií z textov v prirodzenom jazyku. Je realizovaný jej detailným popisom vrátane možností aplikácie v rámci oblasti NLP.

Tretí cieľ, *návrh spôsobu vyhodnotenia kvality*, pozostáva z realizácie experimentov pomocou metódy AUP. Experimenty by mali potvrdiť alebo vyvrátiť funkčnosť metódy AUP. Rovnako by mali byť schopné vyhodnotiť kvalitu metódy AUP.

Štvrtý cieľ, *identifikácia slabých miest*, prináša zhodnotenie metódy AUP podporenej experimentmi. Obsahuje návrhy na zlepšenie a rozšírenie metódy AUP.

2 Sémantické siete

Sémantické siete umožňujú popisovať zložitejšie štruktúry znalostí, kedy jednotlivé objekty majú viac vlastností a sú v určitých vzájomných vzťahoch s inými objektmi. V takýchto prípadoch je vhodné spojiť takéto vlastnosti dohromady do tvaru jediného popisu zložitého objektu. Použitie sémantických sietí je v týchto situáciách výhodnejším spôsobom reprezentácie znalostí, než aké poskytujú logické formalizmy (výroková a predikátová logika), ktoré sú veľmi užitočné pre reprezentáciu jednoduchých faktov, no pre popis zložitých štruktúr nepostačujúce.

Sémantická sieť je grafickým zápisom použitým na reprezentáciu znalostí pomocou hranami navzájom prepojených uzlov. Počítačové implementácie sémantických sietí boli spočiatku vyvíjané v rámci umelej inteligencie a strojového prekladu, ale ich predchodcovia (nesoftvérové implementácie) boli známi ešte skôr z filozofie, psychológie a lingvistiky [67].

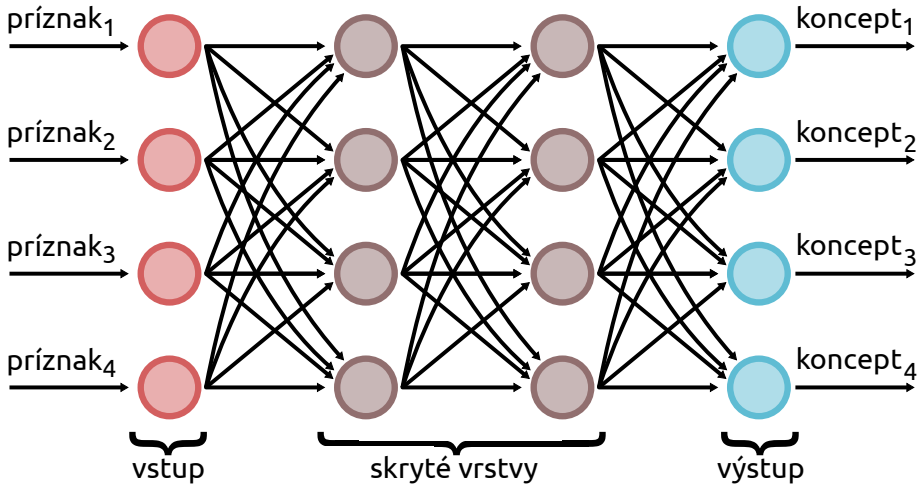
Sémantická sieť v úzkom slova zmysle je orientovaným grafom pozostávajúcim z vrcholov, ktoré reprezentujú koncepty, a hrán, ktoré reprezentujú relácie medzi danými konceptmi. Koncepty (resp. pojmy) sú viac alebo menej abstraktné konštrukty, pomocou ktorých sa buduje model relevantnej časti sveta. Sú kognitívnou jednotkou významu. Koncept je abstraktnou ideou, alebo mentálnym symbolom, niekedy definovaným ako „jednotka znalosti“.

> **Entita** *je objekt, ktorý existuje (nemusí byť materiálnej povahy) a je rozlíšiteľný od iných objektov [70]. V súvislosti s konceptmi sa pod pojmom entita rozumie inštancia konceptu [31].*

> **Koncept** *je idea alebo mentálny obraz, ktorý sa vzťahuje k určitej entite alebo triede entít, alebo ich základným príznakom, alebo určuje použitie termu (zvlášť predikátu), a tak hrá úlohu v rozhodovaní alebo jazyku [49].*

V širšom slova zmysle sa dá sémantická sieť chápať ako sieť, ktorá obsahuje

koncepty. Koncept nemusí byť reprezentovaný práve jedným uzlom a relácia tiež nemusí byť reprezentovaná hranou. Príkladom sietí, kde sú relácie popísané uzlom, sú konceptuálne grafy, ktoré sú bližšie popísané v časti 2.1.1. Koncepty môžu byť priamo súčasťou siete ako napríklad uzly v konceptuálnych grafoch, ale môžu mať aj formu výstupu, ako je to u vybraných neurónových sietí. Sowa ako príklad takejto neurónovej siete uvádza sieť na rozpoznávanie písmen. Sieť je znázornená na obr. 2 – 1. Vstupom do siete sú nájdené príznaky snímaného písmena ako úsečky, krivky a uhly. Na výstupe sa nachádzajú koncepty reprezentujúce písmená, ktoré daným príznakom zodpovedajú.



Obr. 2 – 1 Schematická ukážka neurónovej siete rozpoznávajúcej písmená abecedy na základe jednoduchých príznakov.

Rozoznáva sa niekoľko typov konceptov podľa ich charakteristiky. Sú to najmä *triedy*, *relácie*, *funkcie*, *procedúry*, *objekty*, *premenné* a *konštanty*. Tvoria zväčša komplikovanú sieťovú štruktúru – doménový model danej problémovej oblasti.

Je potrebné zdôrazniť, že nech je doménový model akokoľvek rozsiahly, nemôže si nárokovať univerzálnu platnosť. Vždy modeluje svet iba z určitého hľadiska. Ak majú byť koncepty použité zmysluplne, sú vo veľkej väčšine závislé od rámca tejktorej domény. Inými slovami, pojmy a vzťahy podstatné napríklad pre poisťovaciu spoločnosť sa takmer určite nebudú dať použiť povedzme pre vzdelávaciu inštitúciu

(hoci je možné, že istá časť modelu môže byť aj v tomto prípade spoločná pre obe domény) [21].

Z pohľadu NLP sú veľmi dôležité sémantické relácie, ktoré popisujú vzťah medzi dvoma konceptmi.

> **Sémantická relácia** *je významovým vzťahom medzi dvoma alebo viacerými konceptmi, entitami alebo množinami entít. Koncepty alebo entity tvoria integrálnu časť relácie, ktorá nemôže existovať samostatne, ale musí ich dávať do vzťahu [31].*

Medzi dôležité sémantické relácie medzi dvoma konceptmi A a B patria:

- **Meronymá** – A je časťou B.
- **Holonymá** – B je časťou A.
- **Hyponymá** – A je podriadený B.
- **Hypernymá** – A je nadriadený B.
- **Synonymá** – A označuje to isté ako B.
- **Antonymá** – A označuje presný opak toho čo B.

Ako klasický príklad sémantických sietí býva často označovaný lexikón anglického jazyka – WordNet [36, 17]. Sémantické siete vytvoril Richard H. Richens z Cambridgskej univerzity v roku 1956 [50] a mali slúžiť ako medzijazyk pri strojovom preklade (z ang. machine translation).

Podľa Sowu [67] je sémantickou sieťou grafický zápis reprezentácie znalostí v podobe prepojených uzlov a hrán. Aj napriek tomu, že Sowa to vo svojej práci nedefinuje explicitne, znalostné štruktúry, ktoré nazýva sémantickými sieťami, majú stále nejaké priame napojenie na koncepty. Buď sú koncepty prvkami grafu, alebo sú generované na výstupe grafu. To ukazuje na vybraných neurónových sieťach, kde ich numerickým výstupom je reprezentovaný nejaký koncept. Z nášho pohľadu tvoria

sémantické siete podľa Sowu konceptuálne znalostné štruktúry - konceptuálne znalostné siete.

Sowa delí sémantické siete na 6 základných typov:

- **Definičné siete** (generalizačné siete) zvýrazňujú reláciu podtypu taktiež známu ako *is-a* reláciu. Podporujú dedenie a to tak, že vlastnosti definované pre nad-typ sa automaticky propagujú do všetkých jeho podtypov – prvkov, ktoré od neho dedia. Vzhľadom na to, že definície sú už zo svojej podstaty pravdivé, automaticky považujeme aj informáciu obsiahnutú v týchto sieťach za pravdivú. Medzi tieto siete sa radia Porfýriov strom, KL-ONE a iné.

- **Rozhodovacie siete** sú navrhnuté na prácu s výrokmi. Na rozdiel od definičných sietí sa pokladá informácia v nich obsiahnutá za platnú len v prípade, že nie je explicitne označená modálnym operátorom. Niektoré asociačné siete boli navrhnuté ako modely konceptuálnych štruktúr tak, aby odzrkadľovali sémantiku prirodzeného jazyka.

  Medzi tieto siete sa radia Existencionálne a konceptuálne grafy, SNePS [63], Tesnièrove grafy závislostí a iné. Príkladom implementácie takýchto sietí sú aj projekty CYC a WordNet.

- **Implikačné siete** používajú na prepájanie vrcholov ako primárnu reláciu reláciu implikácie. Bývajú používané ako vzory pre domnienky, kauzalitu alebo inferencie. Medzi tieto siete sa radia Bayesovské siete.

- **Vykonávacie siete** v sebe zahŕňajú mechanizmus, napríklad „posielanie značky“ alebo pripojené procedúry, ktoré vedia vykonávať inferencie, posielať správy alebo vyhľadávať vzory a asociácie. Medzi tieto siete sa radia grafy toku dát (z ang. data flow diagram) a Petriho siete.

- **Učiace sa siete** stavajú, poprípade rozširujú, svoju vlastnú reprezentáciu získavaním nových znalostí z príkladov. Nové znalosti môžu zmeniť pôvodnú

sieť tým, že pridajú, respektíve odoberú, nejaké jej uzly, prípadne menia číselné hodnoty nazývané tiež váhy, ktoré vyjadrujú vzťahy medzi uzlami, a tak kvantitatívne označujú ich relácie. Medzi tieto siete sa radia neurónové siete.

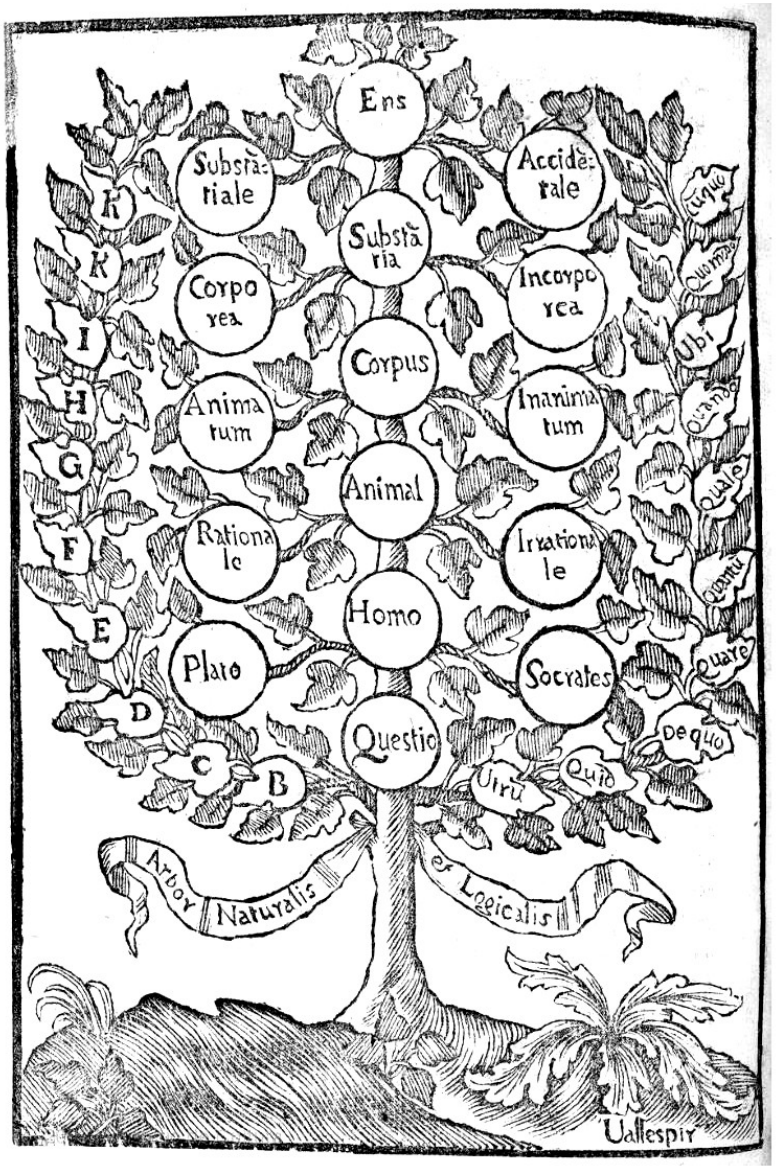
- **Hybridné siete** kombinujú dva alebo viac vyššie spomenutých prístupov. Bežne sa nemusí jednať len o spojenie týchto prístupov do jednej konkrétnej siete. Za hybridné siete sa často označujú aj množiny navzájom úzko prepojených sietí. Typickým príkladom hybridnej siete je UML (z ang. Unified Modeling Language).

Definičné a rozhodovacie siete boli špeciálne vytvorené za účelom reprezentácií znalostí z reálneho sveta. Kým definičné siete slúžia len na modelovanie *is-a* relácie, rozhodovacie siete slúžia na modelovanie sémantiky prirodzeného jazyka. Vzhľadom na zameranie tejto práce sa preto budeme venovať bližšie práve definičným a rozhodovacím sieťam, ktoré sa dajú považovať za ich rozšírenie.

2.1 Definičné a rozhodovacie siete

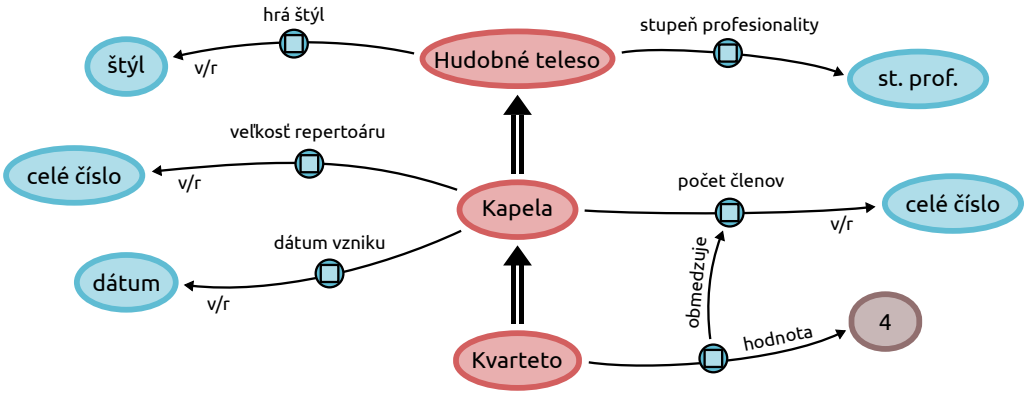
Definičné siete sú štruktúry veľmi blízke programátorom objektovo orientovaných jazykov, keďže vyjadrujú jednu zo základných vlastností objektovo orientovaného programovania (OOP), a to dedenie. Ako typický príklad definičných sietí sa zvyčajne predkladá Porfýriov strom, zobrazený na obrázku 2 – 2.

Obrázok 2 – 2 aj napriek svojmu vysokému veku (tento strom vznikol už v treťom storočí nášho letopočtu) má všetky potrebné vlastnosti, ktoré sa dnes používajú pri definovaní typov konceptov. Slúžil na zobrazenie Aristotelovej kategorizácie rodov, druhov a ich vzájomných vzťahov. Na ilustráciu modernejšieho prístupu je možné spomenúť napríklad deskriptívnu logiku, ktorá rozširuje vlastnosti Porfýriovho stromu. Deskriptívna logika je odvodená od prístupu predstavenom Woodsom v roku 1975 [75] a implementovaným Brachmanom v roku 1979 ako systém na reprezentáciu znalostí - Knowledge Language One (KL-ONE) [9]. Príklad reprezentácie znalostí pomocou KL-ONE je na obr. 2 – 3.



Obr. 2 – 2 Porfýriov strom [33].

Na obr. 2 – 3 je v systéme KL-ONE definované hudobné teleso - „kvarteto“. Červené a modré ovály znázorňujú generické koncepty a hnedý ovál znázorňuje individuálny koncept. Koncept „kvarteto“ dedí svoje vlastnosti z konceptu „kapela“, ktorý dedí opäť vlastnosti z konceptu „hudobného telesa“. Každý z konceptov má vlastné atribúty a ich obmedzenia ako napríklad „počet členov“ alebo „dátum vzniku“ a tieto atribúty sa propagujú do ich podtypov.



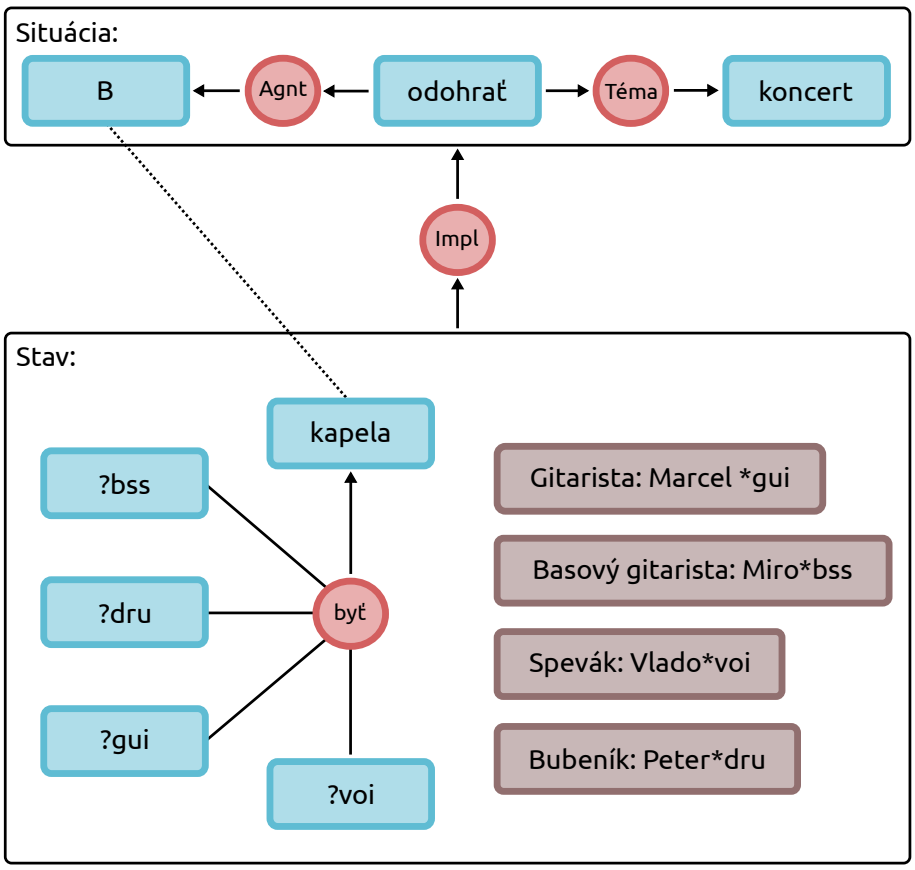
Obr. 2 – 3 Príklad konceptuálnej siete v KL-ONE.

Rozhodovacie siete sú navrhnuté na prácu s výrokmi, napríklad ich definíciu, dokazovanie a odvodzovanie. Vzhľadom na ich schopnosť dobre reflektovať znalosti z reálneho sveta bude ich vybraným typom venovaná v tejto práci osobitná kapitola.

2.1.1 Konceptuálne grafy

Konceptuálne grafy sú jednou z možností zápisu konceptuálnych modelov. Konceptuálne modely (popisy domén) sa vyskytujú hlavne v oblasti návrhu softvéru na podporu mentálnych modelov užívateľa. Mayer [35] preštudoval niekoľko konceptuálnych modelov z oblasti databázových systémov či fyziky a uvádza, že konkrétny konceptuálny model podporuje konceptuálne a znižuje doslovné pamätanie si informácií, čo zlepšuje riešenie problémov. Znázornenie konceptuálnych vzťahov tak poskytuje človeku ďaleko rýchlejšie pochopenie problematiky. Podľa Mineau [37], zvolenie konceptuálnych grafov na zápis konceptuálneho modelu databázových systémov prináša výhody v podobe ľahšieho dolovania znalostí pri zachovaní rovnakej prehľadnosti ako pri použití UML.

Konceptuálne grafy sú rozšírením existencionálnych grafov a vznikli spojením existencionálnych grafov a sémantických sietí. Prvýkrát sa spomínajú v publikácii od John F. Sowu [68], ktorý ich používal na reprezentáciu konceptuálnych schém databázových systémov. Vyjadrovanie pomocou konceptuálnych grafov je logicky presné, pre človeka ľahko čitateľné a vhodné na spracovanie počítačmi. Vzhľadom



Obr. 2 – 4 Konceptuálny graf znázorňujúci koncertujúcu kapelu.

na ich priame prepojenie s prirodzeným jazykom by sa dali považovať za medzivrstvový jazyk na prekladanie medzi formálnymi počítačovými jazykmi a prirodzeným jazykom. Ich grafická reprezentácia ich robí veľmi dobre čitateľnými.

Príklad konceptuálneho grafu je znázornený na obr. 2 – 4. Znázorňuje koncertujúcu kapelu o veľkosti štyroch konkrétnych členov. Vykresľuje stav, v ktorom definuje štyroch členov: gitaristu, hráča na basovú gitaru, speváka a bubeníka a následne ich zoskupuje do konceptu kapely. Znak „*“ je definujúca značka, ktorá vytvára inštanciu konceptu a znak „?“ je značka referencie, ktorá na túto inštanciu odkazuje. Zároveň uvádza danú kapelu do situácie, kde odohráva koncert.

2.2 Ontológie

Ontológia je jednou z reprezentácií znalostí z oblasti modelovania sémantiky. Pojem ontológia sa často používa aj ako zámena za pojmy klasifikácia alebo taxonómia.

V oblasti starogréckej filozofie reprezentoval filozofickú disciplínu zaoberajúcu sa „*bytím*“. Systematické vymedzenie ontológie v podobe vedy v súvislosti s najvšeobecnejšími určeniami bytia, významom bytia a pojmami bytia, sa nachádza v diele Ch. Wolffa „*Ontológia*“ z roku 1730, ale základná otázka tejto filozofickej disciplíny, a to „*Čo je bytie?*“, bola položená Aristotelom v 4. storočí pred Kr. Pojem „*ontológia*“ je v oblasti umelej inteligencie (resp. znalostného inžinierstva) vágny a vedecké kruhy sa plne nezhodujú na určení jeho presného významu.

V súčasnosti sa ontológie používajú v širokom doménovom spektre popísanom napríklad Fridmanom [19]. Od projektu GENSIM[1] z oblasti medicíny, cez projekt PLINIUS [47] z oblasti mechaniky, až po projekt CYC[2], ktorý sa snaží popísať bežné znalosti z reálneho sveta.

Existuje množstvo definícií, ktoré sa rôzne prelínajú. Zoznam relevantných definícii podľa Guarina [23] tvorí:

1. Ontológia ako filozofická disciplína.
2. Ontológia ako neformálny konceptuálny systém.
3. Ontológia ako formálny sémantický význam (account).
4. Ontológia ako špecifikácia konceptualizácie.
5. Ontológia ako reprezentácia konceptuálneho systému teóriou logiky:
   (a) Charakterizovaná špecifickými formálnymi vlastnosťami.
   (b) Charakterizovaná len svojím špecifickým významom.

[1] Dostupné on-line (24.6.2012): http://radimrehurek.com/gensim/
[2] Dostupné on-line (24.6.2012): http://www.cyc.com/

6. Ontológia ako slovník využívaný teóriou logiky.

7. Ontológia ako (meta-úrovňová) špecifikácia teórie logiky.

V mierne pozmenenej podobe sa vo vyššie uvedenom zozname nachádza aj definícia od T. R. Grubera [22] prijímaná s menšími pripomienkami v komunite UI: „*Ontológia je explicitnou špecifikáciou konceptualizácie*“.

N. Guarino predkladá [23] aj ďalšie definície, stavia sa k nim kriticky a odhaľuje ich nedostatky z viacerých hľadísk. Za uspokojivú nakoniec pokladá túto definíciu:

> **Ontológia** *je explicitná, čiastočná špecifikácia konceptualizácie, ktorá sa dá vyjadriť ako meta-úrovňový náhľad na množinu možných doménových teórií za účelom modulárneho návrhu, prestavby a opätovného použitia systémových častí súvisiacich so znalosťami [61].*

Konceptualizácia hovorí, že sa jedná o taký abstraktný model výseku reálneho sveta, ktorý identifikuje relevantné koncepty daného výseku. Explicitná znamená, že je jednoznačne definovaný typ konceptu i podmienky jeho použitia [64].

Podľa [64] je možné ontológie ďalej deliť do štyroch základných skupín podľa zdrojov konceptualizácie:

- Generické ontológie alebo ontológie vyššieho rádu usilujúce o zachytenie bežných zákonitostí.

- Doménové ontológie sú najčastejším typom a ich predmetom je stále nejaká určitá špecifická oblasť.

- Úlohové ontológie označujú generické modely znalostných úloh a metód ich riešení.

- Aplikačné ontológie sú najšpecifickejšie a v ich prípade sa jedná o konglomerát modelov adaptovaných na konkrétnu aplikáciu spravidla zahrňujúc doménovú aj úlohovú časť.

Podľa [73] nachádzajú svoje uplatnenie v troch hlavných oblastiach:

- Komunikácia.
- Inter-operabilita.
- Systémové inžinierstvo: špecifikácia, spoľahlivosť a možnosť opakovaného použitia.

Veľmi jednoduchú a pochopiteľnú definíciu poskytuje aj Web Ontology Working Group[3]. Tá definuje ontológiu takto: *Ontológia je strojovo zrozumiteľná množina definícií, ktorá vytvára taxonómiu tried, podtried a relácie, ktoré medzi nimi existujú.*

Na definovanie ontológií sa používajú ontologické jazyky. Jedná sa často (CycL, KIF) o deklaratívne jazyky založené na predikátovej logike prvého rádu. KG sa dajú tiež chápať ako spôsob zápisu ontológií [66].

V nasledujúcich podkapitolách je priblížená dvojica ontológií: CYC a WordNet.

2.2.1 CYC

Cyc bol projektom, ktorý si kládol za cieľ zhromaždiť komplexnú ontológiu a databázu tvorenú každodennými praktickými znalosťami s cieľom umožniť UI aplikáciám rozhodovanie podobné ľudskému.

Projekt vznikol v roku 1984 pod vedením Douga Lenata ako časť Microelectronics and Computer Technology Corporation. Názov Cyc (odvodený z anglického slova „encyclopedia“) je registrovanou obchodnou značkou firmy Cycorp, Inc. v Texase. Báza znalostí, ktorú systém využíva, je majetkom spoločnosti, ale jej menšia verzia vydaná pod názvom OpenCyc je uverejnená pod open source licenciou[4].

Typickým príkladom znalostí reprezentovaných v databáze je napríklad veta typu „Každá gitara je strunový hudobný nástroj“ a „Strunové hudobné nástroje sa dajú ladiť“. Pri otázke, či sa dá ladiť gitara, dospeje inferenčný systém k logickej odpovedi, že gitara sa ladiť dá. Báza znalostí obsahuje vyše milióna ručne napísaných

[3]Dostupné on-line (24.6.2012): http://www.w3.org/2001/sw/WebOnt/

[4]Dostupné on-line (14.02.2012): http://www.cyc.com/opencyc

výrokov, pravidiel a heuristík pre rozhodovanie o objektoch a udalostiach z každodenného sveta. Tie sú formulované v jazyku CycL, založenom na predikátovom počte so syntaxou veľmi podobnou Lispu.

Názvy konceptov sa v Cyc označujú ako *konštanty*. Konštanty začínajú znakom *#$* a môžu označovať individuálne položky, množiny, pravdivostné funkcie a funkcie. Pravdivostné funkcie sa skladajú z logických spojok, kvantifikátorov a predikátov. Funkcie vracajú nové termy na základe vstupných parametrov. Najdôležitejšími predikátmi sú *#$isa* a *#$genls*. Kým *#$isa* popisuje, že položka je prvkom nejakej kolekcie, *#$genls* vyjadruje, že množina je podmnožinou inej množiny. Fakty o konceptoch sú vyjadrené pomocou špecifických CycL viet. Predikáty sú zapísané pred argumentami v zátvorkách:

```
(#$isa #$Marcel #$Gitarista)
```

„Marcel patrí do množiny Gitaristi“ a

```
(#$genls #$Gitarista #$Hudobnik)
```

„Všetci gitaristi sú hudobníci“.

Vety môžu obsahovať aj premenné označené reťazcom začínajúcim znakom „?“:

```
(#$implies
    (#$and
        (#$isa ?OBJ ?SUBSET)
        (#$genls ?SUBSET ?SUPERSET)
    )
    (#$isa ?OBJ ?SUPERSET)
)
```

čo interpretujeme ako „ak *OBJ* je inštanciou množiny *SUBSET* a *SUBSET* je podmnožinou množiny *SUPERSET*, potom *OBJ* je inštanciou množiny *SUPERSET*“ [51].

2.2.2 WordNet

WordNet [36, 17] je sémantický lexikón anglického jazyka. Zoskupuje anglické slová (a slovné spojenia) do množín synoným nazývaných synsety a poskytuje krátke definície a záznamy rôznych sémantických relácií medzi týmito synsetmi. Má to dvojaký účel: poskytnúť slovník a tezaurus, ktorý by bol viac intuitívne použiteľný a umožniť podporu automatickej analýzy textov a aplikácií pre UI. Synsety sú základné lexikálne koncepty. Pri pohľade na WordNet ako na sémantickú sieť, sú to práve synsety, ktoré zastupujú úlohu konceptov. Databáza a softwarové nástroje sú vydané pod licenciou typu BSD a môžu byť voľne stiahnuté a používané. Databázu je možné prezerať aj on-line[5].

WordNet vznikol v roku 1985 a je udržiavaný na pôde Princetonskej univerzity v Laboratóriu kognitívnych vied pod vedením profesora psychológie George A. Millera. V roku 2003 obsahovala databáza okolo 140 000 slov organizovaných do viac ako 110 000 synsetov s celkovým počtom 195 000 slovných párov.

WordNet inšpirovaný predpokladom, že aj v ľudskom mozgu sa tieto štruktúry ukladajú rozdielne, rozlišuje medzi podstatnými menami, slovesami, prídavnými menami a príslovkami. Každý synset obsahuje množinu synonymických slov alebo slovných spojení, ktoré produkujú špecifický význam (napr. vysoká škola). Slová sa zvyčajne vyskytujú vo viacerých synsetoch naraz. Význam synsetov je ďalej doladený krátkymi poznámkami. Typickým príkladom synsetu s poznámkou je: správny, vhodný, akurátny – (najvhodnejší pre dané špecifické účely; „správny čas orať“, „vhodný moment na ústup“, „akurátny prístup k danej problematike“). Každý synset je relačne prepojený na iné synsety a špeciálne relácie existujú aj medzi synsetmi a slovami a tiež slovami medzi sebou. Tieto relácie závisia od typu synsetu alebo slova (od slovného druhu slova alebo slov v synsete). Jedná sa o rovnaké typy sémantických relácií, aké boli vymenované na začiatku kapitoly, rozšírené o množstvo iných relácií. Dokumentácia k všetkým reláciám WordNetu 3.0 je dostupná na stránkach

[5]Dostupné on-line (15.2.2012): http://wordnetweb.princeton.edu/perl/webwn

projektu[6]. Medzi tieto relácie patria napríklad:

- **Sesterské pojmy** – prepájajú synsety zdieľajúce spoločného rodiča v hierachii konceptov. Napr. synset obsahujúci slovo „January“ („Január“) a synset obsahujúci slovo „March“ („Marec“) zdieľajú spoločného priameho rodiča, a to synset obsahujúci slovné spojenie „Gregorian calendar month“ („Gregoriánsky kalendárny mesiac“).

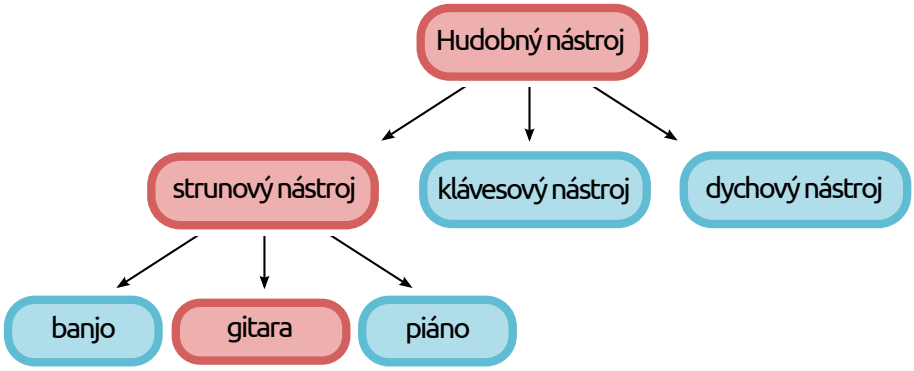
- **Doména** – prepája synset označujúci doménu a synset, ktorý bol v danej doméne klasifikovaný. Napr. synset „tuning“ („ladenie“) je klasifikovaný v synsete „music“ („hudba“).

- **Derivačne odvodená forma** – prepája slová, ktoré sú morfologicky príbuzné a súčasne sémanticky príbuzné. Napr. „guitar“ („gitara“) a „guitarist“ („gitarista“).

  Príklad: Ako príklad môže poslúžiť anglické slovo „*tuning*“ myslené vo význame „*ladenie*“ (napríklad ladenie gitary). Použitím on-line nástroja[7] sa po zadaní daného slova zobrazí zoznam možných výsledkov. Pre možnosť „*(music) calibrating something (an instrument or electronic circuit) to a standard frequency*“ je možný dotaz „*domain category*“, ktorý reláciou doménovej kategórie dáva strunu do súvislosti s „*music (an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner)*“, čiže hudbou.

Na obrázku 2 – 5 je ilustrovaný príklad relácie hyponým v PWN. Uzly grafu reprezentujú synsety (pre zjednodušenie je synset popísaný slovenským prekladom jedného zo slov, ktoré obsahuje) a orientované hrany reprezentujú reláciu hyponým. Z obrázku sa dá vyčítať napríklad to, že gitara je hudobný nástroj patriaci do kategórie strunových nástrojov.

[6]Dostupné on-line (15.2.2012): http://wordnet.princeton.edu/wordnet/documentation/
[7]Dostupné on-line (15.2.2012): http://wordnetweb.princeton.edu/perl/webwn



Obr. 2 – 5 Ilustrácia relácie hyponým v PWN.

WordNet poskytuje hodnotu „polysemy count“, ktorá vyjadruje počet synsetov, ktoré dané slovo obsahujú. Ak sa dané slovo vyskytuje vo viacerých synsetoch (t.z. má viacero významov), tak je zrejmé, že niektoré významy sú bežnejšie ako iné. WordNet to kvantifikuje pomocou „frequency score“. V niekoľkých vzorkách textov sú všetky slová sématicky označené korešpondujúcim synsetom a neskôr je spočítané, koľkokrát sa vyskytli s daným špecifickým významom. Databázové rozhranie vie z užívateľského vstupu dedukovať koreňovú formu slova a do databázy sa ukladajú iba tieto formy [51].

WordNet (PWN) slúži mnohým aplikáciám umelej inteligencie, ktoré sú závislé na lexikálnych zdrojoch. Jeho uplatnenie sa nachádza v širokom rozsahu domén od rozlišovania viacvýznamových slov [77], cez klasifikáciu textov [60], zhlukovanie textov [18] až po rozširovanie dotazov [65] alebo strojový preklad [42]. Medzi jeho silné stránky patrí prístupnosť strojového spracovania a pomerne dobrá licenčná prístupnosť. K WordNetu existuje veľké množstvo aplikačných rozhraní pre mnoho bežne používaných programovacích jazykov (Java, python, C, .NET, . . .). Samotný projekt slúžil ako inšpirácia iným podobným projektom ako VerbNet [62], sústreďujúci sa na slovesá alebo FrameNet [11, 29], ktorý sa sústreďuje na popis konceptov pomocou sémantických rámcov. Vďaka svojim pozitívnym vlastnostiam sa projekt PWN stal veľmi využívaným, o čom svedčí napríklad aj zoznam k nemu relevantných priznaných publikácií zo stránok projektu, ktorý vysoko presahuje počet 400

kusov[8].

2.2.3 Neanglické WordNetové lexikóny

Na PWN nadväzuje množstvo projektov, ktoré si kladú za snahu vytvorenie lexikónov WN typu pre ďalšie jazyky. Zoznam všetkých lexikónov WN typu je prístupný na stránkach organizácie Global WordNet association[9]. Medzi vybrané projekty tohto typu patria napríklad:

- **Japonský WN** (2006-) - aktuálny projekt pokrývajúci japonský jazyk.
- **EuroWordNet** (1997-1999) - projekt pokrývajúci vybrané európske jazyky.
- **BalkaNet** (2001-2004) - projekt pokrývajúci vybrané balkánske jazyky.
- **Český WN** (1998-) - projekt pokrývajúci český jazyk, ktorý je súčasťou EuroWordNetu.

Japonský WN[10] sa začal vytvárať v roku 2006 a jeho prvá verzia bola vydaná v roku 2009. Potreba súčasného výskumu v rámci tohto projektu [7, 74] potvrdzuje, že budovanie lokálnych WN lexikónov je stále aktuálnou a živou oblasťou. V súčasnosti obsahuje 57 238 synsetov, ktoré zhlukujú 93 834 slov. Vznikal postupným prekladom (poloautomatickým i manuálnym) WN z iných jazykov (anglického, francúzskeho a španielskeho) spôsobom, aby pokryl preklad základných 4 959 synsetov z PWN. Postupným rozširovaním [5] a dolaďovaním [6] dosiahol až súčasnú veľkosť s uvádzanou približnou chybovosťou v 5% záznamov.

EuroWordNet[11] je systém WN sémantických sietí pre európske jazyky (holandský, taliansky, španielsky, nemecký, francúzsky, český a estónsky). Projekt bol rozdelený do dvoch častí EuroWordNet 1 (1997-1998 pre holandčinu, taliančinu a

[8]Dostupné on-line (15.2.2012): http://wordnet.princeton.edu/wordnet/publications/
[9]Dostupné on-line (23.06.2012): http://www.globalwordnet.org/gwa/wordnet_table.html
[10]Dostupné on-line (23.6.2012): http://nlpwww.nict.go.jp/wn-ja/index.en.html
[11]Dostupné on-line (15.2.2012): http://www.illc.uva.nl/EuroWordNet/

španielčinu) a EuroWordNet 2 (1998-1999 pre francúštinu, nemčinu, estónčinu a češtinu). Slovníky pre každý jazyk vznikali samostatne v národnom prostredí. Keďže bol EuroWordNet založený na WN 1.5, obsahuje viac (a rôznych) sémantických relácií ako WN 3.0 [48]. Na tvorbu slovníkov bola prijatá spoločná metodika automatického a poloautomatického prevodu už existujúcich jazykových zdrojov a strojovo čitateľných slovníkov do podoby synsetov a sémantických relácií. Kvôli jednotnosti WN bola vytvorená tzv. vrcholová ontológia výrazov (TOC - Top-Ontology Concepts), ktorá pokrýva najzákladnejšie výrazy nezávisle na jazyku. Tá mala byť zastúpená v každej plánovanej jazykovej mutácii. V TOC sa na vrchole nachádzajú tri entity:

- *1stOrderEntity* - fyzické veci napr.: vozidlo, zviera, substancia, objekt apod.
- *2ndOrderEntity* - situácie, napr.: stať sa, byť, začať, pokračovať apod.
- *3rdOrderEntity* - nepozorovateľné entity napr.: myšlienka, informácia, teória, plán apod.

Kým *1stOrderEntity* a *2ndOrderEntity* sa ďalej delia do podskupín (*1stOrderEntity* → forma, kompozícia, pôvod, funkcia; *2ndOrderEntity* → situačný typ, situačná komponenta), *3rdOrderEntity* sa nedelí. Každá podskupina má množstvo vlastných podskupín atď.

Spoločným podkladom pre všetky nové slovníky sa stal PWN vo verzii 1.5. V ňom dostal každý synset jedinečný identifikátor, ktorý bol použitý na vytváranie ekvivalencií medzi nimi a synsetmi iných jazykov. Vďaka tomu mohol vzniknúť tzv. medzijazykový index (Interlingual Index, ILI), ktorý prepája synsety rôznych jazykov a zachycuje tak prekladové ekvivalenty. Na zaručenie minimálnej miery prepojenia WN rôzných jazykov vznikla ďalšia množina, referenčná množina, nazvaná základné koncepty (Base Concepts - BC). Tá viac-menej reprezentovala povinnú slovnú zásobu WN pre každý jazyk [14, 26].

BalkaNet je projektom ktorý trval medzi rokmi 2001-2004. Jeho úlohou bolo vytvorenie WN slovníkov pre vybrané balkánske jazyky: gréčtinu, turečtinu, rumunčinu, bulharčinu a srbčinu. Spôsob návrhu, vytvárania jednotlivých databáz a

očakávané výsledky viac-menej kopírovali metodiku použitú v projekte EuroWordNet.

Český wordnet začal vznikať v roku 1998 na Fakulte Informatiky Masarykovej Univerzity v Brne[12] (FI MUNI) ako súčasť projektu EuroWordNet a ďalej sa rozšíril a zdokonalil v dobe vzniku BalkaNetu (na vývoji ktorého sa FI MUNI podieľala). Na budovanie slovníka boli okrem iných využité i Slovník českých synonym, k vytvoreniu ekvivalencií medzi jazykmi česko-anglický slovník Lingea Lexicon a k pokrytiu kolokácií boli použité gramaticky značkované korpusy DESAM [56] a ESO. Automaticky prevádzané dáta boli ručne kontrolované. Po prekročení hranice 15 000 pojmov bol slovník rozširovaný už iba ručne. V súčasnosti obsahuje český WN 30 000 pojmov [14].

[12]Dostupné on-line (23.6.2012): http://www.fi.muni.cz/

3 Tvorba konceptuálnych znalostných sietí s minimalizáciou vstupu človeka

Znalostné siete sa využívajú pri riešení širokej škály úloh. Od pôvodného zámeru konceptuálnych grafov reprezentácie databázových systémov až po dolovanie zdrojového kódu [38] alebo modelovanie sémantiky agentových systémov [24]. Otázka ich automatickej tvorby je preto veľmi aktuálnou témou nielen v oblasti NLP. V nasledujúcich podkapitolách sú bližšie popísané možnosti tvorby a využitia konceptuálnych znalostných štruktúr so zameraním na automatické a poloautomatické prístupy [43]. Konkrétne prístupy boli vyberané tak, aby ilustrovali spektrum využitia, a zároveň sa upriamili na získavanie relácií medzi konceptmi z textových dokumentov, prípadne ilustrovali zaujímavé postupy tvorby daných znalostných štruktúr.

3.1 Automatické budovanie neanglických WN

Automatizáciou budovania neanglických slovníkov WN typu sa zaoberá väčšie množstvo prác [16, 58, 3, 20]. V súčasnosti boli tieto slovníky vytvorené pre viac ako 50 jazykov [58] (napr. nemčinu, češtinu, maďarčinu). Budovanie takýchto slovníkov vyžaduje obrovské úsilie v podobe času, výpočtových a ľudských zdrojov. Napríklad vytvorenie základov slovníkov (kde jeho časť pre rumunský jazyk obsahovala 20 000 synsetov) pre balkánske jazyky (BalkaNet) trvalo tri roky [3]. Preto akákoľvek forma automatizácie procesov výstavby týchto slovníkov môže priniesť signifikantný posun v rýchlosti ich tvorby a rozširovania.

Vzhľadom na fakt, že mnoho metód automatického budovania slovníkov wordnetového typu je založených na strojovom preklade, bude v tejto stati predstavený jeden z týchto prístupov. Tieto metódy hľadajú spôsoby ako prepísať alebo preložiť zdrojový (anglický) PWN na cieľový WN (v inom ako anglickom jazyku). Tento konkrétny prístup automatickej konštrukcie wordnetov s využitím strojového prekladu a jazykového modelovania [58] bol ako prvý použitý na tvorbu macedónskeho

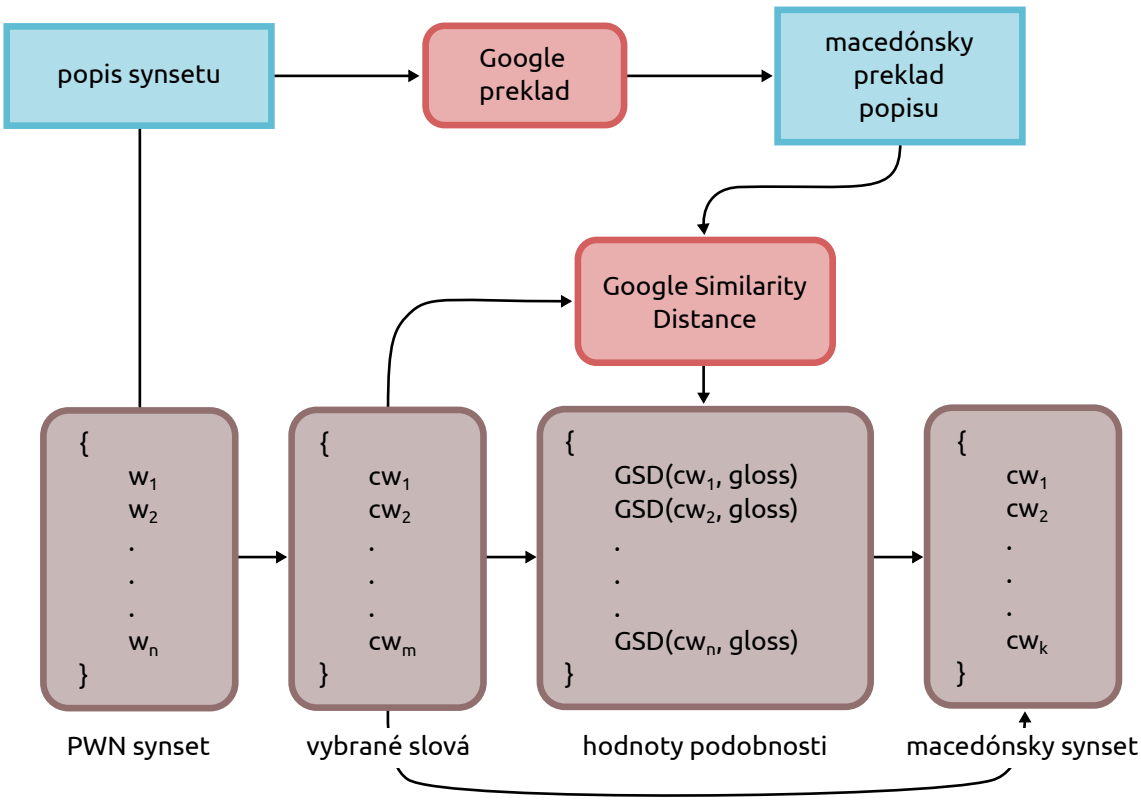
wordnetu. Stavia na predpoklade, že konceptuálny priestor modelovaný PWN slovníkmi nie je závislý na konkrétnom jazyku. Predpokladá, že väčšina konceptov (teda synsetov) sa nachádza rovnako v zdrojovom ako v cieľovom WN a líšia sa hlavne svojou anotáciou. Cieľom daného prístupu sa tak stáva hľadanie vhodného prekladu popisu prvkov synsetov pre cieľový jazyk. Pre každý synset zo zdrojového PWN by mal tento prístup nájsť vhodnú skupinu macedónskych slov, ktoré budú popisovať ten istý synset v macedónskom jazyku.

Prístup je závislý na troch základných prvkoch:

- Na strojovo čitateľnom anglicko-macedónskom slovníku (MRD - akronym spojenia *machine readable dictionary*), ktorý je schopný nájsť všetky preklady slov v zdrojových synsetoch.
- Na službe Google translate (GT), pomocou ktorej je možné prekladať frázy a bloky textov medzi rôznymi jazykmi.
- Na funkcii Google Similarity Distance (GSD), pomocou ktorej je možné kvantifikovať podobnosť významu.

MRD bol špeciálne vyvinutý pre účely tohto prístupu [59]. Obsahuje 181 987 prepojení, pre 61 118 pojmov v macedónskom a 79 956 pojmov v anglickom jazyku. Google translate[13] je prístupný ako on-line služba s existujúcimi rozhraniami pre rôzne programovacie jazyky. GSD je metóda výpočtu sémantickej podobnosti medzi slovami a frázami [13]. Je založená na fakte, že slová a frázy získavajú svoj význam na základe toho, ako sú v spoločnosti používané. WWW je najväčším prístupným zdrojom ľudských znalostí a obsahuje kontextové informácie zadané obrovským množstvom rôznych užívateľov. Autori GSD tvrdia, že s využitím vyhľadávača Google[14] je možné sémantickú podobnosť medzi slovami alebo frázami automaticky

[13]Dostupné on-line (15.2.2012): http://translate.google.com/
[14]Dostupné on-line (15.2.2012): http://www.google.com/



Obr. 3 – 1 Proces tvorby macedónskeho wordnetu s využitím externých zdrojov.

kvantifikovať. GSD medzi dvoma slovami alebo frázami x a y je definovaná takto:

$$GSD(x,y) = \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - \min(\log f(x), \log f(y))}, \tag{3.1}$$

kde x a y sú nejakou vstupnou frázou alebo slovom, $f(x)$ a $f(y)$ je počet vrátených odkazov pre dané vstupy samostatne a $f(x,y)$ je počet vrátených odkazov pre dané vstupy dohromady. N je normalizačný faktor, ktorého minimálna hodnota musí byť väčšia ako maximálny počet vrátených odkazov.

Základný popis prístupu na tvorbu macedónskeho WN je znázornený na obrázku 3 – 1. Vstupom je zdrojový synset zo zdrojového PWN definovaný vymenovaním slov, ktoré do daného synsetu patria $\{w_1, w_2, w_3, \ldots, w_n\}$. Tie sú pomocou MRD preložené do macedónskeho jazyka na skupinu slov $\{cw_1, cw_2, cw_3, \ldots, cw_m\}$. Zároveň je získaný textový popis daného synsetu v jeho zdrojovom jazyku spolu s ukážkovým použitím jeho pojmov. Táto popisná fráza je pomocou GT preložená

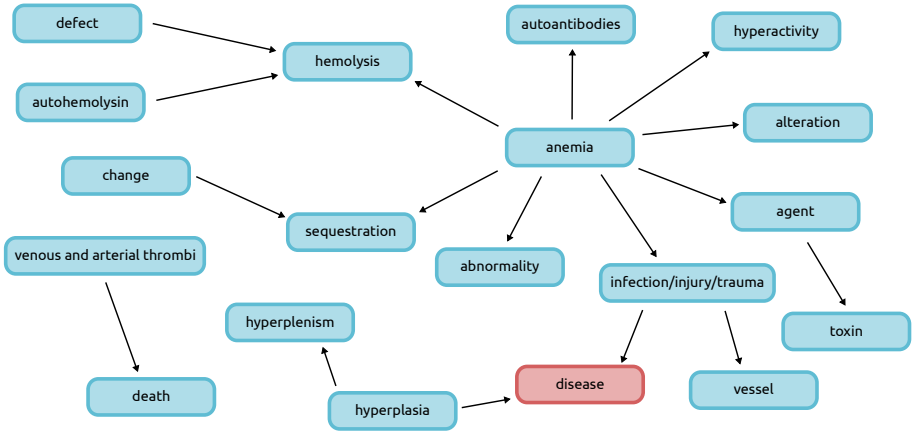
Tabuľka 3 – 1 Klady a zápory prístupu tvorby BalkaNetu.

klady	zápory
kvalitné a reálne výsledku	závislosť na PWN
využitie existujúcich funkcií (GSD)	závislosť na macedónsko-anglickom slovníku

do macedónskeho jazyka. Preklad spolu s preloženými slovami patriacimi do daného synsetu poslúžili ako vstup do funkcie GSD. Bola vypočítaná podobnosť každého z preložených slov k popisnej fráze. Slová, ktorých hodnota podobnosti k popisnej fráze bola vyššia ako 0.2, prešli prvou podmienkou. Po zistení maximálnej hodnoty podobnosti z množiny všetkých podobností, boli výsledky funkcie GSD porovnané s 0.8 násobkom tejto hodnoty. Pokiaľ bola výsledná hodnota násobku vyššia ako táto hodnota, prešli slová aj druhou nutnou podmienkou a boli zaradené k macedónskemu synsetu $\{cw_1, cw_2, cw_3, \ldots, cw_k\}$.

Takto vytvorený slovník bol prvým svojho druhu a nedal sa porovnať s existujúcim. Kontrola celého slovníka človekom by bola príliš časovo náročná. Motiváciou tvorby tohto slovníka bolo jeho využitie v iných oblastiach NLP. Autori sa preto rozhodli zmerať jeho kvalitu porovnaním výsledkov textovej klasifikácie s využitím a bez využitia daného slovníka. S pomocou takto vytvoreného slovníka sa podarilo autorom zvýšiť presnosť klasifikácie oproti prístupu založenom na výpočte kosínovej podobnosti o 6%. Kosínová podobnosť je bežný prístup porovnania dvoch dokumentov. Podobnosť týchto dokumentov je definovaná ako kosínová hodnota veľkosti uhla ich vektorov [58]. To podľa autorov síce nie je dostatočným meradlom kvality ich prístupu, ale odzrkadľuje schopnosť prístupu modelovať znalosti o skutočnom svete.

Zhrnutie: Prístup zvolený pri budovaní BalkaNetu dokázal efektívne využiť existujúce lexikálne a technologické zdroje. S pomocou anglicko-macedónskeho slovníka, on-line automatickej prekladateľskej služby GT a funkcie GSD bol vytvorený kvalitný neanglický WN. V prístupe zohrávala kľúčovú úlohu funkcia GSD, ktorá slúži na výpočet podobnosti významu slov. Výčet kladov a záporov tohto prístupu je v tab. 3 – 1.



Obr. 3 – 2 Implikačná sieť získaná z textu.[57].

Tabuľka 3 – 2 Pravdepodobnostná tabuľka pre „disease“.[57]

hyperplasia	infection/injury/trauma	P(disease=T)	P(disease=F)
T	T	0.83	0.17
T	F	0.5	0.5
F	T	0.33	0.67
F	F	0	1.0

3.2 Dolovanie implikačných sietí z textov

Pomerne aktuálnym problémom je dnes dolovanie kauzálnych relácií z textov v prirodzenom jazyku. Existencia takého automatického procesu by uľahčila prehľad trebárs v doméne medicínskych dokumentov [69] týkajúcich sa vzájomného pôsobenia génov.

V [57] je predstavená metóda dolovania kauzálnych vzorov z textov. Za kauzálny vzor sa v tomto prípade považuje myšlienka, prípadne časť vety, ktorá špecifikuje kauzálnu reláciu medzi príčinou a následkom. Tie sa v textoch vyhľadávajú na základe slovných spojení s funkciou spojok. Systém predpokladá, že príčina a následok sú oddelené v myšlienke nejakou spojkou, ktorá by mohla vykazovať funkciu relačného operátora.

Systém bol testovaný na množine textových súborov v anglickom jazyku a medzi príklady takýchto spojení v anglickom jazyku patria napríklad „*because*“, „*the cause*

Tabuľka 3 – 3 Klady a zápory prístupu dolovania implikačných sietí z textov.

klady	zápory
minimálne znalosti o jazyku	závislý na PWN
ohodnotenie získaných relácií a možnosť inferencie	testovaný iba na angličtine

of“, „*causes*“ alebo „*is a result of*“. Po akumulácii kauzálnych vzorov v texte ďalej nasleduje ich analýza. Tá spočíva v bližšom preskúmaní príčin a následkov na základe spojok „*a*“ a „*alebo*“, interpunkčných znamienok a zisťovaní existencie negačného operátora pri slovnom spojení, ktoré kauzálnu reláciu delí na príčinu a následok (v anglickom jazyku je to slovo „*not*“).

Nasleduje zovšeobecňovanie kauzálnych vzorov na základe podobností slov definovaných vo WN [36] (teda koncepty sú získané z WN). Ak sú napríklad v príčinnej (kauzálnej) časti nejakých rozdielnych vzorov slová, ktoré sú prvkami jedného synsetu, systém to berie do úvahy a môže ich vnímať ako jednu entitu. Do úvahy sa berú okrem synoným aj hypernymá. V poslednom kroku sa generuje implikačná sieť [54]. Uzlami sú v tomto prípade udalosti príčiny a následku. Pravdepodobnostné hodnoty hrán, ktoré ich prepojujú, sa získavajú z početností výskytov kauzálnych vzorov, v ktorých sa dané udalosti vyskytli. Na obr. 3 – 2 a tab. 3 – 2 je ilustrácia výstupu daného prístupu. Z tabuľky vyplýva, že pri výskyte javu „hyperplasia“ a zároveň absencii javu „infection/injury/trauma“ je oveľa vyššia pravdepodobnosť výskytu javu „disease“ ako pri výskyte javu „infection/injury/trauma“ a zároveň absencii javu „hyperplasia“.

V tomto prístupe bolo na riešenie úlohy využitých niekoľko sémantických sietí (implikačné siete, WordNet), taktiež ako metódy boli využité prístupy z rôznych oblastí (dolovanie v textoch učenie pravdepodobnostných modelov).

Zhrnutie: Prezentovaný prístup dolovania implikačných sietí z textov využíva na tvorbu sémantickej siete vlastný postup dolovania kauzálnych zdrojov z textov. Ten pracuje nad textami v anglickom jazyku a

je založený na minimálnych znalostiach o danom jazyku, ktoré spočívajú v rozpoznaní slov slúžiacich ako kauzálne operátory, prípadne operátory spojok a slov, ktoré môžu slúžiť ako operátory negácie. V poslednom kroku prístup využíva PWN ako zdroj výpočtu podobností kauzálnych zdrojov. Na výstupe ponúka sieť konceptov vzájomne previazaných reláciou kauzality. Súhrnný výčet kladov a záporov tohto prístupu je v tab. 3 – 3.

3.3 DIPRE

DIPRE (Dual Iterative Pattern Relation Expansion) je systém na extrakciu relácií vyvinutý Sergeyom Brinom [10], spoluzakladateľom firmy Google. Na ilustráciu systému autor ako názorný príklad používa reláciu: (*autor*, *dielo*). Vstupom do systému DIPRE je malá množina inštancií tejto relácie, konkrétne dvojíc (*autor*, *dielo*). Predpokladajme, že systém na vstup dostal iba jednu dvojicu, a to („Arthur Conan Doyle“, „The Adventures of Sherlock Holmes“). Systém v prvej fáze prehľadáva na internete tie webové stránky, ktoré danú dvojicu obsahujú. Zo stránok, kde danú dvojicu našiel, sa pokúsi vytvoriť všeobecné šablóny vo forme zjednodušených regulárnych výrazov, ktoré plne pokrývajú vstupnú množinu dvojíc (ktorá je v tomto konkrétnom prípade jednoprvková). Tieto šablóny považuje systém za šablóny pokrývajúce reláciu (*autor*, *dielo*) a s ich pomocou extrahuje z webových stránok nové dvojice. Tieto môžu byť priamo výstupom systému alebo môžu slúžiť ako nový vstup a v ďalšej iterácii opäť extrahovať nové šablóny a ďalšie dvojice.

1. $R \leftarrow Vzorka$

   Začína s malou vzorkou R (relations) ukážkových inštancií hľadanej relácie. Táto vzorka je zadaná užívateľom a môže byť veľmi malá. V [10] bola použitá vzorka iba piatich inštancií.

2. $O \leftarrow FindOccurrences(R, D)$

   V nasledujúcom kroku hľadá všetky výskyty všetkých dvojíc z R v nejakej

databáze textových zdrojov D (database). Pre každú dvojicu si ukladá kontext (URL adresa a okolitý text), v ktorom sa daná dvojica nachádzala. Výsledok uloží do množiny výskytov O (occurrences).

3. $P \leftarrow GenPatterns(O)$

   Neskôr systém vytvára šablóny na základe výstupu z predchádzajúceho kroku. Táto časť systému je kľúčová, pretože príliš všeobecné šablóny by pokrývali irelevantné relácie a príliš špecifické šablóny by mohli naopak pokrývať len veľmi málo relevantných relácií, čo sa dá kompenzovať zväčšením prehľadávaného priestoru. Šablóny sú pridané do množiny šablón P (patterns).

4. $R \leftarrow M_D(P)$

   Systém ďalej prehľadáva databázu a na základe šablón z predchádzajúceho kroku objavuje nové relevantné dvojice.

5. R

   Pokiaľ je množina doteraz nájdených dvojíc dostatočná, proces sa ukončí. Ak nie je, môže poslúžiť ako nový vstup do algoritmu.

Šablónu definuje ako usporiadanú päticu: *(order, urlprefix, prefix, middle, suffix)*, kde *order* je typu boolean a zvyšné atribúty sú typu string. Ak má atribút *order* hodnotu true, to znamená, že dvojica $(concept1, concept2)$ zodpovedá danej šablóne ak sa v prehľadávanej kolekcii (WWW) nachádza dokument s URL, ktorá zodpovedá regulárnemu výrazu $urlprefix*$, a ktorý obsahuje text zodpovedajúci nasledujúcemu regulárnemu výrazu: **prefix, (concept1), middle, (concept2), suffix**.

Ak má premenná *order* hodnotu false, poradie konceptov je vo výraze prehodené $(concept1, concept2) \rightarrow (concept2, concept1)$.

5 príkladov z trénovacej množiny uvedených v tabuľke 3 – 4 bolo schopných nájsť na menšej doméne internetových adries 3 šablóny uvedené v tabuľke 3 – 5. Na rovnakej doméne tieto šablóny na seba naviazali 4047 dvojíc konceptov $(autor, dielo)$.

Tabuľka 3 – 4 Príklad trénovacej množiny pre systémy DIPRE alebo SnowBall.

Autor	Dielo
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gelick	Chaos: Making a New Science
Chales Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

Tabuľka 3 – 5 Príklad šablón nájdených prístupom DIPRE.

order	urlprefix	prefix	middle	suffix
false	www.sff.net/locus/c.*	<LI><B>	</B> by	(
false	dns.city-net.com/ mann/awards/hugos/1984.html		by	(
true	dolphin.upenn.edu/ dcummins/texts/sf-award.htm		||	|| (

Obsah premennej *autor* je obmedzený na:

```
[A-Z][A-Za-z .,&]{5,30}[A-Za-z.]
```

Obsah premennej *dielo* je obmedzený na:

```
[A-Z0-9][A-Za-z0-9 .,:'#!?;&]{4,45}[A-Za-z0-9?!]
```

V [10] prebiehala kontrola získaných údajov náhodným výberom 20 diel a ich následným vyhľadaním na rozsiahlych internetových kníhkupectvách. 19 z 20 takto vybraných dvojíc bolo nájdených ako knižné tituly a 1 titul bol nájdený ako článok. Autori boli k dielam priradení správne. Na adrese implementácie mierne upraveného systému DIPRE[15] je možné vidieť príklad výstupu, ktorý obsahuje obrovské množstvo relevantných dvojíc $(autor, dielo)$.

Jedným z hlavných problémov tohto jednoduchého prístupu je problém priameho porovnávania konceptov ako textových reťazcov (navyše s rozlišovaním veľkých

[15]Dostupné on-line (14.7.2012):
http://www.alexmayers.com/q/?q=85a4c8671be552e72bc8388eddda2b5d,0,0,0,80„,1„

Tabuľka 3 – 6 Klady a zápory systému DIPRE.

klady	zápory
ľahko implementovateľný	potreba semištruktúrovaného vstupu
absencia znalostí o jazyku	závislý na veľkosti písmen
kvalita výstupu	
ľahko rozšíriteľný	

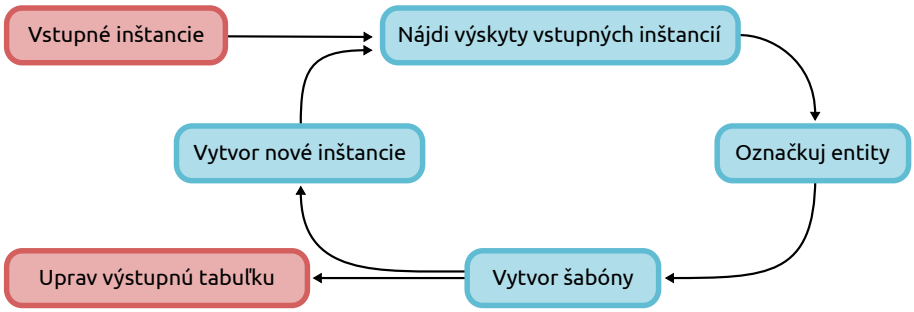
a malých písmen). Ak napríklad autor nebol uvedený plným menom, ale jeho krstné meno bolo skrátené na iniciály, šablóna, v ktorej sa nachádzal, nebola detekovaná. Ďalšie silné obmedzenie tohto prístupu predstavuje fakt, že je plne závislý na semištrukturácii textu, ktorý poskytuje značkovací jazyk HTML, a na neštrukturovanom texte je nepoužiteľný. Zjednodušene by sa dalo povedať, že sa DIPRE sústreďuje na vyhľadávanie silne štruktúrovaných dát v HTML stránkach, akými sú napríklad dlhé zoznamy alebo tabuľky s dvojicami (*autor*, *dielo*) na stránkach kníhkupectiev. V každom prípade tento jednoduchý prístup priniesol vynikajúce výsledky a stal sa základným kameňom pre vznik mnohých ďalších systémov [27, 12, 1].

Zhrnutie: DIPRE je systém na extrakciu relácií z WWW. Vstupom do systému je malá množina inštancií hľadanej relácie. Prednosť systému, ktorou je absencia potreby znalostí o jazyku, je vyvážená potrebou formátu vstupných dokumentov v podobe webových stránok. Výstup prístupu prezentuje množina hľadaných relácií, ktorá disponuje vysokou kvalitou. Tento systém je veľmi jednoducho implementovateľný. Súhrnný výčet kladov a záporov tohto prístupu je v tab. 3 – 6.

3.4 Snowball

Snowball [1] je systém na extrakciu relácií z veľkých kolekcií textových dokumentov v *plain/text* podobe zovšeobecňujúci a rozširujúci prístup DIPRE popísaný vyššie. Snowball, na rozdiel od DIPRE, nie je závislý na semištruktúrovaných dá-

tach, akými sú napríklad HTML dokumenty, a pracuje na predspracovaných textoch v prirodzenom jazyku. Rovnako ako DIPRE si nekladie za cieľ objaviť vo vstupnom texte všetky relevantné relácie, no kladie dôraz na presnosť výsledku [2].



Obr. 3 – 3 Základná architektúra systému Snowball.

Základná architektúra je zobrazená na obr. 3 – 3 a je skoro identická s architektúrou DIPRE. V prvom kroku sú systému ponúknuté usporiadané dvojice entít, pre ktoré je potrebné nájsť relácie. Systém potom vyhľadáva ich spoločné výskyty v danej množine dokumentov a snaží sa identifikovať textové kontexty, v ktorých sa dvojice daného typu nachádzajú. Z týchto kontextov je schopný získať šablóny. Podľa získaných šablón ďalej nájde nové dvojice entít, z ktorých po vyhodnotení tie najvhodnejšie znova vstupujú do procesu prehľadávania dokumentov a indukcie nových vzorov a proces sa zopakuje. Na rozdiel od DIPRE sa Snowball snaží vyhodnocovať presnosť svojich šablón aj získaných výsledkov pomocou vlastných algoritmov. Takto dosahuje v porovnaní s DIPRE oveľa kvalitnejšie výsledky.

Pre Snowball je kľúčovým krokom predspracovanie textov, s ktorými pracuje. Predspracovanie textu je závislé na externých nástrojoch, v označení (z ang. „tag“) entít vo vstupných textoch. V prípade dvojíc $(autor, dielo)$ je úlohou externého nástroja označiť mená ľudí a mená produktov. V [2] sú ako príklady takýchto systémov spomenuté Alembic[16] a LingPipe[17]. Jedná sa o tzv. NER (named entity recognition) systémy, ktoré sú silne závislé na vstupe človeka, alebo sú silne doménovo obmedzené. Úloha Snowball potom spočíva v nájdení medzi označenými menami ľudí a

[16]Dostupné on-line (14.2.2012): http://timeml.org/site/terqas/alembic/
[17]Dostupné on-line (14.2.2012): http://alias-i.com/lingpipe/

Tabuľka 3 – 7 Príklad trénovacej množiny pre systém Snowball.

Organizácia	Mesto
Microsoft	Redmond
Exxon	Irving
IBM	Armonk
Boeing	Seattle
Intel	Santa Clara

Tabuľka 3 – 8 Klady a zápory systému Snowball.

klady	zápory
vstupom je formát *plain/text*	závislý na predspracovaní vstupu
absencia znalostí o jazyku	potreba externých nástrojov
samo-vyhodnocovanie kvality	závislý na veľkosti písmen
kvalita výsledkov prevyšuje DIPRE	

produktov takých dvojíc, ktoré zodpovedajú pravidlu, že daný človek je autorom daného produktu.

Zhrnutie: Systém Snowball je rozšírením systému DIPRE a v porovnaní s DIPRE dosahuje omnoho kvalitnejšie výsledky. Na rozdiel od DIPRE vyhodnocuje kvalitu šablón a výsledkov pomocou vlastných algoritmov. Nie je obmedzený na semištrukturáciu vstupných textov v podobe HTML, ale pracuje nad predspracovanými textami vo formáte *plain/text*. Predspracovanie spočíva v označkovaní textu pomocou externého NER systému. Súhrnný zoznam kladov a záporov tohto prístupu je v tab. 3 – 8.

3.5 Porovnanie uvedených prístupov

Pre názornejšie porovnanie prístupov uvedených v tejto kapitole vznikli tabuľky 3 – 9 a 3 – 10. Tab. 3 – 9 popisuje podobu konceptu, relácie pre každý prístup a či bola daná relácia ohodnotená (vážená). Tab. 3 – 10 uvádza závislosti každého prístupu, teda či je závislý na WN, výpočte podobnosti slov, znalostiach o jazyku, nad ktorým pracuje alebo iných externých nástrojoch.

Tabuľka 3 – 9 Koncepty a relácie vo vybraných prístupoch tvorby SN.

Prístup	Koncept	Relácia	Ohodnotenie relácie
BalkaNet	WN synset	WN relácia	-
Implikačné siete	Slovo	relácia implikácie	áno
DIPRE	Reťazec	implicitne daná	nie
Snowball	Reťazec	implicitne daná	nie
AUP	Slovo	sémantická príslušnosť	nie

Tabuľka 3 – 10 Závislosti pre vybrané prístupy tvorby SN.

	Závislosť			
Prístup	WN	Podobnosť slov	Znalosti o jazyku	Ext. nástroje
BalkaNet	+	GSD	+	WN, GSD
Implikačné siete	+	WN	minimálne	-
DIPRE	-	-	-	-
Snowball	-	-	+	NER
AUP	-	AUP	-	-

Všetky prístupy opísané v tejto kapitole vykazujú závislosť na znalostiach o jazyku, nad ktorým pracujú buď priamo, alebo prostredníctvom externých nástrojov, na ktorých sú závislé. Fakt, že súčasné prístupy sú často závislé na tejto znalosti bol silnou motiváciou pre vytvorenie prístupu, ktorý by dokázal dolovať relácie z textov aj bez nej. Jedná sa o AUP, ktoré je bližšie popísané v nasledujúcej kapitole. Ako koncept je v AUP chápané slovo (token) a reláciou je príslušnosť do jednej sémantickej triedy, ktorá je nehodnotená.

4 Asociatívne učenie pojmov

Vzhľadom na fakt, že metóda asociatívneho učenia pojmov (AUP) sa v minulosti ukázala ako sľubný prístup pri automatickom získavaní lexikálnych zdrojov [51], existuje predpoklad, že by jej rozvoj pomohol rozšíriť aj prístupy extrakcie relácií priblížené v predchádzajúcej kapitole. Táto kapitola poskytuje detailný a kompletný popis AUP. Zároveň je vlastným prínosom do oblasti dolovania konceptov bez závislosti na lexikálnych zdrojoch, či externých nástrojoch.

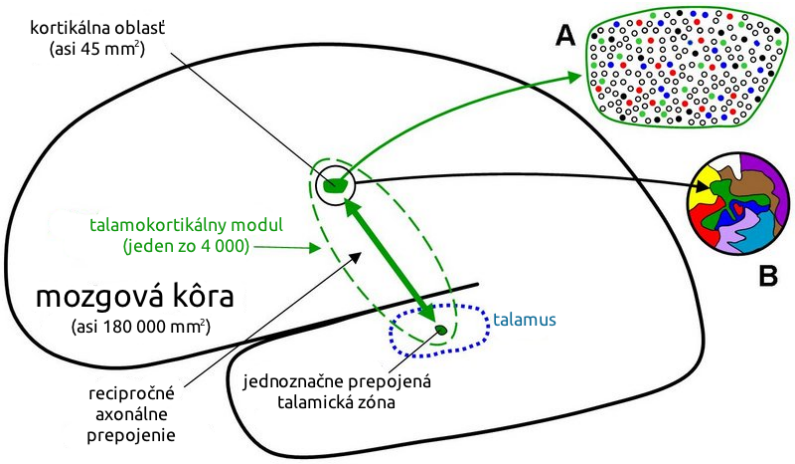
AUP je prístup inšpirovaný neuro-fyziologickým modelom spracovania talamokortikálnej informácie od R. Hecht Nielsena [45]. Model predpokladá existenciu fixného lexikónu, tzv. „symbolov“ v ľudskom talame. Lexikón sa vytvorí v skorom vývinovom štádiu jedinca a neskôr sa už nemení. Učenie pozostáva z vytvárania asociácií medzi neurónmi, odrážajúc tak neurónové spojenia medzi kortikálnymi regiónmi. Najnovšie znalosti o teórii talamu sa nachádzajú v samostatnej publikácii od R. H.-Nielsena [46]. Dynamika učenia je modelovaná podľa Hebbovej metódy [25]. AUP je teda nekontrolované učenie nad prúdom symbolov. Cieľom je naučenie sa znalostnej reprezentácie štruktúry do podoby sémantických sietí, prípadne hierarchií konceptov. Metóda je založená na indukcii asociácií medzi symbolmi vzhľadom na ich spoločné výskyty v kontextovom okne, cez ktoré samotné učenie prebieha.

4.1 Teória konfabulácie

Podľa Nielsena [46] mnoho dôkazov nasvedčuje tomu, že „spracovanie informácií“ podieľajúcich sa na všetkých oblastiach kognície (zrak, sluch, plánovanie, jazyk, rozhodovanie, ovládanie pohybu a premýšľanie...) sa uskutočňuje v mozgovej kôre a v talame. Tvrdí tiež, že existujú dôkazy, že „kognitívne znalosti“ využité v tomto procese sú uložené v mozgovej kôre. V súčasnosti je oblasť skutočného fungovania procesu vnímania (kognície) veľmi slabo preskúmaná rovnako ako pochopenie, čo vlastne tvorí kognitívne znalosti.

Teória konfabulácie je pokusom o prvú konkrétnu a detailnú (a falzifikovateľnú)

vedeckú teóriu popisujúcu proces myslenia. Teória predpokladá existenciu istých neuro-anatomických štruktúr a ich funkcií, ktoré priamo súvisia s fungovaním ľudskej kognície. Jedná sa o talamo-kortikálne moduly a znalostné bázy (obr. 4 – 1 [46]), ktoré sa nachádzajú v talame a mozgovej kôre. Podľa Nielsena [46] je v ľudskom mozgu okolo 4 000 talamo-kortikálnych modulov a zhruba rovnaký počet znalostných báz.



Obr. 4 – 1 Talamo-kortikálne moduly a znalostné bázy podľa teórie konfabulácie.

Teória predkladá štyri základné funkcionálne prvky, ktoré spolu vytvárajú neurónový „hardvér“ na informačné spracovanie vnemov:

- **Lexikón** - každý talamo-kortikálny modul popisuje jeden atribút mentálneho (napr. senzorického, motorického, myšlienkového a iných typov) objektu.
- **Znalostné bázy** - znalostné spoje spájajú páry talamo-kortikálnych modulov, ktoré obsahujú spolu sa vyskytujúce symboly.
- **Konfabulácia** - je operácia informačného spracovania myšlienky.
- **Záver** - výsledkom je podstata akcie, pôvod správania sa.

R.-H. Nielsen podkladá teóriu sériou experimentov [46] založených na jeho vlastnej implementácii. Systém mal za úlohu doplniť nedokončené vety v anglickom jazyku, a to v dvoch alternatívach: bez kontextu a so zadanou kontextovou frázou.

Prezentované výsledky ukazujú, že systém podáva silne relevantné výsledky aj v prípadoch, ak sa odvodené odpovede nenachádzali explicitne v trénovacom korpuse. Vybrané príklady dokončenia viet sú uvedené v tomto tvare:

- Začiatok vety...

  ...doplnenie vety... (kontextová fráza, v prípade pomlčky sa jedná o príklad bez kontextu)

Trojica vybraných príkladov doplnenia viet pomocou teórie konfabulácie [46]:

- The New York...

  ...Times' computer model collapses... (-)

  ...markets traded lower yesterday... (Stocks proved to be a wise investment)

  ...City Center area where... (Downtown events were interfering with local traffic)

- When the United...

  ...Center Party leader urged... (-)

  ...Auto Workers union representation... (The car assembly lines halted due to labor strikes)

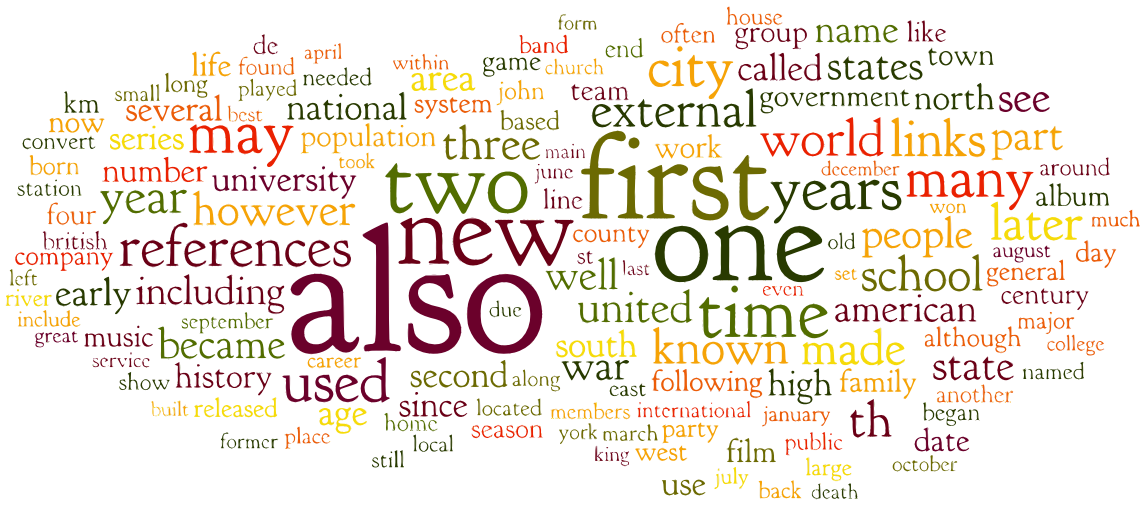
  ...Arab Emirates bought the... (The price of oil in the Middle East escalated)

- I was very...

  ...nervous about my ability... (-)

  ...concerned that they chose... (Democratic citizens voted for their party's candidate)

  ...hungry while knowing the... (Restaurant diners ate meals that were served)



Obr. 4 – 2 Vizualizácia najfrekventovanejších slov lexikónu získaného z Wikipedie (tag cloud).

4.2 Lexikón

Výber vhodného lexikónu známych symbolov je veľmi dôležitou časťou AUP. Často odráža aj kvalitu výstupu systému a jeho veľkosť silne ovplyvňuje náročnosť implementácie na výpočtové zdroje. Vzhľadom na zameranie tejto práce budú symboly lexikónu reprezentovať indexy slov, zjednodušene priamo slová v texte, ktoré sú v implementácii modelu AUP reprezentované ako tokeny.

> **Token** je reprezentáciou symbolu v AUP. V kontexte tejto práce je tokenom index slova (alebo zjednodušene priamo slovo). Slovo je tokenom reprezentované v tvare, v akom bolo použité v texte, napr. pre slová „gitara“ a „gitare“ existujú dva rozdielne tokeny. Token v širšom slova zmysle nemusí nutne zodpovedať iba slovám. Pri rozšírení AUP napr. na doménu hudby mohol token zastupovať jeden tón alebo krátku hudobnú frázu.

> **Lexikón** L je množina tokenov $t_1, t_2, t_3, \ldots, t_n$. Vymedzuje skupinu tokenov, s ktorými dokáže AUP pracovať. Lexikón musí obsahovať vyhradený token reprezentujúci neznámy symbol (reprezentujúci všetky

neznáme symboly). Formálny zápis lexikónu je:

$$L = \{t_1, t_2, t_3, \ldots, t_n\} \tag{4.1}$$

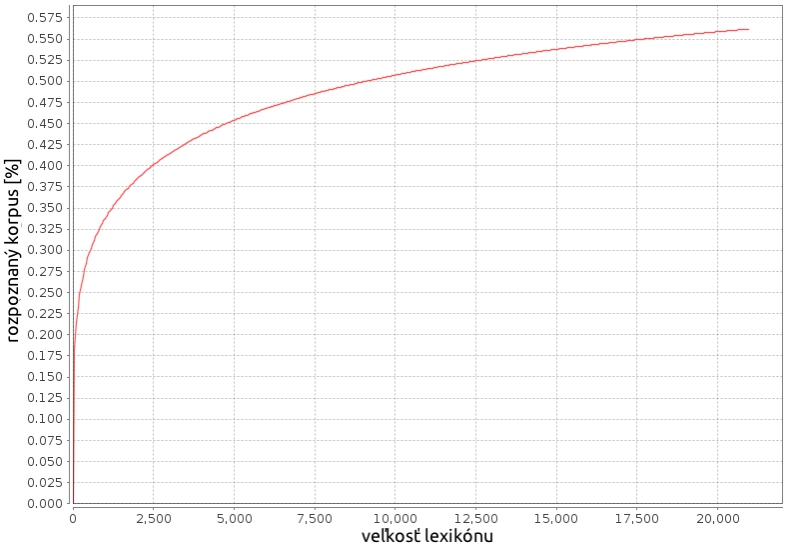
Najjednoduchším spôsobom výberu lexikónu môže byť použitie n najfrekventovanejších slov z korpusu. Pri hľadaní vhodnej hodnoty parametra n nastáva niekoľko problémov. Pri rozsiahlom lexikóne môže implementácia AUP naraziť na problémy spojené s nedostatkom výpočtových zdrojov. V lexikónoch menšej veľkosti môžu naopak chýbať dôležité slová a systém začne produkovať irelevantné výsledky. Na obrázku 4 – 2 sú znázornené konkrétne slová (tag cloud) z lexikónu vytvoreného z 6 000 najpoužívanejších slov z 3 GB korpusu článkov anglickej wikipedie[18]. Obrázok bol vytvorený[19] z reálneho korpusu pre potreby tejto práce. Veľkosť slova je priamo úmerná jeho frekvencii v korpuse. Pre zlepšenie prehľadnosti boli z korpusu pred vizualizáciou odstránené najfrekventovanejšie slová (tzv. stop-words).

Podľa predchádzajúcich výskumov [51], 3 500 najfrekventovanejších slov vo veľkom (napr. 200 MB[20]) súbore textových dokumentov v prirodzenom jazyku (angličtine) tvorí 90 % zo všetkých tokenov v texte. Podľa Murphyho štúdie [40] (vykonanej na inom korpuse v anglickom jazyku) je na pokrytie 90 % použitého korpusu potrebných 10 000 slov, ale už 5 000 slov sa považuje za dostatočné na používanie modelu analogického k AUP. Voľba veľkosti lexikónu je silne závislá na použitom korpuse. Frekvencia slov a ich distribúcia sa môže v korpusoch rôznych doménových oblastí silne odlišovať, preto je potrebné sa otázkou výberu vhodného lexikónu zaoberať pre každý použitý korpus zvlášť. Na obrázkoch 4 – 3, 4 – 4 a 4 – 5 je znázornená závislosť miery rozpoznaného korpusu od veľkosti lexikóna pre tri rôzne korpusy (lineárne úseky na začiatku grafu sú spôsobené nižším rozlíšením grafu, keďže miera

[18]Takýto korpus zodpovedá množine dokumentov o celkovej veľkosti približne 660 miliónov slov, z toho 2.5 milióna jedinečných.

[19]Obrázok bol vytvorený pomocou nástroja Wordle. Ten je dostupný on-line (15.2.2012): http://www.wordle.net/

[20]Takýto korpus zodpovedá množine dokumentov o celkovej veľkosti približne 45 miliónov slov, z toho 500 tisíc jedinečných.



Obr. 4 – 3 Pomer veľkosti rozpoznaného korpusu k veľkosti lexikónu v textoch v slovenskom jazyku.

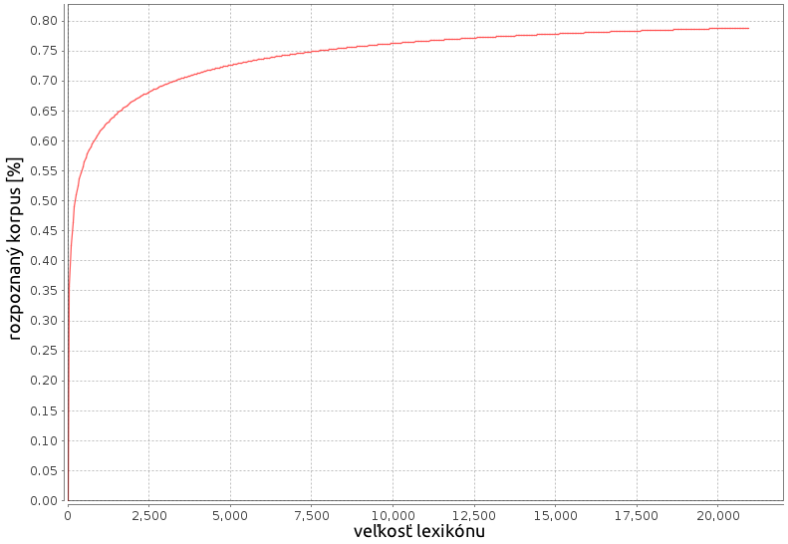
rozpoznaného korpusu bola meraná pri zvyšovaní lexikóna vždy o 1000 tokenov). Os y na týchto grafoch udáva podiel počtu známych tokenov k počtu všetkých rôznych tokenov. Na obrázku 4 – 3 bol graf vytvorený nad korpusom[21] tvorenom textami v slovenskom jazyku zo slovenskej verzie projektu wikipedia[22]. Korpus má veľkosť 121 MB. Na obrázku 4 – 4 bol použitý korpus rovnakej veľkosti, ale tvoril náhodne vybranú časť textov z anglickej verzie wikipedie[23]. Na obrázku 4 – 5 bol použitý korpus z anglickej wikipedie[24] o veľkosti 3 GB. Po prekročení hranice určitého počtu slov do lexikónu už nezodpovedá zisk (v podobe známych slov) z pridania ďalších slov zvýšeniu výpočtovej náročnosti potrebnej na prácu s takýmto slovníkom. Výber veľkosti lexikónu preto zohľadňuje aj dostupné výpočtové prostriedky.

[21] Takýto korpus zodpovedá množine dokumentov o celkovej veľkosti približne 22.5 miliónov slov, z toho 700 tisíc jedinečných.

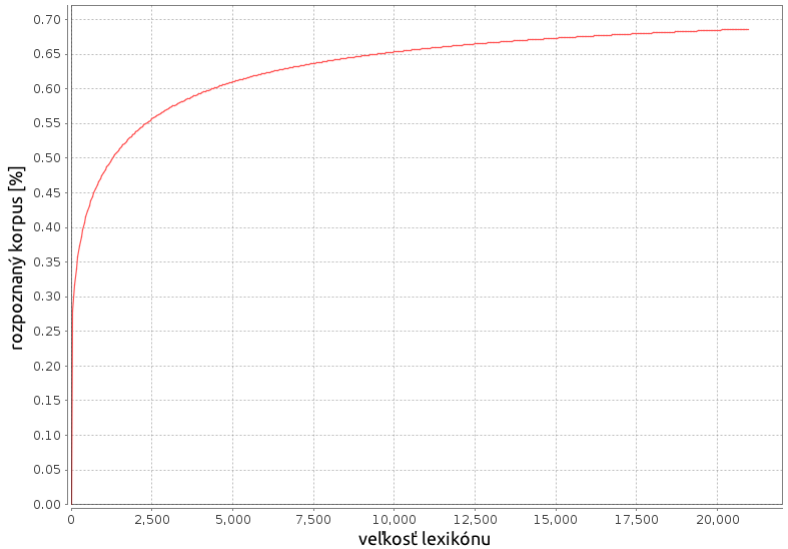
[22] Dostupné on-line (15.02.2012): http://sk.wikipedia.org

[23] Takýto korpus zodpovedá množine dokumentov o celkovej veľkosti približne 25 miliónov slov, z toho 250 tisíc jedinečných.

[24] Dostupné on-line (15.2.2012): http://en.wikipedia.org/



Obr. 4 – 4 Pomer veľkosti rozpoznaného korpusu k veľkosti lexikónu v textoch v anglickom jazyku.



Obr. 4 – 5 Pomer veľkosti rozpoznaného korpusu k veľkosti lexikónu vo veľmi veľkom súbore textov v anglickom jazyku.

4.3 Znalostné bázy

V implementácii teórie konfabulácie od Nielsena [46] je vytvorenie asociácie definované prekročením istého prahu (o veľkosti približne 3) počtu spoločných výskytov dvoch symbolov. Podmienka vytvorenia asociácie v AUP je sprísnená za účelom zníženia pamäťovej náročnosti a výpočtového výkonu možných implementácií. Je založená na vzájomnej informácii dvoch náhodných premenných a jedná sa o často používaný prístup identifikácie asociácií v oblasti NLP [40].

Vzájomná informácia (MI) je mierou prekrytia informácie dvoch náhodných premenných [71]. MI dvoch náhodných premenných X a Y, ktorých hodnoty majú apriórne pravdepodobnosti $p(X = x)$ and $p(Y = y)$ a pravdepodobnosť spoločného výskytu $p(X = x, Y = y)$, je definovaná ako:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(X = x, Y = y) \ln \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \tag{4.2}$$

Bodová vzájomná informácia (PMI) je miera toho, ako veľmi sa líši pravdepodobnosť spoločného výskytu dvoch javov $p(x, y)$ od hodnoty, ktorú by sme očakávali ak by boli dané javy na sebe nezávislé (teda od hodnoty $p(x)p(y)$) [8]. PMI je definovaná ako:

$$PMI(x, y) = \ln \frac{p(x, y)}{p(x)p(y)} \tag{4.3}$$

V AUP je teda nutnou podmienkou pre vytvorenie asociácie medzi párom tokenov ich štatisticky nenáhodný spoločný výskyt. Je definovaný ako prekročenie prahu veličinou definovanou ako signifikancia asociácie. Signifikancia asociácie je zjednodušením PMI, vzhľadom na funkciu porovnávania s prahom a vlastnosti logaritmickej funkcie ako monotónne stúpajúcej funkcie. Definuje ju vzťah:

$$S(x, y) = \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)}, \tag{4.4}$$

kde X a Y sú diskrétne náhodné premenné. Vzájomná signifikancia $S(x, y)$ je definovaná ako podiel vzájomnej pravdepodobnosti symbolov x, y a pravdepodobnosti ich

náhodného výskytu. Ak hodnota $S(x, y)$ prekročí vopred určenú prahovú hodnotu, považujeme dané symboly za asociované.

Kontextová vzdialenosť vyjadruje vzdialenosť medzi slovami tak, že predstavuje počet medzier medzi zdrojovým slovom a cieľovým slovom. Ak sa cieľové slovo nachádza v texte napravo od zdrojového slova, kontextová vzdialenosť nadobúda kladné hodnoty, v opačnom prípade nadobúda záporné hodnoty. V prípade, že je zdrojové slovo zároveň aj slovom cieľovým, je hodnota kontextovej vzdialenosti rovná 0.

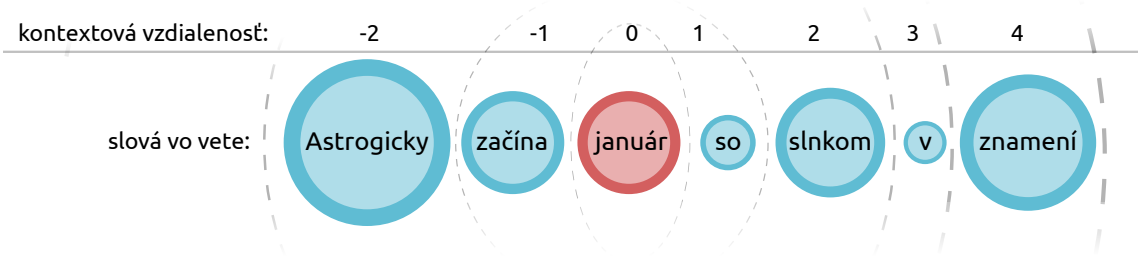
Príklad kontextových vzdialeností na krátkom úryvku textu je zobrazený na obrázku 4 – 6. Dolným indexom pri výpočte signifikancie značíme, pre ktorú kontextovú vzdialenosť platí. Signifikanciu pre kontextovú vzdialenosť i zapisujeme $S_i(x, y)$. Môže nadobúdať hodnoty z intervalu $\langle 0, \infty)$.

Signifikancia v kontextovej vzdialenosti: $S_i(x, y)$ sa počíta podľa vzťahu 4.4 za daných podmienok: ak $i > 0$, tak sa príslušné javové pole skladá z $(i + 1)$-tic po sebe idúcich tokenov, pričom premenná X predstavuje token najviac vľavo z tejto $(i+1)$-tice a premenná Y predstavuje naopak token najviac vpravo z tejto $(i+1)$-tice. Pre $i < 0$ sa javové pole skladá z $(|i| + 1)$-tic po sebe idúcich tokenov, kde premenná X predstavuje naopak token najviac vpravo a Y je token najviac vľavo.

Najvyššie hodnoty signifikancie (pre daný korpus) sú príznačné pre slovné spojenia, ktoré sú silne špecifické a používajú sa v texte výlučne ako dvojica, teda samy o sebe nemajú význam. Najnižšie hodnoty signifikancie naopak označujú dvojice, kde sa najmenej jedno zo slov často asociuje aj s inými slovami. V tabuľke 4 – 1 je vypísaných 20 slovných spojení (v kontextovej vzdialenosti 1) s najvyššími hodnotami a 20 s najnižšími hodnotami signifikancií (pre všetky signifikancie väčšie ako 1). Najvyššiu hodnotu signifikancie dosahujú dvojice slov, ktoré sa vyskytujú spolu oveľa častejšie ako v spojení s inými slovami. V textoch v prirodzenom jazyku

Tabuľka 4 – 1 20 slovných spojení s najvyššími a najnižšími hodnotami signifikancie.

pár tokenov	hodnota signifikancie	pár tokenov	signifikancia
al bah	359895.151515151	rank that	1.0000047433
sheriff elvsted	122438.556701031	a grasp	1.0000169712
gee gee	89973.7878787879	another fast	1.0000481645
al mon	67480.3409090909	words said	1.0000495455
non est	60905.3333333333	what trade	1.0000590555
don pedro	59406.462585034	way bent	1.0000613011
binary transfer	56825.5502392345	each soul	1.0000687116
ali baba	53913.7244652315	sent either	1.0000722485
il est	52340.5208333333	severely be	1.0000862278
contributing scanning	51285.0590909091	therein be	1.0000862278
employee identification	49880.6886142404	woman full	1.0000879963
vanilla electronic	45717.112029733	her price	1.0000914351
mon dieu	43248.9243027889	most fine	1.000093358
gratefully accepts	41178.8211382114	the duty	1.0001052606
asta allmers	40953.5862068966	lips look	1.0001060185
non te	40719.5657142857	view and	1.0001064477
et exeunt	40396.3945578231	doth his	1.0001088726
accepts contributions	39840.3308270677	as belongs	1.0001190163
zip corrected	39811.4105658353	him honor	1.0001236203
internal revenue	38514.1252773511	him injustice	1.0001236203



Obr. 4 – 6 Kontextové okno a kontextová vzdialenosť pre slovo „január“.

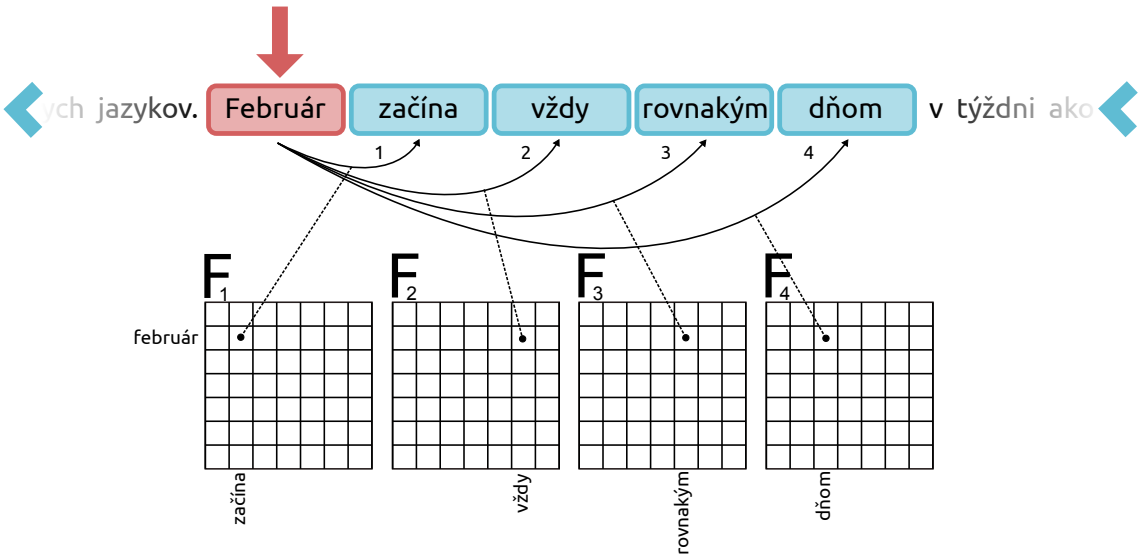
sa v tomto prípade často jedná o dvojslovné výrazy ako napríklad mená ľudí (ali baba, don pedro, asta allmers, ...) alebo ustálené slovné spojenia (ako napr. „binary transfer“).

4.4 Proces učenia

Učenie prebieha nad predspracovaným textovým korpusom. V prvom kroku sa vytvorí lexikón známych symbolov.V druhom kroku sa v textovom korpuse nahradia všetky symboly, ktoré sa nevyskytujú v lexikóne za značky označujúce neznámy symbol. Takto predspracovaný korpus slúži ako vstupný parameter pre učiaci algoritmus. Učenie prebieha v kontextovom okne, ktorého veľkosť je ďalším vstupným parametrom učenia.

V kontextovom okne sa zrátavajú početnosti spoločných výskytov slov, na ktoré počas učenia algoritmus narazil a tie poslúžia neskôr na výpočet pravdepodobností ich spoločných výskytov pre n kontextových vzdialeností, ktoré ďalej slúžia na výpočet signifikancií (vzťah 4.4) a neskôr váh (vzťah 4.13). Matice váh pre danú kontextovú vzdialenosť i a pre daný prah (pre vytvorenie asociácie) sa nazývajú fascikle F_i. Učenie systému cez kontextové okno je znázornené na obr. 4 – 7.

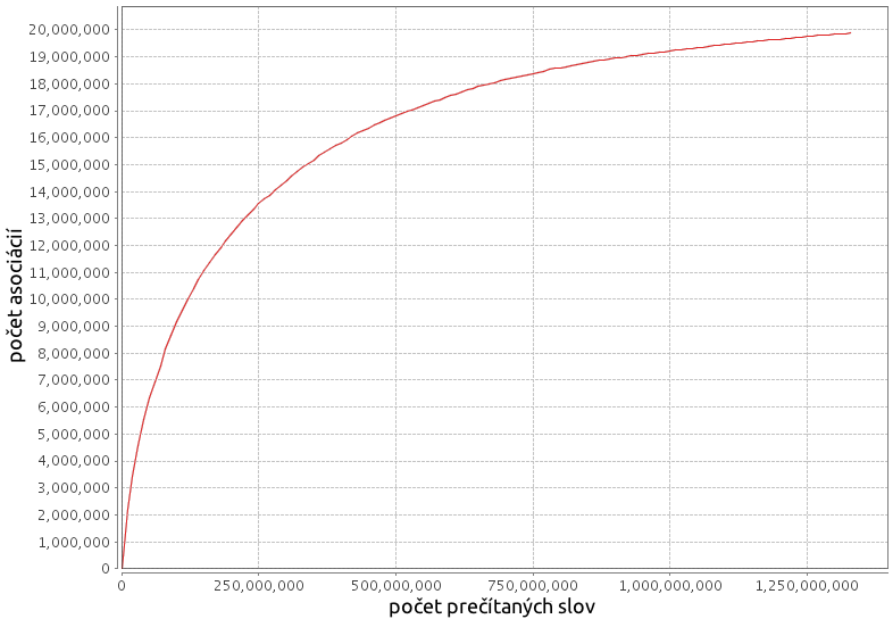
Účelom procesu učenia je indukcia asociácií medzi symbolmi. V prípade, že učenie prebieha nad textom v prirodzenom jazyku a symboly reprezentujú jednotlivé slová v texte je prínosné poznať priebeh rastu počtu asociácií medzi slovami vzhľa-



Obr. 4 – 7 Učenie v kontextovom okne.

dom na veľkosť korpusu. Pri rozlišovaní slov sa neberú do úvahy ich rôzne tvary (skloňovanie, časovanie, atď.) a preto sú slová „*gitara*“ a „*gitare*“ reprezentované dvoma rozdielnymi symbolmi. Za ideálnych podmienok by bolo potrebné proces učenia zastaviť až v momente, keď prestane počet asociácií stúpať. Na obr. 4 – 9 je znázornený rast asociácií pri prahu 1.5, nad korpusom vytvorený umelou gramatikou pomocou programu SLG[25](Simple Language Generator) generujúcou jednoduché vety v prirodzenom jazyku s lexikónom o veľkosti 35 slov. Vzhľadom na nižší počet prečítaných slov je možné na grafe vidieť aj úseky, v ktorých počet asociácií klesá. Tento jav nastáva keď sa v texte po vytvorení asociácie medzi dvoma slovami tieto slová vyskytujú často, ale nie spolu (v danej kontextovej vzdialenosti). Po 10 000 vetách sa počet asociácií ustálil a ďalej nerástol. V reálnych korpusoch, obsahujúcich texty v prirodzenom jazyku by táto ideálna situácia nastala až po prečítaní enormného množstva dát (rádovo niekoľko desiatok GB). Priebeh rastu celkového počtu asociácií vzhľadom na veľkosť spracovaného korpusu (pri veľkosti lexikóna 6 000 slov) je

[25]Program SLG generuje text na základe vlastného formátu definície gramatiky. Ten sa do SLG zadáva ako textový súbor v plain/text formáte a je v textovej podobe jednou z príloh tejto práce. Dostupné on-line (2. 10. 2012): http://tedlab.mit.edu/ dr/SLG/



Obr. 4 – 8 Nárast počtu asociácií v závislosti na počte prečítaných symbolov v 6 GB korpuse.

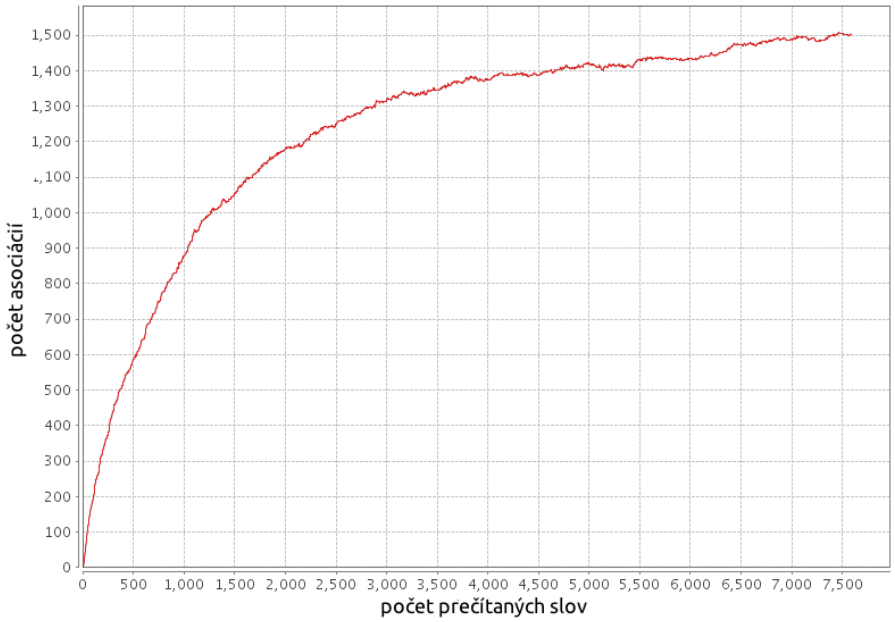
možné vidieť na obr. 4 – 8. Tento rast asociácií sa dá interpolovať funkciou:

$$f(x) : p_1(1 - e^{-p_2 x^{p_3}}), \tag{4.5}$$

kde x označuje počet prečítaných tokenov, a p_1, p_2 a p_3 sú konkrétne parametre s hodnotami $p_1 = 21046251.180057187$, $p_2 = 0.000003894983657307634$ a $p_3 = 0.6468320176772836$. Korpus, v ktorom rast asociácií prebieha podľa funkcie 4.5, by musel mať veľkosť približne 16GB, aby AUP mohlo akumulovať 99.7% všetkých možných asociácií.

4.5 Sémantická príbuznosť

Ľuďom nerobí najmenší problém povedať, či dve slová spolu súvisia. Napríklad ak sa stretnú s dvojicou slov - „auto“ a „bicykel“, tak s istotou vedia povedať, že tieto slová spolu súvisia, obe totiž označujú dopravné prostriedky. Taktiež ľuďom v bežnom živote nerobí najmenší problém posúdiť, či dvojica slov spolu súvisí viac alebo menej. Napríklad bez pochýb určia, že slovo „bicykel“ má bližšie k „autu“



Obr. 4 – 9 Nárast počtu asociácií v závislosti na počte prečítaných symbolov v korpuse vytvorenom nad umelou gramatikou.

ako k „žehličke“ aj bez presnej definície pojmov ako „súvisieť“ alebo „mať bližšie“. Je možné nejakým spôsobom priradiť tejto príbuznosti dvoch slov istú konkrétnu hodnotu? S použitím znalostí nahromadených počas procesu učenia bol vytvorený postup na výpočet sémantickej príbuznosti. Tá sa snaží kvantifikovať sémantickú príbuznosť dvoch slov.

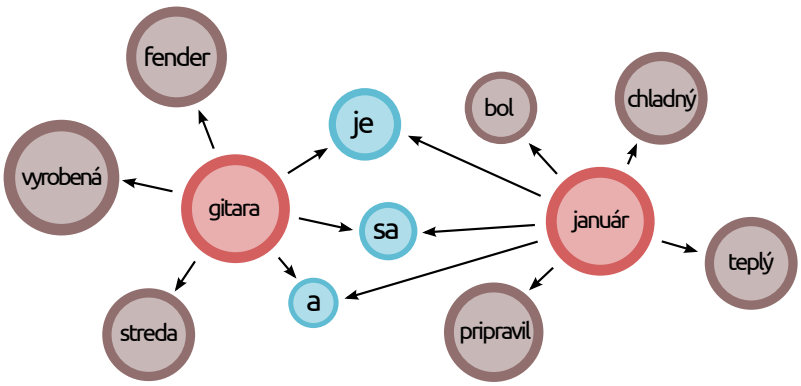
Sémantická príbuznosť *dvojice slov je mierkou ich sémantickej podobnosti na základe is-a relácie.*

4.5.1 Výpočet sémantickej príbuznosti práve dvoch symbolov

Prístup je založený na Jaccardovom indexe [28] počítania podobnosti alebo diverzity dvoch množín:

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{4.6}$$

Popisuje výpočet podobnosti dvoch množín A a B. Obor hodnôt Jaccardovho indexu je <0,1>, takže nie je potrebné ho ďalej normalizovať. V AUP je každý symbol



Obr. 4 – 10 Ukážka zdieľania asociovaných symbolov pre dve rôzne slová v jednej kontextovej vzdialenosti.

množinou asociácií, teda svojim asociovaným sémantickým okolím v rôznych kontextových vzdialenostiach. Sémantické okolie symbolu x v kontextovej vzdialenosti i definujeme ako $O_i(x)$:

$$O_i(x) = \{t \in L | S_i(x,t) > prah\} \tag{4.7}$$

$O_i(x)$ je teda množinou všetkých takých symbolov z množiny známych symbolov lexikónu L, pre ktoré platí, že sú so symbolom x asociované v kontextovej vzdialenosti i. Príbuzné slová sa vyskytujú v jazyku v podobných kontextoch. To znamená, že ich sémantické okolia sú si podobné. Ilustračný príklad sémantických okolí dvoch slov spolu s prienikom ich sémantických okolí ilustruje obr. 4 – 10. Hnedou a modrou farbou sú vyznačené slová patriace do sémantického okolia slov označených červenou farbou. Modrá farba označuje slová, ktoré patria do sémantických okolí oboch slov, teda tvoria ich prienik. Podobnosť dvoch symbolov a a b vzhľadom na kontextovú vzdialenosť i môžeme definovať ako:

$$Rel_i(a,b) = \frac{|O_i(x) \bigcap O_i(y)|}{|O_i(x) \bigcup O_i(y)|} \tag{4.8}$$

Váženou sumou podobností symbolov cez všetky kontextové vzdialenosti dostávame vzťah na výpočet podobnosti dvoch symbolov $Rel^v(x,y)$:

$$Rel^v(x,y) = \sum_{i=1}^{n} v_i Rel_i(x,y) \tag{4.9}$$

Tabuľka 4 – 2 Zoznam prvých desiatich najpodobnejších slov k slovu „yellow“.

slovo	$Rel^v(\text{„yellow“}, slovo)$
yellow	1.0
blue	0.31062254494090974
brown	0.3048039234971298
green	0.30468840697394944
red	0.3003874393901687
grey	0.3002948818491731
golden	0.2896206253346786
white	0.2851556702295284
black	0.2841024251982456
gray	0.28190413094802924

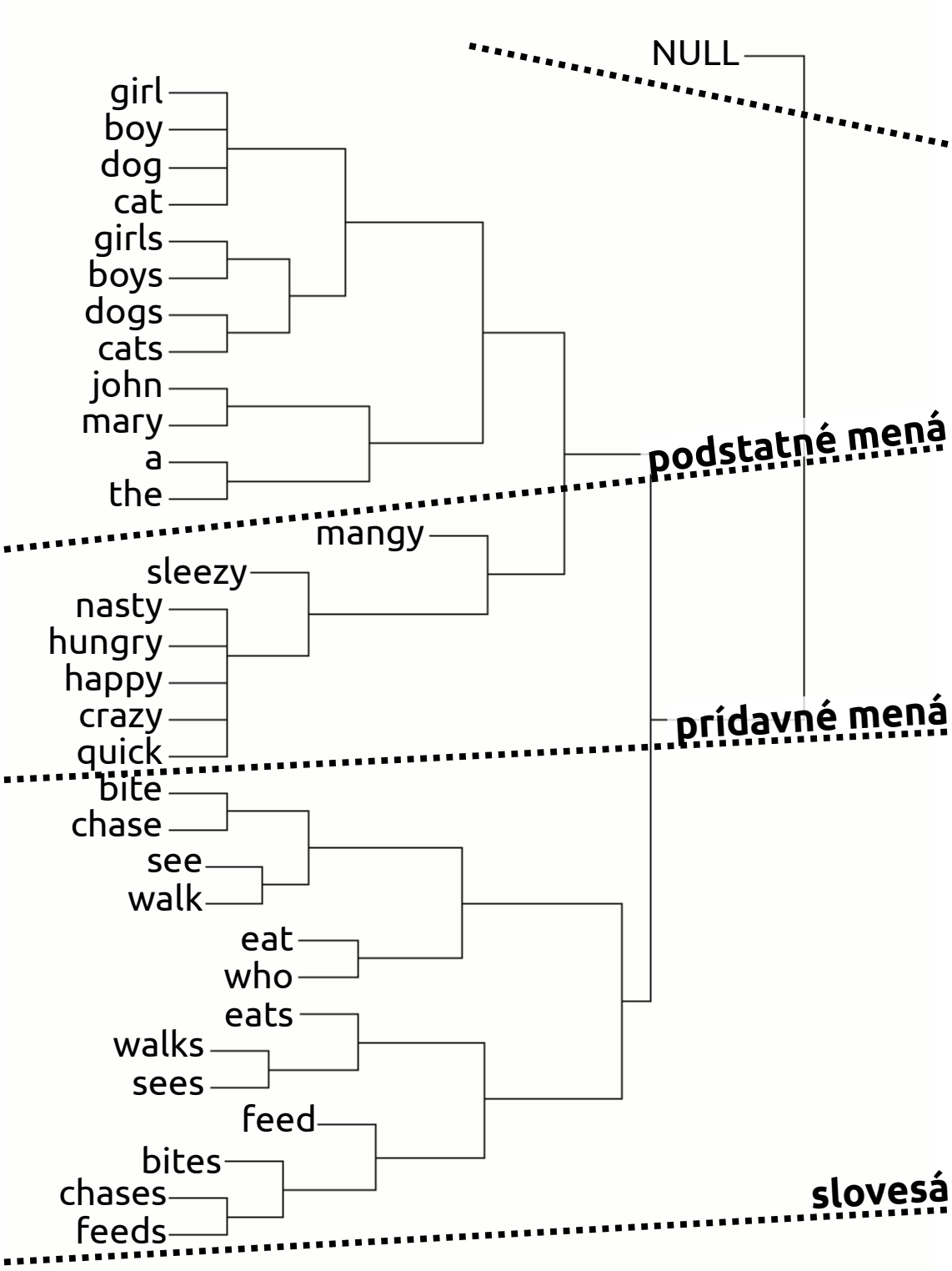
Oborom hodnôt $Rel^v(x, y)$ je množina $\langle 0, \sum_{i=1}^{n} v_i \rangle$. Ak sú váhy pre jednotlivé kontextové vzdialenosti volené tak, že ich súčet dáva hodnotu 1, je oborom hodnôt interval $\langle 0, 1 \rangle$.

Výpočtom podobností vybraného symbolu so všetkými symbolmi z lexikónu L dostávame zoradený zoznam podobností. Príklad časti takého zoznamu pre token „yellow“ je obsiahnutý v tabuľke 4 – 2.

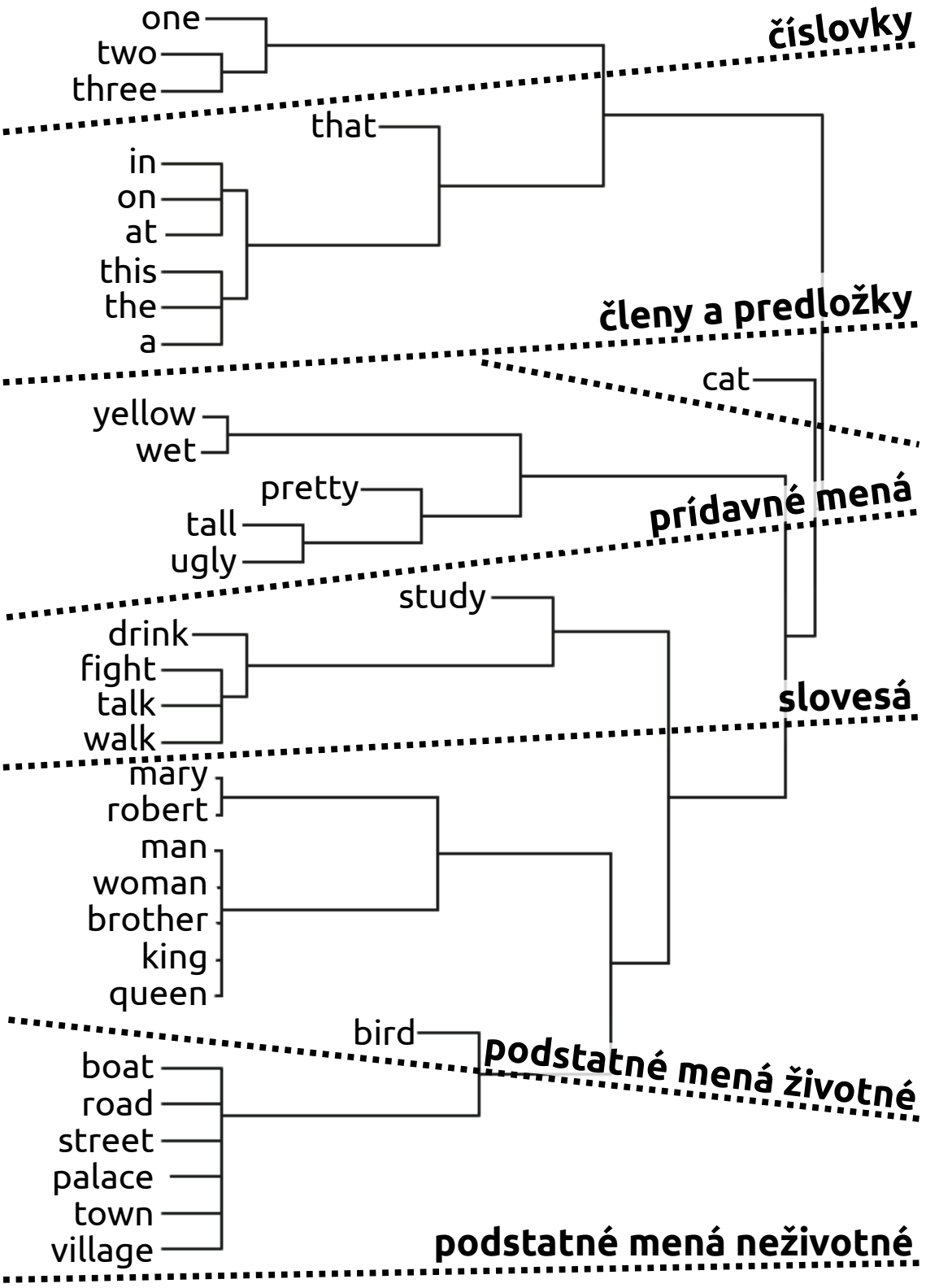
Týmto prístupom boli vytvorené [51] matice podobností pre množinu vybraných slov, ktorá poslúžila ako vstup pre zhlukovací algoritmus hclust[26], funkciu jazyka R[27]. Jeho výstupom bol dendrogram – hierarchická štruktúra konceptov zoradených podľa podobnosti. Zhluky v dendrograme odrážali príslušnosť tokenov do morfologických skupín v anglickom jazyku. Dendrogramy vytvorené nad textom vytvorenom nad umelou gramatikou SLG a množine textov v prirodzenom jazyku (angličtine) sú znázornené na obrázkoch 4 – 11 a 4 – 12.

[26]Dostupné on-line (2. 10. 2012): http://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html

[27]R je programovací jazyk a prostredie určené pre štatistickú analýzu dát a ich grafické zobrazenie. Dostupné on-line (2. 10. 2012): http://www.r-project.org/



Obr. 4 – 11 Ukážka hierarchickej štruktúry vygenerovanej pomocou AUP nad umelou gramatikou.



Obr. 4 – 12 Ukážka hierarchickej štruktúry vygenerovanej pomocou AUP nad korpusom z projektu Gutenberg.

4.5.2 Výpočet sémantickej príbuznosti symbolov a množín symbolov

Výpočet sémantickej príbuznosti, kde je najmenej jedným vstupom množina symbolov je skoro identický s výpočtom sémantickej príbuznosti práve dvoch symbolov. Na to, aby sa dal okamžite používať s doteraz popísanými postupmi je potrebné doplniť iba jeden vzťah, a to sémantické okolie množiny symbolov. Sémantické okolie množiny symbolov definujeme ako:

$$O_i^S = \bigcap_{O_i \in S} O_i, \tag{4.10}$$

kde O_i^S označuje sémantické okolie množiny symbolov a S je množina sémantických okolí daných symbolov v kontextovej vzdialenosti i. V prípade, že počítame sémantickú príbuznosť symbolu k množine symbolov, jednoducho zameníme vo vzorci 4.8 okolie symbolu O_i za okolie množiny symbolov O_i^S.

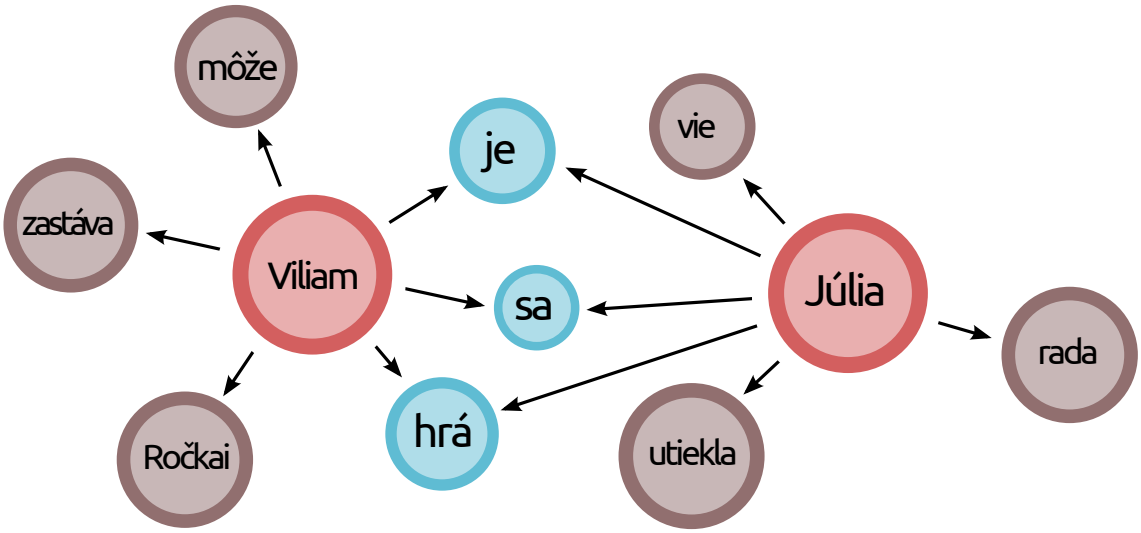
Prvých n symbolov zoradených podľa hodnoty sémantickej príbuznosti k trom slovám „January“, „February“ a „March“ odráža tabuľka 4 – 3.

4.5.3 Príklad výpočtu sémantickej príbuznosti pre jednu kontextovú vzdialenosť

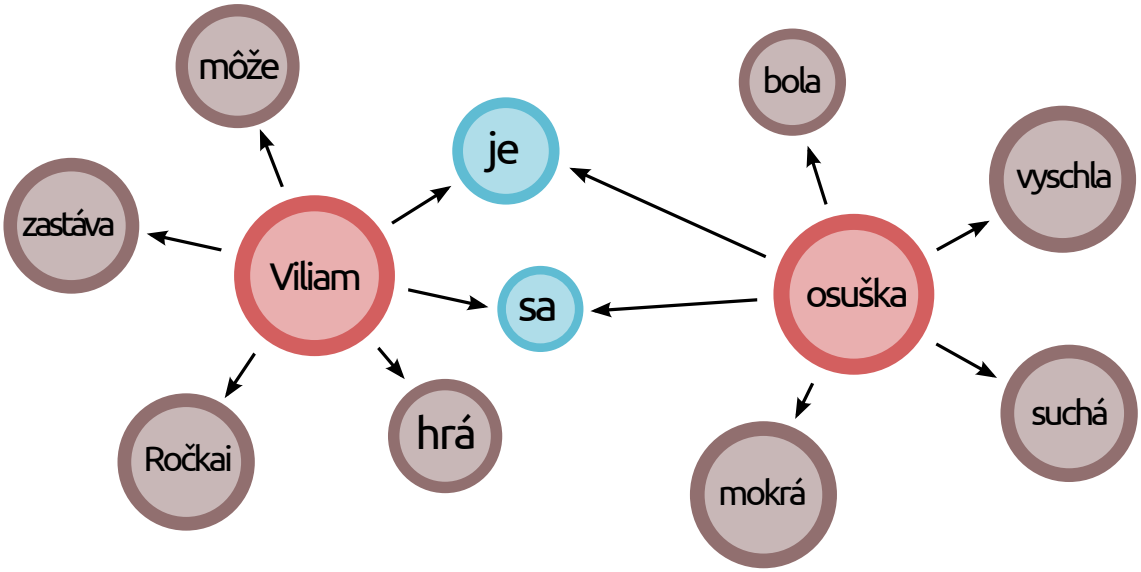
Majme slová „Viliam“, „Júlia“ a „osuška“ také, že poznáme ich sémantické okolie. Chceme vypočítať sémantickú príbuznosť slova „Viliam“ k ostatným slovám. Predpokladajme, že asociácie vyzerajú takto:

- Viliam → môže, zastáva, Ročkai, je, sa, hrá
- Júlia → vie, rada, utiekla, je, sa, hrá
- Osuška → je, sa, suchá, bola, mokrá, vyschla

Počet spoločných asociácií tokenov „Viliam“ a „Júlia“ je 3, počet všetkých rôznych tokenov asociovaných s týmito tokenmi je 9. Kvantitatívne vyjadrenie ich sémantickej príbuznosti je teda $Rel_1(\text{„Viliam“}, \text{„Júlia“}) = 3/9 = 0.333$.



Obr. 4 – 13 Zobrazenie sémantického okolia pre slová „Viliam“ a „Júlia“.



Obr. 4 – 14 Zobrazenie sémantického okolia pre slová „Viliam“ a „osuška“.

Tabuľka 4 – 3 Zoznam prvých dvadsiatich najpodobnejších slov k množine pozostávajúcej zo slov „January“, „February“ a „March“.

slovo	$Rel^v(\{$„January“, „February“ a „March“$\}, slovo)$
january	2.8619116151864556
february	2.7270807776738524
march	2.243796338085861
december	2.1562326106107244
november	2.1520753924288396
september	2.1486218392007594
october	2.106681514366388
april	2.099955014402455
june	2.065955999024085
july	2.042209024936516
august	1.8269473651020265
monday	0.811377994214037
autumn	0.7147185887567146
friday	0.66482290707757
week	0.6403213579159409
month	0.6276177185006208
late	0.6198000353756496
saturday	0.6102371034447043
months	0.5971979292401444
may	0.5915609347899611

Počet spoločných asociácií slov „Viliam“ a „osuška“ je 2, počet všetkých rôznych slov asociovaných s týmito slovami je 10. Kvantitatívne vyjadrenie ich sémantickej príbuznosti je teda $Rel_1($„Viliam“, „osuška“$) = 2/10 = 0.2$. Pre tieto tri slová postupným výpočtom získame výslednú tabuľku príbuzností 4 – 4.

Tabuľka 4 – 4 Matica príbuzností.

	Viliam	Júlia	osuška
Viliam	1	0.333	0.2
Júlia	0.333	1	0.2
osuška	0.2	0.2	1

Z čoho je možné vyvodiť záver, že pojem „Viliam“ má vzhľadom na AUP bližšie k pojmu „Júlia“ ako k pojmu „osuška“.

4.6 Konfabulácia

Podľa Nielsena [46] je konfabulácia charakterizovaná ako kľúčová operácia v rámci kognitívneho spracovania informácií. Predpokladá váhu každej asociácie. Váhy asociácií sa využívajú pri výpočtoch P-slov, ktoré budú popísané neskôr. Nielsen váhu definuje takto:

$$log_c(p(\alpha \mid \lambda)/p_0) + B, \tag{4.11}$$

kde $p(\alpha \mid \lambda)$ je podmienená pravdepodobnosť, že zdrojový symbol α bude aktívny za predpokladu, že λ je aktívny symbol a c, p_0 a B sú pozitívne konštanty.

Vzhľadom na využitie váhy, ktoré predpokladá iba porovnávanie hodnôt s inými váhami a na fakt, že logaritmus je monotónnou stúpajúcou funkciou, môžeme vzťah zjednodušiť až na $p(\alpha \mid \lambda)$.

V AUP sa váha asociácie (vzťah 4.13) medzi dvoma symbolmi x a y v kontextovej vzdialenosti i počíta iba v prípade, že sú dané symboly asociované podľa 4.4 a je daná vzorcom vychádzajúcim zo vzorca výpočtu podmienenej pravdepodobnosti:

$$P(A \mid H) = \frac{P(A, H)}{P(H)} \tag{4.12}$$

$$w_i(x, y) = \frac{p(X = x, Y = y)}{p(Y = y)} \tag{4.13}$$

Vzťah pre výpočet váhy vychádza od vzťahu na výpočet signifikancie 4.4 a líši sa v menovateli. Kým vzťah 4.4 vyjadruje mierku pravdepodobnosti spoločného

výskytu dvoch symbolov, k ich náhodnému výskytu, vzťah 4.12 vyjadruje mierku pravdepodobnosti spoločného výskytu dvoch symbolov, k pravdepodobnosti výskytu cieľového symbolu.

4.7 Výpočet konsenzu - konfabulácia

Konsenzus sa využíva na rozpoznávanie fráz a jeho popis je potrebný pre pochopenie procesu generovania P-slov (popísaných v 4.8) [52].

> **Fráza** je pre potreby tejto práce definovaná ako syntakticky validný vstup rozpoznaný systémom.

Výpočet sa vykonáva nad oknom (niekoľkých slov) a je hľadaním odpovede, ktorá najviac potvrdzuje najslabšiu hypotézu. Je dané okno n slov $t_1, t_2, \ldots, t_n$. Maximálna kontextová vzdialenosť v takom okne je $n-1$. Posledné slovo v okne, t_n, sa nazýva dotazovacím slovom. Všetky ostatné slová hlasujú cez relevantné fascikle $(F_1, F_2, \ldots, F_{(n-1)})$ určené ich kontextovou vzdialenosťou a vytvárajú množiny odpovedí $O_1(t_{n-1}), O_2(t_{n-2}), \ldots, O_{(n-1)}(t_1)$. Prienik týchto množín potom tvorí množinu možných odpovedí R v danom okne a nazýva sa plná množina konsenzov:

$$t_n \in R; R = O_1(t_{n-1}) \bigcap O_2(t_{n-2}) \bigcap \ldots \bigcap O_{(n-1)}(t_1) \tag{4.14}$$

Množina R obsahuje l slov $r_1, r_2, \ldots, r_l$, ktoré sú asociované s každým slovom v okne cez príslušnú kontextovú vzdialenosť. Tieto asociácie majú zvyčajne rozdielne váhy w_i (index váhy vyjadruje fascikel, nad ktorým bola vypočítaná). Aby sme podporili najslabšiu hypotézu, musíme vyjadriť silu konsenzu $s(m_k)$:

$$s(r_k) = min(w_{(n-1)}(t_1, r_k), w_{(n-2)}(t_2, r_k), \ldots, w_1(t_{(n-1)}, r_k)); k = 1, 2, \ldots, l \tag{4.15}$$

Odpoveď, ktorá najviac podporuje najslabšiu hypotézu, je t_{answer}:

$$t_{answer} = r_k; s(r_k) = max(s(r_1), s(r_2), \ldots, s(r_l)) \tag{4.16}$$

4.8 Identifikácia sémanticky a syntakticky blízkych slov - P-slov

Pre potreby tejto práce sú P-slová definované ako slová patriace do nejakej konkrétnej sémantickej triedy. Vstupom do algoritmu na výpočet sémanticky blízkych slov je zvolené známe slovo. Výstupom algoritmu je množina slov, ktoré sú k slovu zadanému syntakticky blízke, zároveň rozšírené aj o slová z jednej sémantickej triedy, čo implicitne definuje samotnú triedu.

> **P-slová** *sú slová patriace do nejakej konkrétnej sémantickej triedy rozšírené o slová z jedného sémantického poľa.*

Algoritmus na výpočet množiny P-slov:

1. Pre zadané slovo **t** vyber všetky asociácie z fasciklov F_{-2}, F_{-1}, F_1, F_2, teda vytvor množiny $O_{-2}(t), O_{-1}(t), O_1(t), O_2(t)$.

2. Nájdi úplnú množinu konsenzov t_{answer} k dotazovanému slovu **t** pre všetky kombinácie $O_{-2}(t) \times O_{-1}(t) \times t_{answer} \times O_1(t) \times O_2(t)$. Výsledok pridaj do množiny **R**.

3. V množine **R** sa nachádza množina všetkých P-slov k dotazovanému slovu.

Systém sa učil nad množinou anglických textov pozostávajúcich z náhodne vybraných titulov beletrie z elektronickej knižnice gutenberg.org[28]. Korpus pozostával z 1 GB textu v anglickom jazyku. Veľkosť lexikónu bola 3 000 slov a prah (potrebný pre vytvorenie asociácie) pre jednotlivé príklady bol použitý v rozsahu 6.0 - 10.0, vzhľadom na mohutnosť výslednej množiny odpovedí s váhou konsenzu väčšou ako 0.001. Príklad výstupu zo systému vyzeral napríklad takto:

- moon → star, stars, moon, lights, bright

[28]Projekt Gutenberg je digitálnou knižnicou obsahujúcou viac ako 40 000 knižných titulov voľne k stiahnutiu. Dostupné on-line (14.7.2012): http://www.gutenberg.org/

- white → brown, grey, black, white, red, yellow, blue, silk, gray, green, purple
- william → joseph, james, william, edward, smith, thomas
- tree → trees, plants, tree, cloud, plant, pine, branches, green, grass
- two → two, fifteen, eight, twelve, seven, three, forty, twenty, six, four, five, thirty

Pomocou prístupu vytvárania P-slov sa AUP podarilo vytvárať rôzne množiny relevantných slov patriacich do jednej sémantickej triedy. Po zadaní vstupného slova bola po dostatočne vysokom prahu dosiahnutá úplná presnosť výsledku.

4.9 Kontext symbolov

Jedným zo základných problémov v oblasti spracovania prirodzeného jazyka je viacvýznamovosť. Slová majú rôzny význam v závislosti od okolností. Slovo „myš“ môže popisovať rovnako „drobného cicavca“, ako aj „počítačové vstupné rozhranie“. No za istých špecifických okolností by bolo vhodnejšie „myš“ označiť ako „domáceho miláčika“, „hlodavca“, alebo dokonca „jedlo“. Predpokladáme, že tieto významy rôznych slov sa dajú odvodiť z ich okolia, teda predpokladáme existenciu asociácie medzi slovom a jeho významom v nejakej kontextovej vzdialenosti. Za daného predpokladu by sa mala dať zostrojiť metóda na vyhľadanie a kvantifikáciu všetkých možných významov-kontextov pre skupinu slov.

> **Kontext** je definovaný ako „časti písaného alebo hovoreného tvrdenia, ktoré predchádzajú alebo nasledujú určité slovo, resp. pasáž a zvyčajne ovplyňujú jeho význam“ [80] alebo ako „časť textu alebo tvrdenia, ktoré obklopuje konkrétne slovo alebo pasáž a určuje jeho význam“ [79].

Aby bolo možné vyhľadať relevantné kontexty, je potrebné kvantifikovať hodnotu väzby medzi vstupným slovom a kontextovým slovom. Tú definujeme ako vplyv

kontextu, podiel váženej sumy signifikancií daných slov cez všetky kontextové vzdialenosti a pravdepodobnosti výskytu kontextového slova v texte:

$$c(x,y) = \sum_{i=1}^{n} \frac{v_i S_i(x,y)}{p(y)}, \tag{4.17}$$

kde x označuje vstupné slovo, y označuje jeho kontext, v_i je váha pre kontextovú vzdialenosť i a $S_i(x,y)$ je hodnota signifikancie pre dané symboly v danej kontextovej vzdialenosti i. Vplyv kontextu je vypočítaný pre n kontextových vzdialeností. Delenie hodnotou $p(y)$ oslabí vplyvy tzv. „stop words“, ktoré sa asociujú s veľkým množstvom slov. Signifikancie medzi vstupným slovom a kontextom sú počítané iba v prípade, že sú dané slová v danej kontextovej vzdialenosti asociované. Vplyv kontextu pre množinu slov T a kontextové slovo x je definovaný ako:

$$c(T,x) = min\left(c(t_0,x), c(t_1,x), \ldots, c(t_n,x)\right); T = \{t_0, t_1, \ldots, t_n\} \tag{4.18}$$

Vplyv kontextu množiny slov T a kontextového slova x je minimálna hodnota z množiny kontextových síl pre dané slovo a každé slovo z množiny T.

Kontext bol vypočítaný pre človekom vybrané trojice slov. Výstup prezentoval zoznam všetkých slov z lexikónu zoradený podľa vplyvu kontextu pre danú trojicu slov. Hádaný kontext bol určený človekom a vychádzal z praktických skúseností o svete. Výsledný zoznam zobrazuje tabuľka 4 – 5. Stĺpec pozícia označuje pozíciu tohto hádaného slova v zoradenom zozname kandidátov. Stĺpec víťaz označuje prvé slovo z daného zoznamu. Keďže vplyv kontextu je silne doménovo závislá premenná, môžeme hodnotu víťaza chápať ako najpravdepodobnejší kontext danej trojice pre doménu vstupných dokumentov použitých na učenie systému.

4.9.1 Kontextovo závislá sémantická príbuznosť

Kontext vybraného slova z pohľadu AUP je definovaný ako slovo asociované s vybraným slovom v rôznych kontextových vzdialenostiach. Množina všetkých kontextov C_t pre dané slovo t je teda zjednotenie všetkých sémantických okolí slov t cez

Tabuľka 4 – 5 Určenie kontextového tokenu pre trojicu slov.

hádaný kontext	trojica slov	pozícia	víťaz
dish	chicken,potatoes,fish	3	roast
art	painting,music,poetry	4	Italian
drink	beer,wine,water	4	bottles
plant	tree,flower,wheat	5	grows
literature	prose,roman,poetry	8	poets
instrument	piano,guitar,organ	8	plays
war	fight,battle,death	8	fight
water	lake,river,sea	10	frozen
cloth	skirt,trousers,coat	10	silk
weapon	bow,gun,sword	15	swung
color	red,green,blue	19	velvet
royal	king,queen,prince	39	palace
building	house,palace,church	46	build
fruit	orange,peach,cherry	50	blossoms

Pre každú trojicu slov bolo hľadané (človekom) kontextové slovo. Stĺpec pozícia označuje pozíciu tohto hľadaného slova v zoradenom zozname kandidátov. Stĺpec víťaz označuje prvé slovo z daného zoznamu.

všetky možné kontextové vzdialenosti:

$$C_t = \bigcup_i O_i(t) \tag{4.19}$$

Možným kontextom c tokenu t je potom každý prvok množiny C_t. Silu kontextu c pre token t v kontextovej vzdialenosti i definuje vzťah 4.20:

$$Ctx_i(t,c) = \begin{cases} 1, & \text{ak } S_i(t,c) > 1 \\ 0, & \text{inak} \end{cases} \tag{4.20}$$

Sila kontextu nadobúda hodnotu 1, ak sú tokeny t a c asociované v kontextovej vzdialenosti i alebo hodnotu 0 vo všetkých ostatných prípadoch. Sila kontextu cez všetky kontextové vzdialenosti je definovaná pomocou vzťahu:

$$Ctx(t,c) = \sum_{i=0}^{n} Ctx_i(t,c), \tag{4.21}$$

kde $Ctx_i(t,c)$ znamená silu kontextu c k tokenu t v kontextovej vzdialenosti i a n označuje počet všetkých možných kontextových vzdialeností. Sila kontextu môže byť použitá na posilenie sily sémantickej príbuznosti. Dvojice tokenov tak budú nadobúdať rôzne hodnoty sémantickej príbuznosti vzhľadom na vopred určený kontext. Silu kontextovo závislej sémantickej príbuznosti popisuje vzťah:

$$Rel^{ctx}(a,b,c) = \gamma Rel^{v}(a,b)Ctx(b,c), \tag{4.22}$$

kde a a b je dvojica tokenov, ktorých príbuznosť je potrebné určiť a c je token definujúci kontext, na ktorom by mali byť závislé. γ je konštanta z intervalu $(0,\infty)$ a slúži na určenie dôležitosti sily kontextu pri výpočte kontextovo závislej sémantickej príbuznosti. Ak je $\gamma = 1/n$, tak $Rel^{ctx}(a,b,c)$ nadobúda hodnoty v intervale $< 0,1 >$. Normalizovaný vzťah pre výpočet sily kontextovo závislej sémantickej príbuznosti je:

$$Rel^{ctx}(a,b,c) = \frac{1}{n} Rel^{v}(a,b) \sum_{i=0}^{n} Ctx_i(b,c) \tag{4.23}$$

Výpočtom príbuzností vybraného symbolu so všetkými symbolmi z lexikónu L dostávame zoradený zoznam príbuzností. Príklad časti takého zoznamu pre token „lord“ v dvoch rôznych kontextoch „heaven“ a „england“ je znázornený v tabuľke 4 – 6.

Tabuľka 4 – 6 Zoznam prvých desiatich najpodobnejších tokenov k tokenu „lord“ v dvoch rôznych kontextoch „heaven“ a „england“.

t_a	Rel^{ctx}(„lord“, t_a, „heaven“)	t_b	Rel^{ctx}(„lord“, t_b, „england“)
lord	1	lord	0.5
god	0.46066879562916996	king	0.453792258009611
earth	0.4188675552004455	ladies	0.423216385498654
dear	0.39101520232349735	summer	0.3824877788606924
heaven	0.39063187380270736	village	0.3617736233482074
high	0.39032927961129305	women	0.36021076069185465
hope	0.36692371115218014	george	0.35408304548178426
does	0.3455632088533151	war	0.352911941938982
help	0.34162238061795236	church	0.3528022937209272
father	0.33563960359705725	years	0.34181020038763477

4.10 Zhlukovanie pojmov na základe ich sémantickej príbuznosti

Vizualizácia symbolu a jemu podobných symbolov na základe ich sémantickej vzdialenosti by mohla vyzerať takto: Výpočtom podobností dotazovaného symbolu k všetkým ostatným symbolom z lexikónu sa dajú vytvoriť zoznamy symbolov zoradené podľa podobnosti k dotazovanému pojmu. Základným problémom pri výpočte podobných symbolov k opýtanému symbolu je určenie hranice v zozname zoradenom podľa podobnosti k vstupnému symbolu. Takýto zoznam síce ilustruje schopnosť AUP identifikovať podobné slová k vstupnému symbolu, ale čelí viacerým problémom. Základné problémy takého zoznamu sú:

- Dĺžka zoznamu - dĺžka zoznamu je vo väčšine prípadov len o málo menšia ako je celková veľkosť použitého lexikónu.
- Chyby v usporiadaní symbolov - často sa v zozname medzi relevantnými symbolmi objavia symboly, ktoré sú z hľadiska empirického poznania jazyka menej

relevantné.

- Nejednoznačnosť významu symbolov - slovo „May“ môže mať význam „mesiac“, ale aj „môcť“. Usporiadanie ostatných slov z lexikónu podľa podobnosti je potom silne závislé na korpuse, nad ktorým AUP akumuloval asociácie.

Tieto problémy sa pokúša riešiť prístup zhlukovania symbolov na základe ich príbuznosti, spomínaný v tejto kapitole. Hľadaný zhluk je potrebné pri zhlukovaní pojmov nejako definovať. Pri výpočte jednoduchého zoznamu zoradeného podľa sémantickej podobnosti bol použitý ako vstup práve jeden symbol. Tento symbol môže patriť do viacerých významových množín. Ako príklad významovej množiny symbolov sa dá chápať množina sesterských synsetov Wordnetu [36]. Kým symbol reprezentovaný slovom „January“ patrí podľa Wordnetu len do jedného synsetu, symbol reprezentovaný slovom „March“ už patrí do štrnástich synsetov (okrem významu „mesiac“ má aj významy ako „pochod“ a iné).

Synsety Wordnetu, v ktorých sa nachádza slovo „January“ sú:

- (n) January, Jan (the first month of the year; begins 10 days after the winter solstice)

Synsety Wordnetu, v ktorých sa nachádza slovo „March“ sú:

- (n) March, Mar (the month following February and preceding April)
- (n) march, marching (the act of marching; walking with regular steps (especially in a procession of some kind)) “it was a long march”; “we heard the sound of marching”
- (n) march, (a steady advance) “the march of science”; “the march of time”
- (n) march, (a procession of people walking together) “the march went up Fifth Avenue”

- (n) borderland, border district, march, marchland (district consisting of the area on either side of a border or boundary of a country or an area) “the Welsh marches between England and Wales”
- (n) marching music, march (genre of music written for marching) “Sousa wrote the best marches”
- (n) Master of Architecture, MArch (a degree granted for the successful completion of advanced study of architecture)
- (v) process (march in a procession) “They processed into the dining room”
- (v) (force to march) “The Japanese marched their prisoners through Manchuria”
- (v) march (walk fast, with regular or measured steps; walk with a stride) “He marched into the classroom and announced the exam”; “The soldiers marched across the border”
- (v) demonstrate, march (march in protest; take part in a demonstration) “Thousands demonstrated against globalization during the meeting of the most powerful economic nations in Seattle”
- (v) parade, exhibit, march (walk ostentatiously) “She parades her new husband around town”
- (v) march (cause to march or go at a marching pace) “They marched the mules into the desert”
- (v) border, adjoin, edge, abut, march, butt, butt against, butt on (lie adjacent to another or share a boundary) “Canada adjoins the U.S.”; “England marches with Scotland”

Vo Wordnete sú synsety vymenované množiny termov, ktoré sú definované rovnako vymenovaním termov, ako aj krátkou všeobecnou definíciou ich významu.

Vzhľadom na fakt, že AUP nemá žiadne znalosti o jazyku, nad ktorým sa učil asociácie, odpadá možnosť popisnej definície zhluku. Kým slovo „March“ sa vyskytuje v mnohých synsetoch, dvojica slov „March“ a „January“ existujú v dvoch sesterských synsetoch (podčiarknutých v zoznamoch vyššie), a to v podstrome synsetu označujúcom mesiace gregoriánskeho kalendára:

- (n) Gregorian calendar month (a month in the Gregorian calendar)

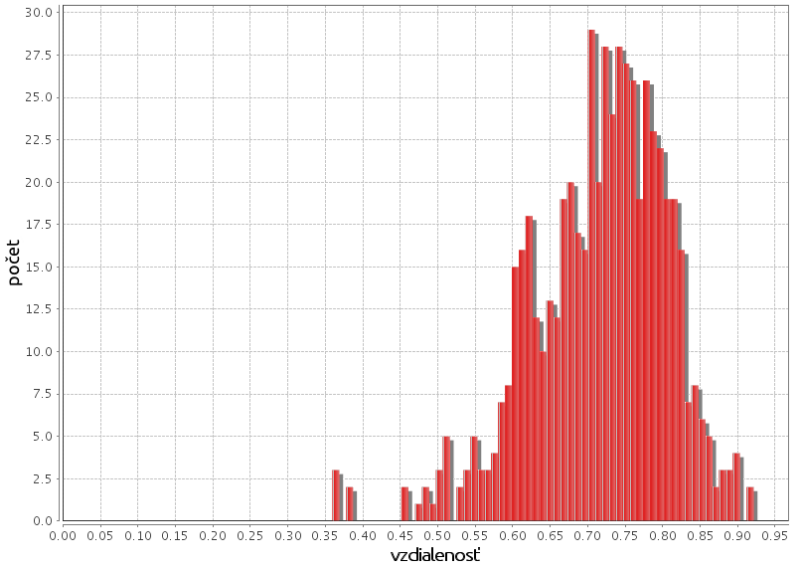
Tieto dva termy teda v rámci Wordnetu môžu poslúžiť ako postačujúce na explicitnú definíciu jedného konkrétneho synsetu, ktorého sú priamymi potomkami. Existuje teda predpoklad, že kombináciou zoznamov kvantifikovaných podobností k dvom alebo viacerým termom by mohlo byť možné automaticky vytvárať množiny konceptov podobných sesterským synsetom vo Wordnete.

Predpokladáme, že pre všetky symboly v rámci jedného synsetu obsahujúceho viac ako dva symboly bude platiť, že si budú približne rovnako podobné. Pre lepšie pochopenie a vizualizáciu problému zavedieme nový pojem sémantickej vzdialenosti dvoch symbolov:

$$D^v(x, y) = 1 - Rel^v(x, y), \tag{4.24}$$

kde $D^v(x, y)$ označuje hodnotu vzdialenosti dvoch symbolov x a y a $Rel^v(x, y)$ označuje ich podobnosť. Každý zoznam termov zoradených podľa podobnosti k dotazovanému termu vieme prepísať pomocou (4.24) ako zoznam termov zoradených podľa vzdialenosti k tomuto termu.

Na obrázkoch 4 – 15 a 4 – 16 sú znázornené distribúcie vzdialeností medzi termami, ktoré sú zároveň podstatnými menami, získané pomocou vzťahu 4.24. Za relevantné podstatné mená sa označili všetky termy, ktoré sa nachádzali aspoň v jednom synsete Wordnetu ako podstatné mená. Učenie prebiehalo nad korpusom o veľkosti 3GB nad článkami z anglickej verzie wikipedie. Ako prah bola zvolená hodnota 1.5. Dáta znázornené na obrázku 4 – 15 pozostávajú zo vzdialeností vypočítavaných medzi jednotlivými podstatnými menami, ktoré sa nachádzajú v rámci jedného synsetu, ktorý obsahoval najmenej dve relevantné podstatné mená. Priemerná hodnota



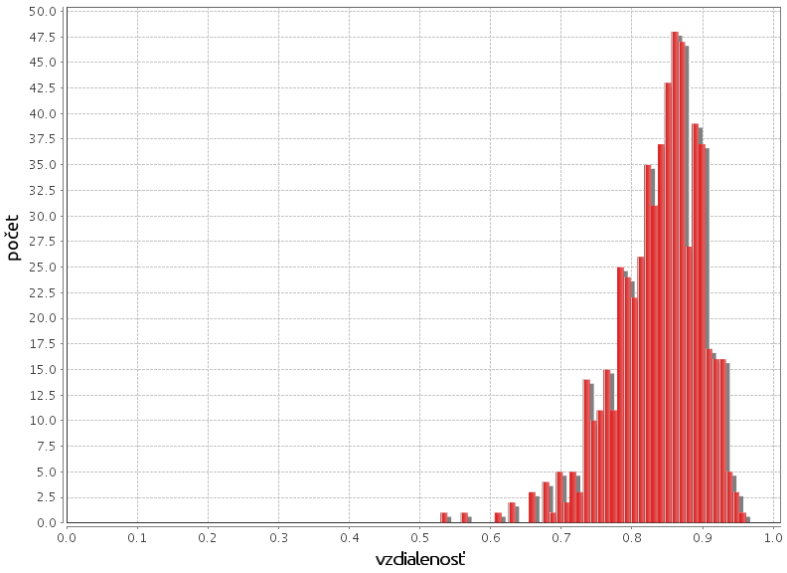
Obr. 4 – 15 Distribúcia vzdialeností vybraných termov v rámci jedného synsetu.

vzdialenosti v tejto konfigurácii bola 0.767. Dáta znázornené na obrázku 4 – 16 pozostávajú zo vzdialeností vypočítavaných medzi rovnakým počtom dvojíc náhodne vybraných podstatných mien z WN. Priemerná hodnota vzdialenosti v tejto konfigurácii bola 0.869. Termy patriace do synsetov boli medzi sebou v tomto prípade o 13% bližšie ako náhodne vybrané termy.

Samotná vzdialenosť od jediného symbolu nie je postačujúcou znalosťou pre vytvorenie zhluku, ktorého prvky by patrili do jednej jasne ohraničenej sémantickej oblasti. Predpokladáme, že všetky termy z jednej ohraničenej sémantickej oblasti budú od seba navzájom približne rovnako vzdialené a táto vzdialenosť by mohla poslúžiť ako fiktívna hranica sémantickej oblasti. Sémantická oblasť je definovaná vymenovaním prvkov:

$$M_c = \{s_1, s_2, \cdots, s_n\}, \tag{4.25}$$

kde M_c je množina označujúca samotnú sémantickú oblasť, symboly $s_1, s_2, \cdots, s_n$ sú jej prvkami a n je počet jej prvkov, teda mohutnosť množiny M_c. Príslušnosť nového symbolu t, nezaradeného v M_c, k tejto sémantickej oblasti je definovaná



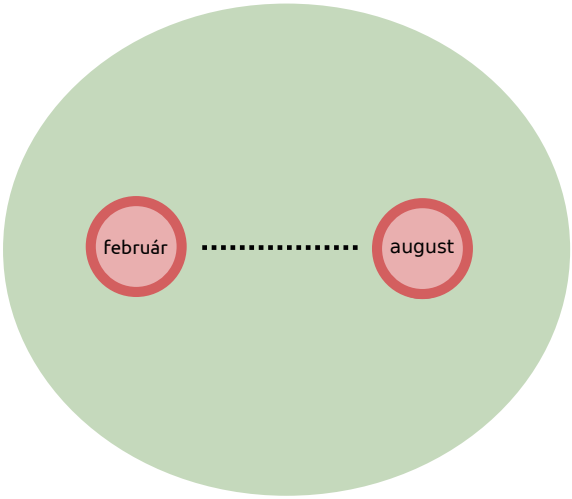
Obr. 4 – 16 Distribúcia vzdialeností náhodne vybraných termov.

podmienkou:

$$\frac{1}{n}\sum_{i}^{n} D_v(s_i, t) \leq \frac{\eta}{C^2\left(n\right)} \sum_{s_i}^{n} \sum_{j=i+1}^{n} D_v(s_i, s_j), \tag{4.26}$$

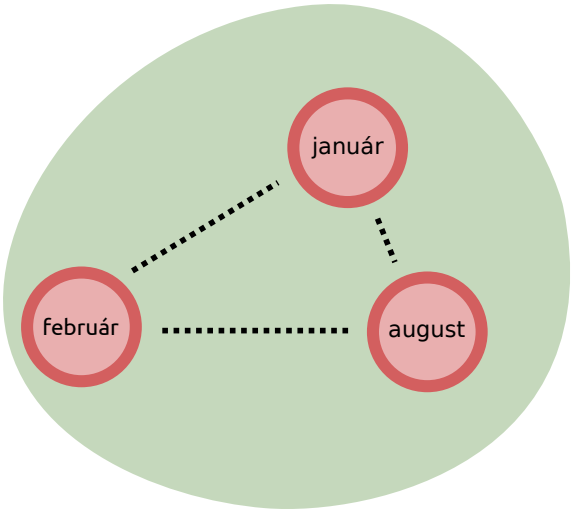
kde s_i a s_j sú prvkami sémantickej oblasti M_c, n je mohutnosťou sémantickej oblasti M_c a $C^2\left(n\right)$ označuje počet dvojprvkových kombinácií všetkých symbolov zo sémantickej oblasti M_c. η označuje parameter, ktorý nadobúda hodnoty z intervalu $(0, \infty)$. Zvyšovaním hodnoty parametra sa zvyšuje návratnosť výslednej množiny na úkor presnosti. Znižovaním hodnoty parametra sa naopak zvyšuje presnosť na úkor návratnosti. Pokiaľ nie je uvedené inak, je $\eta = 1$. Vzťah (4.26) sa dá interpretovať tak, že symbol t patrí do blízkosti sémantickej oblasti M_c vtedy, ak priemerná vzdialenosť termu t od všetkých prvkov sémantickej oblasti M_c je menšia alebo rovná priemernej vzdialenosti medzi všetkými kombináciami dvojíc z danej sémantickej oblasti.

Ak by sme sa pokúsili uložiť dva tokeny, ktoré by boli zároveň prvkami jedinej samostatnej sémantickej oblasti (synsetu, ktorý je priamym rodičom sesterských synsetov), do dvojrozmerného priestoru, vizualizácia možných pozícií spĺňajúcich podmienku 4.26 by mohla vyzerať takto (zelená oblasť):



Obr. 4 – 17 Oblasť príslušnosti do sémantickej oblasti o veľkosti dvoch symbolov.

Každý ďalší token, ktorý by sa dal do zelenej oblasti uložiť tak, aby vzdialenosti medzi tokenmi zodpovedali ich skutočným hodnotám teda spĺňuje podmienku príslušnosti. Pri kombinácii troch symbolov by mohla výsledná vizualizácia vyzerať takto:



Obr. 4 – 18 Oblasť príslušnosti do sémantickej oblasti o veľkosti troch symbolov.

Aj napriek tomu, že tieto vizualizácie slúžia na ilustráciu pre lepšie pochopenie podmienky príslušnosti, boli vytvorené počítačovým programom tak, aby skutočne zodpovedali vzťahu 4.26.

Tabuľka 4 – 7 Výstup zhlukovania symbolov na základe vstupnej trojice troch náhodných mesiacov.

vstup	january, june, november
výstup	april, march, august, february, september, may, june, november, december, july, january, october

Tabuľka 4 – 8 Výstup zhlukovania symbolov na základe vstupnej trojice troch náhodných dní.

vstup	tuesday, wednesday, sunday
výstup	saturday, thursday, monday, sunday, tuesday, wednesday

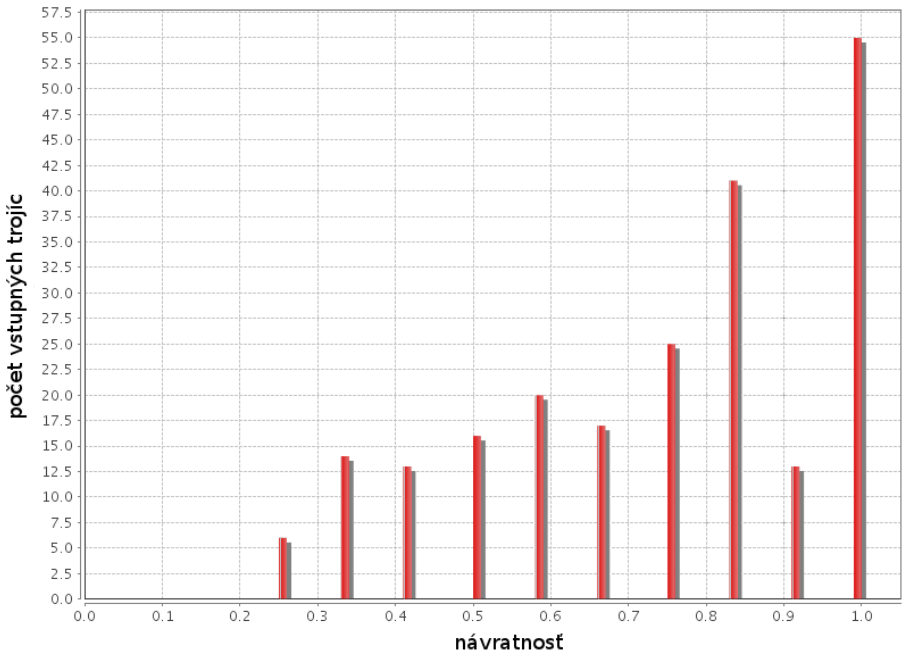
Tabuľka 4 – 9 Výstup zhlukovania symbolov na základe vstupnej trojice troch náhodných číslovieok.

vstup	three, fifty, twenty
výstup	twelve, forty, four, three, twenty, ten, two, seven, five, fifty, fifteen, six, thirty

4.10.1 Ukážka prístupu na sémanticky silne ohraničených množinách

Pre názornú ukážku tohto prístupu je vhodné vybrať množiny, u ktorých je príslušnosť do danej skupiny jednoznačná a zároveň sa jedná o slová, ktoré by mali možnosť dostať sa do ohraničeného lexikónu najpoužívanejších slov. Ako príklad môže poslúžiť dvanásť mesiacov v roku, sedem dní v týždni alebo číslovky. Systému sa zadá na vstup náhodná trojica vybraná z týchto množín a na výstupe sa očakáva správne doplnenie o ďalšie (v ideálnom prípade všetky) prvky z danej množiny. Na akumuláciu asociácií bol použitý korpus textov v prirodzenom jazyku získaný z náhodnej podmnožiny článkov projektu Wikipedia o celkovej veľkosti 3 GB. Ako lexikón bolo použitých 6000 najpoužívanejších slov z rovnakého korpusu. Asociácie boli získavané v štyroch kontextových vzialenostiach (s hodnotami 1, 2, 3, 4) s hodnotou prahu 2 a pri výpočte podobnosti mali všetky váhy hodnotu 0.25. Výsledky pre dané vstupy možno vidieť v tabuľkách 4 – 7, 4 – 8 a 4 – 9.

V ďalšom prípade bolo zo slov reprezentujúcich dvanásť mesiacov v roku vy-



Obr. 4 – 19 Histogram hodnôt návratnosti pre všetky kombinácie vstupných trojíc symbolov.

tvorených všetkých 220 kombinácií trojíc, ktoré poslúžili ako vstup do rovnakého algoritmu. Pre každú z týchto trojíc bola znova vypočítaná príslušnosť ostatných symbolov z lexikónu do blízkosti sémantickej oblasti, ktorú definovali. Ani pri jednej z 220 trojíc nebol žiaden výsledný symbol slovo, ktoré by neoznačovalo mesiac v roku. Úplná množina 12 mesiacov bola nájdená v 25% prípadov. V ostatných prípadoch sa výstup pohyboval od troch (zhodných so vstupnou trojicou) do jedenástich mesiacov. Histogram hodnôt návratnosti prístupu je znázornený na grafe 4 – 19, kde na osi y je znázornený počet vstupných trojíc a na osi x je znázornená im príslúchajúca hodnota návratnosti (úplnosti hľadanej oblasti). Priemerná návratnosť pre všetkých 220 trojíc mala hodnotu 0.738.

5 Beast - implementácia

Beast je implementáciou AUP s rôznymi obmedzeniami v programovacom jazyku Java. Vzhľadom na ciele tejto práce bude základná architektúra a funkcie tejto implementácie uvedená tak, aby k ich pochopeniu neboli potrebné žiadne hlbšie znalosti o konkrétnom programovacom jazyku. Cieľom tejto kapitoly je oboznámenie s fungovaním systému, ktorý bol použitý pri experimentoch (kapitola 6) a so sústredením sa na implementačné prvky, ktoré neboli v kapitole venujúcej sa AUP popísané. Program je poskytovaný ako open-source aplikácia a niektoré fasciklové matice použité v experimentoch (ktorým je venovaná nasledujúca kapitola) sú tiež voľne prístupné. Zverejnenie zdrojového kódu poskytuje lepšiu transparentnosť celého výskumu, možnosť opakovateľnosti experimentov v rôznych prostrediach ako aj ľahšie možnosti odhalenia chýb a nedostatkov. Program je možné si voľne stiahnuť z internetu[29]. Aj keď boli experimenty implementované v podobe JUnit testov, program obsahuje použiteľné grafické prostredie, ktoré je zobrazené na obr. 5 – 1.

Nasledujúce podkapitoly sú venované popisu architektúry systému so základným popisom jej funkčných blokov.

5.1 Token

Token je v modeli pre širšie použitie v budúcnosti reprezentovaný ako objekt zaobaľujúci hodnotu ľubovoľného typu (triedy). Hodnotou tokenu môže byť inštancia ľubovoľnej triedy v jazyku Java. To dáva celému systému možnosť jednoduchej rozšíriteľnosti na inú doménu ako doménu textov. Pre doménu prirodzeného jazyka sme použili na reprezentáciu symbolu triedu Token zaobaľujúci hodnotu triedy String, ktorou sú reprezentované jednotlivé slová. Výnimku tvorí iba tzv. NULL_TOKEN, ktorý označuje token neobsiahnutý v lexikóne. Špeciálnou podtriedou Tokenu, je ExcitedToken, ktorý reprezentuje asociovaný token, a zaobaľuje okrem jeho logickej hodnoty aj váhu, s ktorou bol, resp. je v danej úlohe asociovaný. Tieto typy hod-

[29]Dostupné online (13.2.2012): http://dendrit.fei.tuke.sk/ rockai/index.php?content=school

Obr. 5 – 1 Grafické užívateľské prostredie programu Beast - pôvodnej implementácie AUP

nôt tokenov slúžia iba na lepšiu čitateľnosť pre užívateľa. Vnútorne systém používa na reprezentáciu hodnoty Tokenu celé čísla, indexy daných tokenov v lexikóne.

5.2 Lexikón

Lexikón pozostáva z voliteľne rozsiahlej množiny tokenov a k nim prislúchajúcim celočíselným ID kľúčov. Každému je priradený osobitný a jedinečný ID kľúč. Veľkosť lexikónu musí byť zadaná ešte pred procesom učenia, lebo lexikón rovnako ako typ informácie ukladanej v tokenoch, zostáva počas celého života (učenia a používania) programu nemenný. Tokeny sú v lexikóne uložené v podobe asociatívnej

mapy, v ktorej je každému tokenu (kľúču v danej mape) priradená hodnota frekvencie výskytu daného tokenu v korpuse, nad ktorým sa systém učil. Táto hodnota slúži na zoradenie tokenov podľa ich početnosti, vzhľadom na ďalšie orezanie lexikónu, kde sa do orezaného lexikónu dostane voliteľné množstvo najpočetnejších tokenov. Lexikón musí vždy obsahovať NULL_TOKEN, aby bolo možné priradiť index neznámym slovám. Pri experimentoch bola použitá veľkosť lexikónu 6000 tokenov, ale táto hodnota je voliteľná.

5.3 TokenStream a TokenWindow

TokenStream je trieda, ktorou je kódovaný vstup do učenia systému. Jedná sa o prúd tokenov z textového dokumentu, akúsi pásku, ktorá je čítaná čítacou hlavou o šírke piatich tokenov. Každý token, ktorý sa nenachádza v lexikóne, vrátane interpunkčných znamienok, sa nahradí znakom „#“. Kým hlava stojí nehybne, páska sa môže posunúť o jedno políčko doľava. Proces posunutia pásky smerom doľava nazývame „posunom okna“, pretože je analogický posunu čítacej hlavy smerom doprava. TokenWindow je okno piatich usporiadaných tokenov prečítaných čítacou hlavičkou, poprípade okno piatich usporiadaných tokenov získaných inou cestou (v prípade, že sú zadané manuálne a nie čítané z dokumentu).

5.4 Fascikle

Fascikel je základný objekt obsahujúci znalosti systému. Je implementovaný ako matica N x N tokenov obsahujúca počet spoločných výskytov týchto dvojíc tokenov, vo vzdialenosti danej fasciklom, kde N je veľkosť lexikónu. Znalosti sa uchovávajú v štyroch fascikloch označených veľkými písmenami A, B, C, D. Písmená označujú vzdialenosť medzi tokenmi, ktorých spoločné výskyty si fascikel uchováva a to tak, že A značí vzdialenosť jedného tokenu, B dvoch atď. Všetky fascikle sú uložené v tzv. regióne, ktorý tvorí základnú triedu určenú na uchovávanie znalostí. Pri tomto spôsobe zápisu je možné z jedného fasciklu získať početnosti spoločného výskytu

pre dvojicu tokenov v oboch smeroch (kontextovej vzdialenosti).

Fascikel je reprezentovaný ako objekt obaľujúci dvojrozmerné pole, teda maticu celých čísel. Tento objekt môže byť vďaka využitiu rozhraní aj pripojenie na databázu alebo akýkoľvek iný dátový priestor. Celočíselné matice boli zvolené z dôvodu rýchlosti prístupu k prvkom a jednoduchšej predvídateľnosti pamäťovej náročnosti.

Okrem samotných početností spoločných výskytov táto trieda obsahuje aj pravdepodobnosti samostatných výskytov pre všetky tokeny v kontextovej vzdialenosti, ktorú reprezentuje a základné metódy na výpočet signifikancie a váhy medzi dvoma tokenmi v tejto vzdialenosti.

Fascikle sú serializovateľné objekty a je možné ich po naučení uložiť, resp. pred použitím nahrať.

5.5 Proces učenia

Proces učenia je procesom uchovávania znalostí AUP. V tejto implementácii sa jedná iba o štatistické počty nad textom. Vstupom do tohto procesu je jeden alebo viac textových dokumentov vo formáte „plain/text“. Ten je pomocou balíka LuceneTokenizer od Apachu prevedený na tokenStream.

- 1. Vyber prvé okno z tokenStreamu.
- 2. Otestuj prípustnosť okna.
- 3. Ak je okno neprípustné, posuň okno o jedno políčko, pokračuj krokom 2.
- 4. Pridaj spoločné výskyty tokenov cez príslušné fascikle do regiónu.
- 5. Ak tokenStream pokračuje, posuň okno a pokračuj krokom 2.

TokenStream je prevedený na okno piatich tokenov, ktoré je neskôr indexované za pomoci lexikónu na okno piatich indexov. To vstupuje do metódy kortexu PresentLearningWindow.

5.6 PresentLearningWindow

Funkcia kontroluje prípustnosť okna a ak je okno prípustné, pridá spoločné výskyty tokenov cez príslušné fascikle do regiónu. Okno je prípustné práve vtedy, keď sa jeho posledný token nerovná „#“, to znamená, má svoj index (odlišný od indexu NULL_TOKENU) v lexikóne. Učenie – pridávanie spoločných výskytov do regiónu, potom prebieha sprava doľava po najbližší neznámy (NULL_TOKEN) token alebo v prípade všetkých známych tokenov po začiatok okna.

Príklady prípustných a neprípustných okien:

- a b c d e : okno je prípustné, učíme teda asociácie (a e), (b e), (c e), (d e)
- # a b c d : prípustné je pod-okno a b c d, učíme (a d), (b d), (c d)
- # # b d # : okno nie je prípustné, posledný token je NULL_TOKEN.
- a # # a b : prípustné pod-okno je (a b) učíme (a b)
- # # a b c : prípustné pod-okno (a b c), učíme (a c), (b c)

Pravdepodobnostné funkcie sú v našej implementácii pre prácu s korpusom definované nasledovne:

$$p(x, y) = f(x, y)/c_2 \tag{5.1}$$

$$p(x) = f_L(x)/c_L \tag{5.2}$$

$$p(y) = f_R(y)/c_R, \tag{5.3}$$

kde $f(x, y)$ je počet spoločných výskytov dvoch symbolov (pri textovej reprezentácií napríklad slov) x a y z lexikonu L v korpuse, kde symbol x sa nachádzal v texte naľavo od symbolu y. c_2 je počet všetkých dvojíc všetkých známych symbolov v korpuse. $f_L(x)$ je počet všetkých výskytov symbolu x, kde sa nachádzal naľavo od akéhokoľvek iného známeho symbolu a naopak $f_R(y)$ je počet všetkých výskytov symbolu j, kde sa nachádzal napravo od akéhokoľvek iného známeho symbolu. c_L a

c_R sú hodnoty ktoré označujú celkový počet výskytov symbolov naľavo (c_L) alebo napravo (c_R).

5.7 Cortex

Cortex je trieda obaľujúca základnú funkcionalitu aplikácie Beast. Obsahuje funkcie, ktoré implementujú výpočet konfabulácie. Rovnako obsahuje funkciu PresentLearningWindow, prístup ku všetkým fasciklom a všetky metódy potrebné pre učenie. Od triedy Cortex dedia ostatné funkčné bloky:

- **SynonymCortex** – implementuje funkcionalitu z kapitoly 4.8.
- **SimilarityCortex** - implementuje funkcionalitu z kapitoly 4.5.
- **ContextCortex** - implementuje funkcionalitu z kapitoly 4.9.
- **ClusterCortex** - implementuje funkcionalitu z kapitoly 4.10.

6 Experimenty

Experimenty prezentujúce funkčnosť AUP v úlohe dolovania konceptov prebiehali nad asociáciami akumulovanými nad viacerými odlišnými korpusmi. Korpusy tvorili dlhé texty v prirodzenom jazyku a líšili sa navzájom svojou veľkosťou, zdrojom a jazykom, v ktorom boli texty napísané. Zoznam korpusov je spolu s ich základnými vlastnosťami uvedený v tabuľke 6 – 1. Ich zdroje, Projekt Gutenberg[30], Wikipedia EN[31], Wikipedia SK[32] a Wikipedia CS[33] sú dostupné on-line.

Tabuľka 6 – 1 Zoznam korpusov a ich vlastností použitých v experimentoch.

názov	zdroj	veľkosť	jazyk	predspracovanie
COREN1K	Projekt Gutenberg	1000MB	anglický	-
COREN3K	Wikipedia EN	3000MB	anglický	WP2TXT
CORSK121	Wikipedia SK	121MB	slovenský	WP2TXT
CORCZ370	Wikipedia CS	370MB	český	WP2TXT

Táto kapitola sa na korpusy odkazuje podľa ich názvu uvedeného v danej tabuľke. Pokiaľ nie je uvedené inak, tvorilo lexikón 6000 najpoužívanejších slov z použitého korpusu. Asociácie boli akumulované v ôsmich kontextových vzdialenostiach o veľkostiach $-4, -3, -2, -1$ a $1, 2, 3, 4$ s váhami o veľkosti 1, veľkosťou prahu 1.5 a parameter η mal hodnotu 1. Experimenty prebiehali na implementácii AUP nazvanej Beast, popísanej v predchádzajúcej kapitole.

Korpus získaný z projektu Gutenberg tvorili náhodne vybrané publikácie v anglickom jazyku v *plain/textovej* forme a neprechádzali žiadnym predspracovaním. Korpusy z projektu Wikipedia vznikli spracovaním poslednej databázovej zálohy článkov k dátumu 11.9.2011. Program WP2TXT[34] bol použitý na konverziu databázovej zálohy do *plain/textového* formátu tak, že zachovával iba nadpisy a telá

[30]Dostupné online (13.2.2012): http://www.gutenberg.org/
[31]Dostupné online (13.2.2012): http://en.wikipedia.org/
[32]Dostupné online (13.2.2012): http://sk.wikipedia.org/
[33]Dostupné online (13.2.2012): http://cs.wikipedia.org/
[34]Dostupné online (13.2.2012): http://wp2txt.rubyforge.org/

Tabuľka 6 – 2 Naplnenosť matíc fasciklov pre jednotlivé korpusy.

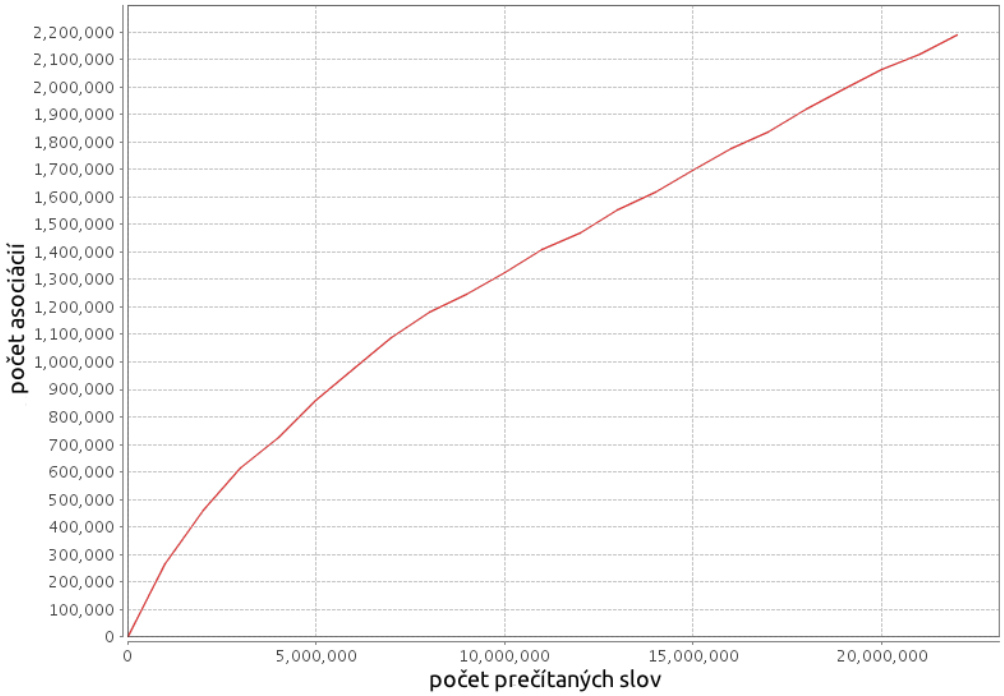
	naplnenie			
názov	f_1	f_2	f_3	f_4
COREN1K	3.75%	5.05%	5.14%	4.28%
COREN3K	25.92%	39.08%	42.46%	37.79%
CORSK121	3.05%	3.43%	2.47%	1.56%
CORCZ370	6.50%	7.89%	6.03%	4.02%

Tabuľka 6 – 3 Počet asociácií v jednotlivých fascikloch pre jednotlivé korpusy.

	počet asociácií				
názov	f_1	f_2	f_3	f_4	Suma
COREN1K	72×10^4	100×10^4	104×10^4	88×10^4	364×10^4
COREN3K	336×10^4	583×10^4	714×10^4	665×10^4	2298×10^4
CORSK121	66×10^4	76×10^4	55×10^4	35×10^4	232×10^4
CORCZ370	133×10^4	167×10^4	131×10^4	89×10^4	520×10^4

článkov. Ukážky korpusov, ktoré boli použité v experimentoch, tvoria prílohu tejto práce.

Sledovaním rastu asociácií sa dá odhadnúť nakoľko bola kvantita textu dostatočná pre zachytenie asociácií (v zmysle AUP) jazyka, v ktorom sú texty písané. Čím nižšia je hodnota zmeny rastu asociácií (rozdiel počtu asociácií) v závislosti na počte prečítaných tokenov v bode, kedy sa učenie zastavilo, tým väčšia je pravdepodobnosť, že počet získaných asociácií v texte zodpovedá počtu asociácií jazyka, v ktorom je písaný. Rast asociácii pre jednotlivé korpusy v procese učenia je obsahom grafov 6 – 1, 6 – 2, 6 – 3 a 6 – 4. Tabuľky 6 – 2 a 6 – 3 sú uvedené pre predstavu závislosti naučených znalostí od veľkosti korpusu a od jazyka, v akom bol korpus písaný. Tabuľka 6 – 2 popisuje naplnenie (teda podiel počtu asociovaných dvojíc tokenov, k počtu všetkých možných dvojíc tokenov) jednotlivých fasciklov $f_1 - f_4$. Tabuľka 6 – 2 popisuje počet asociovaných dvojíc tokenov v jednotlivých fascikloch.



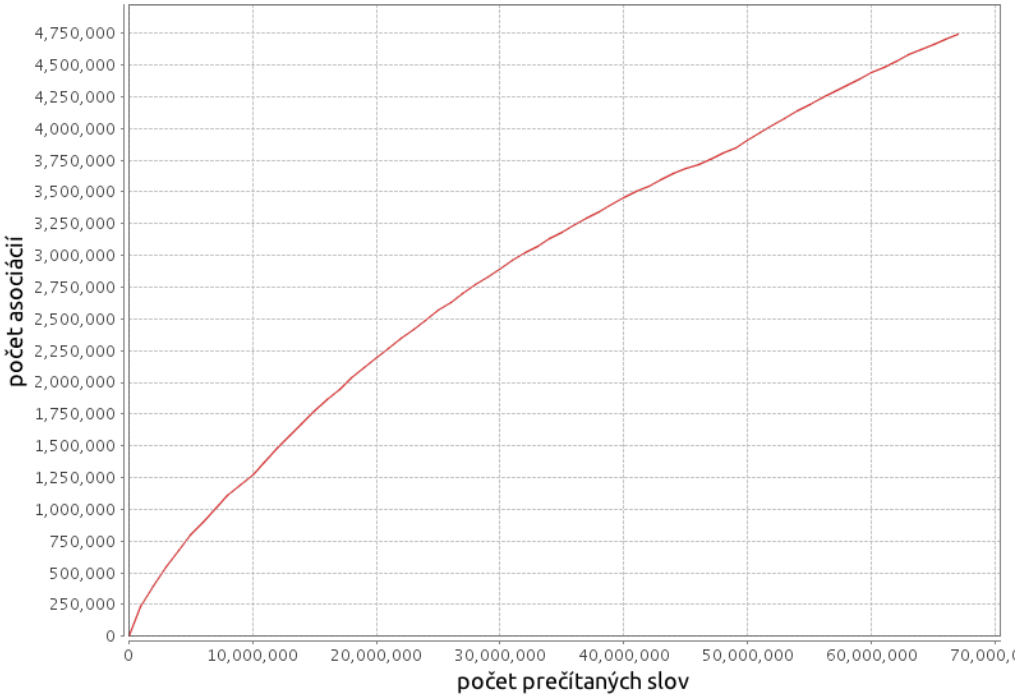
Obr. 6 – 1 Rast asociácií počas učenia korpusu CORSK121

Z grafov je zrejmé, že z daných korpusov boli asociácie jazyka metódou AUP najlepšie získané z korpusu COREN3K. V korpuse 6 – 3 počas učenia, približne medzi 50×10^6 a 120×10^6 prečítanými tokenmi, počet asociácií nerástol. To bolo spôsobené tým, že sa do korpusu nedopatrením dostalo aj pár publikácií vo francúzskom jazyku. Táto chyba nastala pravdepodobne nesprávnym označením francúzskeho textu za anglický v rámci projektu Gutenberg. Aj napriek tomu je korpus v experimentoch použitý z dôvodu otestovania metód na jazykovo nehomogénnych korpusoch.

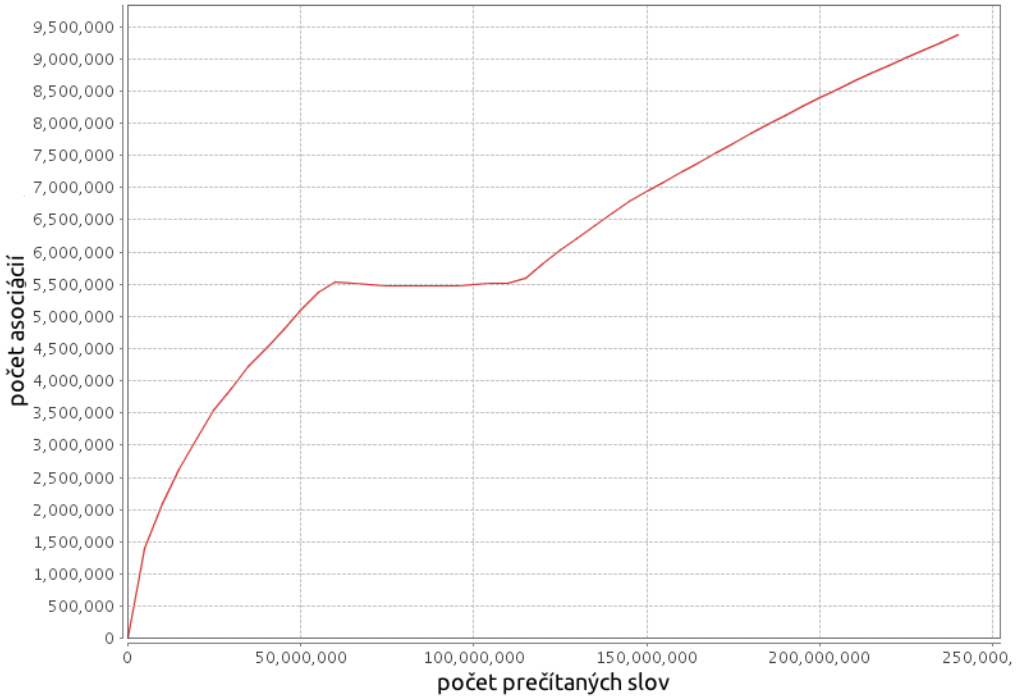
6.1 Počítanie podobných slov

Cieľom experimentu je overenie metódy výpočtu podobných slov, ktorej je venovaná kapitola 4.5.1.

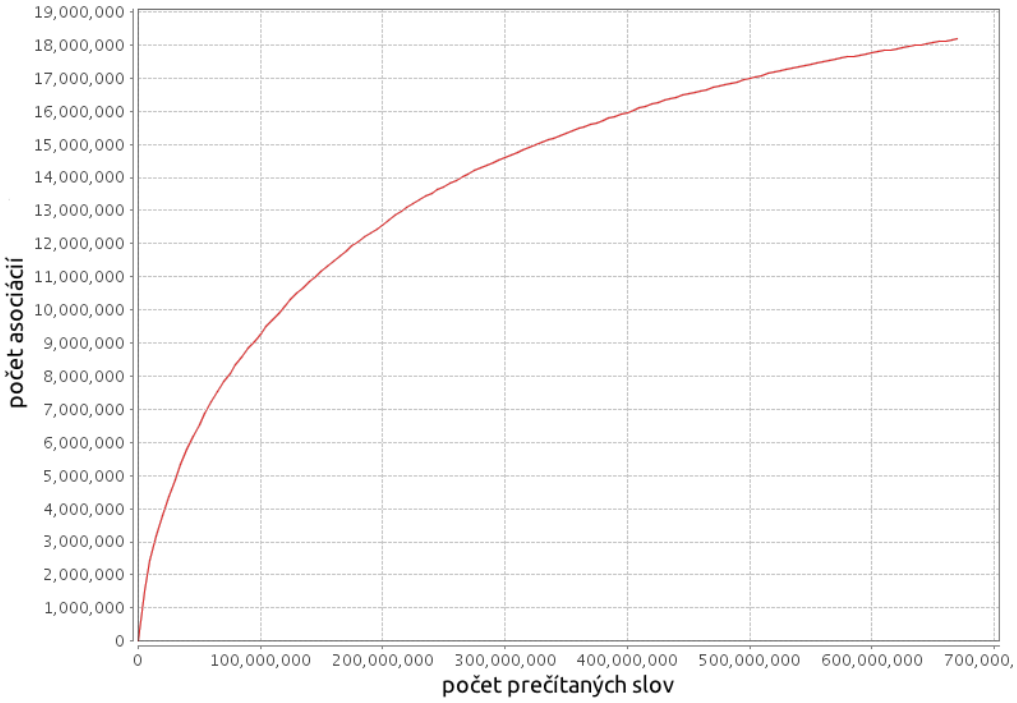
Popis: Nad každým korpusom z tabuľky 6 – 1 bola vypočítaná podobnosť troch vybraných slov (resp. ich prekladu, pokiaľ sa jednalo o odlišné jazyky) k všetkým



Obr. 6 – 2 Rast asociácií počas učenia korpusu CORCZ370



Obr. 6 – 3 Rast asociácií počas učenia korpusu COREN1K



Obr. 6 – 4 Rast asociácií počas učenia korpusu COREN3K

ostatným slovám z lexikónu: „je“, „prezident“, „Alexander“. Podobné slová boli potom zoradené podľa podobnosti. Nad každým korpusom boli teda vypočítané tri zoznamy podobných slov k trom rôznym vstupným slovám. Očakávaným výsledkom bolo, že podobné slová budú spadať do rovnakej sémantickej triedy ako ich vstupy, a teda pomocné slovesá pre slovo „je“, podstatné mená označujúce povolanie alebo titul pre slovo „prezident“ a mená ľudí pre slovo „Alexander“.

Výsledky: Tabuľky 6 – 4, 6 – 5, 6 – 6 a 6 – 7 znázorňujú prvých 20 najpodobnejších slov pre jednotlivé korpusy spolu s hodnotou ich podobností k vybraným slovám. Ak bolo v akomkoľvek jazyku zadané na vstup krstné meno „Alexander“, systém k nemu označil za najpodobnejšie opäť krstné mená. Ak bolo na vstup zadané slovo „prezident“, systém za najrelevantnejšie slová k nemu označil tie, ktoré vo väčšine prípadov označujú nejakú vedúcu pozíciu alebo titul. V prípade slova „je“ AUP za najrelevantnejšie označil jeho rôzne podoby a iné

pomocné slovesá. Vzhľadom na fakt, že sa jedná o veľmi frekventované slovo, teda asociované s obrovským množstvom iných slov, obsahuje výsledok v tomto prípade aj veľa chýb, ktoré by mali byť odstrániteľné zvýšením prahu (a v tom prípade zredukovaním počtu asociácií iba na tie významnejšie). Nájdené slová boli významovo veľmi úzko prepojené k zadaným pojmom. Slová, pri ktorých sa predpokladalo, že budú patriť do výslednej množiny slov sú v tabuľkách podčiarknuté.

Záver: Daný experiment pomohol overiť, že metóda výpočtu podobných slov popísaná v kapitole 4.5.1 dáva relevantné výsledky pre vstupy nielen v anglickom jazyku, ale aj v jazyku slovenskom a českom. Dá sa preto predpokladať, že tento prístup je nezávislý od jazyka a mohol by dosahovať rovnako dobré výsledky aj v iných jazykoch.

Tabuľka 6 – 4 Zoznam najpodobnejších slov k slovám „je“, „prezident“ a „alexander“ z korpusu CORSK121

slovo	podobnosť	slovo	podobnosť	slovo	podobnosť
je	1.00	prezident	1.00	alexander	1.00
bola	0.18	prezidentom	0.12	ľudovít	0.12
bol	0.16	predseda	0.10	karol	0.11
má	0.15	minister	0.10	filip	0.11
sú	0.15	kráľ	0.09	henrich	0.11
alebo	0.14	pápež	0.09	ferdinand	0.11
sa	0.13	prezidenta	0.09	štefan	0.11
bolo	0.13	člen	0.09	ján	0.11
ako	0.13	cisár	0.09	peter	0.10
a	0.13	generál	0.09	eduard	0.10
môže	0.12	karol	0.08	richard	0.10
nie	0.12	predsedom	0.08	františek	0.10
to	0.12	veliteľ	0.07	jozef	0.10
byť	0.12	biskup	0.07	michael	0.10
pre	0.11	panovník	0.07	rudolf	0.09
však	0.10	architekt	0.07	cisár	0.09
nachádza	0.10	súd	0.07	juraj	0.09
s	0.10	sám	0.07	caesar	0.09
aj	0.10	parlament	0.07	kráľ	0.09
boli	0.10	tím	0.07	ivan	0.09

Tabuľka 6 – 5 Zoznam najpodobnejších slov k slovám „je“, „prezident“ a „alexandr“ z korpusu CORCZ370

slovo	podobnosť	slovo	podobnosť	slovo	podobnosť
je	4.00	prezident	1.00	alexandr	1.00
jsou	0.97	předseda	0.17	jindřich	0.14
má	0.89	ministr	0.13	ludvík	0.14
byla	0.88	prezidentem	0.12	fridrich	0.13
není	0.85	král	0.12	eduard	0.13
byl	0.78	císař	0.12	ferdinand	0.13
může	0.76	prezidenta	0.11	vilém	0.13
nebo	0.74	člen	0.11	filip	0.13
bylo	0.69	velitel	0.11	konstantin	0.13
patří	0.67	ředitel	0.10	richard	0.12
být	0.67	předsedou	0.10	rudolf	0.12
se	0.65	premiér	0.10	napoleon	0.12
či	0.63	profesor	0.10	petr	0.12
jako	0.63	václav	0.10	karel	0.12
tvoří	0.62	kapitán	0.10	robert	0.12
bývá	0.62	vůdce	0.10	václav	0.11
pro	0.62	generál	0.10	kníže	0.11
bude	0.62	soud	0.10	císař	0.11
a	0.61	papež	0.09	františek	0.11
byly	0.60	starosta	0.09	jan	0.11

Tabuľka 6 – 6 Zoznam najpodobnejších slov k slovám „is“, „president“ a „alexander“ z korpusu COREN1K

slovo	podobnosť	slovo	podobnosť	slovo	podobnosť
is	1.00	president	1.00	alexander	1.00
was	0.28	proprietor	0.09	cliges	0.10
are	0.22	secretary	0.08	anthony	0.09
has	0.22	bishop	0.08	goethe	0.09
were	0.19	ruler	0.08	papa	0.09
be	0.18	governor	0.08	joseph	0.09
had	0.18	leaders	0.08	gloria	0.09
which	0.17	editor	0.08	adam	0.08
been	0.16	painter	0.08	charley	0.08
would	0.16	agent	0.08	leopold	0.08
being	0.16	member	0.08	olof	0.08
of	0.16	housekeeper	0.08	guy	0.08
have	0.15	philosopher	0.07	socrates	0.08
will	0.15	marquis	0.07	sophy	0.08
am	0.14	judges	0.07	kagig	0.08
or	0.14	scholar	0.07	anne	0.08
by	0.14	magician	0.07	aladdin	0.08
so	0.14	preacher	0.07	ellen	0.08
who	0.14	dean	0.07	ernest	0.08
in	0.14	revolution	0.07	jones	0.08

Tabuľka 6 – 7 Zoznam najpodobnejších slov k slovám „is“, „president“ a „alexander“ z korpusu COREN3K

slovo	podobnosť	slovo	podobnosť	slovo	podobnosť
is	1.00	president	1.00	alexander	1.00
was	0.60	chairman	0.36	james	0.44
has	0.48	governor	0.35	thomas	0.43
are	0.40	director	0.32	henry	0.43
being	0.37	mayor	0.32	charles	0.42
becomes	0.36	leader	0.30	martin	0.40
remains	0.35	minister	0.29	edward	0.40
were	0.35	senator	0.29	william	0.40
be	0.31	professor	0.28	joseph	0.39
had	0.31	founder	0.28	richard	0.39
itself	0.29	ceo	0.27	paul	0.39
been	0.29	commissioner	0.27	george	0.39
which	0.29	lieutenant	0.27	francis	0.39
gets	0.27	secretary	0.27	david	0.38
does	0.27	colonel	0.26	peter	0.38
seems	0.27	politician	0.26	philip	0.38
but	0.27	son	0.26	scott	0.37
and	0.27	commander	0.26	robert	0.37
initially	0.26	historian	0.25	john	0.37
instead	0.26	attorney	0.25	lee	0.37

6.2 Počítanie podobných slov v kontexte

Experimenty počítania podobných slov v kontexte boli v minulosti publikované v [53] a neboli prevádzané nad korpusmi popísanými v tejto kapitole. Korpus použitý v týchto experimentoch bol analogický korpusu COREN1K. Líšil sa iba veľkosťou korpusu, ktorý tvoril v tomto prípade 1GB textových dokumentov a niektoré (náhodne vybrané) texty boli inou podmnožinou textov z projektu Gutenberg. Váhy použité v experimente sú popísané v tabuľke 6 – 8. Boli zvolené tak, aby boli pri výpočtoch uprednostnené kratšie kontextové vzdialenosti. Zvyšné vlastnosti systému a korpusu ostávajú v tomto experimente zachované ako v predchádzajúcich prípadoch.

Tabuľka 6 – 8 Váhy použité pri výpočte kontextu

Kontextová vzdialenosť (i)	±1	±2	±3	±4
Váha (v_i)	1	0.25	0.125	0.065

Cieľom experimentu je overenie metódy výpočtu podobných slov v kontexte popísanej v kapitole 4.9.1.

Popis: Pre 6 slov bola vypočítaná podobnosť vybraného slova v závislosti na dvoch odlišných kontextoch ku všetkým ostatným slovám z lexikónu. Rovnako bola vypočítaná hodnota ich bezkontextovej podobnosti ku všetkým slovám z lexikónu (viď. predchádzajúci experiment). Za očakávaný výsledok sa pokladá, ak najvyššie hodnoty podobnosti budú dosahovať slová z jednej sémantickej triedy (viď podkapitolu 6.1), ktoré zároveň majú v skutočnom svete významové napojenie na kontextové slovo.

Výsledky: Výpočet podobných slov v kontexte je demonštrovaný na šiestich príkladoch v tabuľkách 6 – 9, 6 – 10, 6 – 11, 6 – 12, 6 – 13 a 6 – 14. Každá z tabuliek obsahuje 9 najpodobnejších slov k vybranému slovu (vybranému na základe empirických znalostí jazyka) v troch rôznych prípadoch. V prvom

prípade boli vygenerované najpodobnejšie slová bez akéhokoľvek kontextu rovnako ako v predošlom experimente. V ďalších dvoch prípadoch boli vygenerované podobné slová v závislosti od kontextu, ktorý bol opäť vybraný na základe empirických znalostí jazyka tak, aby sa pomocou neho dali jasne rozlíšiť jednotlivé významy.

Záver: Daný experiment pomohol overiť, že metóda výpočtu podobných slov popísaná v kapitole 4.9.1 zohľadňuje pri výpočte podobných slov kontext. Aj napriek tomu, že subjektívne možno výsledky vnímať ako pozitívne (napr. uprednostnenie tokenov „god“, „jesus“ alebo „father“ v kontexte „heaven“, a uprednostnenie tokenov „george“, „henry“ alebo „charles“ v kontexte „england“), výsledky vo všetkých štyroch prípadoch obsahujú príliš mnoho irelevantných tokenov.

Príklady ilustrujú schopnosť AUP posilniť silu relevantných konceptov na základe určenia slova, ktoré udáva kontext. V tabuľkách sú správne posilnené slová zvýraznené podčiarknutím. Základným problémom tohto prístupu je posilnenie veľkej množiny nerelevantných konceptov len na základe ich silného prepojenia s kontextovým slovom. Tento prístup teda nie je, na rozdiel od jednoduchého výpočtu podobnosti slov, univerzálne funkčný a mal by slúžiť na rozlíšenie príslušnosti vybraných slov, získaných napríklad z výpočtu jednoduchej pravdepodobnosti, do jednotlivých kontextových podskupín.

Tabuľka 6 – 9 Hľadanie podobných slov pre slovo „lord“ v kontextoch „England“ a „god“

Slovo	Sila	Kontext	Sila	Kontext	Sila
-		england		god	
lord	1.44	lord	4.31	lord	4.31
god	0.35	king	1.01	god	1.05
king	0.34	george	0.94	father	0.99
father	0.33	poet	0.9	mother	0.89
brother	0.33	son	0.87	creature	0.8
captain	0.32	gentleman	0.84	heaven	0.79
lady	0.32	henry	0.81	christ	0.78
uncle	0.32	england	0.79	grace	0.75
john	0.31	charles	0.77	happiness	0.73

Tabuľka 6 – 10 Hľadanie podobných slov pre slovo „end“ v kontextoch „life“ a „story“

Slovo	str	Kontext	sila	Kontext	sila
-		life		story	
end	1.44	end	4.31	end	2.88
out	0.18	business	0.53	story	0.47
house	0.17	beginning	0.53	ends	0.44
way	0.17	away	0.49	him	0.4
point	0.17	life	0.48	beginning	0.4
work	0.17	back	0.48	finished	0.39
place	0.16	started	0.45	ended	0.38
side	0.16	duty	0.45	known	0.38
away	0.16	death	0.44	me	0.37

Tabuľka 6 – 11 Hľadanie podobných slov pre slovo „dog“ v kontextoch „house“ a „creature“

Slovo	str	Kontext	sila	Kontext	sila
-		house		creature	
dog	1.44	dog	2.88	dog	4.31
horse	0.38	wood	1.03	man	1.11
man	0.37	guard	1.02	boy	1.1
boy	0.37	stone	0.97	bird	0.94
child	0.36	street	0.94	animal	0.94
girl	0.36	guest	0.82	simply	0.94
woman	0.35	hill	0.78	merely	0.78
fellow	0.35	man	0.74	horse	0.77
cat	0.35	child	0.73	pony	0.72

Tabuľka 6 – 12 Hľadanie podobných slov pre slovo „instrument“ v kontextoch „work“ a „music“

Slovo	str	Kontext	sila	Kontext	sila
-		work		music	
instrument	1.44	instrument	4.31	instrument	2.88
mirror	0.27	forces	0.84	organ	0.86
instruments	0.26	needle	0.84	concert	0.84
operation	0.26	contract	0.82	instruments	0.79
buildings	0.26	committee	0.82	band	0.71
machinery	0.26	shops	0.81	piano	0.71
structure	0.25	machinery	0.77	dance	0.71
arguments	0.25	brain	0.76	gallery	0.68
weapon	0.25	thread	0.75	passion	0.68

Tabuľka 6 – 13 Hľadanie podobných slov pre slovo „blue“ v kontextoch „color“ a „sad“

Slovo	str	Kontext	sila	Kontext	sila
-		color		sad	
blue	1.44	blue	1.44	blue	1.44
white	0.46	white	0.46	white	0.46
green	0.46	green	0.46	green	0.46
red	0.45	red	0.45	red	0.45
black	0.44	black	0.44	black	0.44
yellow	0.44	yellow	0.44	yellow	0.44
grey	0.38	dark	0.36	dark	0.36
brown	0.37	bright	0.34	bright	0.34
dark	0.36	soft	0.31	soft	0.31

Tabuľka 6 – 14 Hľadanie podobných slov pre slovo „head“ k kontextoch „body“ a „England“

Slovo	str	Kontext	sila	Kontext	sila
-		body		england	
head	1.44	head	4.31	arms	0.76
hands	0.43	face	1.22	earth	0.76
face	0.41	arm	1.19	lay	0.73
hand	0.41	arms	1.14	land	0.72
arm	0.4	body	1.06	king	0.67
feet	0.38	legs	0.96	world	0.64
arms	0.38	neck	0.9	again	0.62
heads	0.36	shoulders	0.89	them	0.62
eyes	0.36	heart	0.88	city	0.61

6.3 Zhlukovanie podobných slov na základe podobnosti

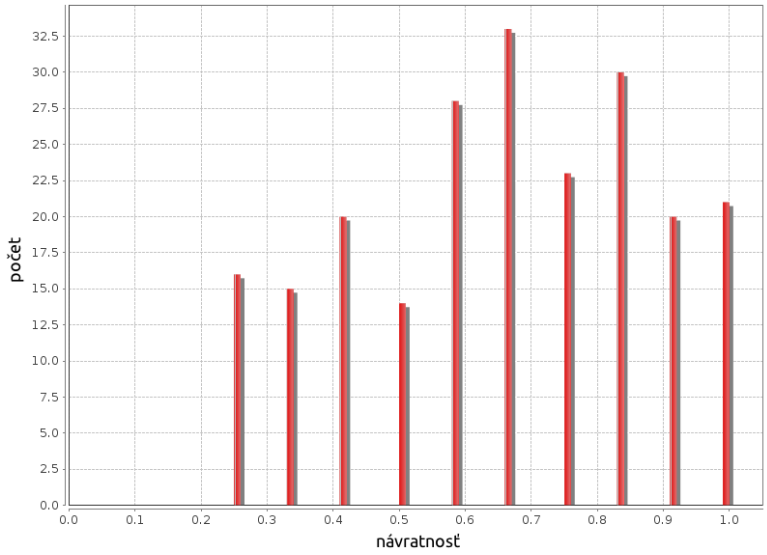
Cieľom experimentu je overenie prístupu zhlukovania mesiacov na základe sémantickej podobnosti slov (kapitola 4.10) na korpusoch písaných v neanglických jazykoch. Experiment je analogický tomu, ktorý je popísaný v kapitole 4.10. Overuje skutočnosť, či vysoké hodnoty presnosti a návratnosti neboli silne viazané na jazyk alebo korpus, nad ktorým boli výpočty vykonané.

Popis: Experiment prebiehal nad množinou dvanástich mesiacov v anglickom jazyku nad korpusom COREN1K. Pre objektívnejšie posúdenie presnosti znalostí získaných nad korpusmi CORSK150 a CORCS300 bol nad nimi tento kompletný výpočet zopakovaný.

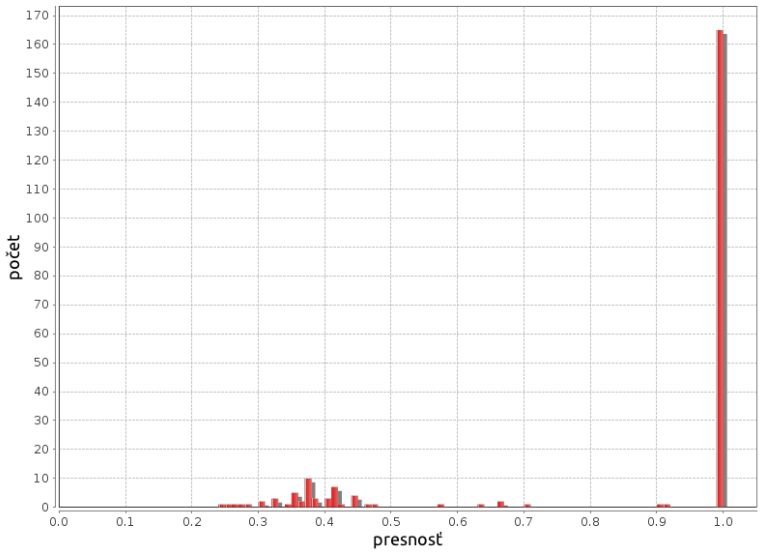
Výsledky: Aj napriek malej veľkosti korpusov dosiahlo AUP pri zhlukovaní slov povzbudzujúce výsledky. Pre korpus CORSK150 bola pri zhlukovaní dosiahnutá úplná presnosť. Histogram hodnôt návratnosti prístupu je znázornený na grafe 6 – 5, kde na osi y je znázornený počet vstupných trojíc a na osi x je znázornená im prislúchajúca hodnota návratnosti (úplnosti hľadanej oblasti). Priemerná návratnosť pre všetkých 220 trojíc mala hodnotu 0.65. Pre korpus CORCS150 nebola pri zhlukovaní dosiahnutá maximálna presnosť a histogram hodnôt presností pre jednotlivé trojice je zobrazený na grafe 6 – 6. Presnosť bola počítaná porovnávaním s úplnou množinou dvanástich mesiacov v tvare nominatívu jednotného čísla. To znamená, že slová v inom tvare ako napríklad „januári“ (lokál) alebo „januárom“ (inštrumentál) boli považované za nesprávny výsledok (čo bol v tomto prípade jediný typ chyby). Priemerná presnosť pre všetkých 220 trojíc mala hodnotu 0.854. Histogram hodnôt návratnosti prístupu je znázornený na grafe 6 – 7. Priemerná návratnosť pre všetkých 220 trojíc mala hodnotu 0.685.

Záver: Prístupu zhlukovania na základe sémantickej podobnosti slov dosiahol vysoké hodnoty presnosti a návratnosti v angličtine, slovenčine a češtine. Dá

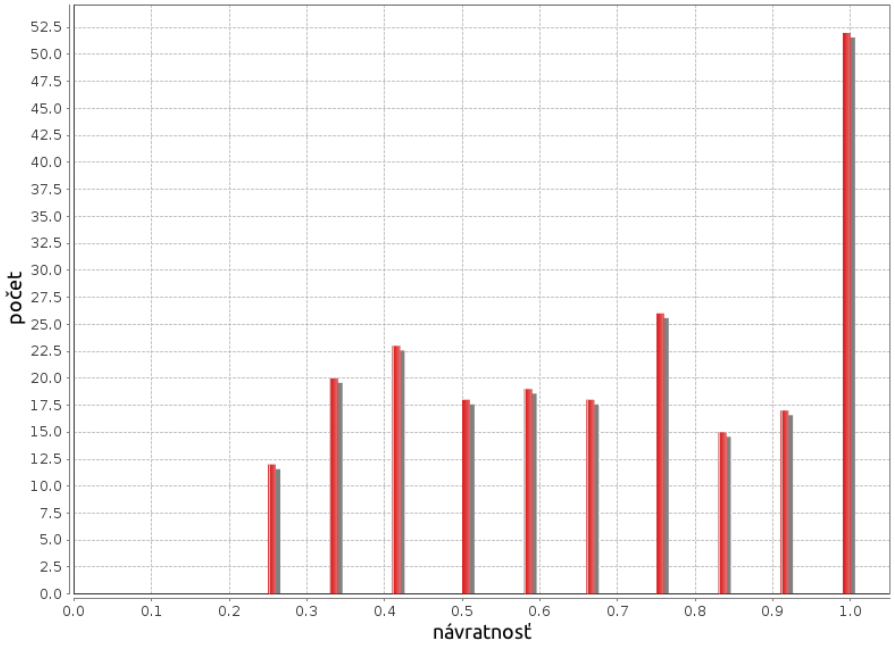
sa preto predpokladať, že tento prístup je nezávislý od jazyka a mohol by dosahovať rovnako dobré výsledky aj v iných jazykoch.



Obr. 6 – 5 Histogram hodnôt návratnosti pre všetky kombinácie vstupných trojíc „mesiacov“ pre korpus CORSK121.



Obr. 6 – 6 Histogram hodnôt presnosti pre všetky kombinácie vstupných trojíc „mesiacov“ pre korpus CORCZ370.



Obr. 6 – 7 Histogram hodnôt návratnosti pre všetky kombinácie vstupných trojíc „mesiacov“ pre korpus CORCZ370.

6.4 Tvorba zhlukov podobných slov na základe podobnosti

Cieľom experimentu je overenie prístupu zhlukovania slov na základe sémantickej podobnosti slov (kapitola 4.10) na rôznych korpusoch.

Popis: Nad všetkými korpusmi bolo vykonaných niekoľko výpočtov zhlukov podobných slov pre náhodne vybrané slová z úzko významovo ohraničených oblastí. Vstupy boli vyberané tak, aby boli pre jednotlivé korpusy analogické, no s prihliadnutím na rozdiely medzi jazykmi, použitím prekladov slov. Nie vo všetkých korpusoch sa do lexikónu prepracovali svojou frekvenciou rovnaké slová. V prípadoch, keď sa preklad slova v danom korpuse nenachádzal, boli použité iné slová z úzko významovo ohraničených oblastí.

Výsledky: Výsledky týchto experimentov sa nachádzajú v tabuľkách 6 – 15,6 – 16,6 – 18 a 6 – 17. Slová chybne zaradené do zhluku sú prešikrtnuté.

Systém bol schopný nájsť slová patriace do zhluku vo väčšine prípadov s vy-

sokou presnosťou. Najhoršie výsledky dosiahli vstupy pre slová označujúce mesiace pre korpus COREN1K a nápoje pre korpus COREN3K. V prípade mesiacov bolo výrazné zníženie presnosti spôsobené slovom „may“, ktoré v angličtine obsahuje okrem významu názvu mesiaca „máj“ aj význam „môcť“. V prípade nápojov v korpuse COREN3K bola znížená presnosť spôsobená pravdepodobne nedostatočnou veľkosťou korpusu a slovnej zásoby.

Záver: AUP preukázalo schopnosť výpočtu podobných slov patriacich do jednej úzko ohraničenej sémantickej triedy v troch rôznych jazykoch. Dá sa preto predpokladať, že tento prístup je nezávislý od jazyka a mohol by dosahovať rovnako dobré výsledky aj v iných jazykoch.

Tabuľka 6 – 15 Tvorba zhluku podobných slov nad kopursom COREN1K

mesiace	january, february, may	april, ~~can~~, december, ~~could~~, july, ~~might~~, tuesday, august, wednesday, ~~has~~, march, january, ~~greece~~, ~~cannot~~, ~~may~~, ~~should~~, ~~must~~, ~~venice~~, february, ~~friday~~, november, ~~isaac~~, ~~we~~, ~~autumn~~, ~~will~~, ~~would~~, october, ~~shall~~, june, september, ~~goethe~~
nápoje	water, beer, wine	milk, tea, ~~stream~~, ~~bottle~~, ~~meat~~, ~~boat~~, coffee, ~~food~~, ~~glass~~, wine, ~~river~~, beer, water, ~~fish~~
číslovky	one, two, three	~~some~~, one, three, two
jazyky	english, french, german	english, german, french
mestá	paris, london, rome	london, paris, rome
farby	red, green, blue	black, white, red, blue, green

Tabuľka 6 – 16 Tvorba zhluku podobných slov nad kopursom COREN3K

mesiace	january, february, may	april, march, august, february, september, may, june, november, december, july, january, october
nápoje	water, beer, wine	~~soil~~, ~~plants~~, ~~fuel~~, ~~fruit~~, ~~energy~~, milk, coffee, beer, ~~fish~~, ~~sugar~~, water, ~~plant~~, tea, wine, ~~meat~~, ~~heat~~, ~~gas~~, ~~oil~~, liquid, ~~coal~~, ~~food~~
číslovky	one, two, three	five, six, two, four, three, one
jazyky	english, french, german	spanish, english, ~~british~~, italian, german, french, russian
mestá	paris, london, rome	paris, berlin, rome, london
farby	red, green, blue	green, red, blue

Tabuľka 6 – 17 Tvorba zhluku podobných slov nad kopursom CORSK121

mesiace	január, február, marec	február, jún, september, október, január, november, apríl, marec, december, máj, júl
planéty	zem, mars, saturn	~~planét~~, mesiac, ~~boh~~, ~~slnka~~, ~~planéty~~, jupiter, ~~rýchlosť~~, saturn, ~~loď~~, zem, ~~apollo~~, mars, slnko, planéta, teleso
číslovky	jeden, dva, tri	jeden, tri, dva, štyri, dve
jazyky	nemecký, anglický, slovenský	francúzsky, slovenský, český, anglický, nemecký
mestá	bratislava, košice, prešov	bratislava, zvolen, žilina, košice, trenčín, prešov
farby	modrá, zelená, červená	žltá, červená, zelená, modrá

Tabuľka 6 – 18 Tvorba zhluku podobných slov nad kopursom CORCZ370

mesiace	červenec, březen, duben	leden, říjen, únor, březen, duben, květen, červenec
hudobné nástroje	kytara, klavír, bicí	housle, kytara, zpěv, klavír, ~~kytaru~~, bicí, varhany
číslovky	jeden, dva, tři	dvě, jeden, tři, čtyři, dva
jazyky	anglicky, česky, rusky	francouzsky, řecky, latinsky, rusky, německy, česky, hebrejsky, anglicky
mestá	berlín, bratislava, praha	berlín, liberec, bratislava, moskva, třebíč, hradec, zlín, olomouc, praha, vídeň, kolín, ostrava, brno, kladno, londýn, pardubice, plzeň, paříž
farby	červená , modrá, žlutá	žltá, červená, zelená, modrá

6.5 Rozširovanie zhlukov podobných slov na základe podobnosti pomocou parametra

Pri výpočte zhlukov podobných slov na základe podobnosti sa dá presnosť výslednej množiny ďalej korigovať pomocou parametra η. Zvyšovaním hodnoty parametra sa zvyšuje návratnosť výslednej množiny na úkor presnosti. Znižovaním hodnoty sa parametru naopak zvyšuje presnosť na úkor návratnosti.

Cieľom experimentu je overenie vplyvu parametra η na výsledný zhluk podobných slov. Tento parameter bol popísaný v kapitole 4.10.

Popis Nad dvoma korpusmi (COREN3K, CORCS300) boli zvolené trojice tokenov, pre ktoré boli vypočítavané zhluky podobných slov (viď. predchádzajúci experiment). Tieto boli vypočítavané pri postupne zvyšujúcej sa hodnote parametra η.

Výsledky Vplyv veľkosti hodnoty parametra η na výsledný zhluk je možné vidieť v tabuľkách 6 – 19, 6 – 20 a 6 – 21.

v tabuľke 6 – 19 vidieť ako sa postupným zvyšovaním hodnoty parametra η zvyšovala návratnosť výsledného zhluku a znižovala presnosť. V poslednom prípade pri hodnote 1.15 už bola presnosť nepostačujúca, a aj drobným zvýšením hodnoty parametra o hodnotu 0.01 sa nachádzalo vo výslednom zhluku obrovské množstvo slov, ktoré už nepopisovali iba farby.

Tabuľka 6 – 20 prezentuje ako sa postupným zvyšovaním hodnoty parametra η zvyšovala návratnosť výsledného zhluku. Pri hodnote 1 sa pri plne naučenom systéme predpokladá úplná presnosť.

Tabuľka 6 – 21 uvádza ako sa postupným zvyšovaním hodnoty parametra η zvyšovala návratnosť výsledného zhluku.

Záver Zmeny hodnoty parametra η zodpovedajú jeho predpokladanej funkcii pre oba testované jazyky.

Tabuľka 6 – 19 Vplyv parametra η pri vytváraní zhlukov nad korpusom COREN3K pre slová „green“, „red“, „blue“ označujúce farby

η	výsledný zhluk
1.0	green, red, blue
1.01	green, white, red, blue
1.03	black, green, white, red, blue
1.04	black, yellow, green, white, red, blue
1.08	black, yellow, ~~dark~~, green, white, red, blue
1.09	black, yellow, ~~dark~~, green, white, red, orange, blue
1.1	brown, black, silver, yellow, ~~dark~~, green, white, red, orange, blue
1.11	brown, black, silver, yellow, ~~dark~~, green, white, red, orange, blue, golden
1.13	wood, gold, grey, ~~flat~~, ~~iron~~, yellow, ~~dark~~, ~~light~~, green, ~~stone~~, orange, brown, ~~metal~~, black, silver, ~~big~~, red, white, blue, golden
1.14	~~wood~~, ~~small~~, gold, ~~flat~~, pink, yellow, ~~dark~~, ~~steel~~, green, ~~stone~~, orange, ~~upper~~, ~~ice~~, ~~metal~~, black, ~~mountain~~, white, grey, ~~double~~, ~~iron~~, ~~light~~, purple, brown, silver, ~~big~~, ~~glass~~, red, color, golden, blue
1.15	~~wood~~, ~~small~~, gold, ~~flat~~, pink, yellow, ~~dark~~, ~~steel~~, green, ~~stone~~, orange, ~~upper~~, ~~ice~~, ~~tree~~, ~~water~~, ~~metal~~, black, ~~mountain~~, ~~wild~~, white, ~~old~~, ~~deep~~, grey, ~~hill~~, ~~iron~~, ~~double~~, ~~rock~~, ~~light~~, gray, ~~bright~~, ~~ring~~, purple, brown, ~~forest~~, silver, ~~electric~~, ~~large~~, ~~big~~, ~~glass~~, red, ~~color~~, blue, golden, ~~smaller~~

Tabuľka 6 – 20 Vplyv parametra η pri vytváraní zhlukov nad korpusom CORCZ370 pre slová „září“, „říjen“, „srpen“ označujúce mesiace

η	výsledný zhluk
0.97	září, říjen, srpen
0.98	září, říjen, červen, únor, listopad, duben, srpen, červenec
0.99	leden, září, říjen, červen, únor, listopad, březen, listopadu, duben, květen, srpen, červenec
1.00	březen, prosinec, ~~června~~, ~~května~~, ~~srpnu~~, leden, září, říjen, únor, červen, ~~srpna~~, listopad, ~~července~~, ~~října~~, duben, ~~listopadu~~, květen, srpen, červenec

Tabuľka 6 – 21 Vplyv parametra η pri vytváraní zhlukov nad korpusom CORCZ370 pre slová „modrá“, „červená“, „zelená“ označujúce farby

η	výsledný zhluk
1.0	modrá, červená, zelená
1.01	modrá, červená, zelená
1.02	modrá, červená, zelená, žlutá
1.04	černá, bílá, modrá, červená, zelená, žlutá
1.07	černá, bílá, modrá, zlatá, červená, zelená, žlutá
1.08	černá, bílá, modrá, zlatá, ~~turistická~~, červená, zelená, žlutá
1.1	červený, černá, ~~stezka~~, zlatá, červená, zelená, ~~trasa~~, žlutá, ~~červené~~, bílá, modrá, ~~lesní~~, ~~barva~~, ~~dobrá~~, ~~turistická~~, ~~dlouhá~~, ~~potok~~

7 Záver

Táto práca sa zaoberala možnosťou príspevku do problematiky automatickej extrakcie relácií medzi konceptmi v textoch v prirodzenom jazyku.

V úvode bola priblížená problematika, zadefinované ciele práce a popísaná štruktúra práce. Kapitola 2 predstavuje základné delenie a vybrané druhy konceptuálnych sietí a vybrané metódy tvorby týchto konceptuálnych sietí uvádza kapitola 3. Pri vybraných metódach poukazuje na problém predspracovania textov a silnú závislosť na lexikálnych zdrojoch. Keďže AUP slúžilo už vo svojich počiatkoch na automatické získavanie lexikálnych zdrojov, ďalšie smerovanie práce vedie k samotnému popisu AUP s prihliadnutím na jeho možnosti v tejto oblasti, čo tvorí obsah kapitoly 4. Kapitola prináša okrem samotného podrobného popisu aj informácie o využití AUP a ľahko pochopiteľné príklady fungovania jeho základných funkčných blokov. V kapitole 5 je stručne popísaná implementácia AUP - program Beast.

Kapitola 6 obsahuje ukážky a experimentálne overenia možnosti využitia AUP v doméne úloh budovania konceptuálnych sietí z textov v prirodzenom jazyku pomocou jeho implementácie Beast. V tejto kapitole boli prezentované vlastnosti vybraných funkčných blokov AUP na korpusoch z rôznych doménových oblastí. Korpusy tvorili čo do veľkosti veľmi veľké množiny textov v prirodzenom jazyku a líšili sa svojou veľkosťou a jazykom, v ktorom boli texty písané. AUP sa vzhľadom na výsledky v tejto kapitole javí ako spoľahlivý prístup na extrakciu relácií z textov na základe ukážkovej relácie a to aj napriek nedostatočným veľkostiam korpusov, nad ktorými bol učený, alebo šumom v týchto korpusoch. Ukázalo sa ako vhodný prístup na rozpoznávanie sémantickej podobnosti slov. Naopak, prístup rozpoznávania sémantickej podobnosti slov v kontexte (rozšírenie rozpoznávania sémantickej podobnosti slov), sa experimentálne neosvedčil. S využitím AUP sa podarilo nájsť plné zhluky konceptov patriacich do jednej sémantickej triedy ako napríklad mesiace v roku, číslovky, mená osôb alebo farby, a to rovnako pre angličtinu ako aj pre slovenčinu a češtinu. Takto získané znalosti o reláciách alebo zhlukoch konceptov z jednej sémantickej

triedy môžu v ďalšom kroku poslúžiť na predspracovanie textov, ktoré budú použité ako vstup do systémov ako je DIPRE alebo Snowball. AUP sa ukázal ako sľubný prostriedok, ktorý by v budúcnosti mohol poslúžiť na odstránenie lexikálnych závislostí týchto dvoch systémov.

7.1 Prehľad splnenia cieľov práce

Táto podkapitola sumarizuje splnenie cieľov tejto práce. Vedecké a technologické prínosy práce sú popísané v podkapitole 7.1.1.

- **Identifikácia medzier** - V kapitole 2 bol poskytnutý základný popis sémantických sietí podľa Sowu. Popis bol bližšie zameraný na vybrané sémantické siete ontologického typu. Vybrané prístupy tvorby týchto sémantických sietí popisuje kapitola 3. V závere boli tieto prístupy kvalitatívne porovnané a vyhodnotené. Za spoločné obmedzenie týchto prístupov sa označuje **závislosť na znalostiach o jazyku**, nad ktorým pracujú, a to priamu, alebo nepriamu, v podobe externých nástrojov, ktoré sú pri daných prístupoch potrebné.

- **Návrh konkrétnej metódy** - Celá kapitola 4 sa venuje **pôvodnému prístupu AUP**. Návrh a implementácia sú **hlavným prínosom práce**. Jedná sa o nový prístup, ktorý je schopný bez znalostí o jazyku identifikovať v jazyku koncepty a vybrané relácie. Je založený na identifikácii pravdepodobností spoločných výskytov dvojíc slov v korpuse. Obsahuje metódy, ktoré sú na základe týchto hodnôt pravdepodobností schopné:

  - Výpočtu bezkontextovej sémantickej príbuznosti (viď 4.5).
  - Identifikovať sémanticky a syntakticky blízke slová - P-slová (viď 4.8).
  - Výpočtu kontextovo závislej sémantickej príbuznosti (viď 4.9.1).
  - Zhlukovať pojmy na základe ich sémantickej príbuznosti (viď 4.10).

- **Experimentálne vyhodnotenie** - Kapitola 4 obsahuje pri popise vyššie spomínaných metód experimenty overujúce ich funkčnosť na anglickom korpuse. Kapitola 6 prináša **súbor experimentov** pre vybrané metódy:
  - Výpočet bezkontextovej sémantickej príbuznosti (viď 4.5).
  - Výpočet kontextovo závislej sémantickej príbuznosti (viď 4.9.1).
  - Zhlukovanie pojmov na základe ich sémantickej príbuznosti (viď 4.10).

  Experimenty prebiehajú **na rôznych korpusoch** (líšiacich sa svojou veľkosťou, zložením a jazykom, v ktorom boli písané). Potvrdili funkčnosť daných metód na korpusoch v slovenskom, českom a anglickom jazyku.

- **Identifikácia slabých miest** - Časť 4.2 popisuje obmedzenia metódy **vzhľadom na vybraný korpus a počet známych slov**. **Pomocou experimentov** (kapitola 6) boli odhalené slabé miesta vybraných metód prístupu AUP (kontextovo závislej sémantickej príbuznosti). AUP obsahuje možnosti ďalšieho rozvoja v budúcnosti, ktorý by mohol množstvo limitácií odstrániť. **Súpis ďalšieho možného rozvoja** AUP sa nachádza v závere tejto kapitoly (viď 7.2).

7.1.1 Prínosy práce

Za hlavné vedecké prínosy tejto práce autor pokladá:

- Návrh prístupu schopného identifikácie sémanticky a syntakticky blízkych slov k zadanému slovu.
- Návrh prístupu schopného výpočtu kontextovej a bezkontextovej sémantickej príbuznosti.
- Návrh prístupu schopného zachytávať v texte v prirodzenom jazyku koncepty, ktoré zdieľajú rovnaké relácie, bez znalostí o jazyku nad ktorým pracuje.

- Experimentálne overenie daných prístupov nad korpusmi v troch rôznych jazykoch.

Za hlavný technologický prínos tejto práce autor pokladá:

- Zverejnenie zdrojového kódu implementácie, využívanie voľne stiahnuteľných korpusov je veľkým prínosom pre možné nadviazanie výskumu AUP tretími stranami.

7.2 Zameranie vývoja AUP v budúcnosti

Ďalší vývoj AUP môže byť realizovaný vo viacerých navzájom nezávislých vetvách. Z mnohých možných rozšírení sa dajú spomenúť tie najvýznamnejšie:

- Rozšírenie prístupu z jednotlivých slov na dvojice až n-tice slov.
- Preskúmanie súčasných vlastností prístupu na dostatočne veľkých a čistých korpusoch.
- Rozšírenie a preskúmanie prístupu s korpusmi v iných ako indoeurópskych jazykoch ako napríklad perzský jazyk alebo japončina.
- Návrh a realizácia implementácie spojenia AUP so systémom Snowball.

Rozšírenie prístupu z jednotlivých slov na dvojice až n-tice slov by prinieslo možnosť širšieho využitia AUP v praxi. Kým súčasná implementácia je obmedzená iba na prácu s jednotlivými slovami, v bežnom jazyku sa entity označujú aj viacslovnými spojeniami. Toto rozšírenie by odstránilo silné obmedzenie súčasného systému, ktorý by tým pádom mohol zhlukovať napríklad celé mená osôb, názvy produktov alebo miest a mohol by poslúžiť ako plnohodnotný nástroj v oblasti NER.

Kým AUP dosahuje veľmi dobré výsledky aj na menších korpusoch, ako to bolo prezentované napríklad na korpuse v slovenskom jazyku, jeho preskúmanie

na dostatočne veľkých korpusoch je stále potrebné. Pri väčších korpusoch môžu nastať problémy (spôsobené veľkým počtom asociácií), ktoré sa pri malých korpusoch neprejavia. Tieto problémy by mohli byť analogické k problémom spojeným s preučením neurónových sietí, keď sa tieto stávajú príliš silne zameranými na oblasť vymedzenú trénovacou množinou.

AUP sa javí ako prístup, ktorý dosahuje dobré výsledky nad korpusmi písanými v jazykoch z rodiny indoeurópskych jazykov. Je potrebné preskúmať, či sa zároveň nejedná o obmedzenie systému. Na preskúmanie jazykov akým je napríklad japončina alebo perzský jazyk by bolo potrebné doplniť implementáciu o možnosť čítania týchto jazykov. Vzhľadom na fakt, že súčasná implementácia je slobodným softvérom s otvoreným kódom, je možné, že sa o toto rozšírenie pričiní niektoré zo zahraničných pracovísk.

Návrh a realizácia implementácie spojenia AUP so systémom Snowball by mohla vyriešiť silnú závislosť systému Snowball na predspracovaní textov, s ktorými pracuje. Ak by sa jednalo priamo o implementáciu AUP, ktorá by dokázala pracovať s viacslovnými entitami, tento systém by mohol byť úplne oslobodený od externých lexikálnych zdrojov. Zo vstupnej množiny relácií by si systém pomocou AUP sám vytvoril a označil množiny konceptov, nad ktorými chce pracovať. Po extrakcii potrebného počtu relácií by mohol následným prehľadávaním hodnôt parametra η množinu konceptov rozširovať tak, aby ostala zachovaná kvalita extrahovaných relácií.

Aj vďaka pozitívnym výsledkom je vo vývoji AUP motivujúce pokračovať. Vzhľadom na možnosť slobodného stiahnutia, používania a rozširovania jeho implementácie, dáva do rúk všetko potrebné pracoviskám, ktoré by sa na jeho rozšírení chceli akokoľvek podieľať alebo overiť experimentálne výsledky prezentované v tejto práci.

7.2.1 Problém viacvýznamovosti slov

V spracovaní prirodzeného jazyka patrí riešenie problému nejednoznačnosti slovného významu k dlhodobým výzvam. Mnoho slov má viacero významov a pre počítače je zložité určiť, v ktorom význame sa dané slovo nachádza. Ako príklad môže poslúžiť vyššie spomínané slovo označujúce v angličtine mesiac máj, „May“, ktorého druhý význam by sa dal do slovenčiny preložiť ako „môcť“. Preskúmanie možností AUP v tejto oblasti bolo motivované skutočnosťou, že prístupy bez učiteľa sa javia byť v tejto oblasti sľubným trendom [78].

Ľudia rozlišujú význam slov na základe slov, ktoré sa vyskytujú naľavo a napravo od daného slova. Keďže práve s touto informáciou AUP pracuje a je to jediná znalosť, ktorú si o texte ukladá, je pravdepodobné, že by sa s použitím jeho metód dal zostrojiť systém na rozlišovanie významu slov. Tento systém bude stručne popísaný a ilustrovaný na jednoduchom príklade.

Ako príklad opäť poslúži viacvýznamovosť slov označujúcich v anglickom jazyku mesiace. Slovo „May“ sa nachádza práve v dvoch významoch. Každý z týchto významov je možné popísať podobne ako pri zhlukovaní podobných slov, nejakou trojicou slov. Príklad popisu sa nachádza v tabuľke 7 – 1.

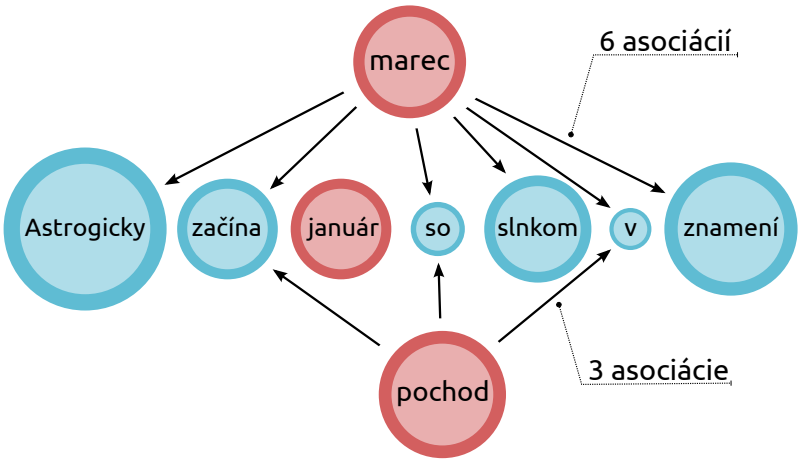
Tabuľka 7 – 1 2 možné významy anglického slova „May“ popísané trojicou slov.

Význam	Popisná trojica slov
Mesiac	May, January, February
Môcť	may, might, could

Predpoklad, že AUP obsahuje informáciu o týchto významoch, je založený na experimentálnych výsledkoch pri zhlukovaní sémanticky podobných slov. S pomocou slovníka je možné vybrať také trojice slov, ktoré budú popisovať 2 významy slova „may“, bez toho, aby sa v nich samé vyskytovalo. Pre dané trojice je možné dohľadať pomocou tejto metódy zhlukovania zvyšné slová patriace do tried, ktoré samotné trojice definujú. Tabuľky 4 – 7 a 7 – 2 uvádzajú výsledok tejto operácie.

Tabuľka 7 – 2 Výstup zhlukovania symbolov na základe vstupnej trojice troch náhodných dní.

vstup	should,could,might
výstup	will, must, could, should, may, would, can, might



Obr. 7 – 1 Spojenie dvoch zoznamov vypočítaných pre 2 symboly.

Uvedenými výsledkami bolo experimentálne overené, že AUP má dostatočné znalosti na rozlíšenie daných dvoch významov slova „may“. Význam rozlišuje jeho kontext, slová v jeho okolí, konkrétne napr. jeho použitie vo vete. Vzhľadom na fakt, že výpočet podobnosti je možný práve vďaka znalosti okolia slov, mohol by pri náhrade sledovaného slova, za iné slová z jeho významovej množiny, počet asociácií odpovedať miere istoty, že sa dané slovo nachádza vo vete v danom význame. K významom z tabuľky 7 – 1 sa dajú priradiť tieto dva rôzne kontexty, v ktorých sa v daných významoch slovo „may“ nachádza:

- *...you* may *leave the table when...*
- *...the week in which* May *begins in one year...*

Výmenou slova „may“ za iné slovo z jeho významového zhluku, existuje možnosť spočítať počet jeho asociácií k slovám z oboch kontextov. Počet asociácií pre rovnako početné n-tice by mal byť v pozitívnych prípadoch omnoho vyšší ako v prípadoch

Tabuľka 7 – 3 2 možné významy anglického slova „March“ popísané trojicou slov.

Význam	Popisná trojica slov
Mesiac	March, January, February
Pochod	march, walk, move

negatívnych (ak je význam n-tice zhodný s významom kontextu). Na obr. 7 – 1 je zozbrazený ilustračný príklad pre slovenskú vetu. Pokiaľ by bolo slovo „január“ nahradené slovom „marec“, tak by bolo cez príslušné kontextové vzdialenosti s ostatnými slovami vo vete asociované v šiestich rôznych asociáciách. Ak by bolo nahradené slovom „pochod“, to by bolo v danej konkrétnej vete prepojené s ostatnými slovami len cez tri asociácie. Slovo „marec“ sa teda ako náhrada slova „január“ do tejto vety hodí viac ako slovo „pochod“. Pre trojice z tabuľky 7 – 3 a kontexty uvedené vyššie je počet asociácií v pozitívnom prípade oproti negatívnemu prípadu dvojnásobný. Pre podobný prípad s mesiacom marec je výsledok analogický.

- *...to a long distance* march *carrying full kit as...*
- *...the week in which* March *begins in one year...*

Predbežné výsledky návrhu prístupu k riešeniu problémov nejednoznačnosti pomocou slov sa javia byť sľubné, no bez širšieho experimentálneho overenia nie je možné danému návrhu plne dôverovať. Pri jednoduchých príkladoch bolo možné určiť z dvoch významov vo vete ten správny, a to s veľkou mierou istoty. Ak by sa daný prístup podarilo v budúcnosti rozšíriť a experimentálne overiť, stále by čelil problému nutnosti znalosti všetkých významových hladín dotazovaného slova.

Zoznam použitej literatúry

[1] Agichtein, E., Gravano, L.: Snowball: Extracting Relations from Large Plain-Text Collections., *DL '00 Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 85–94, New York, NY, USA: ACM, 2000.

[2] Agichtein, E.: *Extracting Relations From Large Text Collections.*, Eugene Agichtein, Ph.D. Thesis, 2005.

[3] Barbu, E., Mititelu, B. V.: Automatic Building of Wordnets., *In Proceedings of Recent Advances in Natural Language Processing IV*, John Benjamins, pp. 217–226, Amsterdam, 2007.

[4] Benuskova, L.: Kognitivna neuroveda., In Rybar, J., Benuskova, L., Kvasnicka, V. (eds) *Kognitivne vedy.* Kalligram, Bratislava, pp. 47-104. 2002, ISBN 80-7149-515-8.

[5] Bond, F., Isahara, H., Uchimoto, K., Kuribayashi, T., Kanzaki, K.: Extending the Japanese WordNet. *In 15th Annual Meeting of The Association for Natural Language Processing*, Tottori, C1-4., 2009.

[6] Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., Kanzaki, K.: Enhancing the Japanese WordNet.,*In Proceedings of the 7th Workshop on Asian Language Resources (ALR 2009)*, ACL- IJCNLP, Singapore, 2009.

[7] Bond, F., Baldwin, T., Fothergill, R., Uchimoto, K.: Japanese SemCor: A Sense-tagged Corpus of Japanese., *In The 6th International Conference of the Global WordNet Association (GWC-2012)*, Matsue, 2012.

[8] Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction., *In Proceedings of the Conference of the German Society for Computational Linguistics*, Tübingen, 2009.

[9] Brachman, R. J.: On the Epistemological Status of Semantic Networks., *Associative Networks: Representation and Use of Knowledge by Computers*, Academic Press, 3-50, 1979.

[10] Brin, S.: Extracting patterns and relations from the World-Wide Web., *In Proceedings of the 1998 International Work-shop on the Web and Databases (WebDB'98)*, 1998.

[11] Boas H. C.: „From Theory to Practice: Frame Semantics and the Design of FrameNet.“, *In Langer, S., Schnorbusch D.(eds.), Semantisches Wissen im Lexikon*, Tübingen: Narr. pp. 129-160. 2005.

[12] Bobrovnikoff, D.: *Semantic Bootstrapping with a Cluster-Based Extension to DIPRE*, Stanford University, 2000.

[13] Cilibrasi, R., Vitanyi, M. B.: The Google Similarity Distance., *In the Proceedings of IEEE Trans. on Knowl. and Data Eng., vol. 19*, pp. 370–383, 2007

[14] Čapek, T.: *Systém pro částečné sémantické značkování volného textu.*, Diplomová práce, Masarykova univerzita, Fakulta informatiky, vedoucí práce Lukáš Svoboda, 2006.

[15] Desel, J.,Juhás, G.: „What Is a Petri Net? Informal Answers for the Informed Reader“., *Hartmut Ehrig et al. (Eds.): Unifying Petri Nets*, LNCS 2128, pp. 1–25, 2001.

[16] Farreres, X., Rigau, G., Rodriguez, H.: Using WordNet for Building WordNets., *In Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 1998.

[17] Fellbaum, Ch.: *WordNet: An Electronic Lexical Database.*, Cambridge, MIT Press, 1998.

[18] Feng, C., Copeck, T., Szpakowicz, S., Matwin, S.: „Semantic clustering: acquisition of partial ontologies from public domain lexical sources“., *In Proceedings of the 7th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*, SRDG Publications, 1993.

[19] Fridman, N.N., Hafner, C.D.: *The State of the Art in Ontology Design: A Survey and Comparative Review.*, *AI Magazine 18(3)*, Association for the Advancement of Artificial Intelligence, United States, pp. 53–74, 1997.

[20] Fiser, D., Sagot, B.: Combining Multiple Resources to Build Reliable Wordnets., *In Proceedings of TSD'2008*, pp.61–68, Springer-Verlag Berlin, Heidelberg 2008.

[21] Furdík, K.: *Získavanie informácií v prirodzenom jazyku s použitím hypertextových štruktúr.*, Doktorandská dizertačná práca, Katedra kybernetiky a umelej inteligencie FEI, Technická univerzita v Košiciach, Košice, 2003.

[22] Gruber, T. R.: *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, International Journal Human-Computer Studies, 43(5-6), 907-928, 1995

[23] Guarino, N. 0 Giaretta P.: *Ontologies and Knowledge Bases: Towards a Terminological Clarification*, In N. Mars (Ed.), Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, Amsterdam. 1995

[24] Harper L. W., Delugach H. S.: Using Conceptual Graphs to Capture Semantics of Agent Communication.,p392-404,ICCS

[25] Hebb D.O. ,The Organization of Behavior. ,John Wiley, New York, USA,1949

[26] Hlaváčová D.: *Databáze slovesných valenčních rámců VerbaLex* [online]. 2008 [cit. 2012-06-23]. Disertační práce. Masarykova univerzita, Filozofická fakulta. Vedoucí práce Karel Pala.

[27] Hu Y., Lu R., Chen Y., Pei B.: Text Retrieval Oriented Auto-construction of Conceptual Relationship

[28] Jaccard P.: (1901), „Étude comparative de la distribution florale dans une portion des Alpes et des Jura“, Bulletin de la Société Vaudoise des Sciences Naturelles 37: 547–579.

[29] Johnson C., Fillmore C., Wood E., Ruppenhofer J., Urban M., Petruck M., Baker C. (2001): The FrameNet Project: Tools for lexicon building. Manuscript. International Computer Science Institute, Berkeley, CA.

[30] Jones T. M. (2008). Artificial Intelligence: A Systems Approach, Infinity Science Press LLC Hingham, Massachusetts New Delhi.

[31] Khoo C. S. G., Na J. C.: (2006). Semantic relations in information science. Annual Review of Information Science and Technology 40: 157-229.

[32] Král R., *Jaký to má význam?*, Dizertační práce. Masarykova univerzita, Fakulta informatiky, 2004

[33] Llull, R., Lulli, B. R.: *Logica nova: jam Valentiae impressa anno 1512 et nunc Palmae cum libris Logica parva.*, De quinque praedicabilibus & decem praedicamentis et De natura, ex typis Michaëlis Cerda & Antich & Michaèlis Amoròs, 1744.

[34] Madsen, M.: *The limits of machine translation.*, Masters thesis, Copenhagen School of Bussiness, 2009.

[35] Mayer, R. E.: Models for understanding., *Review of Educational Research, 59 (1)*, SAGE Publications, pp 43–64, 1989.

[36] Miller, G. A.: WordNet: A Lexical Database for English., *In Communications of the ACM Vol. 38 No. 11*, ACM, pp 39–41, 1995.

[37] Mineau, G. W., Missaoui, R., Godin, R.: Conceptual Modeling Using Conceptual Graphs., *In Proceedings of KRDB'2000*, CEUR-WS, pp.73–86, 2000.

[38] Mishne, G.: *Source code retrieval using conceptual graphs.*, Master's thesis, University of Amsterdam, 2004.

[39] Moncecchi, G., Wonsever, D., Minel, J.-L.: „A survey of kernel methods for relation extraction"., *In Proceedings of the Workshop in Natural Language Processing and Web-based Tecnhologies*, IBERAMIA, Argentina, pp. 90–98, 2010.

[40] Murphy, R. C.: Phrase Detection and the Associative Memory Neural Network., *In the Proceedings of the 2003 International Joint Conference on Neural Networks*, Portland Oregon, 2003.

[41] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification., *In Linguisticae Investigationes 30(1)*, John Benjamins Publishing Company, pp. 3–26, 2007.

[42] Navigli, R., Velardi P., Gangemi, A.: „Ontology learning and its application to automated terminology translation"., *In IEEE Intelligent Systems, vol. 18 n. 1*, pp. 22-31, 2003.

[43] Nguyen, B., Badaskar, S.: *A Review of Relation Extraction*, Language Technologies Institute, Carnegie Mellon University, 2007.

[44] Nepil, M.: *Relational Rule Induction for Natural Language Disambiguation.*, Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno, 2003.

[45] Nielsen, R. H.: *A Theory of the Cerebral Cortex.*, ICONIP, 1998.

[46] Nielsen, R. H.: *Confabulation Theory: The Mechanism of Thought.*, Springer, 2007.

[47] Paul Van Der, Vet, P. E., Speel P., Mars, N. J. I., The Plinius ontology of ceramic materials., *In Proceedings of ECAI94's Workshop on Comparison of Implemented Ontologies*, 1994.

[48] Pazienza, M. T. Stellato, A. Tudorache, A.: A Bottom-up Comparative Study of EuroWordNet and WordNet 3.0 Lexical and Semantic Relations., *In 'LREC' European Language Resources Association*, ELRA, 2008.

[49] Pearsall, J. E.: *The New Oxford Dictionary of English.*, Oxford, Clarendon Press, 1998.

[50] Richens R.H.: Preprogramming for mechanical translation., *Mechanical Translation, Vol. 3(1)*,Massachusetts Institute of Technology, pp. 20–25, 1956.

[51] Ročkai, V., Kende, R.: Associative learning of concepts., *AI METH 2007*, CAMES, Gliwice, 2007.

[52] Ročkai, V.: Automatické generovanie slov patriacich do jednej syntaktickosémantickej triedy., *In Proceedings of 2nd Workshop on Intelligent and Knowledge oriented Technologies*, Centre for Information Technologies, FEI TU Košice, Slovakia, 2008.

[53] Ročkai, V.: Context for concepts., *SCYR 2009 - 9th Scientific Conference of Young Researchers*, FEI TU, Košice, 2009.

[54] Ruggeri, F., Faltin, F., Kenett, R.: Bayesian Networks., *In Ruggeri F., Faltin F. & Kenett R., Encyclopedia of Statistics in Quality & Reliability*, Wiley & Sons, 2007.

[55] Rybár, J., Kvasnička, V., Farkaš, I.: *Jazyk a kognícia.*, Kalligram, Bratislava, pp. 235–261, 2005, ISBN 80-7149-716-9.

[56] Rychlý, P., Šmerk, P., Pala, K.: *DESAM – morfologicky označkovaný korpus českých textů.* 2010.

[57] Sanchez, O., Poesio, M.: Acquiring Bayesian Networks from Text., *In the 4th LREC Proceedings*, European Language Resources Association, Lisboan, Portugal, 2004.

[58] Saveski, M., Trajkovski, I., Automatic Construction of Wordnets by Using Machine Translation and Language Modeling., *In Proceedings of the 13th International Multiconference Information Society*, Ljubljana, Slovenia, 2010.

[59] Saveski, M., Trajkovski, I.: Development of an English-Macedonian Machine Readable Dictionary by Using Parallel Corpora., *Communications in Computer and Information Science*, Springer Berlin Heidelberg, 2011.

[60] Scott, S., Matwin, S.: Text classification using WordNet hypernyms., *In Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Association for computational linguistics, Montreal, 1998.

[61] Schreiber, G., Wielinga, B., Jansweijer, W.: The KAKTUS View on the 'O' Word., *In Proceedings of IJCAI95 Workshop on Basic Onto-logical Issues in Knowledge Sharing*. Montreal, Canada, 1995.

[62] Schuler, K. K., Korhonen, A., Brown, S.: VerbNet overview, extensions, mappings and applications., *HLT-NAACL (Tutorial Abstracts)*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 13–14, 2009.

[63] Shapiro, S. C.,Rapaport, J. W.: The SNePS family., *In Computers & Mathematics with Applications*, Pergamon Press, Oxford, UK, pp. 243–275, 1992.

[64] Sklenák, V., *Sémantický web*, INFORUM, Praha, 2003.

[65] Smeaton, A., Kelledy, F., O'Donnell, R.: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish., *In The Fourth Text REtrieval Conference (TREC-4)*, School of Computer Applications, Dublin City University, Dublin, 1995.

[66] Sowa, J. F.: *Conceptual graphs tutorial.*, Dostupné on-line (18.12.2010): <http://www.huminf.aau.dk/cg/index.html>.

[67] Sowa, J. F., Semantic networks., *Encyclopedia of Artificial Intelligence.*, edited by S. C. Shapiro, Wiley, New York, 1987.

[68] Sowa, J. F.: *Conceptual Structures: Information Processing in Mind and Machine.*, Addison-Wesley, Reading, MA, 1984.

[69] Strehovský, M.: *Moderní metody získávání a zpracování biomedicínských dat.*, Bakalárska diplomová práca, Masarykova univerzita, Brno, 2008.

[70] Sumathi, S., Esakkirajan, S.: *Fundamentals of Relational Database Management Systems.*, Springer, 2007, ISBN 978-3-540-48397-7.

[71] Teukolsky, S.A. Vetterling, W.T., Flannery: Conditional Entropy and Mutual Information., *Numerical Recipes: The Art of Scientific Computing (3rd ed.)*, New York: Cambridge University Press, 2007, ISBN 978-0-521-88068-8.

[72] Turing, A.: Computing Machinery and Intelligence., *Mind Vol. LIX (236)*, pp. 433–460, 1950.

[73] Uschold, U. , Gruninger, M.: Ontologies: Principles, Methods, and Applications., *Knowledge Engineering Review, Vol. 11, No. 2.*, 1996.

[74] Wen, H. M. S., Eshley, G. H., Bond, F.: (2012) Using Wordnet to predict numeral classifiers., *In Chinese and Japanese in The 6th International Conference of the Global WordNet Association (GWC-2012)*, Matsue., 2012.

[75] Woods, B., Syntax, Semantics, and Speech., *In D. R. Reddy (ed.), Speech Recognition*, New York: Academic Press, 1975.

[76] Wooldridge, M.: *An Introduction to MultiAgent Systems.*, John Wiley & Sons, 2002, ISBN 0-471-49691-X.

[77] Xiaobin, L., Szpakowicz, S., Matwin, S.: A WordNet-based algorithm for word sense disambiguation., *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 1995.

[78] Yarowsky, D.: Unsupervised word-sense disambiguation rivaling supervised methods., *In Proceedings of the 33rd Annual Meeting of the Association for Computational Lainguistics (ACL '95)*, pp. 189–196, Cambridge, MA, 1995.

[79] Dictionary.com, context, in The American Heritage® Dictionary of the English Language, Fourth Edition. Source location: Houghton Mifflin Company, 2004. http://dictionary.reference.com/browse/context. Dostupné on-line (28.10.2009): <http://dictionary.reference.com>.

[80] Dictionary.com, context, in Dictionary.com Unabridged. Source location: Random House, Inc. http://dictionary.reference.com/browse/context. Dostupné on-line (28.10.2009): <http://dictionary.reference.com>.

Zoznam obrázkov

Zoznam tabuliek

8 Prílohy

8.1 Definícia gramatiky programu SLG

```
S  : SP VI . \(.25\) | SP VT OP . |
 {sub-intr, SP NP N, VI} | {sub-trns, SP NP N, VT} |
 {trns-obj, VT, OP NP N} | {sub-obj, SP NP N, OP NP N} |
 {intrans-ref, VI, SP RC VI};

SP | OP : NP | NP RC (.3) |
 {sub-intr, NP N, RC VI} | {sub-trns, NP N, RC VT} |
 {trns-obj, RC VT2, NP N};

RC : who VI | who VT OP | who SP VT2 | {trns-obj, VT, OP NP N} |
 {sub-trns, SP NP N, VT2};

NP : ART ADJ N | {noun-art, N, ART} | {noun-adj, N, ADJ};

ART: "" | the | a;

ADJ: "" (0.6) | quick | happy | hungry | nasty | mangy | crazy |
 sleazy;

N  : boy | boys | girl | girls | Mary | John | cat | cats |
 dog | dogs;

VI : walks | walk | bites | bite | eats | eat | barks | bark;

VT | VT2 : chases | chase | feeds | feed | sees | see | walks |
 walk | bites | bite;
```

```
sub-intr {
 boy  | girl  | Mary  | John  : walks | eats;
 boys | girls : walk  | eat;
 cat  | dog   : walks | bites | eats  | barks;
 cats | dogs  : walk  | bite  | eat   | bark;
 cat  | cats  ! bark  | barks;
}

sub-trns {
 boy  | girl  | Mary   | John : chases | feeds | sees(.1) | walks;
 boys | girls : chase  | feed     | see(.1) | walk;
 cat  | dog   : chases | sees(.2) | bites;
 cats | dogs  : chase  | see(.2)  | bite;
}

trns-obj {
 walk | walks : cat | cats | dog | dogs;
 see  | sees  : cat | cats;
}

sub-obj {
 Mary ! Mary;
 John ! John;
}

intrans-ref {
 walks | walk ! walks | walk;
 bites | bite ! bites | bite;
```

```
 eats  | eat  ! eats  | eat;
 barks | bark ! barks | bark;
}

noun-adj {
 boy  | boys | girl | girls | Mary | John ! mangy;
 John | cat  | cats | dog   | dogs ! sleazy;
}

noun-art {
 Mary | John : "";
 boys | girls | cats | dogs ! a;
 boy  | girl  | cat  | dog  ! "";
}
```

8.2 Ukážka z korpusu vygenerovaným programom SLG

Ukážka z korpusu tvorenom textovým dokumentom vytvoreným pomocou programu SLG pomocou gramatiky popísanej vyššie. Ukážku tvorí náhodne vybraný úsek tohto textového dokumentu:

the girl walks. quick dogs bite the girls. the quick dogs see the cat. dogs see the cats. the nasty dog walks. a happy girl feeds dogs. the dogs eat. happy John feeds boys. a girl who eats walks a cat who the cats who walk see. the nasty dogs bite the happy cat. hungry cats who quick girls chase chase dogs. the hungry cat who bites sees nasty cats. the girl feeds the dog. the happy boy walks the crazy dogs. crazy boys walk the cats who crazy Mary who walks feeds. Mary chases girls. the dog who walks sees the mangy cats who the nasty dogs see.

8.3 Ukážky z Gutenberg korpusu

Ukážky z korpusu tvorenom množinou dokumentov získaných z elektronickej knižnice Gutenberg. Ukážky tvoria úryvky náhodne vybraných súborov z tejto množiny:

názov súboru: 7abpt10.txt

"Two children came to the hillside. The one, older than his comrade, was Dimas, the son of Benoni. He was rugged and sinewy, and over his brown shoulders was flung a goat-skin; a leathern cap did not confine his long, dark curly hair. The other child was he whom they called the little Master; about his slender form clung raiment white as snow, and around his face of heavenly innocence fell curls of golden yellow. So beautiful a child I had not seen before, nor have I ever since seen such as he. And as they came together to the hillside, there seemed to glow about the little Master's head a soft white light, as if the moon had sent its tenderest, fairest beams to kiss those golden curls.

názov súboru: 7abpt10.txt

"Two children came to the hillside. The one, older than his comrade, was Dimas, the son of Benoni. He was rugged and sinewy, and over his brown shoulders was flung a goat-skin; a leathern cap did not confine his long, dark curly hair. The other child was he whom they called the little Master; about his slender form clung raiment white as snow, and around his face of heavenly innocence fell curls of golden yellow. So beautiful a child I had not seen before, nor have I ever since seen such as he. And as they came together to the hillside, there seemed to glow about the little Master's head a soft white light, as if the moon had sent its tenderest, fairest beams to kiss those golden curls.

názov súboru: 7abpt10.txt

"Two children came to the hillside. The one, older than his comrade, was Dimas, the son of Benoni. He was rugged and sinewy, and over his brown shoulders was flung a goat-skin; a leathern cap did not confine his long, dark curly hair. The other child was he whom they called the little Master; about his slender form clung raiment white as snow, and around his face of heavenly innocence fell curls of golden yellow. So beautiful a child I had not seen before, nor have I ever since seen such as he. And as they came together to the hillside, there seemed to glow about the little Master's head a soft white light, as if the moon had sent its tenderest, fairest beams to kiss those golden curls.

8.4 Ukážky z anglického wikipedia korpusu

Ukážky z korpusu tvorenom množinou dokumentov (článkov) získaných z anglickej verzie elektronickej encyklopédie wikipedia. Ukážky tvoria úryvky náhodne vybraných článkov z tejto množiny:

Joseph Papp (June 22, 1921 – October 31, 1991) was an American theatrical producer and director. Papp established The Public Theater in what had been the Astor Library Building in downtown New York (still located there as of 2010). "The Public,äs it is known, has many small theatres within it. There, Papp created a year-round producing home to focus on new creations, both plays and musicals. Among numerous examples of these creations were the works of David Rabe, Ntozake Shange's "For Colored Girls Who Have Considered Suicide When the Rainbow Is Enuf", Charles Gordone's No Place to Be Somebody (the first off-Broadway play to win the Pulitzer Prize), and Papp's production of Michael Bennett's Pulitzer-Prize winning musical, Ä Chorus Line.Ät Papp's death, The Public Theatre was renamed The Joseph Papp Public Theatre.

The corn earworm is considered to be a major agricultural pest, with a large host range encompassing not only corn, but also numerous other crop plants. Pesticides are one method by which corn earworm populations are controlled; however, since they have been widely used, the insects are resistant to many pesticides. The use of biological controls such as the bacterium Bacillus thuringiensis and various forms of nematodes is also common, although not without its own problems. Corn earworms are only variably vulnerable to the bacterium, and nematodes are only effective once the larvae have pupated and dropped to the ground.

In July 1948, the Squad learned of a plan to steal £250,000 of bullion from a warehouse at Heathrow Airport by drugging the guards. Squad officers replaced the guards and pretended to be drugged, with other officers stationed around the warehouse. When the thieves removed the keys to the safe from Detective Sergeant Charles Hewett the Squad announced their presence and a violent struggle ensued with many on both sides suffering serious injuries. The offenders received an average sentence of 10 years' imprisonment.

8.5 Ukážky zo slovenského wikipedia korpusu

Ukážky z korpusu tvorenom množinou dokumentov (článkov) získaných zo slovenskej verzie elektronickej encyklopédie wikipedia. Ukážky tvoria úryvky náhodne vybraných článkov z tejto množiny:

Jeden z najzvýznamnejšách slavistov Vatroslav Jagič (1838-1923) dokonca tvrdil, že Moravania ešte koncom 19. storočia v podstate hovorili takým istým jazykom ako Slováci. Do polovice 20. storočia bolo aj v Česku a na Slovensku bežné považovať v odborných textoch prinajmenšom obyvateľov Moravy na juhovýchodnej Morave za Slovákov, ich jazyk za nárečie slovenčiny - "moravskoslovenské nárečiaä ich územia aj označovať ako Slovensko či moravské Slovensko - dnes Slovácko (pozri napríklad

Ottuv slovník naučný heslo Slovensko či podrobne Slovenský náučný slovník heslo Slovenský jazyk, časť Slovenské nárečia).

O porozumenie vesmíru v najväčších možných mierkach sa snaží kozmológia, veda, ktorá vznikla z fyziky a astronómie. Počas druhej polovice 20. storočia viedol vznik pozorovacej kozmológie, tiež známej ako fyzikálna kozmológia, k rozdeleniu významu slova vesmír medzi pozorovacích kozmológov a teoretických kozmológov; tí prví opustili snahy pozorovať celé časopriestorové kontinuum, tí druhí sa o to stále pokúšajú v snahe nájsť najlogickejšie domnienky na vymodelovanie celého časopriestoru, navzdory extrémnym ťažkostiam v určení si empirických (založených na skúsenosti) obmedzení týchto špekulácií a vyhnúť sa tak skĺznutiu do metafyziky.

V roku 1601 už Herľany patrili panstvu Trebišov. Už v 17. storočí boli známe ako kúpeľné mesto, kam chodili ľudia z Košíc a Zemplína. Avšak na prelome 17. a 18. storočia sa obec opäť vyľudnila, jej obyvatelia sa vysťahovali do susedných dedín. V súpisoch z rokov 1715 a 1720 sa Herľany vôbec nespomínajú. Neuvádza ich ani konskripcii cirkví a farárov z roku 1746, ani v lexikóne obcí z roku 1773. Napriek tomu niektoré súčasné texty tvrdia, že v 18. storočí boli Herľany veľmi známe kúpeľné miesto, ktoré navštevovali aj cudzinci. Mohlo tak byť v závere storočia, pretože z roku 1808 sa dochoval písomný záznam, kde sú uvedené ako kúpele.

8.6 Ukážky z českého wikipedia korpusu

Ukážky z korpusu tvorenom množinou dokumentov (článkov) získaných z českej verzie elektronickej encyklopédie wikipedia. Ukážky tvoria úryvky náhodne vybraných článkov z tejto množiny:

Triton (nebo také Neptun I) je největší z měsíců planety Neptun. Byl objeven 10. října 1846 britským astronomem Williamem Lassellem. Je to jediný známý velký

měsíc ve Sluneční soustavě s retrográdním pohybem, což znamená, že obíhá v protisměru rotace své planety. S 2700 km v průměru se řadí na pozici sedmého největšího měsíce ve Sluneční soustavě. Kvůli retrográdní dráze a složení podobnému Plutu se předpokládá, že pochází z Kuiperova pásu. Zhruba 15–35 % Tritonu tvoří led. Jeho povrch se skládá ze zmrzlého dusíku a vrstvy ledu, která zřejmě skrývá pevné jádro z hornin a kovů. Jádro tvoří až dvě třetiny jeho celkové hmotnosti.

Obě verze 303 vycházely z verze 302 a skládaly se z rámu (1), tvořeného bočnicemi a vanou, a horním krytem (není zobrazen). Kryt verze 303 se od verze 302 lišil nízkou nástavbou kryjící otvor pro sání a cirkulaci vzduchu chlazení motoru. Pojezdové pásy (2), levý a pravý, každý se 48 články byly v horní části, na každé straně, vedeny přes 3 vratné kladky (3). Hmotnost Goliáše spočívala na malých pojezdových kolech (4), které byly odpružené. Hnací síla byla na pásy přenášena přes rozety (5) poháněné řetězovými převody od motorové části, umístěné pod kryty (6).

Nejstarší klínopisné záznamy tvořily hospodářské záznamy, účetní výkazy, inventáře majetku a seznamy zaměstnanců nebo obětí. Původní nepsané právo bylo s rozvíjející se společností potřeba přeměňovat na právo psané. První známé normativní dílo nechal vydat sumerský vládce Urukagina ve 24. stol. př. n. l. Mimo jiné např. zakázal, aby žena náležela dvěma mužům, a také zakázal vydírání. V díle se též nacházelo heslo „Nechť mocný neubližuje vdovám a sirotkům“, které později převzalo několik panovníků. Toto dílo většinou není považováno za zákoník. Do dnešní doby se nedochovalo.[13]