

Latviešu valoda semantiskajā tīmeklī

Guntis Bārzdiņš, Normunds Grūzītis, Renārs Kudiņš,
Gunta Nešpore, Andrejs Spektors

Anotācija. LU MII uzsāktais projekts SemTi-Kamols ir veltīts semantiskā tīmekļa tehnoloģiju un latviešu valodas mijiedarbības izpētei un attīstīšanai, ar mērķi sekmēt vismodernāko informācijas un valodas tehnoloģiju strauju un sekmīgu attīstību Latvijā. Izklāstītas ir gan teksta semantikas automatiskas izgūšanas principiālās nostādnes, gan arī jau šobrīd projektā paveiktais, attīstot labākās leksisko ontoloģiju *WordNet* un *OntoSem* arhitektūras idejas, pusautomātiski formalizējot latviešu valodas leksikonu (skaidrojošo vārdnīcu) un sastatot to ar ontoloģijas konceptiem. Rezultātu novērtēšanai ir izstrādāts eksperimentāls analizators, kas, izmantojot gramatikas likumus, leksikonu un ontoloģiju, spēj formāli reprezentēt ierobežotas sintakses latviešu valodas teikumu nozīmi.

Atslēgas vārdi: semantiskais tīmeklis, dabīgās valodas analīze, ontoloģija, leksikons, teksta nozīmes reprezentācija.

1. Ievads

Šobrīd tīmekļa pasaulē viena no aktuālākajām tēmām ir semantiskā tīmekļa (*Semantic Web*) tehnoloģiju un tā servisu teorētiska un praktiska attīstība [1; 15]. Tas ir turpinājums interneta un tā pārlūkprogrammu aizsāktajai sabiedrības datorizācijai un daudzu sabiedrības procesu pārejai uz tīklu tehnoloģijām. Semantiskā tīmekļa tehnoloģiju mērķis ir padarīt tiešsaistē pieejamo decentralizēto un lielākoties nestrukturēto informāciju saprotamu ne tikai cilvēkiem, bet arī automatizētām datorprogrammām (aģentiem), tādējādi paverot ceļu masveidīgai informatīvo procesu automatizācijai visdažādākajās tautsaimniecības nozarēs un sabiedrībā kopumā.

Šo mērķi pilnībā sasniegt šobrīd vēl nav iespējams; tam būtu vajadzīgs pilnvērtīgs mākslīgais intelekts. Tāpēc semantiskā tīmekļa ietvaros tiek mēģināts formalizēt tās informācijas attēlošanas un apstrādes jomas, kurās zinātne jau piedāvā piemērotus risinājumus. Viena no centrālajām problēmām ir dabiskās valodas tekstos novērst daudznozīmību, katram tekstā lietotajam vārdam (vai lielākai teksta vienībai) piekārtojot identifikatoru, URI¹, viennozīmīgi norādot uz lietoto vārda nozīmi skaidrojošajā vārdnīcā (ontoloģijā). Ja visi dabiskās valodas teksti tīmeklī būtu šādi anotēti, vismaz teorētiski, pavisam reāla kļūtu efektīva un precīza informācijas meklēšana, kā arī automatizēta vienkāršu secinājumu izdarīšana no tekstos attēlotajām zināšanām.

Aprakstīto metodi praktiski realizēt pagaidām ir ļoti grūti — nav izstrādātas ne piemērotas datorizētas skaidrojošās vārdnīcas (ontoloģijas), ne rīki, kas teksta autoram vai anotētājam ļautu pietiekami viegli piekārtot atbilstošos URI. Tomēr, lai arī pastāv praktiskas grūtības, semantiskā tīmekļa tehnoloģijas jau šobrīd tiek lietotas tādās nozarēs kā gēnu inženierija un farmakoloģija, kur lietoto terminu nozīmes precizitāte ir ļoti svarīga.

Semantiskā tīmekļa tehnoloģijas balstās uz ontoloģiju izstrādi dažādām cilvēka un sabiedrības darbības nozarēm [6]. Datorzinātnē, atšķirībā no ontoloģijas jēdziena filozofiskās izpratnes, ar to saprot (ierobežotas) pasaules modeli, kuru reprezentē strukturēts konceptu koks. Koncepti ir no valodas neatkarīgi jēdzieni, nevis vārdi. Atšķirībā no dabīgās valodas vārdiem katram konceptam ir tikai viena nozīme. Nosacīti ontoloģiju var salīdzināt ar mašīnlasāmu attiecīgās nozares terminu skaidrojošo vārdnīcu, kas ļauj viennozīmīgi lietot nozares informāciju. Lai arī daļa ontoloģiju ir veidotas universālas (tādējādi vispārīgas un virspusējas), reālai praktiskai lietojamībai ir nepieciešamas detalizētas, valodai un lietojumam (domēniem) specifiskas ontoloģijas.

LU Matemātikas un informātikas institūtā ir uzsākts Valsts pētījumu programmas projekts², tālāk saukts *SemTi-Kamols* (*Semantiskā tīmekļa projekts „Kamols”*), kura galvenais uzdevums ir nodrošināt to, lai Latvijā veidotās un sabiedriskā apritē ieviestās ontoloģijas būtu iespējami modernākas un kvalitatīvākas, sekmējot semantiskā tīmekļa tehnoloģiju plašu un strauju ieviešanu. Pakāpeniski tiek attīstītas ar semantiskā tīmekļa praktisko ieviešanu saistītās jomas, sākot ar latviešu valodas datornodrošinājumu un beidzot ar praktiskiem semantiskā tīmekļa izstrādes un lietošanas rīkiem. Viens no pirmajiem uzdevumiem ir iemācīties izveidot tādu latviešu valodas skaidrojošo vārdnīcu (ontoloģiju), kas būtu saprotama ne tikai cilvēkiem, bet arī mašīnai, tuvākajā nākotnē sniedzot ierobežotas valodas analīzes un sintēzes iespējas. Rezultāti pavērs jaunas pētnieciskas un praktiskas iespējas kā datorzinātnē, tā arī valodniecībā un citās saistītās nozarēs. Kā praktisku iespējamo lietojumu var minēt kvalitatīvu tulkošanas sistēmu izveidi, kā arī formālo (juridisko) tekstu semantisku pārbaudi un anotēšanu.

2. Zināšanu attēlošana datorsistēmās

Nedaudz precizēsim, kas tiek saprasts ar diviem centrālajiem valodas un pasaules zināšanu avotiem — leksikonu un ontoloģiju — un kā šīs zinību bāzes tiek realizētas.

Leksikons

Leksikons tradicionāli tiek definēts kā vārdu un izteicienu krājums, kas raksturīgs valodai, kādai sociālai grupai, atsevišķam indivīdam, arī tekstam. Leksikons ir arī vārdnīca: vārdu saraksts ar informāciju par šo vārdu nozīmi un lietošanu. Alternatīvu definīciju piedāvā zināšanu inženierijas klasiķis Dž. F. Sova: leksikons ir tilts starp valodu un zināšanām, kas ir izteiktas šajā valodā [14]. Tātad leksikons nav vien vārdu saraksts, tas reprezentē zināšanas — katru vārdu cilvēks uztver savu leksisko un pasaules zināšanu kontekstā. Līdz ar to par leksikonu var runāt kā par strukturētu vārdu krājumu, kas tiek organizēts, klasificēts cilvēka prātā ar viņa zināšanu palīdzību. Turklāt zināšanu modeļi cilvēkiem ir līdzīgi.

Leksikona reprezentācijā principā ir izšķiramas divas pieejas:

1. Tradicionālo vārdnīcu organizācijā par pamatu tiek ņemti vārdi (leksiskas vienības), tiem piekārtojot nozīmju definīcijas (t. sk. attieksmes ar citiem vārdiem) dabīgās valodas formā. Šādi leksikona struktūra, zināšanas, tiek reprezentētas netieši — tās ir uztveramas cilvēkam, bet ne mašīnai. Cilvēks, lasot definīcijas un piemērus, ar prāta induktīvajām un deduktīvajām spējām apzināti vai neapzināti būvē nozīmju taksonomiju: hierarhiski vai citādi saistītu leksikalizētu jēdzienu tīklu. Lai mašīna varētu veikt to pašu intelektuālo procesu, zināšanas nepieciešams aprakstīt tieši un formāli.
2. Semantisku, leksisku tīklu organizācijas pamatā ir vārdu nozīmes un to saistība ar citām nozīmēm. Dažādu vārdu nozīmes var izteikt vienu un to pašu jēdzienu, nozīmes var grupēt pēc šiem jēdzieniem. Līdz ar to var teikt, ka relācijas pastāv starp jēdzieniem, nevis atsevišķām vārdu nozīmēm.

Ja vārdnīcā vārda nozīme tiek meklēta, zinot vārda formu, tad šādā semantiskā tīklā, zinot interesējošā jēdziena semantiskās īpašības, iespējams atrast visus atbilstošos vārdus. Savukārt nozīmju skaidrojumi tiek „aprakstīti” ar relāciju (relāciju vērtību) palīdzību.

Ontoloģija

Vienā no biežāk citētajām ontoloģijas jēdziena definīcijām teikts, ka tā ir kopējas konceptualizācijas formāla (tieša, precīza) specifikācija [13], kur konceptualizācija ir parādības (priekšmetu apgabala) abstrakts modelis, kurā identificēti parādības jēdzieni;

tiešs — jēdzieni un to lietošanas ierobežojumi ir tieši definēti; formāls — ontoloģijai jābūt mašīnlasāmai; kopējs — ontoloģija aptver objektīvas zināšanas, kas ir pieņemamas kādas grupas ietvaros. Tiesa, attiecībā uz zināšanām (semantiku) ir jāņem vērā, ka absolūta objektivitāte un līdz ar to vienprātīga informācijas interpretēšana praktiski nav iespējama. Uzskatāmi to parāda dažādie atšķirīgie augšējo līmeņu un domēnspecifisko ontoloģiju standarti, kas ir viena no lielākajām problēmām zināšanu inženierijas un semantiskā tīmekļa kontekstā. Nedaudz jāpapildina arī dotās definīcijas skaidrojums: ontoloģijai ir jābūt ne tikai mašīnlasāmai, bet arī mašīnai saprotamai — jāvar veikt spriedumus un izvedumus.

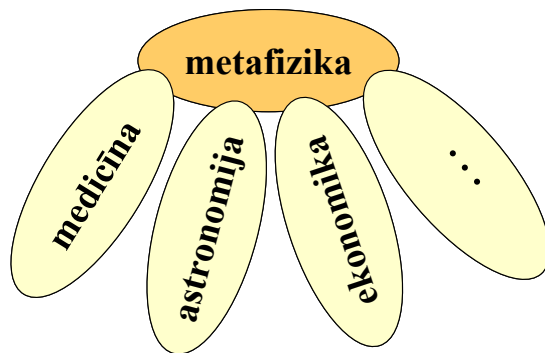
Tipiskie ontoloģijas komponenti ir šādi:

1. Klases (koncepti), kas parasti tiek organizētas taksonomijā, izmantojot *IS-A* (apakšklase-virsklase) relācijas. Matemātiski šādu taksonomiju var dēvēt par koku, tikai ar izņēmumu, ka ir pieļaujama tā zaru saaugšana — daudzkārsā mantošana. Klasei nereti var būt nepieciešamība mantot īpašības no konceptiem, kas atrodas vairākos ontoloģijas zaros, piemēram, jēdziens *ĀBOLS* manto īpašības gan no jēdziena *AUGLIS* (šķirne), gan no *PĀRTIKAS-PRODUKTS* (garša, kalorijas).
2. Relācijas (īpašības), kas atspoguļo klašu mijiedarbību: konceptu vispārināšanas, agregācijas, kauzativitātes, aģentativitātes utt. attieksmes. Principā ar relāciju un to ierobežojumu palīdzību tiek definēti koncepti. Ierobežojumi tiek izteikti, norādot definīcijas un vērtību apgabalu, kuri savukārt ir klases. Piemēram, īpašībai *DARBĪBAS-IZJUTĒJS*³ definīcijas apgabals ir koncepts *NOTIKUMS*, bet vērtību apgabals — *DZĪVA-BŪTNE*, jo mēs pieņemam, ka izjust var tikai kādu notikumu un izjutējam ir jābūt dzīvai būtnei. Īpašību vērtību apgabalus var vēl tālāk ierobežot, piemēram, *KONCERTS* ir *NOTIKUMS* apakšklase un šajā gadījumā īpašības *DARBĪBAS-IZJUTĒJS* vērtību apgabalu var ierobežot ar konceptu *CILVĒKS*, kas ir *DZĪVA-BŪTNE* apakšklase.
3. Aksiomas — modelē vienmēr patiesus izteikumus. Izmantojot kopu teorijas elementus, visu ontoloģijas formālo struktūru var definēt ar aksiomu palīdzību, kas dod iespēju ontoloģijā veikt loģiskus spriedumus, meklēt pretrunas u. tml.
4. Instances — konceptu pārstāvji: konkrēti reālās pasaules objekti.

Ontoloģijas, kurās izmantots tikai 1. un 2. komponents, tiek sauktas par „vieglsvara” ontoloģijām [13]. Lai ontoloģija būtu praktiski lietojama valodas apstrādes vai mākslīgā intelekta problēmu risināšanā, tai jāsaturs lielu apjomu (vismaz ar kārtu 100 000) konceptus un semantiskās relācijas [4].

Šobrīd ir pieejama virkne dažādu gatavu ontoloģiju. Dažas no tām tiek praktiski izmantotas nozīmīgos projektos, bet citas arhitektūras u. c. trūkumu dēļ nav guvušas atsaucību un ir faktiski „mirušas”. Esošās ontoloģijas atšķiras ne tikai pēc atkarības vai neatkarības no valodas, konceptualizācijas niansēm un apjoma, bet arī pēc to mērķa un izstrādātāju uzskatiem (subjektivitātes). Tajā pašā laikā mēģinājumi radīt maksimāli izsmeļošu, universāli lietojamu „viss vienā” ontoloģiju praksē nav devuši cerētos rezultātus, jo dažādu domēnu (mikropasaļu) modeļu apvienošana kopējā makropasaules modelī veicina sākotnējam mērķim pretēju efektu — jēdzienu daudznozīmības rašanos.

Secinājums: ontoloģijas struktūra ir subjektīva un atkarīga no risināmās problēmas rakstura. Vispārīgāko ikdienas ontoloģiju var salīdzināt ar metafiziku, no kuras tālāk var atvasināt konkrētus priekšmetu apgabalus: apakšontoloģijas medicīnai, astronomijai, ekonomikai utt. (sk. 1. attēlu).



1. attēls — no vispārējās un vispārīgas ontoloģijas var atvasināt domēnspecifiskas ontoloģijas (mantojot konceptus un to īpašības).

Apakšontoloģijas pārsvarā lieto šaurākus jēdzienus, piemēram, [pilsētas] plāns, [kara] plāns, [uzņēmējdarbības] plāns. Jēdziena robežas ir patvaļīgas, tās nosaka lietojuma noderība. Turklāt tuvojoties pa ontoloģijas koku konkrētākām nozīmēm, parādās arvien vairāk no katras valodas atkarīgu jēdzienu. Arī tuvu radniecīgās valodās jēdzienu (arī vārdu nozīmju) robežas mēdz atšķirties, kaut arī tie atspoguļo vienu un to pašu reālo īstenību, piemēram, latviešu valodas vārdam *zils* krievu valodā atbilst vārdi *синий* un *голубой*; bet latviešu valodas vārdiem *nākt* un *iet* atbilst tikai viens lietuviešu valodas vārds *eiti*.

SemTi-Kamola projektā tiek attīstītas labākās jau pieminēto *WordNet* un *OntoSem* arhitektūru idejas un iestrādes, tādēļ turpinājumā nedaudz sīkāk par katru no šo leksisko ontoloģiju arhitektūrām.

WordNet

Leksisku tīklu idejas pirmsākumi ir meklējami vismaz pirms 20 gadiem. Pirmais ievērojamais resurss — *WordNet* —, kura arhitektūras iedvesmas avots ir psiholingvistikas teorijas par cilvēka leksisko atmiņu, tika izstrādāts Prinstonas universitātē. *WordNet* tiek definēts kā leksikalizētu jēdzienu semantisks tīkls: leksiskie koncepti tiek reprezentēti ar sinonīmu kopu (*synsets*) palīdzību, savukārt dažādas semantiskās relācijas saista sinonīmu kopas: hiponīmija/hiperonīmija, meronīmija/holonīmija, ietveršana u. c. [17]. Līdz ar to jēdzienu var uztvert kā „konteineru”, kuru izsaka tā saturs un saites uz citiem jēdzieniem. Matemātikas terminos runājot, *WordNet* var aplūkot kā orientētu grafu, kas nesatur ciklus, līdz ar to šādas datubāzes ekspluatācijā iespējams izmantot grafu teorijas metodes.

Bez pamatrelācijām dažādos *WordNet* paplašinājumos ir ieviesta virkne citu relāciju, taču to relāciju skaits, ar kurām parasti operē cilvēks, nav liels. Turklāt viena no vēlamajām zinību bāzes īpašībām ir maksimāli augsta konceptu savienojumu pakāpe, vienlaikus saglabājot minimālu (būtisko) relāciju veidu kopu. Vēl vairāk — lai zinību bāze būtu vispārēji lietojama (leksiskas ontoloģijas mērķis), relācijām vienmēr ir jābūt spēkā [4].

Lietvārdi parasti veido lielāko daļu (70–80%), tiem seko verbi (15–25%). Tas ir pašsaprotami, jo šīs ir galvenās vārdšķiras, kas atspoguļo cilvēka zināšanas. Ar lietvārdiem tiek izteiktas zināšanas priekšmetiskā formā (t. sk. vairums terminu), bet verbi ir īpaši izceļami ar spēju veidot izteicējus — teikuma uzbūves pamatelementus.

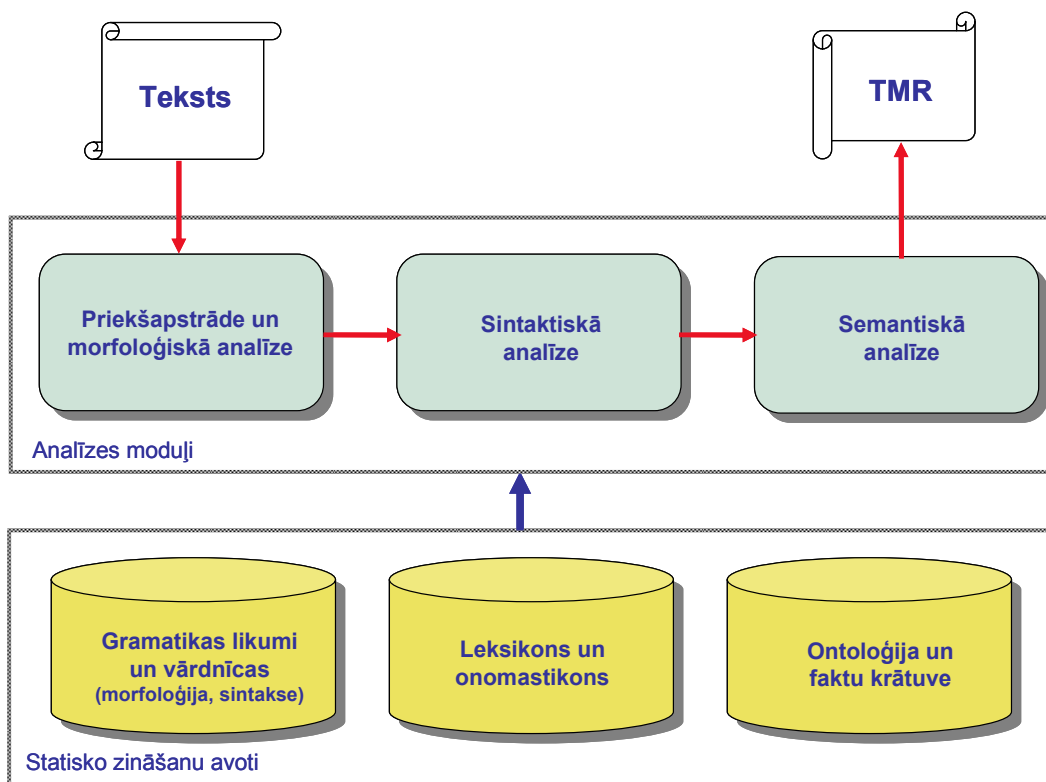
Šobrīd *WordNet* „revolūcija” ir sasniegusi plašus apmērus⁴, un tas tiek visnotaļ aktīvi izmantots dabīgās valodas apstrādes pētniecībā un praksē. *WordNet* tipa semantisku tīklu ir pamats uzskatīt par „vieglsvara” leksisku ontoloģiju, kurā no valodas neatkarīgās un atkarīgās daļas atbilst vienam modelim un ir vienkopus — noteikta robeža starp tām nav novilkta. Tas šādu modeli kopumā padara ļoti valodatkarīgu. Turklāt ar taksonomiju vien nepietiek efektīvai valodas apstrādei un nozīmes analīzei/reprezentācijai.

OntoSem

Ontoloģiskā semantika ir teorija par dabīgās valodas pierakstā izteiktu jēgu un valodas apstrādes tehnikām, kas patvaļīga teksta nozīmes izgūšanai, tās formālai reprezentēšanai (TMR — *Text Meaning Representation*) un spriešanai par zināšanām, izteiktām vai atvasināmām šajā tekstā, kā centrālo resursu izmanto formalizētu pasaules modeli [12].

Ļoti virspusēji *OntoSem* arhitektūru var raksturot šādi (sk. 2. attēlu):

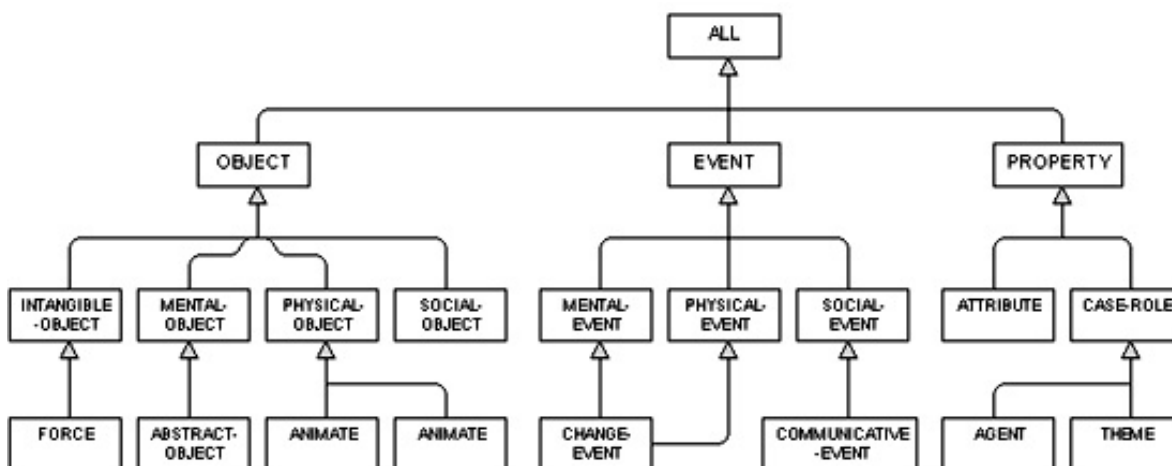
1. Statisku zināšanu avoti: ontoloģija, valodspecifisks leksikons (saista ontoloģiju ar dabīgo valodu), onomastikons (leksikona apakškopa — īpašvārdi un to apzīmētie fiziskās pasaules objekti) un faktu krātuve.
2. Tekstu apstrādes moduļu kopa: morfosintaktiskais un semantiskais analizators, teksta ģenerators u. c.



2. attēls — *OntoSem* tekstu analīzes modelis; valodneatkarīgie avoti ir ontoloģija un faktu krātuve.

OntoSem modelis ar plašu sintaktisko un semantisko struktūru un nozīmju procedūru palīdzību ļauj aprakstīt un izskaitļot valodas (leksikona) dinamiskos aspektus, kas ir ļoti būtiski reālā, efektīvā tekstu analīzē un daudznozīmības risināšanā. Šobrīd ontoloģijas apjoms ir ~6 000 konceptu (sk. 3. attēlu), no kuriem katrs ir aprakstīts ar vidēji 16 īpašībām (*properties*), taču iespējamo pazīmju skaits ir mērāms simtos [10]. Ontoloģiskās semantikas

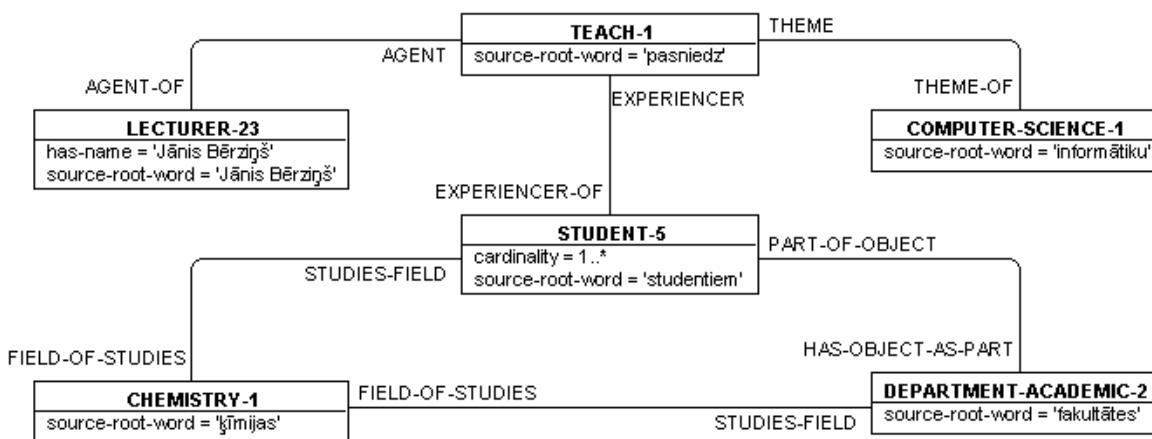
teorija bez angļu valodas tiešā veidā ir realizēta arī spāņu un poļu valodai, kas deva papildu uzticību tās izvēlē pielāgošanai latviešu valodai.



3. attēls — neliels, vienkāršots *OntoSem* ontoloģijas koka fragments. *OntoSem* ontoloģijā angļu valoda tiek izmantota semantiskas metavalodas funkcijā.

TMR var uzskatīt par teikumā izteikto konceptu un īpašību instanču diagrammu (sk. 4. attēlu). Ja ontoloģijā ir informācija par to, kā lietas pasaulē var pastāvēt, tad faktu krātuvē atrodas no teksta(-iem) izgūtās TMR: notikumi, kas ir norisinājušies, norisinās vai norisināsies kādā laika posmā un objekti, kas noteiktās attieksmēs eksistē šajos notikumos.

Kā jau minēts, visas ontoloģijā un līdz ar to arī faktu krātuvē aprakstītās zināšanas var izteikt kā aksiomu kopas. Tas paver iespējas loģisku izvedumu veikšanai faktu krātuves datos un būtībā ir kontekstuālā analīze, kas ir neapšaubāma *OntoSem* arhitektūras priekšrocība.



4. attēls — TMR piemērs teikumam „Jānis Bērziņš pasniedz informātiku ķīmijas fakultātes studentiem” [20].

OntoSem arhitektūrā atšķirībā no *WordNet* vispārīgās (pasaules) un leksiskās zināšanas ir strikti nodalītas; *OntoSem* nav „vieglsvara” ontoloģija, arī relāciju klāsts un izvedumu iespējas ir nesalīdzināmi lielākas. Konteksta analīzē *OntoSem* mehānismiem ir neapšaubāmas priekšrocības. *OntoSem* arī ļauj pietiekami labi leksikonā aprakstīt vārdu darināšanas zināšanas, lai no TMR ļautu ģenerēt tekstu dabīgā valodā, līdz ar to tas ir izmantojams viskomplicētākajām vajadzībām, tādām kā mašīntulkošana.

Zināšanu attēlošana semantiskajā tīmeklī

Zināšanu formalizēšana semantiskā tīmekļa pasaulē balstās uz trīs „vaļiem”: XML (*eXtensible Markup Language*), tās paplašinājumu RDF (*Resource Description Framework*) un RDF paplašinājumu OWL (*Web Ontology Language*). Mērķis: pārveidot vispasaules zināšanas, kas ir decentralizēti un nesavienojami „izkaisītas” tīmeklī ne vien mašīnlasāmā, bet arī mašīnai saprotamā formātā.

XML ļauj definēt savas vārdu telpas (vārdnīcas) un ar strukturālu likumu (shēmu) palīdzību ierobežot pieļaujamo elementu sakārtojumu dokumentā. Piemēram, teikumu „*Jānis Bērziņš pasniedz matemātiku*” XML formā varētu attēlot šādi:

```
<kurss nosaukums="matemātika">  
  <pasniedzējs vārds="Jānis Bērziņš"/>  
</kurss>
```

Taču XML nepiedāvā nekādus formālus līdzekļus vārdnīcas elementu semantikas aprakstīšanai.

RDF pamatā (atšķirībā no XML praktiski neierobežotās sintakses) ir ideja aprakstīt jebkādu informāciju ar trijnieku apgalvojumiem: jebkuras zināšanas ir iespējams izteikt <subjekts, predikāts, objekts> formā. Ikviens no trijnieka elementiem ir resurss; izņēmums: objekts drīkst būt literālis, kas pats sevi definē un nav resurss. Resurss ir lieta, koncepts, abstrakcija, kam var piešķirt un kam ir piešķirts unikāls vārds — identifikators URI (*Unified Resource Identifier*) formā. Subjekts ir resurss, kas tiek aprakstīts, predikāts ir īpašība, bet objekts — tās vērtība [19].

OWL ir šībrīža augstākais sasniegums semantiskā tīmekļa valodu jomā, kas ļauj aprakstīt sarežģītas attiecības starp resursu klasēm, piemēram, apvienojumus, šķēlumus, papildinājumus utt. Tās apakškopa OWL DL (*OWL Description Logic*) definē uz matemātiskās loģikas pamatiem balstītu izteiksmes līdzekļu kopu, kas ir reizē pietiekami

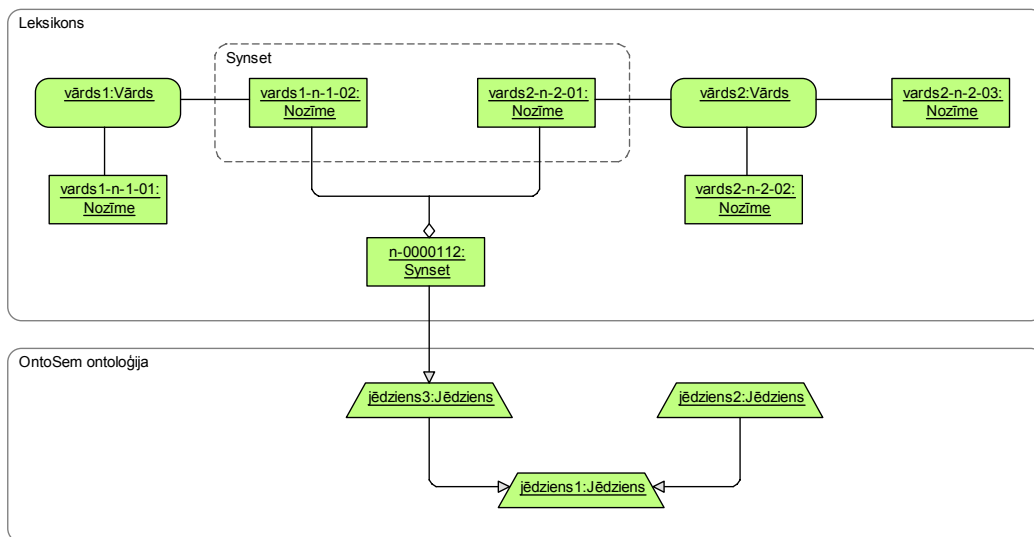
bagāta reālu ontoloģiju aprakstam un pakļaujas arī pilnīgi formāliem matemātiskās loģikas spriešanas likumiem [18]. Tieši šī duālā OWL DL valodas daba rada līdz šim vēl nepieredzētas iespējas gan ontoloģiju pareizības automatizētā pārbaudē, gan to izmantošanā formālu un sarežģītu spriedumu veikšanā par ontoloģijā attēlotajiem datiem. OWL DL balstās uz tā saucamo atvērtās pasaules pieņēmumu, t. i., lietas, kuru (ne)eksistenci nevaram pierādīt, uzskatām par potenciāli eksistējošām, t. i., mums trūkst zināšanu, lai to (ne)eksistenci pierādītu. OWL DL ir veidots, balstoties uz deskriptīvo loģiku (*description logic*), taču programmas, kas pilnībā realizētu tajā iespējamus spriedumus vēl nav izstrādātas. Ir daži rīki (*RACER*, *Pellet*, *FaCT*), kā arī ontoloģiju redaktori (*Protégé*), kas spēj klasificēt ontoloģijā esošās klases, balstoties uz tās aprakstošajiem likumiem, taču asociāciju vērtību ierobežojumus, piemēram, „*visas vērtības no...*”, „*kardinalitāte*” u. c. šie rīki vēl nespēj apstrādāt. OWL DL tiek lietots arī *SemTi-Kamola* projektā, un nodrošina tā rezultātus.

3. Latviešu valodas analīze un semantiskā modelēšana

SemTi-Kamola arhitektūras pamatā ir ņemta *OntoSem* ideoloģija un ontoloģija, taču tā ir būtiski modificēta un kvalitatīvi uzlabota, piemēram, strukturējot īpašības, padarot to elastīgāku un paplašināmāku, kā arī pārveidojot to atbilstoši OWL DL prasībām. *OntoSem*, protams, reprezentē ierobežotu pasaules modeli, taču tas ir samērā veiksmīgi izvēlēts; par to *SemTi-Kamola* īstenotāji pārliecinājās, cenšoties ontoloģijā attēlot latviešu valodas vārdnīcā 1000 biežāk lietotos lietvārdus, darbības vārdus un īpašības vārdus (to nozīmes), kas tika izgūti no nozīmju definīcijām — aptuveni 90% atbilda kādam no *OntoSem* ontoloģijas konceptiem. Atlikušie 10% nozīmju norāda uz ontoloģijas trūkstošajiem zariem (tie pamatā saistās ar dažāda veida grupējumiem un sakārtojumiem — *saistība*, *sakopojums*, *sastāvs*, *uzbūve*, *krājums* u. tml.), pie kuru pievienošanas tiek strādāts. Turklāt ontoloģija tiek ne tikai paplašināta, bet arī „attīrīta”, piemēram, no īpašībām, kuru semantika savstarpēji pārklājas, paturot tikai būtiskās īpašības.

Latviešu valodas leksikons tiek organizēts no ontoloģijas neatkarīgā modelī, veidojot tam iekšēju taksonomiju, lai konceptuālā līmenī nepazustu un „neizlīdzinātos” valodas nianšes. Leksikona modelis ir ievērojami vienkāršots un pielāgots fleksīvajai latviešu valodai, informācija par sintaktiskajām struktūrām, kurās vārds kādā nozīmē varētu

iesaistīties, ir pārnesta uz ontoloģiju, lai leksikonā nav jāiekļauj locījumu semantika, padarot to smagnēju un grūti paplašināmu. Turklāt leksikona modelis tiek projektēts un īstenots, izmantojot arī labākās *WordNet* tipa „ontoloģijas” idejas, kas nodrošinās savienojamību gan ar *OntoSem*, gan *WordNet* (sk. 5. attēlu).



5. attēls — leksikons tiek organizēts izmantojot *WordNet* mehānismus; piesaiste ontoloģijai notiks ar sinonīmu kopu palīdzību.

Gan ontoloģijas konceptu un īpašību, gan leksikona nozīmju reprezentēšanai tiek izmantota OWL ontoloģiju valoda. Līdz ar to vārdu nozīmes ir resursi, kurus nepieciešams unikāli identificēt. Pilns URI identifikatora šablons un piemērs:

$$\text{URI} = \langle \text{leksikona URL}^5 \rangle + \langle \text{vārdšķira} \rangle + \langle \text{vārda identifikators} \rangle + \langle \text{homonīma indekss} \rangle + \langle \text{nozīmes numurs} \rangle + \langle \text{pamatforma} \rangle + \langle \text{sinonīms} | \text{hiponīms} | \text{hiperonīms} \rangle$$

<http://www.semti-kamols.lv/lexicon/sense#1-12786-2-1-zeme-planeta>

Pēc būtības nozīmi unikāli identificējošā daļa ir: vārdšķira, vārda identifikators, homonīma indekss un nozīmes numurs.

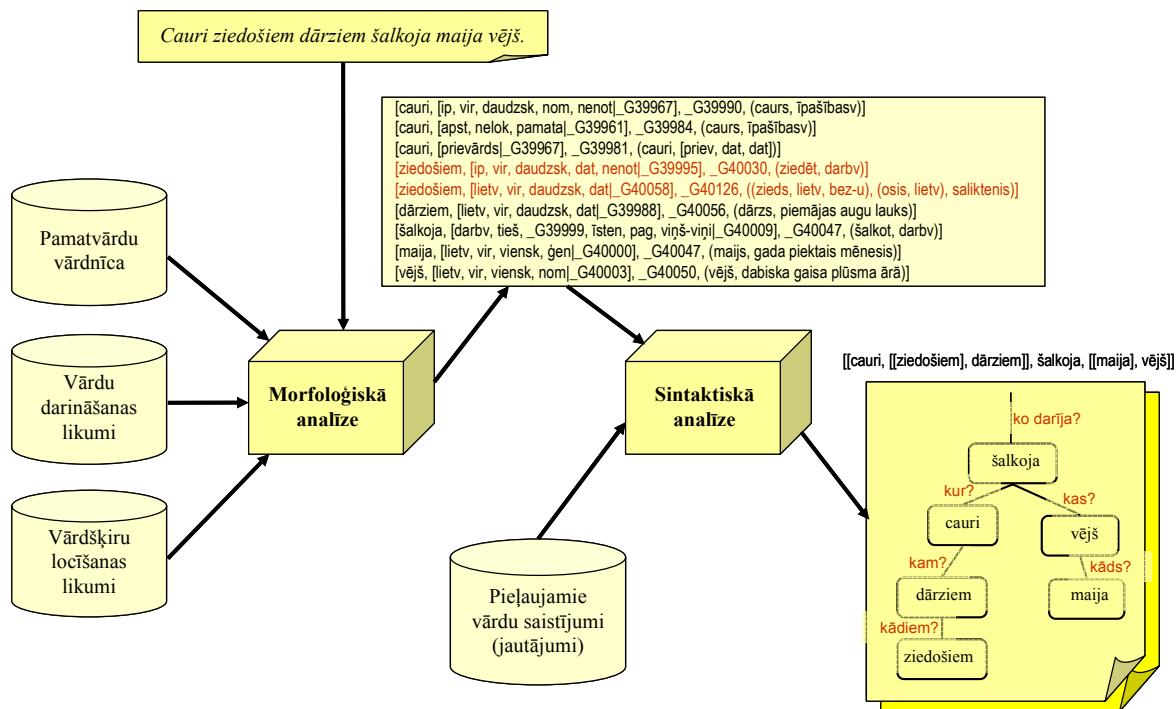
Šādā formalizācijas līmenī dabiskās valodas un OWL valodas tekstu struktūra jau kļūst ļoti līdzīga. Principiālā atšķirība ir tikai tā, ka cilvēki dabiskajā valodā ērtības labad optimizē gan nozīmju URI, gan sintakses pierakstu: no formālā viedokļa tie tiek pārāk saīsināti, radot daudznozīmību. Tātad, ar ieviesto URI palīdzību dekonstruējot vārdus (daudznozīmību), ikvienu latviešu valodas teikumu (kas atbilst pašreizējiem sintaktiskajiem un leksiskajiem ierobežojumiem) ir iespējams pārvērst vienā vai vairākās viennozīmīgās TMR, pierakstītās OWL formātā (sk. 6. attēlu).

„Māte māca meitas.”
<pre> <teach rdf:ID="teach_726"> <tense> <present rdf:ID="present_967"/> </tense> <agent> <parent rdf:ID="parent_155"> <gender> <female rdf:ID="female_718"/> </gender> <count> <singular rdf:ID="singular_263"/> </count> </parent> </agent> <experiencer> <offspring rdf:ID="offspring_918"> <gender> <female rdf:ID="female_719"/> </gender> <count> <plural rdf:ID="plural_200"/> </count> </offspring> </experiencer> </teach> </pre>

6. attēls — teikuma vienkāršots TMR OWL formā.

LU MII līdzšinējās iestrādes latviešu valodas morfoloģiskajā un sintaktiskajā analīzē [11] tiek sekmīgi attīstītas semantiskā tīmekļa vajadzībām: izmantojot ārēji norādītus (papildināmus) gramatikas likumus (teikumu konstrukciju vārdnīcu), kas uzdoti bezkonteksta gramatikas formā, analizators prot ierobežotas sintakses teikumus pārvērst konceptuālo grafu formā, kas ir ekvivalenta RDF trijnieku apgalvojumiem (sk. 7. attēlu). Tālāk, izmantojot ontoloģiju un pretstatīto leksikonu, sākotnējais dabiskās valodas teikums tiek automātiski translēts TMR OWL formā.

Kā jau minēts, no potenciāli daudznozīmīgiem teikumiem (kāda ir lielākā daļa dabiskās valodas teikumu) automātiskas analīzes rezultātā tiek iegūtas, iespējams, vairākas TMR. Daudznozīmības risināšana dabīgajā valodā jebkurā gadījumā paliek aktuāls jautājums, taču pakāpeniskā analīze ar katru nākamo soli ievērojami samazina iespējamo TMR skaitu: morfoloģisko daudznozīmību ierobežo sintaktiskā analīze, savukārt no potenciālajiem sintaktiskajiem kociem, nederīgos „atmet” ontoloģijas ierobežojumi. Piemēram, teikumam „kapteinis māca koku” morfoloģiski un sintaktiski pilnīgi pieņemamas ir divas interpretācijas („kapteinis māca koka pagali” vai „kapteinis māca kuģa pavāru”), taču *OntoSem* ontoloģiskie ierobežojumi pieļauj tikai otro interpretāciju (jo mācīšanas notikuma DARBĪBAS-IZJUTĒJS relācijai vērtību apgabals var būt tikai dzīva būtne).



7. attēls — teikumu gramatiskās analīzes process.

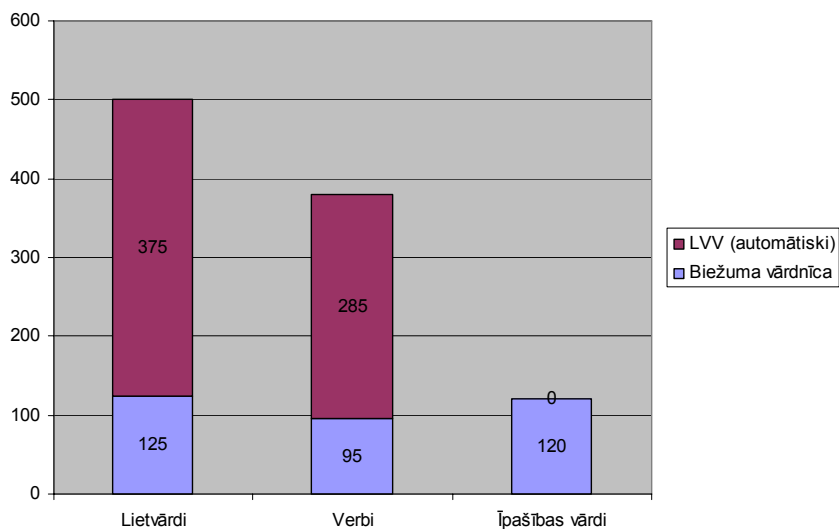
Viens no projekta mērķiem ir sakārtot un latviskot „metafizikas” līmeni (sk. 1. attēlu) — radīt elastīgu pamatu tālākai attīstībai un domēnu paplašinājumiem. Tā kā *OntoSem* pretendē uz šādu līmeni (ņemot vērā minētās nepilnības, ar kuru novēršanu *SemTi-Kamols* nodarbojas), uzdevums principā reducējas uz biežāk lietoto latviešu valodas vārdu nozīmju adekvātu attēlošanu ontoloģijā. Tomēr ir jāatceras, ka ontoloģija ir metavaloda, kas apraksta jēdzienus, nevis valodu. Līdz ar to latviskošana nozīmē latviešu valodas leksikona formalizēšanu un sastatīšanu ar ontoloģijas jēdzieniem (sk. 5. attēlu). Viens no visērtāk izmantojamiem avotiem, kas (ierobežoti) atspoguļo leksikonu, ir skaidrojošās vārdnīcas. *SemTi-Kamola* projektā šim nolūkam tika izraudzīta Latviešu valodas vārdnīca (LVV) [9], jo tā ir pieejama mašīnlasāmā formā, tā pamatā atspoguļo mūsdienu valodu, turklāt LVV apjoms (~25 000 šķirķļu) darba sākumam šķiet optimāls. Tiesa, latviešu valodas skaidrojošo vārdnīcu izvēles iespējas ir niecīgas: Latviešu literārās valodas vārdnīca [7], kuras atvasinājums ir LVV, kā arī dažādas terminoloģijas un svešvārdu vārdnīcas.

Jāpiebilst, ka katra vārda nozīme tiek uztverta kā atsevišķa analīzes vienība, tātad atšķirīgas viena vārda nozīmes var tikt piesaistītas atšķirīgiem konceptiem, citādi nebūtu iespējams runāt par sinonīmijas, hiponīmijas vai citādām semantiskām attiecībām starp

leksikona vienībām. Tāpēc tiek runāts par 1000 biežāk lietoto vārdu nozīmēm (leksiski semantiskajiem variantiem), nevis vārdiem.

Izsmeltošas plaša pārklājuma leksiskas ontoloģijas manuāla izveide (sastatīšana) ir ļoti laikietilpīgs process, kas prasa ievērojamus cilvēkresursus. Lai arī procesa automatizācijas iespējas ir daudzējādā ziņā ierobežotas un apgrūtinātas, pat necīga automatizācija te ļauj ietaupīt simtiem cilvēkstundu (sk. 4. nodaļu).

Jāpiemin — lai iegūtu objektīvu valodas vārdu biežuma sarakstu, ir nepieciešams apjomīgs, sabalansēts un morfoloģiski marķēts tekstu korpuss; tāds latviešu valodai (vismaz publiski) nav pieejams. Lai sabalansētu vispārīgo vārdu kopu, papildus biežuma noteikšanai LVV tika izmantoti dati no 1973. gada Latviešu valodas biežuma vārdnīcas [8], atmetot tajā laikā sastopamos padomju ideoloģijai raksturīgos vārdus: tika noteikts, ka 25% no katrai vārdšķirai atvēlētās vārdu daļas ir jānāk no biežuma vārdnīcas, kamēr pārējie 75% tika ņemti no LVV definīciju tekstos biežāk sastopamajiem vārdiem. Rezultātā apstiprinājās pieņēmums: vārdu biežums valodā un vārdnīcas definīcijās daļēji pārklājas (mūsu gadījumā — vidēji par 50%), tomēr vārdnīcā biežāk lietoto vārdu sarakstā sastopami daudzi tādi vārdi, kas valodā kopumā lietoti daudz retāk (piemēram, *kopums*, *grupa*, *atbilstība*, *spēja*). Par sākuma etalonu tika noteikta 1000 vārdu nozīmju kopa, kuras sadalījums ir proporcionāls vārdnīcā skaidroto vārdu sadalījumam pa vārdšķirām (sk. 8. attēlu).



8. attēls — atlasāmo vispārīgo vārdu sadalījums pa vārdšķirām un izguves avotiem.

Subjektivitātes, kļūdu un citu problēmu novērtēšanai kopas elementi (nozīmju URI un atbilstošās definīcijas) tika sadalīti starp dalībniekiem tā, lai katru vārda nozīmi ar leksikonu sastatītu vismaz divi cilvēki. Iegūtā rezultāta fragments:

Koncepts	Nozīmes URI	Nozīmes definīcija
ALLOY	#1-10041-0-1-sakausējums-savienojums	Vairāku vielu (parasti metālu) savienojums, ko iegūst kausējot.
FORCE	#1-10859-0-1-spēks-lielums	Fizikāls lielums, ar ko raksturo ķermeņu mehānisko mijiedarbību.
PLANET	#1-12838-0-1-zeme-planēta	Planēta, ko apdzīvo cilvēki.
EARTH-SURFACE	#1-12838-0-2-zeme-kārta	Šīs planētas garozas virsējā kārtā.
CREATE-ARTIFACT	#2-3376-0-1-izveidot-izgatavot	Veidojot izgatavot (priekšmetu); veidojot piešķirt formu, veidu.
CHANGE-EVENT	#2-5668-0-1-piepildīt-padarīt	Padarīt pilnu (ar ko); aizņēmot (ko).
WORK-ACTIVITY	#2-5420-0-1-paveikt-izdarīt	Izdarīt, padarīt.
LEARN	#2-3907-0-1-mācīties-izglītoties	Iegūt zināšanas un prasmi; izglītoties; skoloties; studēt.
TEMPORAL-UNIT	#1-7457-0-1-nakts-posms	Laika posms no vakara līdz rītam.

9. attēls — ar ontoloģiju sastatīta leksikona fragments.

Viens no secinājumiem — vispārīgu nozīmju sastatīšana ir sarežģīts un darbietilpīgs uzdevums, automātiski korektus rezultātus kopumā nav iespējams iegūt, tie visi ir manuāli jāpārskata. Tajā pašā laikā šis sastatīšanas rezultāts ir ļoti vērtīgs lingvistisks resurss, kas padara iespējamu gan dabiskās valodas tekstu automatizētu translēšanu atbilstoši *OntoSem* ontoloģijai, gan uz TMR semantiskā tīmekļa valodā OWL, gan uz tekstu jebkurā citā dabiskā valodā, kurai izveidots līdzīgs sastatījums, izmantojot to pašu *OntoSem* ontoloģiju.

4. Zināšanu izguve no mašīnlasāmām vārdnīcām

Pilnīga latviešu valodas leksikas formalizācija prasītu ārkārtīgi lielu manuālu darbu. Idejas par automatizācijas iespējām leksisku taksonomiju konstruēšanā nav nekas jauns; pirmsākumi šīs problemātikas aktīviem pētījumiem ir meklējami jau pagājušā gadsimta 80. gados [2; 5]. Viens no ērtākajiem un bagātākajiem avotiem leksisko un pasaules zināšanu izguvei ir skaidrojošās vārdnīcas un enciklopēdijas. Zināšanas tajās ir dotas nevis tieši (formāli), bet netieši (dabīgās valodas teikumos), tiesa, savā ziņā vienveidīgā formā.

Lai būtu vērts veikt automātisku relāciju izguvi, tai jābūt relatīvi vienkārši realizējamai, bet rezultātam — plaša pārklājuma un maksimāli semantiski korektam. Kompilējot vārdnīcas, tiek vispārēji ievēroti šķirkļu veidošanas principi. Izmantojot heuristiskus

pieņēmumus, kas balstīti uz šķirkļu analīzi, iespējams būtiski atvieglot uzdevumu un padarīt automatizāciju noderīgu.

Angļu un citās valodās ir veikti ievērojami pētījumi šajā jomā, kuru rezultāts ir vairāk vai mazāk vispārīgas, gan no valodas atkarīgas, gan neatkarīgas hipotēzes (piemēram, lietvārdu un verbu definīcijās hiperonīmi parasti ir vārdkopu neatkarīgie komponenti). Tas nodrošina uz heuristiku balstītu zināšanu izgūšanu.

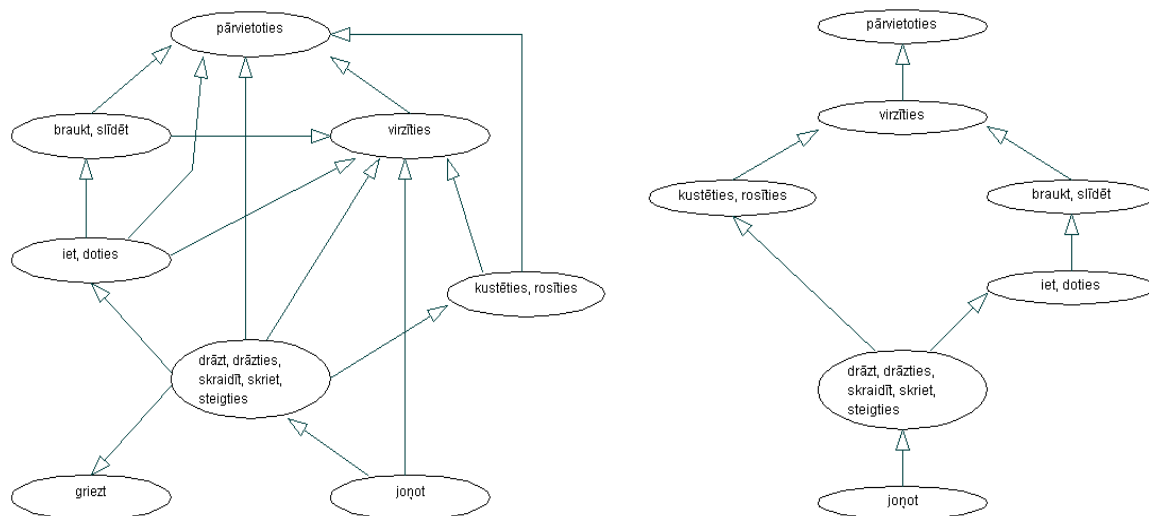
Latviešu valodai trūkst līdzīgu, praktiski izmantojamu leksikogrāfisku un datorlingvistisku pētījumu; šādi pētījumi ir uzsākti *SemTi-Kamola* ietvaros. Tiek pētīts, cik lielā mērā var izmantot citām valodām izstrādātu heuristiku jeb kuri pieņēmumi ir valodatkarīgi, kuri ne, un kādas ir latviešu valodas (leksikogrāfijas) īpatnības. Savukārt leksikosintaktiskie šabloni ir pilnībā valodatkarīgi.

Formāli strukturēta leksikona iegūšanai ir veikti sākotnējie eksperimenti automatizētā latviešu valodas vārdu tīkla konstruēšanā [3], par ieejas datu avotu ņemot LVV, kurai LU MII ir agrāk sagatavota mašīnlasāma versija [11]. Eksperiments tika sākts ar verbu apakškopu, jo skaidrojumos formāli atpazīt to pamatformas ir nosacīti vienkāršāk.

Tika leksikogrāfiski un statistiski analizētas nozīmju definīcijas, formulējot hipotēžu un apgalvojumu kopu semantisko pamatrelāciju izguvei:

1. Ja nozīme ir skaidrota tikai ar vienu verbu (nenoteiksmē) vai verbu uzskaitījumu, bez paplašinātājiem, tā ir definīcija ar sinonīmiem.
2. Hiponīmijas un hiperonīmijas attieksmju radītās mijnorādes ir transformējamas uz sinonīmijas attieksmēm: $subClassOf(x, y) \ \& \ subClassOf(y, x) \Rightarrow sameAs(x, y)$.
3. Iekavās dotie verbi nenoteiksmē, marķēti kā piemēri, ir definētās nozīmes hiponīmi.
4. Palīgteikumos nav meklējami ne hiperonīmi, ne hiponīmi, ne sinonīmi.
5. Pārējie verbi nenoteiksmes formā ir definētās nozīmes hiperonīmi.
6. Verbu analītiskās formas un modifikatori nav nenoteiksmes formas.
7. Hiponīmu indeksu var ģenerēt, invertējot hiperonīmu indeksu un otrādi.
8. Sinonīmu indekss ir iegūstams kā hiperonīmu un hiponīmu indeksu šķēlums.

Grafa fragments, kas iegūts, izmantojot pieņēmumu un atbilstošu šablonu kopu, ir redzams 10. attēlā. Kā tas bija sagaidāms, verbu semantiskā analīze kopumā ir ievērojami grūtāka nekā, piemēram, lietvārdu semantiskā analīze, jo lietvārdu nozīmēs ir precīzāk nosakāmas robežas starp virsklasēm/apakšklasēm.



10. attēls — sākotnēji iegūtas taksonomijas fragments un tā automātiski „attīrīts” ekvivalents. Nav izšķirtas vārdu nozīmes.

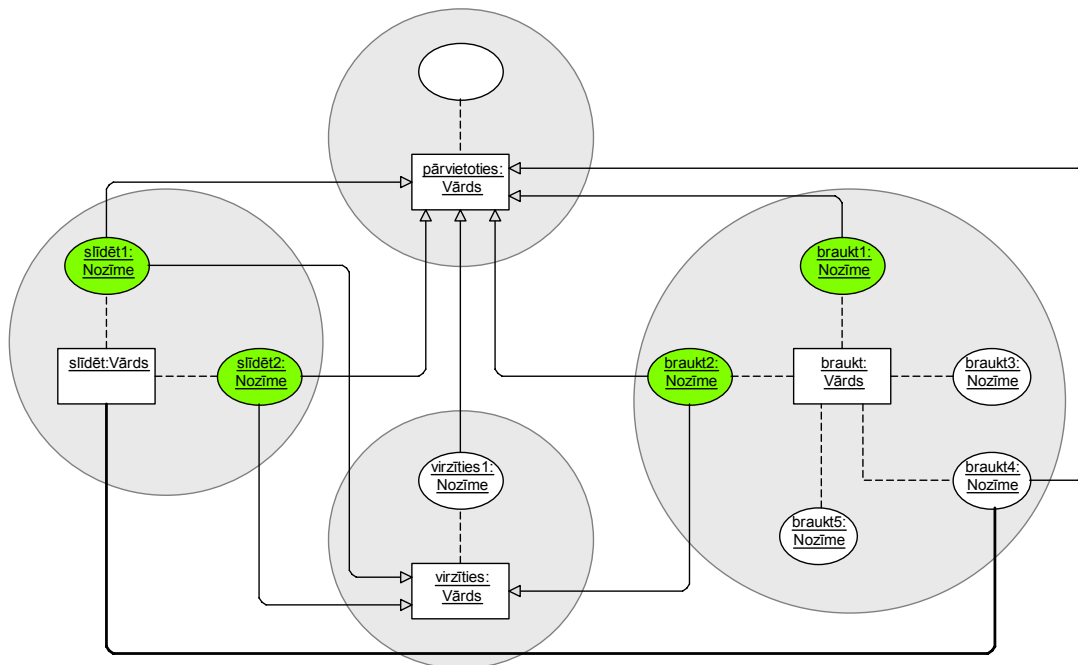
Balstoties uz verbu analīzi, tika attīstīta lietvārdu un īpašības vārdu pamatrelāciju izgūšanas metodika. Lietvārdu definīciju analīze pēc būtības neatšķiras no verbu analīzes, arī šeit galvenās automātiski izgūstamās semantiskās attiecības ir hiponīmija. Toties īpašības vārdu gadījumā jāizmanto cita pieeja, jo atšķiras starp īpašības vārdiem pastāvošās semantiskās attiecības — dominē sinonīmija un antonīmija.

Būtiskas problēmas rada ne tikai daudznozīmība, bet arī vispārīgas nozīmes vārdu definēšana. Sekas tam ir relāciju mijnorādes, un šķautņu „liekvārdība”, kas apgrūtina hierarhiskās struktūras vizuālu uztveršanu un rada problēmas izvedumu veikšanā: lielākajai daļai vārdu ir vienlaicīga hiperonīmijas piesaiste viena zara dažādu līmeņu mezgliem. Šādu problēmu kā ļoti izplatītu automātiski izgūtās hierarhijās min arī N. Ide un Ž. Veronis [5].

Vispārināšanas relācijas (R) ir ne tikai asimetriskas, bet arī transitīvas: $aRb \ \& \ bRc \Rightarrow aRc$. Izmantojot šo īpašību, starp grafa virsotnēm, kas ir saistītas ar hiponīmiskām relācijām, tika meklēti transitīvie slēgumi, reducējot lieko šķautņu skaitu.

Viens no galvenajiem eksperimenta secinājumiem: leksikona augšējo līmeņu taksonomija ar ontoloģiju jāasastata manuāli, savukārt konkrētas nozīmes būs iespējams vairāk vai mazāk kvalitatīvi automātiski sastāt ar šīs vispārīgās taksonomijas palīdzību.

Daudznozīmības novēršana (mazināšana) ir aktuālākā leksikona konstruēšanas problēma, kuras risināšanas metodes šī projekta turpinājumā tiks izstrādātas. Raugoties no atsevišķa šķirkļa pozīcijām, dažādās skaidrotā vārda nozīmes var viegli izšķirt; problēma ir atrast izmantoto vārdu atbilstošās nozīmes (sk. 11. attēlu).



11. attēls — ceļā uz daudznozīmības novēršanu. Vārda *slīdēt* abas nozīmes (to *IS-A* relācijas) pilnībā pārklājas, tādējādi tās var apvienot sinonīmu kopā; *virzīties* ir tikai viena nozīme; *braukt* un *slīdēt* ir sinonīmi tikai *braukt* 4. nozīmē.

Paralēli ir jāuzlabo semantisko pazīmju un relāciju izšķiršanas spējas, paplašinot un detalizējot heuristiku un šablonus. Ļoti svarīga ir arī piemēru analīze (hiponīmija, lietojumu informācija) un stabilu vārdu savienojumu analīze.

5. Rezultāti un secinājumi, nākotnes ieceres

Aprakstītie *SemTi-Kamola* projekta rezultāti ir publiski pieejami projekta tīmekļa lapā¹. Galvenie rezultāti šobrīd ir: uzlabota un attīrīta *OntoSem* ontoloģija, pārveidota OWL-DL formā, latviešu valodas leksikona (skaidrojošās vārdnīcas) formalizēšana semantiskā tīmekļa vajadzībām, nozīmju unikāla identificēšana oriģināli izveidotajā URI formātā, 1000 vispārīgāko un arī biežāk lietoto latviešu valodas vārdu nozīmju manuāla piesaiste ontoloģijas konceptiem. Tāpat ir izveidota un publiski pieejama pilotaplikācija ierobežotu latviešu valodas teikumu automātiskai transformēšanai formālā jēgas reprezentācijā (OWL formātā), starpvalodā, kas demonstrē semantiskā tīmekļa tehnoloģiju potenciālās iespējas.

Eksperimentālās programmatūras iestrādes semantisko relāciju izguvē un leksisko taksonomiju konstruēšanā no mašīnlasāmas skaidrojošās vārdnīcas ir devušas pietiekami daudzsoļus rezultātus, lai attīstītu pētniecību un izstrādi šajā virzienā; ir noteikta virkne

risināmo problēmu un metodisko uzlabojumu. Būtisks ieguvums ir manuāli izveidotā un ontoloģijai piesaistītā leksikona augšējā līmeņa taksonomija, kuru korekti automatizēti piesaistīt būtu neiespējami, bet ar kuras palīdzību būs iespējams automatizēti un ar augstāku precizitāti pievienot un organizēt dziļākus leksikona līmeņus.

Liels darbs ir veikts dažādu esošo semantiskā tīmekļa tehnoloģiju un rīku izmantošanas iespēju izpētē latviešu valodas un projekta situācijas vajadzībām. Tieši loģiskie izvedumi (un līdz ar to automātiskas secināšanas līdzekļu atbalsts) ontoloģijā, leksikonā un faktu krātuve nākotnē sniegs vislielāko ieguvumu un izsmalcinātākās informācijas analīzes iespējas latviešu valodā. Taču no teorētiski pamatotas līdz praktiski lietderīgai latviešu valodas semantiskā tīmekļa sistēmai vēl ir nepieciešams veikt ne tikai kvantitatīvus uzlabojumus, bet arī iekļaut kvalitatīvus leksisko un pasaules zināšanu avotus un komplicētus tekstu strukturālās un gramatiskās analīzes rīkus.

Projekta turpinājumā tiks izstrādāts pēc iespējas pilnīgāks latviešu valodas „precīzo” teikumu sintaktiskais un semantiskais metamodelis, kas sekmēs dažādu semantiskā tīmekļa aplikāciju izveidi un ieviešanu Latvijā. Visas iespējamās lietojuma sfēras šobrīd ir grūti paredzēt, taču kvalitatīva tulkošana no/uz latviešu, angļu u. c. valodām, kvalitatīvāku informācijas meklētājprogrammu attīstīšana semantiskā tīmekļa videi, pretrunu meklēšana dažādos juridiskos tekstos (līgumi, likumi) u. tml., ir tikai daži no piemēriem.

Viena no projekta pēdējās fāzes iecerēm ir uz semantiskā tīmekļa balstītas nākotnes e-Latvijas koncepcijas izstrāde. Tās mērķis ir nodrošināt semantiskā tīmekļa tehnoloģiju pieejamību citos e-Latvijas projektos. Tajā tiks izstrādāti virkne mazāku pilotprojektu, kas palīdzēs uzsākt šo tehnoloģiju praktisku izmantošanu citos projektos. Informācijas apmaiņas līmenī šis process jau ir uzsākts ar citiem Valsts pētījumu programmas „Informācijas tehnoloģijas” projektiem.

Piezīmes

1. *Uniform Resource Identifier* — noteiktas sintakses simbolu virkne, kas veido vārdu vai adresi, kas var tikt izmantota, lai atsauktos uz kādu resursu; fundamentāls tīmekļa arhitektūras komponents [16].
2. Izglītības un zinātnes ministrijas Valsts pētījumu programmas „Informācijas tehnoloģijas” finansējums un Eiropas Sociālā fonda atbalsts (projekts “Doktorantu un jauno zinātnieku pētniecības darba atbalsts Latvijas Universitātē”). Papildus skatīt tīmekļa vietni <http://www.semti-kamols.lv>
3. Šeit un turpmāk konceptu nosaukumi tiek rakstīti ar lielajiem burtiem.
4. http://www.globalwordnet.org/gwa/wordnet_table.htm
5. *Uniform Resource Locator* — vienkāršoti — tīmekļa vietnes adrese.

Vēres

1. Berners-Lee, T., Hendler, J., Lassila, O. The Semantic Web. *Scientific American*. 2001. Sk. internetā (2006.23.01) <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>
2. Chodorow, M. S., Byrd, R. J., Heidorn, G. E. Extracting Semantic Hierarchies from a Large On-Line Dictionary. In: *Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics*. Chicago: 1985. Pp. 299–304.
3. Grūzītis, N. *Ontoloģiska latviešu valodas leksikona datubāze: arhitektūra un izveides problemātika*. Maģistra darbs. Rīga: LU Datorikas nodaļa, 2005. 92 lpp.
4. Harabagiu, S. M., Moldovan, D. I. Knowledge Processing on an Extended WordNet. In: Ed. Fellbaum, C. *WordNet: an Electronic Lexical Database*. Cambridge: MIT Press, 1998. Pp. 379–405.
5. Ide, N., Véronis, J. Refining Taxonomies Extracted from Machine-Readable Dictionaries. In: Eds. Hockey, S., Ide, N. *Research in Humanities Computing 2*. Oxford: Oxford University Press, 1994. Pp. 145–170.
6. Knublauch, H. Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with Protégé/OWL. In: *Proceedings of International Workshop on the Model-Driven Semantic Web*. Monterey: 2004.
7. *Latviešu literārās valodas vārdnīca*. Rīga: Zinātne, 1972–1996. 1.–8. sējums.
8. *Latviešu valodas biežuma vārdnīca*. Rīga: Zinātne, 1973.
9. *Latviešu valodas vārdnīca*. Rīga: Avots, 1987.
10. McShane M., Nirenburg S., Beale S. An Implemented, Integrative Approach to Ontology-Based NLP and Interlingua. Sk. internetā (2006.23.01) http://ilit.umbc.edu/ILIT_Working_Papers/ILIT_WP_06-05_Controlled_Langs.pdf
11. Milčonoka, E., Grūzītis, N., Spektors, A. Natural Language Processing at the Institute of Mathematics and Computer Science: 10 Years Later. In: *Proceedings of the First Baltic Conference “Human Language Technologies — the Baltic Perspective”*. Riga: 2004. Pp. 6–11.
12. Nirenburg, S., Raskin, V. *Ontological Semantics*. Cambridge: MIT Press, 2004.
13. OntoWeb Consortium. Technical Roadmap (2002). Sk. internetā (2006.23.01) http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-1-2.pdf
14. Sowa, J. F. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove: Brooks/Cole, 2000.
15. *Spinning the Semantic Web*. Eds. Fensel, D., Hendler, J., Lieberman, H., Wahlster, W. Cambridge: MIT Press, 2003.
16. The Internet Society. Uniform Resource Identifier: Generic Syntax (2005). Sk. internetā (2006.23.01) <http://www.gbiv.com/protocols/uri/rfc/rfc3986.html>
17. *WordNet: an Electronic Lexical Database*. Ed. Fellbaum, C. Cambridge: MIT Press, 1998.

18. WWW Consortium. OWL Web Ontology Language Overview (2004.10.02). Sk. internetā (2006.23.01) <http://www.w3.org/TR/owl-features>
19. WWW Consortium. RDF Primer (2004.10.02). Sk. internetā (2006.23.01) <http://www.w3.org/TR/rdf-primer>
20. Zvonkova, O. *Latviešu valodas teikumu ontoloģiskā semantika*. Maģistra darbs. Rīga: LU Datorikas nodaļa, 2005. 137 lpp.