LATVIJAS UNIVERSITĀTE
Matemātikas un informātikas institūts

NORMUNDS GRŪZĪTIS

IEROBEŽOTAS LATVIEŠU VALODAS FORMĀLĀ GRAMATIKA UN SEMANTIKA

Promocijas darba kopsavilkums
datorzinātņu doktora (*Dr. sc. comp.*) zinātniskā grāda iegūšanai

Nozare: datorzinātne
Apakšnozare: datoru un sistēmu programmatūra

Rīga 2010

Darbs ir izstrādāts ar Eiropas Sociālā fonda atbalstu.
Projekts „Atbalsts doktora studijām Latvijas Universitātē”.

Zinātniskais vadītājs:

vad. pētnieks, Dr.phys. Andrejs Spektors
Latvijas Universitātes Matemātikas un informātikas institūts

Recenzenti:

profesors, Dr.math. Kārlis Podnieks
Latvijas Universitāte

vad. pētniece, Dr.sc.ing. Alla Anohina-Naumeca
Rīgas Tehniskā universitāte

profesors, Ph.D. Aarne Ranta
Gēteborgas universitāte (Zviedrija)

Darba aizstāvēšana notiks Latvijas Universitātes (LU) Datorzinātnes nozares promocijas padomes atklātā sēdē 2011. gada 2. februārī plkst. 16:00 LU Matemātikas un informātikas institūtā (Raiņa bulvārī 29), 413. auditorijā.

Ar darbu un tā kopsavilkumu var iepazīties LU Bibliotēkā (Kalpaka bulvārī 4).

Padomes priekšsēdētājs: Jānis Bārzdiņš

Anotācija

Promocijas darba pētījuma priekšmets ir latviešu valodas kā izteikti fleksīvas, sintētiskas valodas automātiska gramatiskā un semantiskā analīze. Darbā ir piedāvāts oriģināls, hibrīds gramatikas modelis, kas ir piemērots tādu valodu sintaktiskajai analīzei, kurās vārdu secība ir relatīvi brīva. Izstrādātais modelis ir aprobēts praksē, formalizējot latviešu valodas apakškopu, kas aptver dažādas vienkāršos paplašinātos teikumos sastopamas sintaktiskās konstrukcijas. Tālāk problēma tiek sašaurināta ne tikai sintaktiski, bet arī semantiski, izstrādājot viennozīmīgu, taču iespējami dabisku ierobežotu latviešu valodu (un atbilstošus automātiskas analīzes/sintēzes līdzekļus), kuras semantika ir definēta aprakstošajā loģikā. Darbā ir parādīts, ka teikuma informatīvās struktūras analīze ir pietiekams līdzeklis viennozīmīgai kvantoru un koreferenču noteikšanai ierobežotas sintētiskas valodas formā dotās OWL terminoloģiskajās aksiomās, SWRL izvedumu likumos un SPARQL integritātes vaicājumos. Darbā ir piedāvāta un realizēta divlīmeņu tulkošanas pieeja, demonstrējot ierobežotās latviešu valodas teikumu automātisku, semantiski precīzu translēšanu uz OWL (un otrādi), lietojot esošu ierobežotu angļu valodu kā starpvalodu un atkalizmantojot tās rīkus. Papildus ir piedāvāta pusautomātiska metode sistemātiskas leksiskās daudznozīmības atbalstam un nozīmju precīzai izšķiršanai ierobežotas valodas tekstos, vienlaikus risinot OWL ontoloģiju sastatīšanas problēmu.

Pateicība

Vispirms vēlos pateikties savam vadītājam, Andrejam Spektoram, kurš mani iesaistīja aizraujošajā datorlingvistikas nozarē un vēlāk arī iedrošināja studēt doktorantūrā, atbalstot visā studiju un promocijas darba izstrādes laikā, un Guntim Bārzdiņam, kura vadītā Valsts pētījumu programmas projekta *SemTi-Kamols* laikā ir tapusi liela promocijas darba daļa un iestrādes turpmākajiem pētījumiem un no kura es ļoti daudz esmu mācījies. Liels paldies arī visiem maniem kolēģiem par draudzīgo un auglīgo darba vidi, jo īpaši Guntai Nešporei un Baibai Saulītei par nemitīgajām konsultācijām un diskusijām par valodniecības jautājumiem, Ingunai Skadiņai par vērtīgajiem komentāriem, Kārlim Čerānam par izaicinošiem un iedvesmojošiem testpiemēriem, kā arī Ilmāram Poikānam un Jānim Džeriņam par palīdzību visdažādāko tehnisko jautājumu risināšanā.

Paldies Krasimiram Angelovam un Ārnem Rantam par dažādu būtisku *Grammatical Framework* aspektu izskaidrošanu, kā arī par sākotnējo rezultātu novērtēšanu un iedrošināšanu aktīvi turpināt izstrādi. Paldies arī visiem anonīmajiem publikāciju recenzentiem par konstruktīvo kritiku un vērtīgajiem ieteikumiem, kas ir sekmējuši manu izaugsmi un darba rezultātus.

Būtisku izaugsmi man deva arī studijas *NordForsk* finansētajā Ziemeļvalstu Valodu tehnoloģiju doktorantūras skolā (NGSLT), par ko esmu ļoti pateicīgs. Paldies arī LU Datorikas fakultātei un LU Matemātikas un informātikas institūtam par radošo un stimulējošo gaisotni. Promocijas darba izstrāde ir bijusi iespējama pateicoties Valsts pētījumu programmas un Eiropas Sociālā fonda finansējumam.

Visbeidzot vislielāko paldies vēlos teikt savai ģimenei un visiem tuvajiem cilvēkiem par līdzi jušanu, jo īpaši sievai par sapratni un milzīgo atbalstu, dēliem par iedvesmu, spēku un mācīšanu efektīvāk izmantot laiku un vecākiem par izglītību un visām rūpēm. Šis darbs ir veltīts viņiem.

Satura rādītājs

Promocijas darba tēmas aktualitāte un iegūtie rezultāti 6

Promocijas darba vispārējais raksturojums 8

Ievads 10

1 Kas ir vārda nozīme? 13

2 No valodas līdz ontoloģijai 16

2.1 Atkarību gramatikā balstīta sintaktiskā analīze 16

2.2 Daudzdimensiju ontoloģija 22

3 Mikroontoloģijas kā formāli precīzas skaidrojošās vārdnīcas 25

3.1 Loģikā balstīta dabiskās valodas apakškopa 25

3.2 Mikroontoloģijas 26

3.3 Mikroontoloģiju integrēšana un daudznozīmības risināšana 27

4 Ierobežota latviešu valoda ontoloģiju izstrādei un verbalizēšanai 30

4.1 Teikuma informācijas struktūras analīze 30

4.2 SVO teikuma šablona paplašinājumi 32

4.3 Sintaktiskā pārfrazēšana 33

4.4 Divlīmeņu tulkošana uz/no OWL 37

Nobeigums 42

Atsauces

Autora publikācijas 43

Izmantotā literatūra 44

Pielikums

Autora referāti par darba rezultātiem 50

Autora personiskais ieguldījums promocijas darbā iekļautajās publikācijās 50

Promocijas darba tēmas aktualitāte un iegūtie rezultāti

Tēmas aktualitāte

Mūsdienās, kad lielākā daļa informācijas ir pieejama elektroniskā formā, arvien nozīmīgāku lomu iegūst šādas informācijas automātiskas analīzes līdzekļi. Promocijas darbā, kas ir daļa no plašāka pētījuma, ir aplūkotas formālas analīzes iespējas tāda veida informācijai, kas dota dabiskās valodas (teksta) formā. Darbs galvenokārt ir veltīts tieši latviešu valodā izteiktu zināšanu analīzei (teksta nozīmes reprezentācijai), kā arī latviešu valodas teikumu sintēzei, balstoties uz jau esošiem formalizētu zināšanu avotiem — ontoloģijām.

Latviešu valoda ir izteikti fleksīva, sintētiska valoda: tās bagātā morfoloģija pieļauj sintaktiski brīvu vārdu secību teikumā. Pretstatā sintētiskām valodām ir izteikti analītiskas valodas (piem., angļu), kuru vienkāršā morfoloģija teikumos tiek kompensēta ar stingri noteiktu vārdu secību. Datorlingvistikas pētījumos pasaulē lielākā uzmanība ir tikusi pievērsta angļu valodas un citu salīdzinoši analītisku valodu formālajai analīzei. Savukārt vispāratzītās analītisko valodu analīzes metodes bieži vien tiešā veidā tiek izmantotas arī sintētisku valodu analīzē, līdz ar to (atkarībā no analīzes mērķa) ne vienmēr tiek ņemtas vērā būtiskas sintētisko valodu īpašības.

Pētījuma ietvaros ir izstrādātas oriģinālas latviešu valodas sintaktiskās un semantiskās analīzes metodes, ņemot vērā šīs izteikti sintētiskās valodas īpašības. Izstrādātās metodes var tikt piemērotas arī citu sintētisko valodu analīzē; vēl vairāk — daļa no piedāvātajām metodēm ir valodneatkarīgas un vispārpieņemto pieeju kontekstā piedāvā alternatīvu perspektīvu.

Tomēr neierobežotas latviešu valodas analīze šobrīd vēl ir apgrūtināta — ikviena dabiskā valoda visos tās analīzes līmeņos (no morfoloģijas līdz diskursam) ir ārkārtīgi daudznozīmīga, un ar likumbāzētām metodēm vien vispārīgā gadījumā ir praktiski neiespējami nodrošināt iespējami atbilstošākā analīzes varianta izvēli. Lai to nodrošinātu, likumbāzētās metodes būtu jākombinē ar statistiskām metodēm, taču valodas sintaksi un semantiku objektīvi atspoguļojoša statistiskā modeļa iegūšanai ir nepieciešams apjomīgs, manuāli anotēts tekstu korpuss — latviešu valodai šāds resurss pagaidām nav pieejams.

Tādēļ promocijas darba centrālais pētījuma objekts un praktiski nozīmīgākais rezultāts ir sintaktiski ierobežota latviešu valoda (un tās analīzes līdzekļi) ar skaidri definētu semantiku. Ierobežotie izteiksmes līdzekļi nodrošina precīzu, deterministisku teikumu nozīmes attēlošanu izvēlētajā formālajā valodā — semantiskā tīmekļa ontoloģiju valodā OWL [15].

Viens no semantiskā tīmekļa tehnoloģiju mērķiem ir padarīt tiešsaistē pieejamo, decentralizēto un lielākoties nestrukturēto informāciju ne tikai mašīnlasāmu, bet arī mašīnai „saprotamu", tādējādi paverot ceļu masveidīgai informatīvo procesu automatizācijai un integrācijai visdažādākajās tautsaimniecības nozarēs un sabiedrībā kopumā. Šo tehnoloģiju pamatā ir dažādu priekšmetu apgabalu (nozaru jeb domēnu) ontoloģiju (konceptuālo modeļu) izstrāde, nodrošinot gan zināšanu formālu reprezentāciju, gan automātiskas secināšanas iespējas.

Domēnspecifisko ontoloģiju izstrādē ļoti nozīmīga loma ir ne tikai zināšanu inženieriem (sistēmanalītiķiem), bet arī attiecīgo nozaru ekspertiem (cilvēkiem bez zināšanām informāciju tehnoloģijās un programmēšanā), līdz ar to dabiskās valodas saskarne pretstatā formālajām notācijām piedāvā alternatīvus un ļoti intuitīvus izteiksmes un uztveres līdzekļus. Ne mazāk svarīgs blakusefekts dabiskās valodas kā ontoloģiju saskarnes lietošanā ir semantiski precīzas automātiskas tulkošanas iespējas, veicinot ontoloģiju multilingvālu izstrādi un izmantošanu.

Galvenie darba rezultāti

- Izstrādāts oriģināls, L. Tenjēra atkarību gramatikas pieejā [16] balstīts modelis fleksīvu sintētisku valodu gramatiskajai analīzei. Atšķirībā no „klasiskās" formālās atkarību gramatikas pieejas šajā modelī vienlaikus ar vertikālajām pakārtojuma attieksmēm starp teikuma locekļiem iespējams aprakstīt arī horizontālās (ne pakārtojuma) attieksmes starp analītisko vārdformu sastāvdaļām. Izstrādātais modelis ir aprobēts praksē, formalizējot nosacīti plašu latviešu valodas gramatikas apakškopu: aptverot visdažādākās sintaktiskās konstrukcijas, kas ir sastopamas vienkāršos paplašinātos teikumos (piem., verbu analītiskās formas u.c. kompleksus teikuma locekļus un to iespējamos pakārtojumus).
- Formulēts OWL mikroontoloģiju jēdziens un izveidota metodika šādu ontoloģiju konsekventai izstrādei, ar mērķi nodrošināt dažādu domēnspecifisko ontoloģiju sistemātisku integrēšanu. Izstrādāta valodneatkarīga metode daudznozīmīgo (no OWL viedokļa — pretrunīgo) klašu automātiskai atklāšanai un izšķiršanai (vārdu nozīmju kontrolētai ieviešanai) un daļēji automatizētai vārdu nozīmju noteikšanai tekstos, kas apraksta integrētās ontoloģijas indivīdus un to savstarpējās attieksmes (t.i., faktos).
- Izstrādāta ierobežota, taču dabiska latviešu valoda, radot formālajām notācijām alternatīvu, intuitīvu izteiksmes līdzekli OWL ontoloģiju izstrādei un atspoguļošanai (verbalizācijai). Parādīts, ka teikuma informatīvās struktūras analīze ir pietiekams līdzeklis viennozīmīgai kvantoru (t.sk. zināmās un jaunās informācijas) noteikšanai ierobežotas sintētiskas valodas formā dotās OWL terminoloģiskajās aksiomās, SWRL izvedumu likumos [17] un SPARQL integritātes vaicājumos [18]. Izstrādāta ierobežotās valodas (tās parsētāja un vienlaikus arī ģeneratora) efektīva realizācija *Grammatical Framework* (GF) [19] vidē. Lai nodrošinātu iespēju lietot dažādas teikuma konstrukcijas (un to redukcijas) vienas un tās pašas nozīmes (ontoloģijas aksiomas vai likuma) izteikšanai, kā arī lai dabiskās valodas līmenī pieļautu kontrolētu atkāpšanos no striktā *subjekts-verbs-objekts* (SVO) šablona, ierobežotajai valodai ir izstrādātas paralēlas gramatikas: viena — robustai analīzei, bet otra — iespējami labākai pārfrazēšanai. Visbeidzot ir izstrādāts prototips, kas, izmantojot oriģinālu divlīmeņu tulkošanas pieeju, demonstrē ierobežotas, dabiskas latviešu un angļu valodas teikumu translēšanu uz OWL (un otrādi), lietojot valodas *Attempto Controlled English* [20] vienkāršoto OWL apakškopu (un tās brīvi pieejamo infrastruktūru) kā starpvalodu.

Pētījuma ietvaros sasniegtie rezultāti ir kolektīva darba rezultāts. Promocijas darba autors ir aktīvi piedalījies visos pētījuma posmos un ir devis būtisku ieguldījumu šo rezultātu sasniegšanā. Autora personiskais ieguldījums ir konkretizēts kopsavilkuma turpmākajās nodaļās un pielikumā.

Promocijas darba vispārējais raksturojums

Promocijas darbs „Ierobežotas latviešu valodas formālā gramatika un semantika" ir izstrādāts no 2005. gada līdz 2010. gadam Latvijas Universitātes (LU) Datorikas fakultātē un LU Matemātikas un informātikas institūtā (MII) vadošā pētnieka Andreja Spektora vadībā. Galvenie darba rezultāti ir atspoguļoti 11 publikācijās [1–11], no kurām lielākā daļa ir tapušas profesora Gunta Bārzdiņa vadītā Valsts pētījumu programmas „Informācijas tehnoloģiju zinātniskā bāze" projekta „Semantiskā tīmekļa izpēte, attīstīšana un piemērošana Latvijas vajadzībām" (*SemTi-Kamols*) ietvaros.

Promocijas darbs ir **tematiski vienota publikāciju kopa**, kurā ir atspoguļoti autora pētījumi un to rezultāti dažādos dabiskās valodas automatizētas sintaktiskās un semantiskās analīzes aspektos.

Pētījuma priekšmets

Latviešu valodas kā izteikti fleksīvas sintētiskas valodas automātiska gramatiskā un semantiskā analīze. Precīzi definēta latviešu valodas apakškopa kā zināšanu reprezentācijas valoda. Uz ontoloģijām balstīta leksiskās daudznozīmības risināšana.

Pētījuma mērķis

Pētījuma plašākais mērķis ir izstrādāt oriģinālu formālas gramatiskās analīzes (reprezentācijas) metodi (modeli), kurā būtu ņemtas vērā fleksīvu sintētisku valodu īpatnības (kā paraugu formalizējot praksē plaši lietotu latviešu valodas apakškopu), kā arī attīstīt vispārzināmās, uz ontoloģijām balstītās teksta nozīmes reprezentācijas pieejas.

Pētījuma konkrētais mērķis ir izstrādāt sintaktiski ierobežotu, deterministiski analizējamu, taču vienlaikus iespējami dabisku un intuitīvu latviešu valodas apakškopu, kuras semantika ir definēta aprakstošajā loģikā — pirmās kārtas loģikas apakškopā, kurā balstīta semantiskā tīmekļa ontoloģiju valoda OWL [15] un izvedumu likumu valoda SWRL [17].

Pētījuma posmi

Promocijas darba rezultāti, no vienas puses, atspoguļo noslēgtu pētījumu ciklu, sākot ar pētījumiem sintētiskas valodas formālās gramatikas modelēšanā, kā arī leksiskās un sintaktiskās semantikas modelēšanā un beidzot ar pētījumiem loģikā balstītu valodu jomā, izstrādājot ierobežotu latviešu valodu un tās automātiskas analīzes un sintēzes līdzekļus OWL ontoloģiju uzdošanai un verbalizēšanai. No otras puses, iegūtie rezultāti (ierobežotās latviešu valodas realizācija) ir uzskatāmi par prototipu, kas ir radījis pamatu tālākiem pētījumiem un iedrošinājis aktīvi turpināt pētniecību un izstrādi šajā virzienā.

Promocijas darbā veiktos pētījumus var iedalīt četros posmos, no kuriem pirmajos gūtā pieredze ir devusi būtisku pamatu nākamo posmu sekmīgai realizācijai. Šie pētījumi ir apkopoti tematiskā kārtībā, kas lielākoties atbilst arī to hronoloģiskajai secībai:

- **Pirmā posma** pētījumi un to rezultāti ir aprakstīti kopsavilkuma 1. nodaļā un atbilstošajās publikācijās [1–3]. Autors ir aktīvi piedalījies visos šajā posmā veiktajos pētījumos, un autora personiskais ieguldījums veido vairāk nekā 50% no kopējā pētījumu apjoma. Šajā posmā gūtās atziņas rosināja tālākos pētījumus, kuros ir sasniegti galvenie promocijas darba rezultāti.
- **Otrā posma** pētījumi un to rezultāti ir aprakstīti kopsavilkuma 2. nodaļā un atbilstošajās publikācijās [4–6]. Autors ir aktīvi piedalījies visos šajā posmā veiktajos pētījumos, un autora personiskais ieguldījums veido vairāk nekā 40% no kopējā

pētījumu apjoma. Galvenais šī posma rezultāts ir izstrādātais **hibrīdais atkarību gramatikas modelis**, kas ir īpaši piemērots valodu ar relatīvi brīvu vārdu secību aprakstīšanai un nodrošina, no vienas puses, elastīgu, no otras — detalizētu analīzi. Šajā posmā veiktie pētījumi un iestrādes bija nozīmīgs ideju avots turpmākajai loģikā balstītās latviešu valodas un tās analīzes metožu izstrādei.

- **Trešā posma** pētījumi un to rezultāti ir aprakstīti kopsavilkuma 3. nodaļā un atbilstošajās publikācijās [7, 8]. Autors ir aktīvi piedalījies visos šajā posmā veiktajos pētījumos, un autora personiskais ieguldījums veido vairāk nekā 60% no kopējā pētījumu apjoma. Galvenais šī posma rezultāts ir ieviestais **mikroontoloģijas jēdziens** un izveidotā **metodika šādu ontoloģiju izstrādei un integrēšanai**. Arī šie pētījumi deva svarīgas atziņas pēdējā pētījumu posma veiksmīgai īstenošanai.
- **Ceturtā posma** pētījumi un to rezultāti ir aprakstīti kopsavilkuma 4. nodaļā un atbilstošajās publikācijās [9–11]. Autors ir vadījis visus šajā posmā veiktos pētījumus, un autora personiskais ieguldījums veido vairāk nekā 90% no kopējā pētījumu apjoma. Galvenais šī posma rezultāts ir izstrādātā **ierobežotā, deterministiskā latviešu valoda ar tās analīzes un pārfrazēšanas līdzekļiem**, kas nodrošina šajā valodā dabiski izteiktu komplicētu apgalvojumu un izvedumu likumu semantiski precīzu interpretāciju ontoloģiju valodā OWL un otrādi — ontoloģijas aksiomu un likumu verbalizēšanu precīzā, intuitīvi uztveramā latviešu valodas apakškopā.

Pētījuma teorētiskā un praktiskā nozīme

Pētījumā izstrādātajam atkarību gramatikā balstītajam hibrīdajam sintaktiskās analīzes modelim projektā *SemTi-Kamols* ir izstrādāta arī atbilstoša analizatora paraugrealizācija[1], kas kopā ar vienkāršus paplašinātus teikumus aprakstošo latviešu valodas formālo gramatiku tiek izmantota dažāda veida latviešu valodas tekstu korpusu pusautomātiskā un pat automātiskā morfoloģiskā anotēšanā [21]. Šie eksperimentālie korpusi savukārt tiek izmantoti, piemēram, faktorētu valodas modeļu ieguvē statistiskajā angļu-latviešu mašīntulkošanā [14, 22].

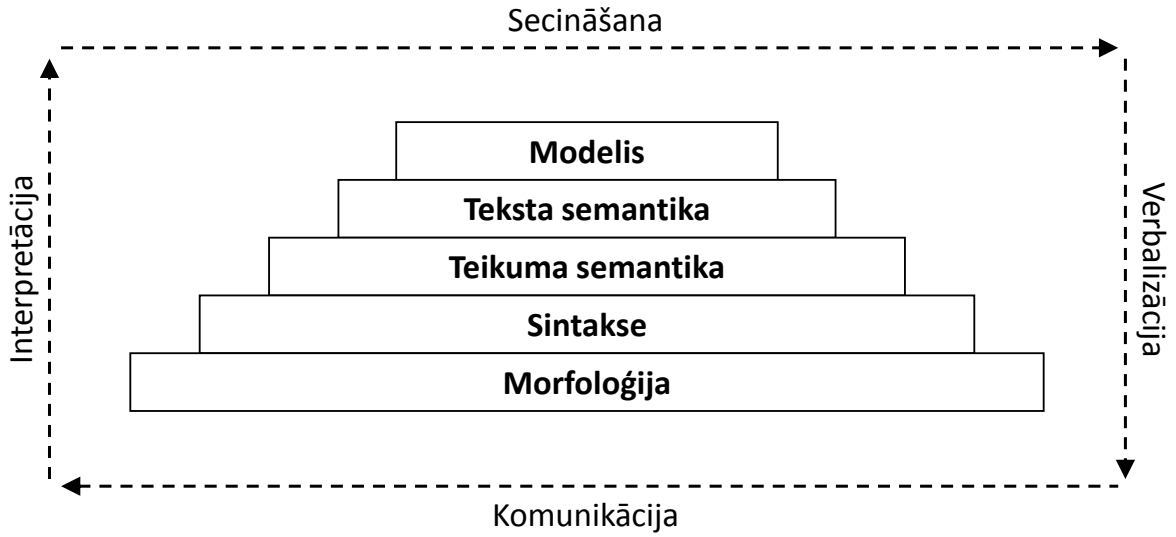
2011. gadā LU MII plānots sākt ERAF lietišķo pētījumu projekta „Semantisko datubāzu platforma nozaru speciālistiem" īstenošanu, kurā kā nozīmīga komponente (dabiskās valodas saskarne) tiks izmantota un attīstīta promocijas darbā izstrādātā ierobežotā latviešu valoda (paralēli ierobežotai angļu valodai un UML tipa grafiskai notācijai). Dabiskā valoda kā ikvienam intuitīvi uztverama zināšanu reprezentācijas valoda ir nozīmīgs papildinājums citām augsta līmeņa OWL notācijām, piemēram, no modelēšanas valodas UML atvasinātajām grafiskajām notācijām [23], kuru mērķauditorija, līdzīgi kā ierobežotai dabiskai valodai, arī lielā mērā ir tieši nozaru eksperti, kas var nebūt tieši saistīti ar IT nozari. Grafiskās valodas palīdz uzskatāmi atklāt modeļa kopainu un dažādas kopsakarības, taču tās nav piemērotas sarežģītu ontoloģijas ierobežojumu definēšanai un uzskatāmai vizualizēšanai. Tas savukārt ir parocīgi un viegli uztverami izdarāms, lietojot ierobežotu dabisko valodu.

Kopsavilkuma ievadā ir īsumā izklāstītas galvenās problēmas dabiskās valodas tekstu automātiskā analīzē. No pirmās līdz ceturtajai nodaļai ir dots konspektīvs autora veikto pētījumu, pamatnostādņu, iegūto rezultātu un secinājumu apraksts, kas plašāk ir izklāstīti atbilstošajās publikācijās. Nobeigumā ir doti vispārēji secinājumi un iezīmēti turpmākās attīstības virzieni. Pielikumā ir uzskaitītas starptautiskās zinātniskās konferences, kurās autors ir ziņojis par pētījumu rezultātiem, kā arī ir konkretizēts autora personiskais ieguldījums katrā no promocijas darbā iekļautajām publikācijām.

[1] Analīzes algoritma paraugrealizācija neietilpst promocijas darba autora ieguldījumā. Autora ieguldījums ir formālās gramatikas izstrāde. Gan analizators, gan gramatika ir brīvi pieejami: http://www.semti-kamols.lv/

Ievads

Promocijas darbs ir veltīts problemātikai starpnozares jomā — datorlingvistikā. Datorlingvistikas centrālais mērķis ir teksta nozīmes formāla reprezentācija, šādas reprezentācijas automātiska iegūšana no dotā teksta un otrādi — teksta ģenerēšana dotajam modelim. Mašīntulkošana ir viens no tipiskākajiem lietojumiem, kas savā būtībā ietver šo ciklu (tulkojot domu pret domu), un arī promocijas darba centrālie rezultāti ir savā ziņā saistīti ar mašīntulkošanas problēmu risināšanu, taču ar šo jēdzienu darbā tiek saprasta tulkošana ne tikai starp dabiskām valodām (to ierobežotām apakškopām), bet arī starp dabiskām un formālām valodām. Lai arī mūsdienu praksē ar dabiskās valodas apstrādi saistītos lietojumos dažādu objektīvu iemeslu dēļ līdz padziļinātai semantiskajai analīzei lielākoties (pagaidām) nemaz nenonāk, arī teksta gramatiskā (morfoloģiskā un sintaktiskā) analīze kā tāda sniedz nozīmīgas iespējas informācijas izguvē, analīzē un tulkošanā.

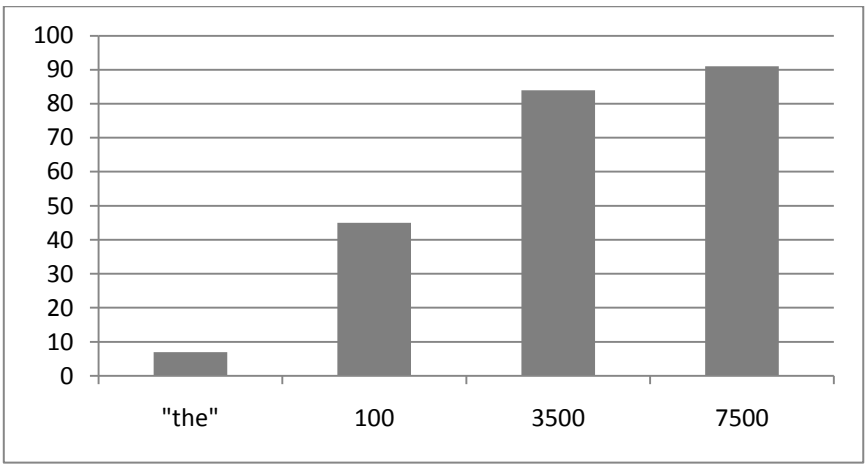


1. attēls. Dabiskās valodas analīzes līmeņi un to loma saziņas procesā.

Galvenā problēma dabiskās valodas tekstu automātiskā analīzē ir tās izteiktā daudznozīmība visos analīzes līmeņos (sk. 1. attēlu): no morfoloģisko pazīmju līdz pat diskursa referentu noteikšanai. Lai arī tas ļoti apgrūtina valodas deterministiskas analīzes iespējas un vispārīgā gadījumā to faktiski padara neiespējamu, daudznozīmība ir fenomenāls līdzeklis, kas valodas lietošanu padara ērtu cilvēkam. Vārdu, vārdformu un sintaktisko konstrukciju klāsts valodā ir galīgs, taču to kombinēšanas un interpretācijas iespējas ir teorētiski bezgalīgas. To uzskatāmi ilustrē vārdu izplatība, piemēram, Britu nacionālajā tekstu korpusā (BNC; sk. 2. attēlu).

Saskaņā ar Zifa likumu [25] jebkura vārda sastopamības biežums ir apgriezti proporcionāls šī vārda pozīcijai valodas vārdu biežumsarakstā (attiecībā pret kādu tekstu korpusu). Tas nozīmē, ka visbiežāk lietotais vārds tekstos sastopams divreiz biežāk nekā otrs biežākais vārds, trīsreiz biežāk nekā trešais biežākais vārds utt. Citiem vārdiem sakot, salīdzinoši maz vārdu tiek lietoti bieži — lielākā daļa tiek lietoti reti. No tā varam secināt divas lietas. Pirmkārt, dažādos kontekstos vārdi tiek intensīvi atkalizmantoti, radot pamatu plašai leksiskajai daudznozīmībai, un, otrkārt, ir nepieciešams ļoti liels[2] tekstu korpuss, lai tas būtu pietiekoši reprezentatīvs attiecībā uz reti lietotiem vārdiem un to nozīmēm.

[2] Salīdzinājumam, minētais BNC korpuss aptver 100 milj. vārdlietojumu, taču mūsdienu korpuslingvistikā tiek runāts jau par miljardus vārdlietojumu lieliem, no tīmekļa resursiem kompilētiem korpusiem [26].



2. attēls. Vārdu izplatība BNC korpusā: 7500 biežāk lietotie vārdi pārklāj 90% vārdlietojumu [24].

Teorētiski katrs nākamais valodas analīzes līmenis palīdz atrisināt daudznozīmību iepriekšējā līmenī: morfoloģisko daudznozīmību lielā mērā novērš sintaktiskā analīze, sintaktisko daudznozīmību — leksiskā un sintaktiskā semantika, bet semantisko daudznozīmību — konteksts un ekstralingvistiskās (pasaules) zināšanas.

Kā piemēru apskatīsim vienkāršu teikumu: *„es ceļu māju”*. Ja nošķirti analizējam katra vārda morfoloģiskās pazīmes, iegūstam šādus vienkāršotas analīzes variantus:

- *„es”* — vienskaitļa pirmās personas vietniekvārds;
- *„ceļu”* — verbs īstenības izteiksmē, tagadnē, pirmajā personā (*„celt”*) vai lietvārds vienskaitļa akuzatīvā (*„ceļš”*), vai lietvārds daudzskaitļa ģenitīvā (*„ceļš”* vai *„celis”*);
- *„māju”* — lietvārds vienskaitļa akuzatīvā (*„māja”*) vai lietvārds daudzskaitļa ģenitīvā (*„māja”*), vai verbs īstenības izteiksmē, tagadnē, pirmajā personā (*„māt”*).

Sintaktiskā analīze samazina morfoloģiskās analīzes variantu un to iespējamo kombināciju skaitu, taču arī tā pati par sevi diemžēl nedod viennozīmīgu rezultātu:

- *„es”*$_{\text{SUBJEKTS}}$ + *„celt”*$_{\text{VERBS}}$ + *„māja”*$_{\text{OBJEKTS}}$
- *„es”*$_{\text{SUBJEKTS}}$ + *„ceļš”*$_{\text{OBJEKTS}}$ + *„māt”*$_{\text{VERBS}}$

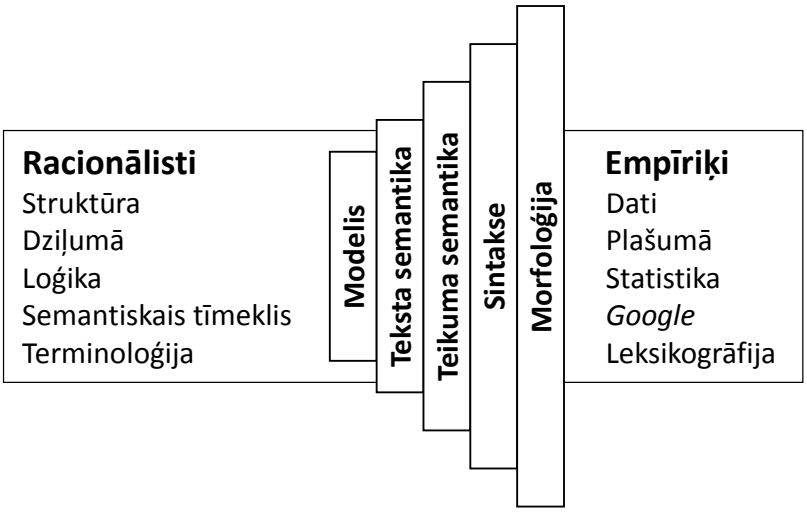
Lai izvēlētos korekto sintaktiskās analīzes variantu, ir nepieciešamas:

- leksiskās semantikas zināšanas, piemēram, vārds *„māja”* ir hiponīms vārdam *„celtne”*, bet vārds *„celt”* ir sinonīms vārdam *„būvēt”*;
- zināšanas par verbu sintaktisko valenci — tipiskajiem paplašinātājiem, piemēram, verbs *„māt”* ir intransitīvs, tātad tiek lietots kopā ar prepozicionālu konstrukciju (piem., atbild uz jautājumu *ar ko?*), nevis tiešo objektu (*ko?*);
- sintaktiskās semantikas (semantiskās valences) zināšanas — kādas semantiskās kategorijas (tipiskās semantiskās lomas) verbs piesaista, piemēram, verbs *„māt”* piesaista `INSTRUMENTU`, bet ne `PACIENSU` (darbības izjutēju);
- pasaules (ontoloģiskās) zināšanas, piemēram, *tas, ko kāds būvē, ir celtne*.

Cilvēks ar šādiem daudznozīmības risināšanas uzdevumiem lielākoties[3] tiek galā ļoti efektīvi un precīzi, taču automātiskā gramatiskajā un semantiskajā analīzē tas ir nopietns šķērslis — lingvistisko un pasaules zināšanu formalizēšanas grūtību dēļ.

[3] Kā rāda pieredze, veidojot sintaktiski [27] un semantiski [28] anotētus tekstu korpusus, pat valodnieki vienu un to pašu teikumu/vārdu mēdz interpretēt dažādi. Nereti iemesls tam ir nepietiekams konteksts vai pārāk detalizēts nozīmju repozitorijs (skaidrojošā vārdnīca), arī atklātie jautājumi sintaksē.

Datorlingvistikā lietotās metodes var iedalīt divās principiālās pieejās: likumbāzētajā, kas saskaņā ar lingvistiski pamatotām teorijām tiecas precīzi aprakstīt valodu kā sistēmu, un statistiskajā, kas, izmantojot salīdzinoši virspusējas statistikas un mašīnmācīšanās metodes, mēģina iegūt valodas modeļa aproksimāciju. Kā plaši zināmu piemēru statistiskajai pieejai atkal var minēt (statistiskās) mašīntulkošanas (SMT) industriju, kas valodas un tulkošanas modeļus iegūst paralēlos tekstu korpusos rēķinot n-grammu varbūtības [29]. Statistika zināmā mērā ļauj izvairīties no nepieciešamības pēc padziļinātas zināšanu formalizēšanas un ļauj vieglāk nodrošināt plašāku valodas pārklājumu. Taču, jo augstāks analīzes līmenis tiek iesaistīts, jo padziļinātākas (strukturālākas) zināšanas ir nepieciešamas[4]. Šī doma uzskatāmāk ir ilustrēta 3. attēlā.



3. attēls. Dabiskās valodas analīzes principiālās pieejas un to mērķi (dalījums racionālistos un empīriķos ir aizgūts no A. Kilgarifa [31]).

Praktiskiem lietojumiem visperspektīvākā ir hibrīda pieeja, apvienojot zināšanās bāzētos un statistiskos līdzekļus. Piemēram, augstas precizitātes un plaša valodas pārklājuma parsētāju var iegūt, gramatikas kopsakarības mācoties sintaktiski anotētā tekstu korpusā (*treebank*) [32], savukārt mašīntulkošanā labākus rezultātus var sasniegt, kombinējot statistiskos un likumbāzētos modeļus [33]. Taču vispirms ir nepieciešams sagatavot pietiekami reprezentatīvus sintaktiski un semantiski anotētus tekstu korpusus. Piemēram, viens no apjomīgākajiem šāda veida resursiem ir vācu valodas korpuss SALSA/TIGER, kas aptver 50 tūkst.[5] sintaktiski anotētu teikumu un kurā semantiskās lomas ir norādītas 20 tūkst. verbu instancēm (to paplašinātājiem) [34]. Uz šī korpusa bāzes ir apmācīts, piemēram, relatīvi augstas precizitātes parsētājs *Shalmaneser* semantisko lomu noteikšanai patvaļīgos tekstos [35]. Jāatzīmē, jo augstāks (dziļāks) analīzes līmenis tiek aplūkots, jo darbietilpīgāka un dārgāka kļūst šāda korpusa izveide, līdz ar to salīdzinoši mazāku manuāli anotētu korpusu ir iespējams izveidot. Turklāt arvien būtiskāka nozīme kļūst domēnspecifiskām zināšanām.

Latviešu valodā šāda līmeņa korpuss nav pieejams, turklāt promocijas darba mērķis ir latviešu valodas struktūras padziļināta analīze, tāpēc darbā statistiskās, virspusējās metodes netiek plašāk apskatītas un izmantotas. Vēl vairāk, aplūkojamā valodas apakškopa tiek ierobežota: iesākumā tikai sintaktiski[6], bet pēc tam — arī semantiski.

[4] Pašreizējo SMT sistēmu kvalitāte to netieši parāda [30].
[5] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
[6] Manuāla visaptverošas gramatikas formalizācija ir ilgstošs un komplicēts process, kā arī latviešu valodniecībā ne viena vien gramatiska parādība vēl nav pietiekami precīzi un sistemātiski kategorizēta.

1 Kas ir vārda nozīme?

Šajā nodaļā ir īsumā ieskicēts sākotnējais pētījums leksiskās un ontoloģiskās semantikas modelēšanā un tajā gūtās atziņas, kas rosināja tālākos pētījumus, kas aprakstīti 2.–4. nodaļā un kuros ir sasniegti galvenie promocijas darba rezultāti.

Sākotnējā pētījuma gaita [1–3] atspoguļo cerības, ka teksta semantiskā analīze ir iespējama virzienā no augšas uz leju — ka primārais uzdevums teksta daudznozīmības risināšanā ir noteikt vārdu nozīmes (attiecībā pret kādu predefinētu skaidrojošo vārdnīcu, t.i., ontoloģiju). Vispopulārākais modelis šādu leksisko ontoloģiju (taksonomiju) izveidei ir Prinstonas *WordNet* [36], pēc kura principiem daudzām pasaules valodām ir izstrādātas analoģiskas taksonomijas[7]. *WordNet* pamatā ir sinonīmu kopas jēdziens: katra sinonīmu kopa reprezentē kādu jēdzienu, apvienojot vārdus (šī jēdziena leksiskās zīmes), kas kādās no to nozīmēm ir savstarpēji sinonīmi[8]. Savukārt starp sinonīmu kopām pastāv dažādas iepriekšdefinētas semantiskās attieksmes: hiponīmija (vispārināšana), meronīmija (daļa-veselais), ietveršana, kauzativitāte u.c.

Uz apjomīgās un detalizētās *WordNet* taksonomijas bāzes ir izstrādāta virkne dažādu algoritmu vārdu nozīmju automātiskai izšķiršanai tekstā (WSD — *word sense disambiguation*) [37], meklējot tuvāko semantisko distanci[9] starp vārdiem (to potenciālajām nozīmēm), kas parādās kādā konteksta „logā", analizējot, starp kurām dotajā kontekstā sastopamo vārdu nozīmēm ir visvairāk savienojumu (gan tiešu, gan netiešu) ar dažādu veidu semantiskajām attieksmēm u.tml. Tiesa, tā brīža labāko WSD sistēmu novērtējums (precizitātes un pārklājuma (*recall*) attiecība kopēji izmantotā paraugkorpusā) bija nedaudz virs 65% [38]. Turklāt jāņem vērā, ka naivā pieeja, vienmēr izvēloties katra vārda biežāk lietoto nozīmi[10], konkrētā paraugkorpusa gadījumā būtu nodrošinājusi 60% precizitāti [38].

Neskatoties uz to, tika atzīta nepieciešamība pēc šāda veida leksikona izveides arī latviešu valodai, taču, ņemot vērā tā manuālas konstruēšanas darbietilpību[11], tika apzinātas pusautomātiskas *WordNet* konstruēšanas iespējas. Viens no atbilstošākajiem leksisko zināšanu izguves avotiem ir (mašīnlasāmas) skaidrojošās vārdnīcas, kas būtībā satur lielu daļu no *WordNet* tipa taksonomijai nepieciešamajām semantiskajām attieksmēm, tikai tās ir dotas netiešā, neformālā veidā — nozīmju definīcijās un lietojumu piemēros. Iespaidojoties no rezultātiem, kas, izmantojot leksiski sintaktiskus šablonus (sk. 4. attēlu), sasniegti, analizējot nozīmju skaidrojumus angļu valodas vārdnīcās [40, 41], situācijas izvērtēšanai latviešu valodā tika veikts eksperiments, iegūstot nosacītas verbu sinonīmu kopas un to savstarpējās hiponīmijas attieksmes [2].

```
such NP0 as {NP}* {(and | or)} NP
".. works by such authors as Herrick, Goldsmith, and Shakespeare .."
⇒  HYPONYM(Herrick,author)
   HYPONYM(Goldsmith,author)
   HYPONYM(Shakespeare,author)
```

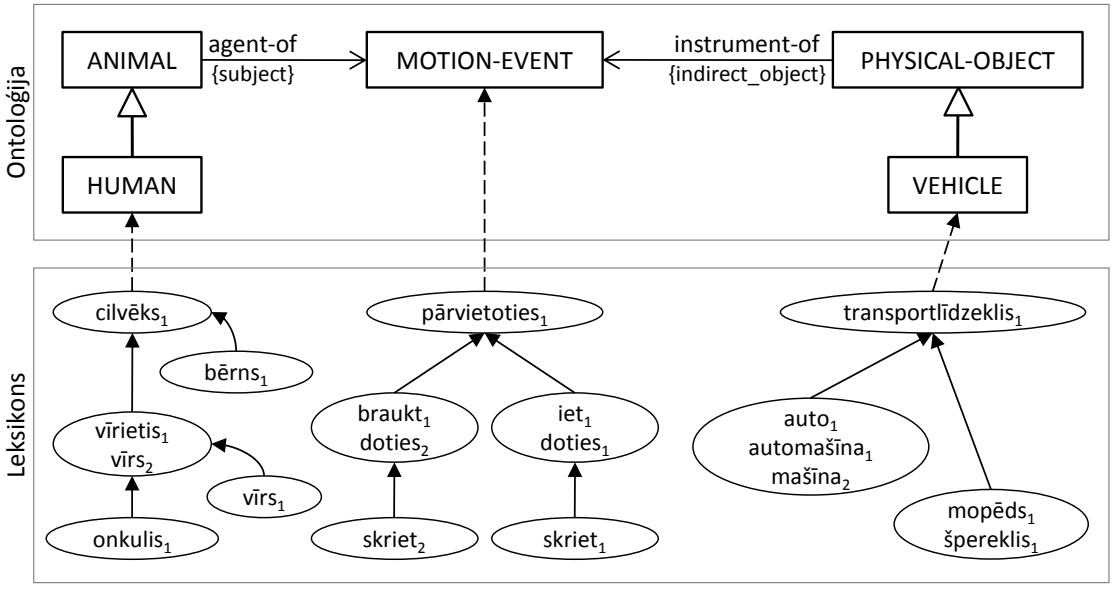
4. attēls. Leksiski sintaktiska šablona piemērs hiponīmijas attieksmju atklāšanai tekstu korpusā [42]. Hipotēze: $\forall$NP:hyponym(NP,NP_0), kur NP ir nominmāla vārdkopa (*noun phrase*).

[7] http://www.globalwordnet.org/gwa/wordnet_table.htm
[8] Jāatzīmē, ka absolūtu sinonīmu dabiskajās valodās nav daudz.
[9] *WordNet* ir orientēts, aciklisks grafs.
[10] Arī uz vārdu nozīmēm un to lietošanas biežumu var attiecināt Zifa likumu [39].
[11] Oriģinālais un arī apjomīgākais Prinstonas *WordNet* tika veidots un attīstīts gandrīz 20 gadus.

Šī eksperimenta laikā, veidojot sinonīmu kopas, netika izšķirtas vārdu nozīmes — ja kaut vienā no nozīmēm divi vārdi izrādījās sinonīmi, tie tika uzskatīti par absolūtiem sinonīmiem. Kā izrādījās, pietiekami precīzi definēti šabloni zināmā mērā spēj kompensēt šādu rupju pieņēmumu: iegūto sinonīmu kopu vidējais izmērs faktiski neatšķiras no Prinstonas *WordNet* rādītājiem[12]. Tas deva pārliecību, ka, paplašinot heiristisko šablonu kopu detalizētākai semantisko attieksmju izgūšanai[13] un izmantojot bilingvālus vai pat multilingvālus, sastatītus resursus (tekstu korpusus un tulkojošās vārdnīcas) precīzākam nozīmju dalījumam [44], kā arī veicot manuālu pirmsapstrādi un pēcapstrādi [1, 45], ir iespējams iegūt kvalitatīvu, praktiski lietojamu rezultātu.

Tomēr oriģinālajai *WordNet* pieejai ir vairāki vispārzināmi trūkumi: pārāk sadrumstalotas nozīmes ar grūti nosakāmām robežām[14], domēnspecifiskas informācijas trūkums, virspusējs formalizācijas līmenis un līdz ar to ierobežotas secināšanas iespējas. Praktiskiem lietojumiem nereti ir nepieciešamas domēnspecifiskas ontoloģijas, kurās ir precīzāk formalizētas nepieciešamās pasaules zināšanas, taču, no otras puses, šādu ontoloģiju sasaistei ar dabisko valodu ir nepieciešamas leksiskās zināšanas, turklāt robežu, kur beidzas leksiskās zināšanas un sākas ontoloģiskās, ir grūti noteikt [46].

Saprātīgs viduscelš ir šādu resursu nodalīšana un vienlaikus sastatīšana, nozīmju nianses atstājot leksiskās taksonomijas līmenī, bet lietoto jēdzienu noteikšanu (t.i., WSD) un teksta nozīmes formālo interpretāciju — ontoloģijas līmenī (sk. 5. attēlu). Šāda taktika tika izvēlēta promocijas darba pētījuma pirmajā posmā [1], un par tā brīža attīstītāko ontoloģiskās semantikas pieeju tika atzīta *OntoSem* [47], kuras izstrāde bija ilgusi vairāk nekā 10 gadus un kas pretendē uz vispārēju pasaules zināšanu aptveršanu.



5. attēls. Ilustratīvs piemērs leksisko un pasaules zināšanu nošķiršanai un sastatīšanai.

OntoSem nav vienīgā vispārējā („augšējā") līmeņa ontoloģija; tāda ir, piemēram, arī nozarē plaši zināmā SUMO ontoloģija, kas turklāt ir sastatīta ar *WordNet* [48], taču papildus leksiskajai semantikai *OntoSem* modelē arī sintaktisko saistāmību. Tomēr arī *OntoSem* pieejā

[12] Prinstonas *WordNet* verbu sinonīmu kopu vidējais izmērs ir aptuveni 1.82 (http://goo.gl/jcFgA), kas sakrīt ar eksperimenta rezultātiem [2]. Tiesa, izmantotā latviešu valodas vārdnīca aptver par 30% mazāk verbu, kas, ņemot vērā nozīmju neizšķiršanu, nedaudz iespaidotu rezultātus.

[13] Veicot padziļinātu, nozīmju definīcijās izmantotajai valodai pielāgotu sintaktisko analīzi [43].

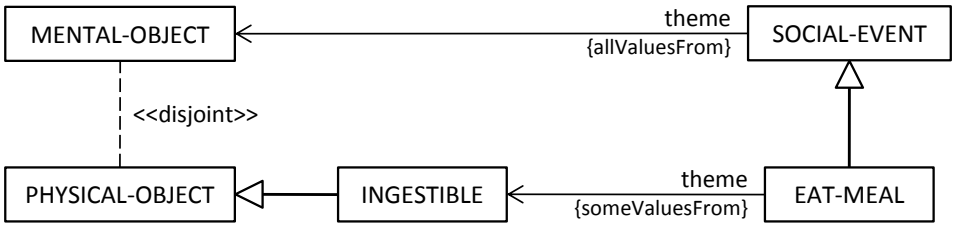
[14] Kā rāda pieredze, pat veicot manuālu *WordNet* nozīmju marķēšanu korpusā, vienprātīga nozīmju izvēle iespējama tikai 70% vārdlietojumu [38].

atklājās vairāki aspekti, kas būtiski apgrūtina atkalizmantošanas un attīstīšanas iespējas. Lai arī tā pretendē uz vispārēju zināšanu un līdz ar to neierobežota teksta nozīmes reprezentāciju, tās nosacīti plašais pārklājums ir virspusējs un padziļinātai analīzei katrs domēns jādefinē atsevišķi [49], ieviešot un detalizēti aprakstot nepieciešamos jēdzienus un leksikonu[15]. Tas ir ļoti darbietilpīgs un komplicēts process: ontoloģijā jādefinē ne tikai jēdzieni, kas reprezentē objektus un notikumus (darbības), bet arī specifiski šos notikumus aprakstoši skripti, savukārt leksikonā bez vārdu nozīmju sasaistes ar ontoloģijas jēdzieniem un sintaktisko šablonu aprakstīšanas jādefinē arī specifiski procedurālās semantikas skripti kontekstatkarīgu parametru rēķināšanai [50]. Tas viss šo pieeju padara atkarīgu no konkrētās *OntoSem* implementācijas. Turklāt *OntoSem* lietotais ontoloģijas formālisms neatbilst vispārpieņemtiem standartiem, konkrēti — OWL[16], kas liedz izmantot efektīvos OWL secinātājus un apgrūtina esošo domēnontoloģiju integrēšanu.

OntoSem ontoloģija ir definēta specifiskā, *freimu* jeb situāciju semantikas pieejā balstītā formālismā, kas, no vienas puses, piedāvā plašākus izteiksmes līdzekļus nekā OWL, bet, no otras puses, nenodrošina dažādu ierobežojumu uzdošanu, tādējādi ierobežojot automātiskas secināšanas iespējas. Turklāt *OntoSem* semantika daļēji ir iekodēta specifiskajos teksta analīzes rīkos. Lai padarītu šo ontoloģiju formālāku un savietojamu ar semantiskā tīmekļa standartiem un rīkiem, pētījuma ietvaros tā tika konvertēta uz OWL [3].

Aprakstošajā loģikā darbojas atvērtās pasaules princips: visas klases pēc noklusēšanas pārklājas, ja vien tiešā veidā nav pateikts, kurām no tām nevar būt kopīgu instanču. Šādi ierobežojumi ir nepieciešami gan leksisko nozīmju izšķiršanai, gan automātiskās secināšanas nodrošināšanai (WSD vajadzībām), tāpēc, manuāli analizējot konvertētās ontoloģijas klases (no augšas uz leju), šādi ierobežojumi tika pakāpeniski ieviesti (sk. `disjoint` 6. attēlā). Taču, parādoties pretrunām, atklājās fundamentāla problēma — neizbēgama daudznozīmība arī ontoloģijas līmenī[17]. Kā sistemātiski izšķirt jēdzienus un tos saistošās attieksmes? Kas formāli ir vārda nozīme un vai vārdiem vispār ir iepriekš definētas (definējamas) nozīmes [51]? Tā nav tikai *OntoSem* problēma, bet attiecas uz vispārējām ontoloģijām kā tādām. Vai ir iespējams izveidot konsekventu vispārēju (multidomēnu) ontoloģiju?

Lai labāk izprastu šos principiālos nozīmes rašanās jautājumus, tika nolemts meklēt atbildes no jauna: analizēt valodu kā **vienotu sistēmu** no gramatikas zemākajiem līmeņiem līdz pat ontoloģijai, formālās likumsakarības nosakot nevis virspusēji, bet iespējami dziļi[18].



6. attēls. Nesaderīgs (pretrunīgs) fragments no OWL formā konvertētās *OntoSem* ontoloģijas, ko atklāja automātisks secinātājs (klasei `EAT-MEAL` nav iespējamas instances).

[15] Līdz ar to *OntoSem* ontoloģija nekādā ziņā nav balansēta — atsevišķu pētniecības projektu rezultātā padziļināti ir aprakstīti tikai daži domēni (piem., politika, ekonomika un biomedicīna).

[16] Promocijas darbā ar OWL tiek saprasta valoda OWL DL (t.sk. jaunākā, OWL 2 valoda), kas balstīta aprakstošajā loģikā (DL) — pirmās kārtas loģikas (FOL) apakškopā, kas nodrošina efektīvas secināšanas iespējas.

[17] Tas attiecas ne tikai uz vārdu nozīmēm, bet arī uz jēdzienu izpratni: paralēli eksistē dažādas skolas un uzskati (piem., par to, vai cilvēks ir dzīvnieks — sk. 5. attēlu).

[18] Jāatzīmē, ka korpusā balstītās WSD sistēmas spēj uzrādīt arī ļoti labus rezultātus, sasniedzot precizitāti virs 80% — ja detalizētā *WordNet* vietā izmanto krietni robustāku nozīmju repozitoriju [52]. Tomēr ar nozīmju uzminēšanu vien nepietiek, jānosaka arī attieksmes starp teikuma locekļiem un jāveic konteksta analīze.

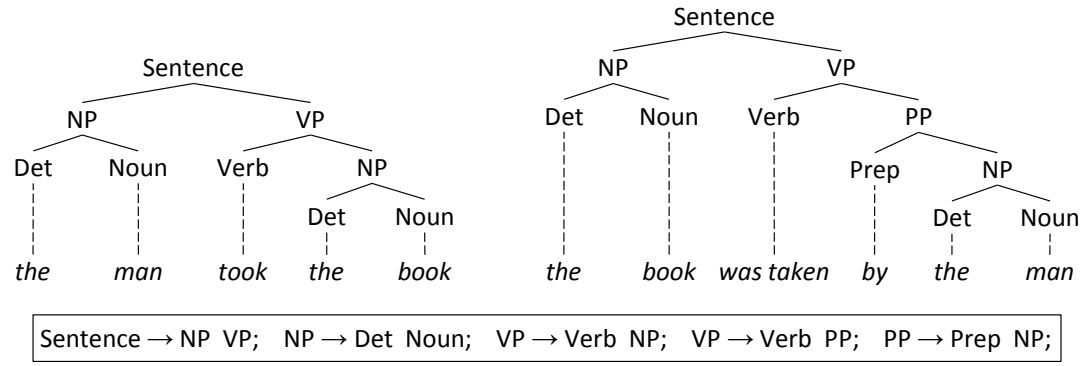
2 No valodas līdz ontoloģijai

Šajā nodaļā konspektīvi izklāstītas pētījuma otrā posma [4–6] nostādnes un galvenie rezultāti. Pretstatā iepriekšējam posmam teksta nozīmes un ontoloģijas izcelsme tika meklēta pretējā virzienā — no lejas uz augšu, t.i., no morfosintaktiskās analīzes līdz situāciju semantikai un diskursam, cenšoties nepalaist garām nevienu formālu detaļu. Lai arī praktisku iemeslu dēļ aplūkojamā latviešu valodas apakškopa tika sintaktiski ierobežota, nekādi konkrēti semantiskie ierobežojumi netika uzlikti.

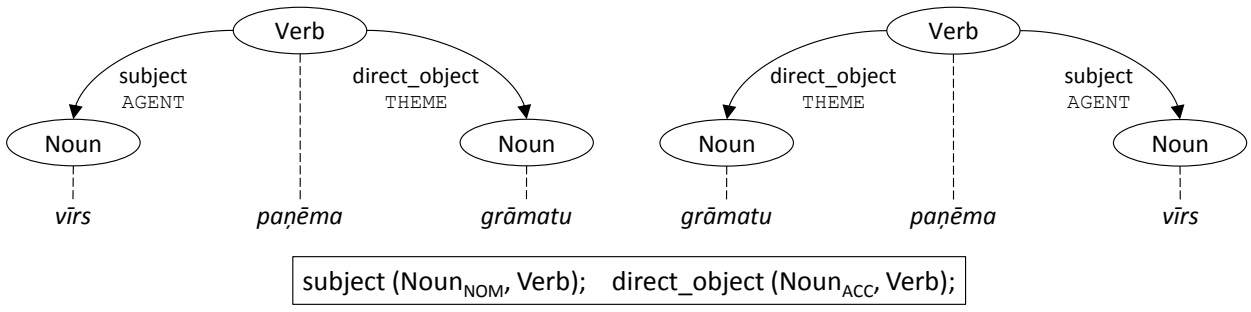
2.1 Atkarību gramatikā balstīta sintaktiskā analīze

Latviešu valoda pieder baltu valodu grupai, kur dominē sintētiskas formas: tā kā gramatiskās nozīmes ir izteiktas galvenokārt ar fleksijām (galotnēm), vārdu secība teikumā ir sintaktiski brīva — teikuma locekļu noteikšanu nodrošina vārdu locīšanas paradigmas un to saskaņošana teikumā. Taču pat sintētiskās valodās vārdu secība ir tikai relatīvi brīva: to ierobežo gan analītisko formu lietošana, gan semantiskais saistījums (sk. 4. nodaļu).

Sintaktiskajā analīzē var izšķirt divas galvenās pieejas: vārdkopu struktūras (PSG — *Phrase Structure Grammar*) jeb datorzinātnē labi zināmā bezkonteksta gramatikas pieeja [53] un atkarību gramatikas (*dependency grammar*) pieeja [16] (sk. 7. un 8. attēlu). Pirmajā tuvinājumā, pieņemot, ka tiek aplūkoti tikai projektīvi (*projective*) analīzes koki, t.i., tādi koki, kuru šķautnes nekrustojas (vienlaikus saglabājot vārdu lineāro secību teikumā), abi formālismi ir vāji ekvivalenti[19] un to gramatikas ir savstarpēji transformējamas [54]. Tomēr atšķirīgās pieejas piedāvā principiāli atšķirīgus skatījumus un metodiku ar attiecīgām priekšrocībām un trūkumiem.



7. attēls. Vienkāršots piemērs angļu valodas teikumu attēlošanai PSG gramatikā (darāmās un ciešamās kārtas analīzes koki atšķiras, kompensējot formālo atšķirību trūkumu morfoloģijas līmenī).



8. attēls. 7. attēlā doto teikumu ekvivalenti latviešu valodā un to attēlojums atkarību gramatikā. Latviešu valodā nav jāmaina kārta, lai mainītu vārdu secību — analīzes koks abos gadījumos sakrīt (tiek lietotas tās pašas funkcijas), atšķiras tikai lineārā secība: verba paskaidrotāju pozīcija teikumā.

[19] To gramatikas akceptē un ģenerē ekvivalentas teikumu kopas, taču nepiešķir teikumiem vienādu struktūru.

Analītiskām valodām (piem., angļu), kurās vārdu secība ir strikti noteikta, tradicionāla ir pieeja no augšas uz leju: izejot no vārdu lineārās secības, teikumi tiek dalīti vārdkopās un to sastāvdaļās (t.sk. rekursīvās) saskaņā ar produkcijas likumu kopu (bezkonteksta gramatiku)[20]. Turpretim fleksīvu, sintētisku valodu gadījumā tradicionāla ir pieeja no apakšas uz augšu: tiek noteikti teikuma locekļi un pakārtojuma attieksmes starp tiem atbilstoši valodā pieļaujamajām sintaktiskajām funkcijām (atkarību gramatikai). Turklāt atkarību gramatika, atšķirībā no PSG, ļauj adekvāti modelēt arī neprojektīvas (*non-projective*) struktūras, t.i., pārtrauktas vārdkopas (*discontinuous constituents*), kas valodās ar relatīvi brīvu vārdu secību ir samērā bieža parādība[21]. Tiesa, praksē analītisko un sintētisko valodu dalījums netiek īpaši ievērots: sintētisku valodu formalizēšanai un sintaktiski anotētu tekstu korpusu izveidei nereti ir tikusi izvēlēta PSG pieeja[22], savukārt pēdējo piecu gadu laikā, parādoties efektīviem, korpusā balstītiem parsētājiem[23], populāra ir kļuvusi atkarību analīzes pieeja — arī analītisko valodu gadījumā, un tiek attīstītas metodes esošo PSG korpusu konvertēšanai atkarību reprezentācijā [59]. Atkarību gramatikas priekšrocības ir atzītas, piemēram, tādos semantiskās analīzes lietojumos kā informācijas izguve [60] un semantisko lomu (freimu un to elementu) automātiska noteikšana neierobežotas valodas teikumos [61, 62].

Vienas vai otras pieejas priekšrocības nenosaka tikai valodas tips. No vienas puses, atkarību gramatika piedāvā salīdzinoši elastīgākas analīzes iespējas: ar ievērojami mazāku vienkāršu likumu (funkciju) kopu, nespecificējot vārdu lineāro secību, iespējams pārklāt vienlīdz plašu valodas apakškopu. No otras puses, atkarību gramatika nav piemērota valodas ģenerēšanai[24], kas ir PSG pieejas stiprā puse. Taču promocijas darba kontekstā svarīgāks aspekts ir sintaktiskās analīzes koku tālākas atainošanas iespējas semantiskajās struktūrās. Šajā ziņā priekšrocība ir atkarību gramatikas pieejai, kuras centrā ir verbs (izteicējs) un tā argumentu struktūra, kas tiek kodēta tiešā veidā — ar pakārtojuma attieksmēm (atkarībām). Turklāt verbs atrodas teikuma centrā ne tikai sintaktiski, bet arī semantiski — kā notikums (situācija, darbība). Līdz ar to šādas pakārtojuma attieksmes tiešā veidā var attēlot semantiskajās attieksmēs (sk. 8. attēlu). Jāatzīmē arī, ka šāds tiešs, divvirzienu attēlojums starp sintaktiskās un semantiskās analīzes līmeņiem atvieglo iespēju sintaktiskās daudznozīmības risināšanā izmantot informāciju par leksisko semantiku (saistāmību) [63].

Eksistē dažādas atkarību gramatikas teorijas, reprezentācijas un implementācijas [58]. Eksistē arī vairāki hibrīdi formālismi, kas PSG pieeju paplašina ar atkarību gramatikas elementiem. Zināmākā no tiem ir HPSG (*head-driven* PSG) pieeja [64], kas cita starpā nodrošina iespēju norādīt vārdkopu centrālos (*head*) locekļus un to valenci (saistāmību). Arī iepriekš minētā TIGER korpusa anotēšanā ir izmantota hibrīda shēma: šķautnēm, kas savieno vārdkopas un to sastāvdaļas, ir piekārtoti sintaktisko funkciju apzīmējumi, kā arī papildus strukturālajām šķautnēm tiek lietotas sekundāras funkcionālās saites [65].

Atkarību pieejā balstītās gramatikas lielākoties tiek realizētas, ļoti vienkāršojot oriģinālo L. Tenjēra pieeju [16], lingvistisko precizitāti un dabiskās valodas īpatnības upurējot par labu efektīviem analīzes algoritmiem [66]. Tādējādi katrs vārds (gan palīgnozīmes, gan pilnnozīmes jeb patstāvīgas nozīmes) faktiski tiek uzskatīts par patstāvīgu teikuma locekli, kas ir

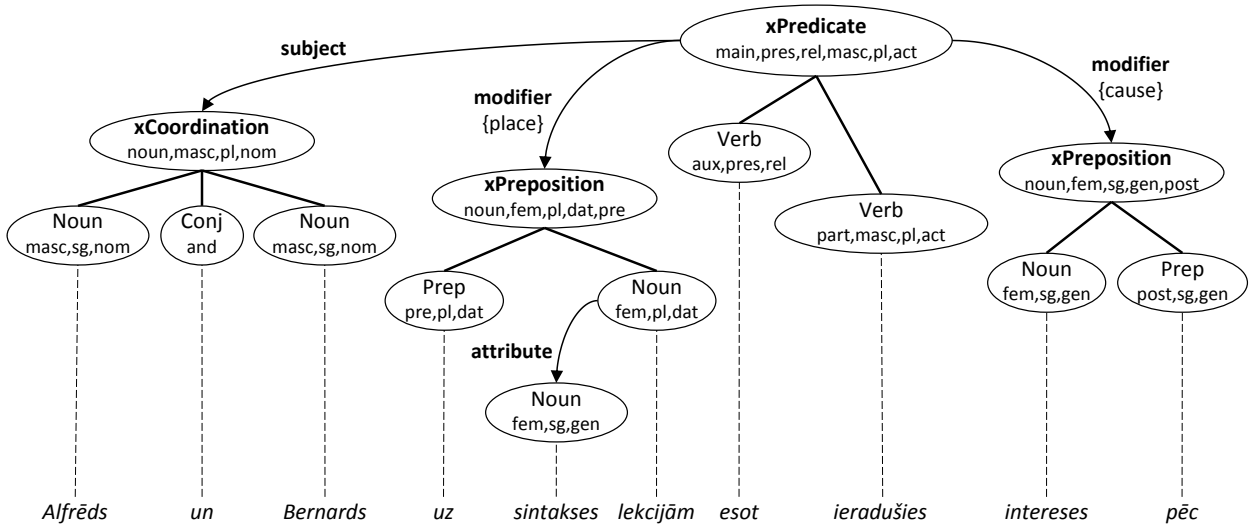
[20] Analīzes koka meklēšanas algoritmu, protams, var implementēt (un parasti implementē) no lejas uz augšu.
[21] Piemēram, Prāgas atkarību korpusā neprojektīvas struktūras ir sastopamas 23% teikumu [55].
[22] Tas daļēji ir skaidrojams ar datorlingvistikas pētījumos ilgu laiku izteikti dominējošās angļu valodas iestrāžu pārņemšanu un savietojamības nodrošināšanu ar standarta PSG algoritmiem, rīkiem un anotēšanas formātiem.
[23] Klasisko PSG algoritmu (piem., Ērlija algoritma [56]) sarežģītība ir $O(n^3)$. Projektīvu atkarību koku konstruēšanai ir tāda pati sarežģītība [57]. Taču neprojektīvu konstrukciju atpazīšana un dažādu globālu nosacījumu apmierināšana analīzes kokā vispārīgā gadījumā ir NP pilna problēma [55, 58].
[24] Lai arī sintaktiski teikuma locekļu secība izteikti fleksīvās valodās ir brīva, pragmatiski tā tomēr ir saistīta.

iesaistīts atsevišķā pakārtojuma attieksmē[25] [57]. Turklāt netiek ņemts vērā, ka valodās ar brīvu vārdu secību vārdu secība pat tīri sintaktiski nav absolūti brīva[26]. Ņemot vērā šos aspektus, pētījuma ietvaros ir izstrādāts hibrīds, atkarību gramatikā stingri balstīts sintaktiskās analīzes modelis, saukts par *SemTi-Kamols* gramatiku, kas nodrošina iespējami tuvu oriģinālās Tenjēra gramatikas formālu reprezentāciju un to paplašina [4, 5].

Tenjēra strukturālajā sintaksē ir izšķirti trīs pamatjēdzieni [16]: sintaktiskās attieksmes (*connexions structurales*), junkcijas (*jonctions*) un transference (*translation*). Analīzes koka virsotņu apzīmēšanai Tenjērs papildus ir ieviesis kodola jēdzienu (*nucléus*). Tās ir sintaktiski nedalāmas vienības, ko veido viens centrālais elements (pilnnozīmes vārds) un, neobligāti, viens vai vairāki palīgvārdi un kas tiek iegūtas morfosintaktiskās transferences rezultātā. Pakārtojuma attieksmes figurē kodolu līmenī, nevis teikuma virsējās struktūras (individuālu vārdu) līmenī. Savukārt junkcijas apvieno tādus elementus, kas atrodas nevis pakārtojuma, bet sakārtojuma (horizontālās) attieksmēs (galvenokārt vienlīdzīgi teikuma locekļi un teikuma daļas). Visbeidzot transference ir operācija, kuras izpildes laikā kodola funkcionālais jeb palīgvārds izmaina pilnnozīmes vārda oriģinālo kategoriju (funkciju).

Piedāvātajā *SemTi-Kamols* hibrīdajā modelī ir realizēts Tenjēra sintaktiskā kodola jēdziens, lietojot mehānismu, kas ir ļoti tuvs transferences operācijai. Turklāt kodola jēdziens ir paplašināts, ietverot tajā arī junkcijas. Lai to īstenotu, ir ieviests komplekso jeb „x-vārdu" jēdziens (sk. 9. attēlu). No formālā viedokļa x-vārds ir līdzeklis, kas, abstrahējot analītiskās formas un vienlīdzīgus teikuma locekļus, analīzes laikā pārtrauc teikuma dalīšanu apakšvirknēs un duālās dabas dēļ kalpo kā tilts starp PSG un vienkāršotu atkarību gramatiku:

- no PSG viedokļa x-vārds ir neterminmāls simbols, kas analīzes laikā aizvieto visus to veidojošos elementus (apraksta horizontālās relācijas);
- no atkarību gramatikas viedokļa x-vārds ir vienkāršs vārds, kas pakārtojuma attieksmē var tikt iesaistīts gan centrālā, gan atkarīgā teikuma locekļa lomā (tiek iesaistīts vertikālajās relācijās).



9. attēls. Vienkāršots teikuma analīzes koks hibrīdajā, atkarību gramatikā balstītajā modelī.

Transferences rezultātā analīzes kokā tiek izveidotas jaunas, mākslīgas virsotnes (kompleksi vārdi) ar savām morfosintaktiskajām pazīmēm, kas lielākoties tiek mantotas no tās

[25] Piemēram, analītiskā verba forma *„bija paņēmis"* tiek mākslīgi sadalīta divos teikuma locekļos (divās analīzes koka virsotnēs), kā rezultātā palīgverbs paskaidro patstāvīgās nozīmes verbu vai otrādi.
[26] Piemēram, prepozicionālajā konstrukcijā *„uz galda"* prievārda pozīcija vienmēr ir pirms lietvārda.

veidojošajiem elementiem, bet var atspoguļot arī papildu informāciju, kas attiecas uz x-vārdu kopumā un kas var būt nepieciešama atbilstošās(-o) pakārtojuma funkcijas(-u) piemērošanā. Rekursīvi aizvietojot visas analītiskās formas un vienlīdzīgos teikuma locekļus ar šādiem x-vārdiem, tiek iegūta vienkārša teikuma struktūra, kuras analīzei var izmantot labi zināmās vienkāršotās atkarību gramatikas metodes, piemēram, klasisko Kovingtona algoritmu [57], kas secīgi akceptē vārdu pa vārdam un katru no tiem mēģina piesaistīt kā centrālo (*head*) vai atkarīgo (*dependent*) locekli kādam (iespējami tuvākajam) no iepriekšējiem vārdiem (katrai virsotnei, izņemot sakni, ir tieši viens vecāks), tādējādi analīzes koku atrodot laikā $O(n^3)$. Turklāt Kovingtona algoritmu iespējams paplašināt, ietverot tajā arī x-vārdu analīzi [67].

Kompleksos vārdus veidojošajiem elementiem ir noteikta secība, kas gramatikā tiek norādīta, taču tā neattiecas uz šo elementu atkarīgajiem locekļiem, kas teikumā var būt brīvi iestarpināti starp x-vārdu elementiem un analīzes kokā parādīties kā vienkāršas atkarības. Piemēram, 9. attēlā analītiskās formas *„uz lekcijām”* elementu *„lekcijām”* paskaidro apzīmētājs *„sintakses”*. Līdz ar to, formalizējot gramatiku, x-vārdi tiek definēti, norādot tikai to primāros elementus, turpretim sekundārie (netiešie) tiek ietverti, izmantojot kopīgās pakārtojuma funkcijas (sk. 10. attēlu). Jāatzīmē, ka gan x-vārdus veidojošie elementi, gan to potenciālie paskaidrotāji var būt ne tikai vienkārši vārdi, bet arī paši par sevi x-vārdi.

Vārdformas (morfoloģiskās analīzes rezultāts):

```
w('Alfrēds',[noun,masc,sg,nom]).
w('un',[conj,and]).
w('Bernards',[noun,masc,sg,nom]).
w('uz',[prep,pre,sg,gen]).
w('uz',[prep,pre,pl,dat]).
w('sintakses',[noun,fem,sg,gen]).
w('sintakses',[noun,fem,pl,nom]).
w('sintakses',[noun,fem,pl,acc]).
w('lekcijām',[noun,fem,pl,dat]).
```

x-Vārdu šabloni (produkcijas likumi):

```
x([[_,[verb,aux,TENSE,MOOD]],[_,[verb,part,GEND,NUM,VOICE]]],
  ['xPredicate',[verb,main,TENSE,MOOD,GEND,NUM,VOICE]])
  :- member(TENSE,[pres,fut]).
x([[_,[noun,_,_,CASE]],['un',[conj,and]],[_,[noun,_,_,CASE]]],
  ['xCoordination',[noun,_,pl,CASE]]).
x([[_,[prep,pre,NUM,CASE]],[_,[noun,GEND,NUM,CASE]]],
  ['xPreposition',[noun,GEND,NUM,CASE,pre]]).
x([[_,[noun,GEND,NUM,CASE]],[_,[prep,post,NUM,CASE]]],
  ['xPreposition',[noun,GEND,NUM,CASE,post]]).
```

Sintaktiskās funkcijas (atkarību likumi):

```
f(subject,_,[noun,GEND,NUM,nom],[verb,main,_,MOOD,GEND,NUM,_])
  :- member(MOOD,[ind,cond,rel]).
f(modifier_place,_,[noun,_,_,_,pre],[verb,main,_,_,_,_,_]).
f(modifier_cause,_,[noun,_,_,_,post],[verb,main,_,_,_,_,_]).
f(attribute,left,[noun,_,_,gen],[noun,_,_,_]).
```

10. attēls. Stipri vienkāršots fragments no realizētās *SemTi-Kamols* gramatikas, kas apraksta 9. attēlā doto teikumu. Morfoloģisko pazīmju kopā[27] ir arī tādas pazīmes/vērtības, kas tieši neattiecas uz morfoloģiju, bet ir noderīgas korektākas un mazāk daudznozīmīgas sintaktiskās analīzes nodrošināšanai (piem., norāda prievārda iespējamo pozīciju, `pre` vai `post`, attiecībā pret vārdkopas centrālo elementu).

SemTi-Kamols modelis var tikt izmantots dažādu sintaktisko konstrukciju adekvātai attēlošanai, t.sk. analītiskām verbu formām (piem., veidojot saliktus laikus un modālus verbus;

[27] http://www.semti-kamols.lv/doc_upl/TagSet.pdf

sk. *„esot ieradušies”* 9. attēlā), prepozicionālām konstrukcijām (sk. *„uz lekcijām”* un *„intereses pēc”*) un sakārtojuma attieksmju aprakstīšanai (sk. *„Alfrēds un Bernards”*). Teorētiski x-vārda jēdziens var tikt paplašināts arī attiecībā uz pakārtotām un sakārtotām teikuma daļām: palīgteikumu (caur tā centrālo locekli — izteicēju) parasti var piesaistīt kādam no virsteikuma locekļiem kā paskaidrotāju, savukārt vienlīdzīgas teikuma daļas var traktēt analoģiski kā vienlīdzīgus teikuma locekļus (arī šeit par saknēm ņemot šo teikuma daļu izteicējus). Taču praksē, automātiski analizējot teikumu, šādi x-vārdi radītu ievērojamu sintaktisko daudznozīmību[28]. Līdz ar to, formalizējot latviešu valodas gramatiku, ir aplūkota tās apakškopa — vienkārši paplašināti teikumi — vienkāršoti pieņemot, ka semantiski katru teikuma daļu var aplūkot atsevišķi un kontekstuāli sasaistīt teksta semantikas (diskursa) līmenī[29]. Vēl vairāk, tā kā formāli precīzas gramatikas manuāla definēšana, vispusīgi pārklājot pat tikai vienkāršus paplašinātus teikumus, ir ļoti darbietilpīgs process[30], projekta ietvaros ir izstrādāts daļējs parsētājs (*chunker*): ja netiek atrasts pilna teikuma analīzes koks, tiek atgriezti iespējami garākie apakškoki, kurus var uzkonstruēt ar dotajiem x-vārdu un atkarību likumiem. Tādējādi tas ir piemērots neierobežota teksta daļējai gramatiskai analīzei[31] ar potenciālu informācijas izguves lietojumos.

Lai izvairītos no nekorektu vai ļoti maz ticamu analīzes koku konstruēšanas un tādējādi ievērojami samazinātu nepamatoti daudznozīmīgu konstrukciju rašanās iespējas (kas parasti noved pie kombinatoriska sprādziena analīzes variantu kopā), *SemTi-Kamols* gramatikā sintaktisko funkciju lietošana bez morfoloģisko pazīmju unifikācijas tiek papildus ierobežota ar dažādiem lokāliem un globāliem nosacījumiem: tiek norādīts, kuriem x-vārdu elementiem ir iespējami paskaidrotāji (apakškoks) un kuriem nav (parasti funkcionālajiem vārdiem), atsevišķiem teikuma locekļiem tiek fiksēta pozīcija attiecībā pret virslocekli, ko tie paskaidro (piem., apzīmētājs latviešu valodā tiek lietots pa kreisi no lietvārda), kā arī tiek ierobežots, kuri teikuma locekļi ir unikāli (piem., subjekts) un kuri nav (piem., apstākļi). Šādi nosacījumi ir arī solis nedaudz tuvāk valodas ģenerēšanas lietojumam (lai arī tas nav šī gramatikas modeļa mērķis). Taču teikuma locekļu atkārtošanās ierobežojumu apmierināšana ir NP pilna problēma [67], kas pašreizējo *SemTi-Kamols* parsētāja prototipa implementāciju padara nepraktisku (meklējot iespējamos analīzes variantus, tiek veikta pārmērīga pilnā pārlase[32]). Problēmai ir iespējams praktisks risinājums, piemēram, visu analīzes variantu atrašanai (neierobežojot teikuma locekļu atkārtošanos) lietojot dinamiskās programmēšanās paņēmienus (analoģiski kā Ērlija algoritmā) un korekto variantu atlasi reducējot uz meklēšanu variantu tabulā [67]. Tomēr promocijas darba uzdevums nav efektīva analīzes algoritma izstrāde, bet gan analīzes modelis (korpusa anotēšanas shēma) kā tāds, nenodarbojoties ar priekšlaicīgu optimizāciju. Turklāt gramatikā balstītu parsētāju agrāk vai vēlāk aizstās korpusā (empīriskos datos) balstīts statistisks parsētājs: vienkāršotu atkarību modeļu gadījumā statistiski algoritmi darbojas lineārā laikā [32], pat pieļaujot neprojektīvu atkarību heiristisku analīzi [55].

---

[28] Piemēram, formāli būtu ļoti grūti nošķirt vienlīdzīgus izteicējus no salikta sakārtota teikuma daļām.

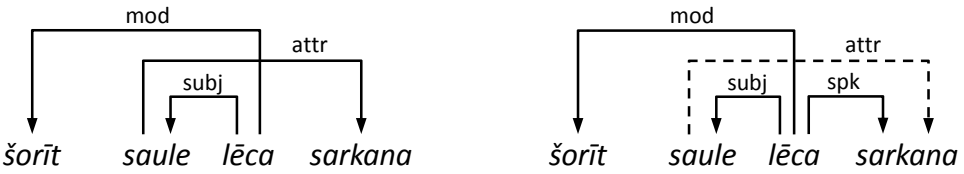
[29] Teorētiski ikvienu saliktu sakārtotu vai pakārtotu teikumu ir iespējams pārveidot par vienu vai (parasti) vairākiem vienkāršiem paplašinātiem teikumiem, pakārtojuma saikļu vietā lietojot, piemēram, anaforiskus vietniekvārdus (sk. 2.2. apakšnodaļu). Par salikta sakārtota teikuma daļām līdzīgi norāda arī Tenjērs [16].

[30] Divu gadu garumā, periodiski papildinot gramatiku ar tipiskiem un gadījuma piemērus aprakstošiem likumiem, ir definēti nepilni 300 x-vārdu šabloni un vairāk nekā 100 atkarību likumi. Optimizācijas rezultātā izdevies ievērojami samazināt sākotnējo likumu kopu (450/200), būtiski paplašinot gramatikas pārklājumu. Tomēr reprezentatīvas gramatikas iegūšana ir iespējama tikai uz līdzsvarota tekstu korpusa pamata.

[31] Izstrādātā gramatika tiek lietota latviešu valodas tekstu korpusa automātiskā morfoloģiskā marķēšanā [21], saglabājot visus analīzes variantus (kuru ir mazāk, nekā ja tiktu izmantots tikai morfoloģiskais analizators).

[32] Latviešu valodas tīmekļa korpusa automātiskā morfoloģiskā marķēšana tika veikta *BalticGRID* [21].

Jāpiemin, ka *SemTi-Kamols* parsētāja (algoritma) pašreizējā implementācija, līdzīgi kā lielākā daļa atkarību gramatiku parsētāju [58], neatbalsta neprojektīvu pakārtojuma attieksmju analīzi (modelis kā tāds šādu ierobežojumu neuzliek). Līdzīgi kā citās sintētiskās, fleksīvās valodās arī latviešu valodā šādu gadījumu nav mazums. Taču mūsdienu latviešu valodniecības pieeja ļauj lielāko daļu no tiem reducēt uz projektīva koka konstruēšanu[33] — balstoties uz sekundāri predikatīva komponenta (SPK) jēdzienu [68]. Tiek uzskatīts, ka atkarīgais loceklis (SPK) šādā gadījumā paskaidro situāciju kopumā, t.i., gan tiešo centrālo locekli, gan izteicēju. Pamatojums tam ir semantisks: kā jau iepriekš tika uzsvērts, sintaktiski brīvā vārdu secība pragmatiski tomēr ir saistīta, t.i., atspoguļo atšķirīgu teikuma nozīmi vai nozīmes niansi. Taču, velkot divas šķautnes, tiek pārkāpts iepriekšminētais princips, ka katrai virsotnei ir ne vairāk kā viens vecāks (sk. 11. attēlu). Tāpēc *SemTi-Kamols* gramatikā ir aprakstītas attieksmes tikai starp SPK un izteicēju, izvairoties gan no neprojektīviem kokiem, gan vairākām vecāku virsotnēm. To, ka netieši tiek paskaidrots arī subjekts (vai cits loceklis — atkarībā no galotnes), var izsecināt heiristiski.



11. attēls. Teikums, kura struktūra tipiskā atkarību gramatikas pieejā tiktu atainota kā neprojektīvs koks (pa kreisi), *SemTi-Kamols* gramatikā tiek traktēts kā projektīvs koks (pa labi), balstoties uz sekundāri predikatīva komponenta jēdzienu un tā vienkāršotu interpretāciju.

Piedāvātajam modelim ir zināma līdzība ar minēto HPSG pieeju, kurā vārdkopas pazīmes tiek nodotas caur vārdkopas centrālo elementu, t.i., vārdkopu reprezentē tās centrālais elements. Šī arī nav pirmā reize, kad ticis norādīts uz vienkāršotās atkarību gramatikas trūkumiem, neadekvāti attēlojot virkni būtisku sintaktisko konstrukciju, kas sastopamas gan analītiskās, gan sintētiskās valodās. Viens no tuvākajiem ir T. Jērvinena un P. Tapanainena piedāvātais gramatikas modelis (un tā vēlākā realizācija *Connexor* parsētājā), kurš ir pa vidu starp pārlieku vienkāršotajām pieejām un Tenjēra oriģinālo gramatiku: iespējami adekvāts gramatikas attēlojums, saglabājot praktiskas analīzes iespējas (lietojot tikai lokālus, heiristiskus nosacījumus) [66]. *SemTi-Kamols* modelī ir iets tālāk, izšķirot vienkāršus un kompleksus vārdus (kodolus) un sintaktiskā kodola jēdzienu attiecinot plašāk, t.sk. vienlīdzīgu teikuma locekļu formalizēšanai[34]. Turklāt, spriežot pēc pašreizējās *Connexor* realizācijas[35], tehniski tiek izmantots vienkāršots atkarību gramatikas modelis, kurā ne tikai junkciju, bet arī kodolu elementu atainošanai tiek izmantotas iepriekš atrunātas relācijas.

Cits Tenjēra pieejai ļoti tuvs formālisms, ir TDS (*Tesnière's Dependency Structure*) [69]. TDS ir izmantoti visi Tenjēra pamatjēdzieni, un šajā reprezentācijā ir izdevies konvertēt *Penn Treebank* — vienu no pirmajiem un lielākajiem sintaktiski (PSG) anotētajiem angļu valodas korpusiem [70]. Tomēr TDS pieeju ir iespējams uzlabot: Tenjēra gramatikā atkarību veidi ir vienkāršoti, un pakārtojuma attieksmes faktiski ir anonīmas — sintaktisko funkciju vietā Tenjērs lieto vispārinātas vārdšķiru klases, kas transferences operāciju laikā tiek

[33] Ņemot vērā iepriekš doto nosacījumu, ka aplūkojamā valodas apakškopa ir ierobežota līdz vienkāršiem paplašinātiem teikumiem. Salikta pakārtota teikuma gadījumā šķautnes var krustoties arī pāri teikuma daļām. Šāda situācija var rasties, piemēram, apzīmētāja palīgteikuma dēļ (piem., *„[..] tādu valodu analīzei, kurās vārdu secība ir brīva."*). Interesanti, ka apzīmētāja palīgteikumus bieži vien ir iespējams sistemātiski konvertēt par apzīmētājiem, iegūstot vienkāršu paplašinātu teikumu (sk. 4.3. apakšnodaļu).

[34] Jērvinens un Tapanainens piedāvā junkciju elementus savstarpēji saistīt ar atrunātām `cc:` relācijām [66].

[35] http://www.connexor.eu/technology/machinese/demo/syntax/

piešķirtas kodoliem. Lai arī tas nodrošina vienkāršāku analīzi, teikuma struktūra TDS modelī tiek reprezentēta salīdzinoši virspusēji. Tāpēc *SemTi-Kamols* modeli var uzskatīt par paplašinājumu arī TDS pieejai. Visbeidzot jāatzīmē, ka Prāgas atkarību gramatikas korpusā PDT (*Prague Dependency Treebank*), kas ir pasaulē vadošais šāda veida korpuss, ir lietota vairāku līmeņu anotācijas shēma, adaptējot FGD (*Functional Generative Description*) teoriju [71]: dziļajā jeb tektogramatiskajā līmenī, kas attēlo teikuma lingvistisko nozīmi (neizmantojot papildu zināšanas par situāciju), analītiskās vārdformas, apozīcijas u.tml. tiek apvienotas kopīgās virsotnēs, neizšķirot to iekšējo struktūru, savukārt virsējā jeb analītiskajā līmenī tiek lietota vienkāršota atkarību gramatika, mākslīgi veidojot pakārtojuma attieksmes starp palīgvārdiem un pilnnozīmes vārdiem.

Izstrādātais un latviešu valodas apakškopai aprobētais modelis ir piemērojams arī citu sintētisko valodu gramatikas reprezentācijai. Baltu valodām pēc gramatiskās struktūras ļoti tuvas ir, piemēram, slāvu valodas. Jāatzīmē arī, ka šobrīd LU MII norit darbs pie šī modeļa aprobēšanas neierobežotiem tekstiem, veidojot sintaktiski anotētu latviešu valodas paraugkorpusu[36], ko cita starpā ar laiku varēs izmantot arī statistiska parsētāja apmācīšanai.

2.2 Daudzdimensiju ontoloģija

Pieņemot, ka izstrādātais hibrīdais sintaktiskās analīzes modelis nodrošina iespējami precīzas un adekvātas latviešu valodas gramatikas formalizēšanas iespējas, jautājums ir, kā rīkoties tālāk, formalizējot leksisko un situāciju semantiku, kā arī teksta diskursu kopumā. Pat iegūstot apjomīgu anotētu tekstu korpusu un uz tā bāzes — augstas precizitātes statistisku parsētāju, teikuma sintaktiskā struktūra tomēr izriet no tā semantiskās struktūras [16]. Tāpēc, ņemot vērā sākotnējās nostādnes, šī pētījumu posma turpinājumā tika meklēti risinājumi, kā iespējami formāli reprezentēt dažādos semantikas un pragmatikas aspektus tā, lai teksta analīzes un reprezentācijas līmeņi nebūtu nošķirti, bet gluži pretēji — ietverti vienotā modelī[37].

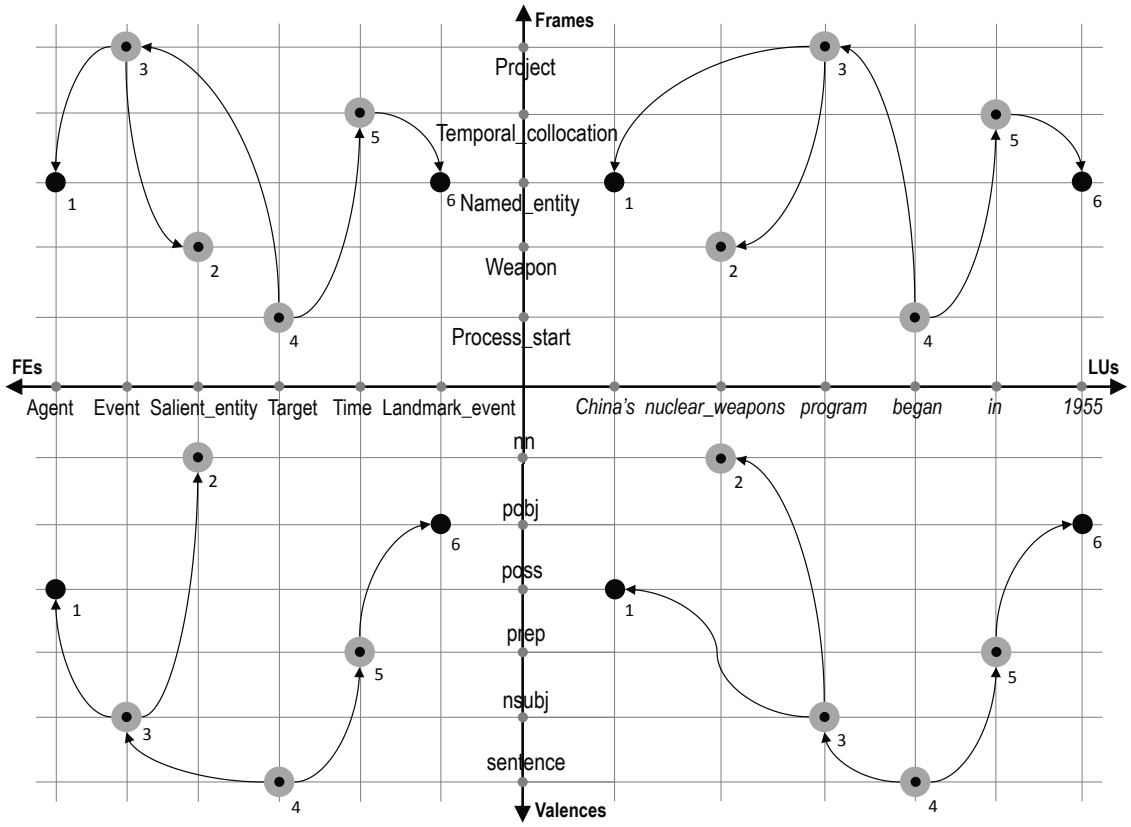
Viena no visattīstītākajām situāciju (*frame*) semantikas pieejām ir *FrameNet* [74], kas piedāvā padziļinātu ieskatu leksiskajā semantikā un tās mijiedarbībā ar teikuma sintaktisko struktūru. *FrameNet* modelī leksiskā semantika tiek formalizēta, vārdus (to nozīmes) piesaistot *freimiem* — vispārinātām situācijām. Atšķirībā, piemēram, no *OntoSem* pieejas (sk. 1. nodaļu), kuras ontoloģija ietver tūkstošiem jēdzienu, *FrameNet* apraksta mazāk nekā 1000 situācijas. Lai arī šo idealizēto situāciju kopa nav pilnīga un lielā mērā ir iegūta, analizējot nelielu, nebalansētu domēnspecifisku tekstu kolekciju[38], *FrameNet* modelis savā būtībā ir pietiekami elastīgs un universāls. To apliecina, piemēram, *FrameNet* veiksmīgā izmantošana, semantiski anotējot apjomīgo SALSA/TIGER korpusu (sk. 1. nodaļu), kā arī nosacīti daudzsološie rezultāti automātiskā (statistiskā) semantisko lomu marķēšanā neierobežotos tekstos [75]. Šeit jāpiemin arī *WordNet* loma, kas nodrošina plašu *FrameNet* pārklājumu, jo oriģinālais *FrameNet* leksikons ir ļoti ierobežots[39] [76].

Tomēr *FrameNet* izstrādē galvenā uzmanība ir pievērsta lingvistiskajiem aspektiem, tāpēc tā formalizācijas pakāpe ir daļēja. Precīzāka ontoloģiska formalizācija uzlabotu šī nozīmīgā resursa izmantošanas iespējas automātiskā semantiskā analīzē. Šādā nolūkā ir tikusi veikta gan *FrameNet* daļēja konvertēšana uz OWL [77], gan sastatīšana ar vispārējo SUMO ontoloģiju [78]. Tas darīts galvenokārt tādēļ, lai nodrošinātu striktu ontoloģisku kontroli pār

[36] Korpusa anotēšanai ir izstrādāts *SemTi-Kamols* PML (*Prague Markup Language*) profils [72], kas ļauj izmantot tādus PML formātu atbalstošus rīkus kā *TrEd* [73], kas ir pārbaudīts praksē, veidojot PDT korpusu.
[37] Semantiskās analīzes līmenī reprezentācijas modeļa izvēle nav atkarīga no valodas tipa.
[38] http://goo.gl/S8Two
[39] Tas joprojām aktualizē 1. nodaļā minētās iestrādes latviešu valodas leksiskās taksonomijas izveidē.

aptuveni 40 predefinētiem *FrameNet* semantiskajiem tipiem, kas paredzēti freimu elementu[40] (FE) ierobežošanai, kā arī freimu mantošanas struktūras formalizēšanai. Taču problēma ir tā, ka pašā *FrameNet* šie līdzekļi nav konsekventi lietoti to vājās formalizācijas un daudznozīmības dēļ. No otras puses, *FrameNet* formalizācijas mēģinājumos nav veltīta īpaša uzmanība tam, ko *FrameNet* atspoguļo precīzi: leksisko vienību (LU — *lexical units*) sasaistei ar freimiem un to valencēm (FE).

Lai nezaudētu dažādās analīzes dimensijas, ko piedāvā *FrameNet*, kā arī lai nodrošinātu iespēju elastīgi ieviest jaunas dimensijas (piem., diskursa anotācijām), pētījuma ietvaros ir piedāvāta netradicionāla pieeja, kas visus analīzes līmeņus un aspektus apvieno vienā, universālā modelī — daudzdimensiju ontoloģijā jeb galīgā punktu kopā, kas izkārtoti daudzdimensiju telpā[41] [6]. Šāds ontoloģijas modelis ļauj uzkrāt visas lingvistiskās un pasaules zināšanas, kuras iespējams nokodēt teksta manuālas anotēšanas laikā. Pieņemot, ka paralēli tiktu anotēti gan vispārīgi, gan dažādu nozaru teksti, uzkrātie fakti ar laiku varētu kļūt par pamatu korelāciju automātiskai izguvei: gan statistiska semantiskā parsētāja apmācībai, gan objektīvākas vispārējas ontoloģijas izveidei.



12. attēls. Četras plaknes vizualizē dotā teikuma 4D anotācijas grafa dažādus 2D aspektus. Neizceltās virsotnes apzīmē vārdā nosauktas instances (*named entities*), kas analīzes laikā tiek lietotas FE aizpildīšanai, bet ne freimu iniciēšanai. Orientētās šķautnes ir pakārtojuma attieksmes.

Piedāvātās pieejas pamatā ir ideja par *FrameNet* anotāciju ģeometrisku izvietošanu daudzdimensiju telpā tā, lai pašas anotācijas ir atliktas uz (daudzdimensiju) koordinātu asīm, savukārt teikumā minētās instances ir izvietotas kā anonīmi šīs telpas punkti, kurus saista anonīmas pakārtojuma relācijas (sk. 12. attēlu). Ja iedomājamies, ka *FrameNet* četru dimensiju (4D) telpā ir izkārtoti punkti un šķautnes, kas reprezentē ne vien individuālu teikumu,

[40] Freimu elementi tiek saukti arī par slotiem un semantiskajām lomām.
[41] Ieviešot daudzdimensiju ontoloģijas jēdzienu, mēs uz brīdi atkāpjamies no OWL un pirmās kārtas loģikas kā zināšanu reprezentācijas valodas.

bet veselu tekstu korpusu, šīs telpas aizpildījums būtu sarežģīts un tomēr veidotu zināmu struktūru: punktu un šķautņu izkārtojums, visticamāk, iezīmētu noteiktas šīs telpas apakštelpas[42]. Respektīvi, valodas un pasaules zināšanu šabloni, pēc kuriem vadījies teksta anotētājs (cilvēks), tiešā un netiešā veidā būtu atspoguļoti ģeometriski: anonīmo instanču un relāciju savstarpējās pārklāšanās vietas un pārklājuma „blīvums" no ontoloģijas viedokļa var tikt interpretētas (vispārinātas) kā klases un to tipiskās īpašības (*properties*).

Tomēr 4D anotācijas vispārīgā gadījumā nav pietiekamas, lai pilnībā izšķirtu teikuma nozīmi, jo bez plašāka konteksta analīzes patvaļīgam teikumam var būt iespējamas vairākas vienlīdz korektas interpretācijas. Koreferenču[43] noteikšana ir viens no primārajiem diskursa analīzes uzdevumiem. Lai to nodrošinātu, pirmkārt, esošā 4D telpa jāpaplašina ar vēl vienu dimensiju, kas dotu iespēju secīgi atzīmēt tekstu veidojošos teikumus un to daļas. Otrkārt, koreferenču norādīšanai ir jāievieš atsevišķa relācija (papildus pakārtojuma relācijai). Tādējādi teikumi un to daļas tiktu sasaistītas ar diskursa referentu un koreferenču palīdzību. Turklāt šādas sasaistes rezultātā tiktu norādīts kuriem diskursa referentiem (punktiem) ir kopīga identitāte.

Tekstos minēto instanču izvietojums (klasteri) daudzdimensiju telpā kopā ar koreferenču anotācijām un no tā izrietošajiem apgalvojumiem *„tas pats"* (`owl:sameAs`), *„līdzīgs tam"* (*`similarTo`*[44]) un *„tāda veida"* (`rdf:type`) būtībā atspoguļo pasaules (ontoloģiskās) zināšanas, kas ietvertas tekstu korpusā un teorētiski var tikt izmantotas freimu elementu semantisko tipu — vismazāk formālās *FrameNet* daļas — objektīvai sistematizēšanai un ierobežošanai. Jāatzīmē, ka piedāvātais modelis ir konceptuāli savietojams arī ar CPA (*Corpus Pattern Analysis*) — *FrameNet* papildinošu pieeju [80], kurā valences tiek definētas katrai verba nozīmei atsevišķi, nevis vispārinātiem freimiem, pieļaujamās valenču vērtības ierobežojot ar sistemātiskāku ontoloģiju [81], salīdzinot ar *FrameNet* semantiskajiem tipiem.

Šīs pētījuma daļas pašreizējais rezultāts ir teorētisks, un praksē tas ir pārbaudīts, manuāli anotējot vien atsevišķus latviešu valodas gadījuma teikumus un rindkopas. Lai arī *SemTi-Kamols* projekta ietvaros ir izstrādāts prototips rīkam (GUI), kas ievērojami atvieglo teksta daudzdimensiju manuālu anotēšanu, un anotēšanas modelis kā tāds ir konceptuāli vienkāršs, kā rāda pieredze, šis process ir ļoti darbietilpīgs[45]. Lai gan ir parādīts, ka, izmantojot angļu valodai pieejamos statistiskos sintaktiskos un semantiskos parsētājus, iespējams automātiski iegūt teksta daudzdimensiju anotācijas uzmetumu [82], kuru ar GUI rīka palīdzību atliek „vien" manuāli rediģēt, latviešu valodas gadījumā pašlaik atduramies pret jau minēto vistas un olas problēmu — korpusā balstīta plaša pārklājuma statistiska parsētāja trūkumu.

[42] Šāda tendence tika novērota pētījuma laikā, veicot atsevišķu, nelielu gadījuma tekstu anotēšanu.

[43] Promocijas darbā ar koreferencēm tiek saprastas anaforas un to antecedenti. Piemēram, tekstā *„Sarkangalvīte devās pie vecmāmiņas. Viņai rokās bija groziņš."* anafora ir *„viņai"*, antecedents — *„Sarkangalvīte"*.

[44] Autors piekrīt P. Heisam [79], ka `owl:sameAs` semantika nereti ir pārāk strikta (precīza), lai to pēc būtības atbilstoši (korekti) lietotu.

[45] Tas attiecas arī uz „klasiskiem" sintaktiski un semantiski anotētiem tekstu korpusiem, tādiem kā minētais SALSA/TIGER.

3 Mikroontoloģijas kā formāli precīzas skaidrojošās vārdnīcas

Ņemot vērā latviešu valodas sintaktiski un semantiski anotēta tekstu korpusa trūkumu, promocijas darba turpmākie divi pētījumi [7–11] ir veltīti valodas deterministiskas analīzes iespējām, aplūkojamo latviešu valodas apakškopu sašaurinot ne tikai sintaktiski, bet arī semantiski, vienlaikus saglabājot ierobežotās valodas praktisku lietojamību noteiktās sfērās.

3.1 Loģikā balstīta dabiskās valodas apakškopa

Eksistē dažādas visnotaļ izsmalcinātas ierobežotas dabiskās valodas (CNL — *controlled natural language*), lielākoties angļu valodas apakškopas, piemēram, ACE (*Attempto Controlled English*) [20], kas piedāvā neformālus, intuitīvus izteiksmes līdzekļus, saglabājot teksta precīzas interpretēšanas iespējas [83]. Šīs valodas var iedalīt divās vispārīgās grupās, atbilstoši formālistu vai naturālistu pieejai[46] [84]. Formālistu pieejā tiek nodrošināta CNL deterministiska divvirzienu attēlošana kādā no formālajām valodām, piemēram, pirmās kārtas loģikā vai tās apakškopā — aprakstošajā loģikā [85], kas dod iespēju izmantot esošos automātiskās secināšanas un modeļu būvēšanas rīkus [7]. Lai gan loģikā balstīta CNL ir šķietami neformāls, augsta līmeņa zināšanu reprezentācijas līdzeklis (salīdzinot ar atbilstošo formālo notāciju), tā tomēr savā būtībā ir formāla valoda, kas ir ne mazāk un ne vairāk izteiksmīga kā atbilstošais formālisms un kuras teikumu interpretācija ir iepriekš (intuitīvi) paredzama. Savukārt naturālistu pieejā daudznozīmības iespēja tiek samazināta, bet ne izslēgta, tādējādi pieļaujot plašāku dabiskās valodas pārklājumu un neformālāku tās lietošanu, vienlaikus saglabājot augstas precizitātes, taču varbūtiskas interpretācijas iespējas. WSD (*word sense disambiguation*) nodrošināšanai tiek izmantota kontekstjutīga heiristika un tiek veikta virspusēja sintaktiskā analīze; respektīvi, tiek meklēta "vislabākā" interpretācija. Promocijas darbā ir izvēlēta formālistu pieeja un tāda dabiskās valodas apakškopa, kuras semantika ir precīzi definējama ontoloģiju valodā OWL.

Lai nodrošinātu loģikā balstītu CNL tekstu viennozīmīgu interpretāciju (diskursa reprezentācijas struktūras konstruēšanu [86] un šīs struktūras translēšanu vēlamajā formālajā notācijā), pamatā tiek izmantoti divu veidu ierobežojumi: neliela, predefinēta likumu kopa potenciāli daudznozīmīgu sintaktisko konstrukciju viennozīmīgai interpretēšanai un viennozīmīgs leksikons — pilnnozīmes vārdi netiek interpretēti, bet tiek uztverti kā predikātu vārdi, kuru nozīmi definē vienīgi loģikas formulas, kuras ir atvasinātas no tajā brīdī analizētā teksta. Tātad lietoto vārdu nozīme netiešā veidā tiek definēta pašā CNL tekstā. Papildus, nosakot diskursa referentus, tiek pieņemts, ka anaforas antecedents ir tuvākā nominālā vārdkopa (NP), kas neatrodas noslēgtā redzamības apgabalā [87]. Turklāt, ja pati anafora arī ir NP (angļu valodā: marķēta ar noteikto artikulu *„the"*), abām NP ir leksiski precīzi jāsakrīt, kas dabiskā valodā, protams, bieži vien netiek ievērots [7].

Sintaktiskie ierobežojumi uzspiež „tikai" precizitāti (tiešumu), konsekvenci un vienmēr (visos kontekstos) vienu un to pašu „pareizo" sintaktiskās analīzes koka variantu, bet leksiskie ierobežojumi rada fundamentālu šķērsli, jo dabiskās valodas leksika ir daudznozīmīga pēc savas būtības. Šī ierobežojuma sekas ir tādas, ka loģikā balstītās valodas galvenokārt tiek lietotas kā CNL saskarne domēnontoloģiju izstrādē un verbalizēšanā, pieņemot, ka viena domēna ietvaros lietotā terminoloģija ir viennozīmīga.

Taču leksiskā daudznozīmība (polisēmija) faktiski ir vienīgais intuitīvais un dabiskais līdzeklis, kas CNL tekstā ļautu integrēt dažādu domēnu pasaules jeb fona zināšanas. Polisēmijas galvenais cēlonis ir tas, ka valodā ir galīga vārdu kopa, bet jēdzienu un kontekstu

[46] Šeit var vilkt zināmas paralēles ar ievadā doto dalījumu racionālistos un empīriķos.

(nosacīti — domēnu) kopa ir potenciāli neierobežota. Lai arī laika gaitā rodas jauni vārdi (termini), tas ir lēns izlases veida process, jo jaunradītajiem vārdiem jākļūst vispāratzītiem un lietotiem. Līdz ar to esošo vārdu atkalizmantošana dažādos kontekstos ir ikdienas parādība un to var iedalīt divos pamatveidos: metaforiska[47] atkalizmantošana, balstoties uz līdzību (piem., *„pele"* attiecībā uz *„kursora pozicionēšanas ierīce"*), un metonīmiska atkalizmantošana, balstoties uz saistību (piem., *„bibliotēka"* attiecībā uz *„bibliotēkas ēka"*) [89]. Par laimi viena vārda dažādas nozīmes parasti tiek lietotas dažādos domēnos (kontekstos) [90]; atbilstošā domēna tieša norādīšana CNL tekstā būtu līdzeklis ierobežotas polisēmijas atbalstam, saglabājot loģikā balstītas CNL īpašības.

Turpinot vārda nozīmes formālo pazīmju meklējumus, promocijas darbā ir piedāvāta metode klašu vārdu (lietvārdu un lietvārdu savienojumu) nozīmju izšķiršanai — leksisku domēnu identifikatoru minimālai, sistemātiskai lietošanai, integrējot un lietojot vairāku domēnu ontoloģijas [8]. Tādējādi formālistu pieejā tiek pieļauta ierobežota (kontrolējama) polisēmija, vienlaikus saglabājot deterministiskas analīzes iespējas. Piedāvātais risinājums var tikt paplašināts arī īpašību (*properties*), t.i., verbu nozīmju formālai izšķiršanai, kombinējot deklaratīvās (statiskās) un procedurālās (dinamiskās) fona zināšanas [8], taču dinamiskie aspekti promocijas darbā netiek sīkāk apskatīti.

3.2 Mikroontoloģijas

Tipisks risinājums daudznozīmīgu klašu izšķiršanai ir mākslīgu, kompleksu leksēmu konstruēšana, vārdformas līmenī norādot konkrēto nozīmi (piem., *„bibliotēka-celtne"* pretstatā *„programmatūras-bibliotēka"*). Tomēr konsekventi principi bieži vien netiek ievēroti, padarot ontoloģijas leksikonu nesistemātisku un galalietotājam nedraudzīgu [91]. Promocijas darbā izstrādātā metode ir līdzeklis šī trūkuma novēršanai (vai vismaz mazināšanai), nodrošinot veidu, kā nepieciešamības gadījumā (un tikai) šādus vairākvārdu savienojumus veidot sistemātiski un lielā mērā automātiski — domēnontoloģiju integrēšanas laikā.

Iekšēji viennozīmīgas, saderīgas domēnontoloģijas, kurās jēdzienu apzīmēšanai ir lietota dabiskās valodas leksika, sauksim par mikroontoloģijām. Tā vietā, lai mēģinātu leksikonu padarīt globāli viennozīmīgu, tiek piedāvāts sašķelt fona zināšanas daudzos, relatīvi mazos, taču leksiski (terminoloģiski) viennozīmīgos mikrodomēnos (sk. 13. attēlu). Šādas pieejas priekšrocība ir tās elastība un mērogojamība[48].

Aizvietojot vairākdomēnu un vispārējas ontoloģijas ar virkni domēnspecifisku mikroontoloģiju, neizbēgami tiek radīta ontoloģiju sapludināšanas, sastatīšanas jeb integrēšanas[49] problēma. Šajā kontekstā ontoloģiju integrēšanas un WSD problēmas ir cieši saistītas, un pārliecinošu panākumu trūkums kā vienas, tā otras problēmas risināšanā lielā mērā ir saistīts ar to, ka šīs problēmas parasti tiek aplūkotas neatkarīgi[50]. Promocijas darbā tiek piedāvāts abas problēmas risināt vienlaikus, izmantojot pusautomātisku metodi vārdu nozīmju dalīšanai OWL mikroontoloģiju integrēšanas procesā. Respektīvi, tiek piedāvāts par nozīmju „repozitoriju" (skaidrojošo vārdnīcu) izmantot tās pašas semantiski precīzās mikroontoloģijas, nevis ārējus leksikonus. Paredzams, ka šāds loģiski pamatots vārdu nozīmju dalījums sekmēs (pus)automātiskas WSD iespējas faktus aprakstošos CNL tekstos.

---

[47] „Valoda ir mirušu metaforu kapsēta" [88].

[48] Piedāvātajam mikroontoloģijas jēdzienam ir zināma līdzība ar J. Vilka u.c. ieviesto apkārtnes (*environment*) jēdzienu [92], kā arī ar *Cyc* mikroteoriju jēdzienu [93], taču mikroontoloģiju pieeja atšķiras ar to, ka tā izmanto standarta OWL DL secināšanas līdzekļus un ir savietojama ar esošajām CNL.

[49] Angļu valodā lietotie termini: *merging*, *matching*, *aligning*.

[50] Lai arī šajā virzienā ir virkne iestrādņu [94], tipiskā pieeja ir virspusēju metožu izmantošana, izmantojot ārējas leksiskas taksonomijas (piem., *WordNet*) [95].

	Domēns	Terminoloģija
T-kaste	General	*Every building is a physical_entity.* *Every collection is an abstract_entity.* *No **physical_entity** is an **abstract_entity**.*
	Building	*Everything that has a roof is a building.* *Every **library** is a building.* *Every green_roof is a roof.*
	Programming	*Everything that contains something is a collection.* *Every **library** is a collection.* *Every function is something.*
A-kaste	**Fakti**	
	*There is a **library**[building] that has a green_roof.*	
	AbsoluteValue is a function. [..] *The **library**[programming] contains AbsoluteValue.*	

13. attēls. Dotās trīs mikroontoloģijas (verbalizētas valodā ACE) un no tām neatkarīgie fakti (arī valodā ACE) ilustrē leksēmas *„library"* formālās daudznozīmības rašanos (lietojot automātisku secinātāju, tiktu atklāta pretruna). Lai izveidotu saderīgu, integrētu ontoloģiju, atbilstošā nozīme (vārdtelpa) būtu jānorāda tiešā veidā katrā no leksēmas *„libray"* lietojumiem.

Teikumus, kurus akceptē un ģenerē ierobežotās valodas, kas ir balstītas OWL semantikā, var iedalīt divās grupās: terminoloģiskajos (ontoloģiskajos) teikumos un faktus aprakstošajos jeb faktuālajos teikumos, kas attiecīgi atbilst aprakstošās loģikas T-kastes (*TBox*) un A-kastes (*ABox*) aksiomām [96]. Ontoloģiskie teikumi definē kategorijas (klases) un iespējamās attieksmes starp tām (piem., *„Ikviena pele*[zooloģija] *ir dzīvnieks."*), savukārt faktuālajos teikumos ir runa par indivīdiem (instancēm), kas pieder noteiktām kategorijām (piem., *„Šī pele*[datortehnika] *ir pieslēgta pie datora."*).

Atšķirībā no līdzšinējās CNL prakses promocijas darbā piedāvātajā pieejā abu veidu teikumi tiek strikti nodalīti ontoloģiskajos un faktuālajos tekstos. Tā kā mikroontoloģiju kontekstā ontoloģiskie teksti pēc definīcijas ir viennozīmīgi, polisēmija ir aktuāla tikai faktuālajos tekstos. Šādam dalījumam ir praktisks iemesls: CNL „vidusmēra" lietotājs tiek atslogots no ontoloģisko zināšanu sniegšanas, ļaujot viņam koncentrēties uz faktu uzdošanu un vaicājumiem, bet mikroontoloģiju izstrāde un integrēšana paliek nozaru ekspertu un zināšanu inženieru pārziņā — saskaņā ar esošo praksi.

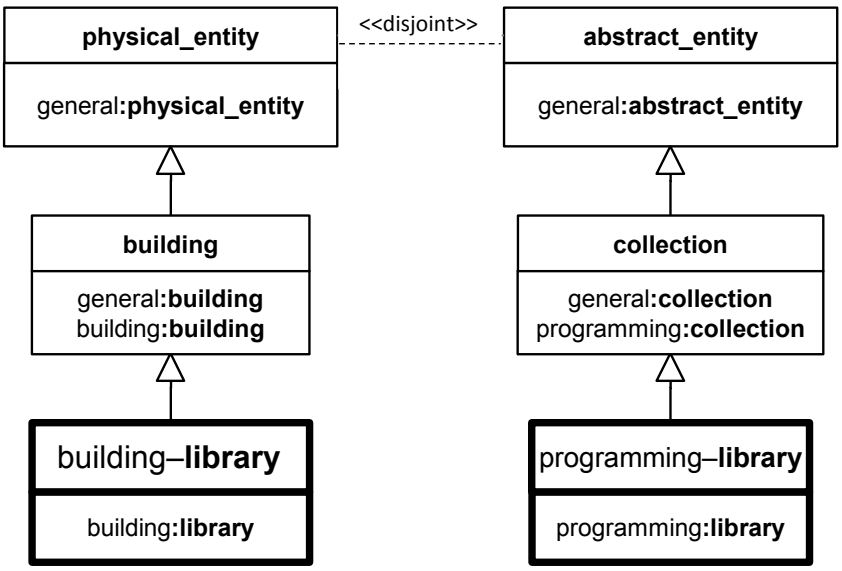
3.3 Mikroontoloģiju integrēšana un daudznozīmības risināšana

Ir būtiski uzsvērt, ka piedāvātajā polisēmijas risināšanas pieejā WSD process ir sadalīts divos soļos: vispirms, mikroontoloģiju integrēšanas laikā, tiek iegūts minimālais nozīmju dalījums[51], bet pēc tam, integrētajai mikroontoloģijai atbilstošo faktuālo tekstu analīzes laikā, tiek noteiktas daudznozīmīgo vārdlietojumu nozīmes, balstoties uz iepriekš iegūto (integrētajā mikroontoloģijā fiksēto) nozīmju dalījumu. Abu soļu īstenošanai tiek izmantots viens un tas pats instruments — OWL DL secinātājs, ar kuru var automātiski pārbaudīt gan iegūtā nozīmju dalījuma (T-kastes), gan faktuālā teksta interpretācijas (A-kastes) loģisko saderību.

Vispārīgais algoritms, pēc kura vadīties, integrējot (sastatot) mikroontoloģijas, pirmajā tuvinājumā ir pavisam vienkāršs. Pirmkārt, ja vairākās mikroontoloģijās figurē klases ar vienādiem vārdiem, izvirzām hipotēzi, ka jēdzieni, kurus šīs klases reprezentē, pārklājas un ka šīs klases tādējādi ir ekvivalentas (`owl:equivalentClass`). Otrkārt, pārbaudām izvirzītās hipotēzes patiesumu: ja sastatītā ontoloģija joprojām ir saderīga (*consistent*), hipotēze

[51] Atgādinājums: promocijas darbā mikroontoloģiju polisēmijas problēma tiek aplūkota ierobežoti — tikai attiecībā uz klašu vārdiem; klašu īpašības (*properties*) tiek uzskatītas par globāli viennozīmīgām [8].

apstiprinās un ekvivalences aksiomu paturam, bet, ja tā ir kļuvusi nesaderīga, hipotēzi (un ieviesto aksiomu) noraidām, savukārt nesaderīgo klašu vārdus pārsaucam (precizējam), ietverot tajos izcelsmes domēnu vārdus[52] (sk. 14. attēlu).



14. attēls. Piemērs integrētai ontoloģijai, kas iegūta, sastatot 13. attēlā dotās mikroontoloģijas. Ilustrācijā ir atainotas ekvivalento klašu kopas un attieksmes starp tām. Katru kopu veido viena vai vairākas klases, kas nākušas no mikroontoloģijām (norādes uz to izcelsmes vārdtelpām ir saglabātas), un viena jaunizveidota klase, kas pieder noklusētajai integrētās ontoloģijas vārdtelpai. Daudznozīmības gadījumā jaunizveidotās klases vārdā ietverts vārdtelpas lokālais identifikators kādai no mikroontoloģijām.

Taču papildus ir jāņem vērā vēl divi aspekti:

- Iespējams, katru vienādi nosaukto klašu kopu var sadalīt tādās apakškopās, kuru elementi ir savstarpēji saderīgi, bet šīs kopas savā starpā ir pretrunīgas.
- Iespējams, viens un tas pats vārds dažādās mikroontoloģijās ir lietots tādu klašu apzīmēšanai, no kurām vienas ontoloģijas klase patiesībā ir citas ontoloģijas apakšklase (piem., astronomijas mikroontoloģijā vārds *„mēness”* varētu tikt lietos jebkuras planētas pavadoņa apzīmēšanai, bet zemkopības mikroontoloģijā ar to visticamāk būs apzīmēts tieši planētas Zeme pavadonis).

Līdz ar to precizēts algoritms mikroontoloģiju automātiskai sastatīšanai, pieņemot, ka dotajai mikroontoloģiju kopai ir iespējams tikai viens „konceptuāli pareizs” nozīmju dalījums, ir šāds:

- Starp visām dotajām mikroontoloģijām:
  (1) Katram vienādi nosaukto klašu pārim
    - katrā no abiem virzieniem
      a) ieviešam relāciju `rdfs:subClassOf` un pārbaudām pašreizējā sastatījuma saderību, izmantojot OWL DL secinātāju;
      b) ja pašreizējais sastatījums ir saderīgs, paturam jaunieviesto relāciju; pretējā gadījumā to atmetam.
  (2) Katram vienādi nosaukto klašu pārim `(X,Y)`
    - ieviešam `<X owl:equivalentClass Y>`, ja ir spēkā abas aksiomas: gan `<X rdfs:subClassOf Y>`, gan `<Y rdfs:subClassOf X>`.

[52] Vārdu savienojumu veidošana terminoloģijā ir tipisks paņēmiens, kā izvairīties no jaunu vārdu radīšanas.

Piedāvātā nozīmju dalīšanas procedūra dos korektu un viennozīmīgu rezultātu tikai tad, ja pirms tās darbināšanas būs izpildīti divi būtiski nosacījumi:

- Dotās mikroontoloģijas satur pietiekošus ierobežojumus, lai rastos pretruna, ja relācija `rdfs:subClassOf` tiek ieviesta starp jebkurām divām vienādi nosauktām klasēm, kas „konceptuāli pareizajā" nozīmju dalījumā patiesībā nepārklājas.
- Dotās mikroontoloģijas satur pietiekošus ierobežojumus, lai rastos pretruna, ja attiecībā uz jebkurām divām vienādi nosauktām klasēm, par kurām „konceptuāli pareizajā" nozīmju dalījumā ir spēkā aksioma `<X rdfs:subClassOf Y>`, bet ne otrādi, tiek ieviesta aksioma `<Y rdfs:subClassOf X>`.

Diemžēl automātiski noteikt, vai iegūtais mikroontoloģiju sastatījums ir konceptuāli pareizs vai nē, nav iespējams — šī pārbaude paliek zināšanu inženiera (un nozaru ekspertu) ziņā. Tas nozīmē, ka iteratīva izvēlēto mikroontoloģiju precizēšana[53] un integrētās ontoloģijas pārbaude, līdz tiek sasniegti abi iepriekšminētie nosacījumi, vispārīgā gadījumā ir darbietilpīgs uzdevums. Neskatoties uz to, šī kopumā pusautomātiskā metode nodrošina sistemātiskāku un elastīgāku pieeju, salīdzinot ar pilnībā manuālu domēnontoloģiju sastatīšanu vai pat vispārējas (makro-) ontoloģijas izstrādi „no nulles". Tā ir arī formāli precīzāks līdzeklis, salīdzinot ar pilnībā automātiskām, virspusējām ontoloģiju sastatīšanas metodēm.

Savukārt CNL galalietotājs šajā pieejā ir pilnībā atbrīvots no ontoloģisku zināšanu nodrošināšanas un var fokusēties vienīgi uz faktuāliem tekstiem, t.i., faktu verbalizēšanu (pārfrazēšanu ierobežotā valodā). Daudznozīmīgo vārdlietojumu nozīmju izšķiršana atkal ir iespējama, izmantojot automātisku secinātāju: apvienojot integrētās ontoloģijas un faktuālā teksta aksiomas un piemeklējot tādu nozīmju (vairākvārdu savienojumu) komplektu, pie kura apvienojums joprojām ir saderīgs. Tiesa, vispārīgā gadījumā (nepietiekama konteksta dēļ) šī metode nodrošina daļēju WSD, izslēdzot kontekstuāli nesaderīgos variantus. Problēmas var radīt arī dabisku un intuitīvu vārdtelpu identifikatoru (mikroontoloģiju prefiksu) izvēle, īpaši, ja tiek integrēti kāda plašāka domēna apakšdomēni [8]. Pieņemot, ka mikroontoloģiju gadījumā veidosies samērā rupjš (*coarse-grained*) nozīmju dalījums, domēnu prefiksu vietā varētu lietot skaitliskus nozīmju identifikatorus, līdzīgi kā tas tiek darīts WSD pieejās, kas balstās uz tradicionālām vārdnīcām un *WordNet*, taču izvairoties no pārlieku smalkā nozīmju dalījuma problēmas (sk. 1. nodaļu). Jebkurā gadījumā lietotājdraudzīgam CNL redaktoram ir nepieciešama priekšāteikšanas funkcionalitāte [87, 97], kas šajā gadījumā ne tikai nodrošinātu CNL gramatikai atbilstošu teikumu ērtu konstruēšanu, bet daudznozīmīgajiem vārdlietojumiem varētu palīdzēt izvēlēties atbilstošo nozīmi.

Lai gan uz reāliem, praktiskiem piemēriem piedāvātā metode nav pārbaudīta, salīdzinot ar vispārējām ontoloģijām (tādām kā *OntoSem*), mikroontoloģiju pieeja teorētiski piedāvā vairākas būtiskas priekšrocības: tā ir mērogojama — reālijas nav „jāsaspiež" ierobežotā skaitā aptuvenu kategoriju; netiek uzspiests viens vienīgs, saderīgs pasaules modelis, ļaujot līdzās pastāvēt dažādiem (pretrunīgiem) viedokļiem un skolām, kam nav jābūt iepriekš paredzētiem. Mikroontoloģijas var uztvert kā momentuzņēmumus, kas nodrošina nepastāvīgu, pagaidu (vēsturisku) reāliju kategorizēšanu. Kā arī vairs nav aktuāls dalījums leksiskajās un pasaules zināšanās, bet tikai dalījums ontoloģiskajos un faktuālajos tekstos.

Mikroontoloģiju pieeja tika demonstrēta uz ierobežotas angļu valodas piemēra, taču šī metode ir pilnībā valodneatkarīga un tieši tāpat būtu lietojama ierobežotas latviešu valodas gadījumā.

[53] Var būt nepieciešams papildus ieviest starpontoloģijas, definējot ierobežojumus (*bridging axioms*) starp dažādu mikroontoloģiju dažādiem jēdzieniem [8].

4 Ierobežota latviešu valoda ontoloģiju izstrādei un verbalizēšanai

Pēdējā pētījumu posmā [9–11] ir aplūkota tāda latviešu valodas apakškopa, kurai ir iespējams nodrošināt deterministisku un tādējādi pilnībā automātisku, precīzu sintaktisko un semantisko analīzi. Salīdzinot ar iepriekšējo posmu (sk. 3. nodaļu), par izpētes centrālo objektu kļūst ontoloģiskie, nevis faktuālie teksti, t.i., mikroontoloģiju izstrāde ar iespējami dabiskiem, intuitīviem un viennozīmīgiem CNL līdzekļiem, kā arī esošu ontoloģiju (gan individuālu, gan integrētu mikroontoloģiju) atspoguļošana jeb verbalizēšana iespējami dabiskā, precīzā CNL (t.i., ontoloģisko tekstu sintēze). Kā jau tika minēts, (mikro-)ontoloģiju izstrāde ir ne tikai darbietilpīgs, bet arī zināšanu ietilpīgs process, kurā bez pieredzējušiem sistēmanalītiķiem (zināšanu inženieriem) jāiesaista arī nozaru eksperti.

Lai OWL ontoloģijas padarītu uztveramākas gan nozaru ekspertiem, gan arī sistēmanalītiķiem, tiek lietotas dažādas augsta līmeņa notācijas (salīdzinot ar aprakstošās loģikas matemātisko notāciju vai OWL serializēšanas formātiem). Šīs notācijas var iedalīt vismaz trīs grupās: grafiskās valodas, tādas kā UML un tās profili [23], ierobežotas dabiskās valodas, tādas kā ACE [20], kā arī „cilvēkam lasāmās" zemāka līmeņa notācijas, tādas kā Mančestras OWL sintakse [98], kas OWL semantiku atspoguļo ļoti tieši un tādēļ prasa ievērojami lielāku treniņu, lai gūtu pietiekamas lasīšanas un rakstīšanas iemaņas.

CNL pieeja piedāvā visneformālāko un, iespējams, visintuitīvāk uztveramo zināšanu reprezentācijas līdzekli, t.i., nav nepieciešama nekāda apmācība, lai cilvēks spētu salīdzinoši precīzi interpretēt automātiski ģenerētu CNL tekstu. Ierobežotās valodas sevi jau ir pierādījušas praksē — priekšmetu apgabalos, kur ekspertu iesaistīšana ir ļoti būtiska, piemēram, veicinot apjomīgas topogrāfijas ontoloģijas izstrādi Lielbritānijas kartogrāfijas aģentūrā [99]. Turpretim grafiskās notācijas nodrošina alternatīvu skatījumu, uzskatāmāk atklājot ontoloģijas struktūru un kopsakarības, un ir veiksmīgi izmantotas, piemēram, Latvijas medicīnas statistikas reģistru reinženierijā un datu piekļuvē [100]. Faktiski šīs abas perspektīvas ir abpusēji papildinošas: grafiska valoda nav īsti piemērota sarežģītu ontoloģijas ierobežojumu (klašu izteiksmju), īpašību ķēžu (*property chains*) un izvedumu likumu uzdošanai, kas savukārt ir parocīgi un viegli uztverami izdarāms, lietojot dabisko valodu.

Promocijas darbā uzsvars ir likts uz zināšanu inženierijā salīdzinoši nepieredzējušiem nozaru ekspertiem un ontoloģiju vidusmēra galalietotājiem, līdz ar to uz iespējami dabisku CNL kā OWL saskarni. Vēl vairāk, uzsvars ir likts uz multilingvālu ontoloģiju izstrādi un verbalizēšanu, pieņemot, ka konceptuālajā līmenī zināšanas kā tādas ir valodneatkarīgas, tādējādi veicinot ontoloģiju atkalizmantošanu. Savā ziņā šo var uzskatīt par mašīntulkošanas problēmu, taču ne tās klasiskajā izpratnē un lietojumā.

Vispārzināmās, attīstītās CNL, kas turklāt atbalsta OWL [85], pamatā ir balstītas angļu valodā — izteikti analītiskā valodā ar noteiktu vārdu secību, vienkāršu morfoloģiju un konsekventu artikulu lietojumu, kas ievērojami atvieglo CNL teikumu tulkošanu to semantiskajā reprezentācijā (aprakstošās loģikas aksiomās). Turklāt angļu valoda bieži vien tiek izvēlēta arī kā „metavaloda" loģisko simbolu (klašu un to īpašību) apzīmēšanai ontoloģijas līmenī.

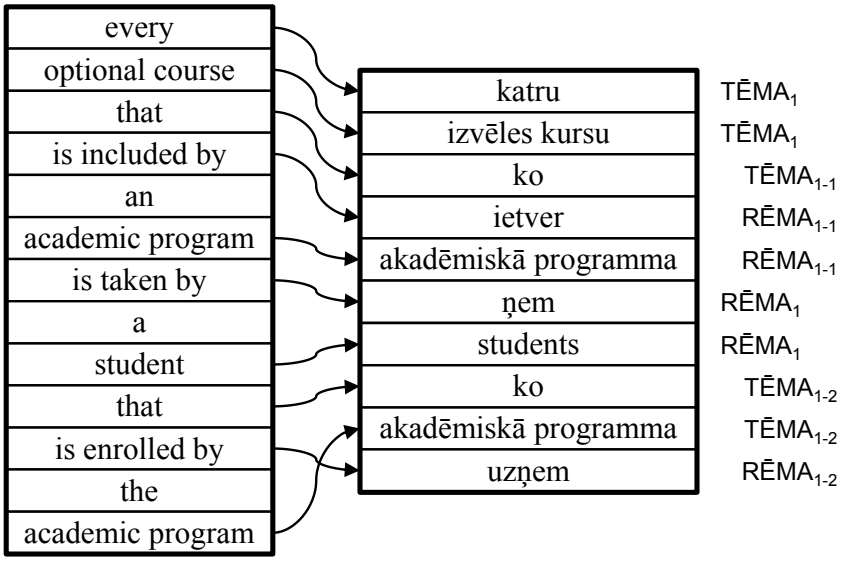
4.1 Teikuma informācijas struktūras analīze

Ir parādīts, ka *Grammatical Framework* (GF) — formālisms, kas nodrošina ērtus līdzekļus paralēlu gramatiku (un, vienlaikus, to analizatoru/sintezatoru) izstrādei — ir īpaši piemērots multilingvālu (paralēlu) CNL gramatiku ātrai un elastīgai izstrādei [101]. Tas paver iespēju atkalizmantot rīkus, kas ir izstrādāti esošajām angļu valodā balstītajām CNL (piem., teksta translēšanai uz/no OWL), arī tām valodām, kurām šādi rīki nav pieejami. Taču

izteikti sintētisko valodu (tādu kā baltu un slāvu) gadījumā[54] tulkošana uz angļu valodu (un otrādi) nav tik vienkārša pat ierobežotas valodas apakškopas gadījumā, it īpaši ja tiek apskatīti ne tikai terminoloģiski teikumi (T-kastes aksiomas), bet arī izvedumu likumi (SWRL [17]). Likumos bieži vien tiek lietotas anaforas (teikuma robežās), kas nominālu vārdkopu (NP) gadījumā angļu valodā ir atpazīstamas pēc noteiktā artikula lietojuma, bet baltu valodās un lielākajā daļā slāvu valodu šādi „marķieri" vispārīgā gadījumā netiek lietoti un nav arī iekodēti lietvārdu galotnēs. Līdz ar to viena no galvenajām problēmām semantiski precīzās tulkošanas laikā ir jaunās informācijas (potenciālie antecedenti) un zināmās informācijas (anaforas) saglabāšana.

Lai arī latviešu valodā teorētiski būtu iespējams uzspiest mākslīgu „artikulu" lietošanu, izmantojot nenoteiktos un norādāmos vietniekvārdus, tas valodu padarītu nedabisku. Tas vēl uzskatāmāk izpaužas lietuviešu valodā, kurai atšķirībā no latviešu valodas vēsturiski nav bijusi tik liela ietekme no analītiskās vācu valodas. Tomēr tas nenozīmē, ka latviešu valodā (un citās sintētiskajās valodās) nav formālu pazīmju, kas palīdz noteikt teikuma informatīvo struktūru; šādas pazīmes ir, bet vispārīgā gadījumā tās teikuma virsējā struktūrā (linearizācijā jeb vārdformu lineārajā secībā) parādās netieši.

Izrādās, ka pastāv sakarība starp zināmo un jauno informāciju un vārdu secību teikumā, ko var formulēt kā teikuma tematiskās (*theme* jeb *topic*) un rematiskās (*rheme* jeb *focus*) daļas mijiedarbību (TFA — *topic focus articulation*) [102]. Tēma ir tas, par ko tiek runāts (zināmais), bet rēma — jaunais, kas tiek pateikts par tēmu. Lai arī neierobežotā latviešu valodā (un ne tikai latviešu valodā) teikuma informatīvā struktūra ne vienmēr tiek atspoguļota ar sistemātiskām izmaiņām vārdu secībā [103], promocijas darbā ir izvirzīta hipotēze, ka ierobežotas, loģikā balstītas latviešu valodas gadījumā, izmaiņas neitrālajā vārdu secībā var tikt izmantotas kā drošas, intuitīvi apmierināmas formālās pazīmes [9], tādējādi kompensējot „trūkstošos" artikulus. Vienkāršotās TFA metodes pamatprincips ir šāds: visas NP, kas atrodas pirms izteicēja, pieder teikuma tematiskajai daļai, bet izteicējs un visas NP, kas seko pēc tā — rematiskajai daļai (sk. 15. attēlu). Izstrādātās ierobežotās latviešu valodas ilustrēšanai turpinājumā tiks pārmaiņus izmantoti (verbalizēti) fragmenti no vienkāršotām universitātes un savvaļas dabas ontoloģijām.



15. attēls. Izveduma likums, kas paralēli verbalizēts angļu un latviešu valodā. Terminu sastatījums ilustrē, kā izmaiņas vārdu secībā ietekmē teikuma informatīvo struktūru; angļu valodā mainās sintaktiskās konstrukcijas (piem., darāmās kārtas vietā tiek lietota ciešamā kārta).

[54] Bagātīga morfoloģija, sintaktiski brīva vārdu secība, nav noteikto/nenoteikto artikulu.

Lai pārbaudītu šo hipotēzi praksē, tika izstrādāta ierobežotas latviešu valodas GF gramatika[55], pārklājot būtiskākās OWL konstrukcijas (izņemot kardinalitātes, literālas vērtības u.c. sekundāras konstrukcijas), kā arī nodrošinot izvedumu likumu un integritātes vaicājumu uzdošanu[56]. Gramatikas analīze pilnībā tika balstīta teikuma informatīvās struktūras analīzē, nodrošinot precīzu tulkošanu uz ACE un daļēji[57] arī uz OWL. Šim nolūkam papildus tika izstrādātas vēl divas GF gramatikas: viena, kas pārklāj attiecīgo ACE apakškopu, bet otra — attiecīgo OWL funkcionālā stila sintakses [15] apakškopu. Izvēloties reprezentatīvu piemēru kopu (aptverot dažāda veida un sarežģītības aksiomas un likumus), tika veikta dažādu nozaru pārstāvju anketēšana (piedalījās gandrīz 80 respondenti) [10]. Aptaujas rezultāti[58] apstiprina, ka vienkāršos vārdu secības nosacījumus dzimtās valodas runātājs apmierina intuitīvi, un TFA metode līdz ar to ir piemērota zināmās un jaunās informācijas (t.i., koreferenču) automātiskai noteikšanai kontrolētas latviešu valodas teikumos. Turklāt aptauja atspoguļoja arī vairākus citus būtiskus aspektus:

- bagātīgā morfoloģija dod visai plašas iespējas, kā vienu un to pašu nozīmi var teikumā sintaktiski dažādi realizēt (piem., apzīmētāja palīgteikumu var aizstāt ar ekvivalentu apzīmētāju: *„kurss, kas iekļauts akadēmiskajā programmā” → „akadēmiskajā programmā iekļautais kurss”*);
- nenoteiktie un norādāmie vietniekvārdi („artikulu” lomā) noteiktās situācijās ir vēlami, jo tie var atvieglot teikuma lasāmību un līdz ar to interpretāciju: nenoteiktais vietniekvārds (*„kāds” ≈ „a”*) parasti tiek lietots kopā ar objektu, ja to nepaskaidro apzīmētāja palīgteikums (vai apzīmētājs), savukārt norādāmais vietniekvārds (*„šis” ≈ „the”*) atvieglo sarežģītu likumu uztveršanu (papildus vārdu secībai);
- daudzskaitļa teikumi bieži vien ir kodolīgāki un dabiskāki;
- valodas OWL izteiksmes līdzekļu ierobežojumus (pieļaujami tikai SVO (subjekts-verbs-objekts) trijnieki, nepastāv laika dimensija u.c.) CNL virsējās struktūras līmenī zināmā mērā var (šķietami) mazināt, saglabājot viennozīmīgas divvirzienu translēšanas iespējas, piemēram, objekta vietā lietojot vietas apstākli, verba (darbības) vietā lietojot lietvārdu (lomu), bet vienkāršās tagadnes vietā — salikto tagadni.

Tādējādi papildus gramatikai, kas ģenerē iespējami labāko (noklusēto) linearizāciju (ņemot vērā informācijas struktūru), ir nepieciešama (un ir izstrādāta) paralēla gramatika, kas akceptē iespējamās (pieļaujamās) sintaktiskās alternatīvas un „artikulu” patvaļīgu lietošanu.

4.2 SVO teikuma šablona paplašinājumi

Teorētiski, visām OWL ontoloģiju aksiomām (predikātiem) sintaktiski ir jāatbilst SVO šablonam (vispārināšanas aksiomas ir speciālgadījums). Tomēr praksē var būt ļoti grūti vai pat neiespējami piemeklēt atbilstošu (t.sk. intuitīvu) verbu, lai apzīmētu noteikta veida attieksmes starp klasēm, vai arī lietot sintaktisku objektu (akuzatīvā) tā, lai teikums būtu (semantiski) dabisks.

---

[55] Ierobežotās latviešu valodas GF gramatika (t.sk. tās paplašinājumi, kas aprakstīti nākamajās apakšnodaļās) ir ekvivalenta bezkonteksta gramatikai un ir vienlaikus izmantojama gan valodas analīzei, gan ģenerēšanai.

[56] Tiešsaistes demonstrācija un izstrādātās gramatikas ir pieejamas šeit: http://valoda.ailab.lv/cnl/

[57] GF nav realizēts anaforu atbalsts, kaut gan teorētiski tas būtu iespējams [104]. Anaforas (t.sk. tiešā veidā lietoti mainīgie) var parādīties ne tikai likumos, bet arī īpašību aksiomās (sk. 21. attēlu). Tādējādi GF ietvaros tulkošanu uz OWL var nodrošināt tikai klašu aksiomām (neatkarīgi no lietoto klašu izteiksmju sarežģītības). Alternatīva CNL gramatiku specificēšanas valoda, kas nodrošina koreferenču analīzi, ir *Codeco* [87], taču tā nav piemērota multilingvālu CNL gramatiku izstrādei.

[58] http://goo.gl/3CdDj (Aptaujas rezultāti ierobežotai lietuviešu valodai: http://goo.gl/2q8Gu)

Pirmajā gadījumā attieksmi starp divu klašu instancēm dabiskajā valodā visticamāk ir iespējams norādīt vismaz ar lomas(-u) palīdzību, verba vietā lietojot nominālu vārdkopu. Vadošās ierobežotās (angļu) valodas, t.sk. ACE, nodrošina atšķirīgus līdzekļus šāda veida attieksmju definēšanai (sk. nākamo sadaļu), bet ierobežotā latviešu valodā tās var tikt izteiktas unificētā formā (t.sk. tiešā un inversā lietojumā): veidojot nominālu vārdkopu, kas sastāv no termina (klases vārda) ģenitīvā un tam sekojoša lomas vārda, kura locījums ir atkarīgs no konteksta. Piemēram, *„ikviens kurss ir kādas akadēmiskās programmas daļa"* (ACE: *„every course is a part of an academic program"*) un *„ikvienas akadēmiskās programmas daļa ir kāds kurss"* (ACE: *„for every academic program its part is a course"*).

Otrajā gadījumā tipisks ontoloģisko teikumu pseido objekts ir vietas apstāklis (izteikts ar lietvārdu), piemēram, *„ikviens kurss ir iekļauts kādā akadēmiskajā programmā"* pretstatā *„ikvienu kursu iekļauj kāda akadēmiskā programma"*. Jāatzīmē, ka pašreizējā realizācijā ir ņemti vērā tikai tādi apstākļi, kurus latviešu valodā var izteikt ar lokatīvu. Perspektīvā ir jānodrošina atbalsts arī tādiem vietas apstākļiem, kas izteikti ar prepozicionālu konstrukciju (piem., *„uz galda"*), taču maz ticams, ka praksē būtu nepieciešams atbalsts citu kategoriju apstākļiem (veida, mēra, nolūka). No otras puses, apstākļa veidam no sintaktiskās analīzes viedokļa faktiski nav nekādas nozīmes, turklāt analoģiski varētu tikt pieļauta arī noteikta veida netiešo papildinātāju lietošana (piem., tādu, kas atbild uz jautājumu *ar ko?*) — paši prievārdi netiek nekā interpretēti un ontoloģijā tiek pievienoti kā predikātu vārdu sufiksi (sk. 4.4. apakšnodaļu). Tomēr jāņem vērā, ka fleksīvā valodā prievārdu sintaktiskā saistāmība var būt daudznozīmīga (piem., *„uz galda*$_{\text{GEN}}$*"* un *„uz lekciju*$_{\text{ACC}}$*"*). Šī problēma promocijas darbā sīkāk nav pētīta, taču, tā kā prievārda saistāmību teikumā nosaka verba saistāmība, potenciāls risinājums būtu šo informāciju (ierobežojumu) ietvert domēna leksikonā. Dabiskajā valodā saistāmība, protams, nav viennozīmīga un var būt atkarīga no verba nozīmes, taču fiksēta domēna (nozīmes) ietvaros visticamāk šāds risinājums derētu. Savukārt virzienā no OWL uz CNL daudznozīmību atrisinātu prievārda sintaktiski atbilstošais tulkošanas ekvivalents (*„on a table"* → `GEN`, *„to a lecture"* → `ACC`).

4.3 Sintaktiskā pārfrazēšana

Ņemot vērā aptaujas rezultātus un respondentu tieši un netieši dotos ieteikumus, sākotnējā ierobežotās latviešu valodas GF gramatika ir būtiski pilnveidota [10, 11], kā arī, balstoties uz GF paralēlo gramatiku pieeju, ir nodrošināts teikumu sintaktiskās pārfrazēšanas mehānisms, lietotāja veidoto teikumu automātiskai pārrakstīšanai (un esošo ontoloģiju verbalizācijai) iespējami dabiskā un vienlaikus precīzā formā (linearizācijā), saglabājot viennozīmīgas analīzes iespējas. Svarīgi, ka pārfrazēšanas laikā tiek saglabāta gan oriģinālā informācijas struktūra, gan GF abstraktās sintakses analīzes koks, tādējādi netiek ieviestas nekādas semantiskas izmaiņas.

```
abstract Ontology ..
fun
  RestrictedClass : C → R → C_Bar ;
  Restriction : Lnk → Pred → E_Bar → R ;
  Link : Lnk ;
  ExistentialQuantification : EC_Bar → E_Bar ;
  ExistentialClass : Ex → C_Bar → EC_Bar ;
  Some : Ex ;
  Class : C → C_Bar ;
```

16. attēls. Fragments no abstraktās gramatikas, kas neatkarīgi no valodas un domēna definē funkcijas analīzes koku konstruēšanai. Abstrakto gramatiku implementē gan ierobežotās latviešu valodas paralēlās gramatikas, gan gramatikas, kas pārklāj ACE un OWL apakškopas, norādot kā kokus linearizēt.

Kā piemēru aplūkosim četras gramatiski atšķirīgas, bet semantiski ekvivalentas NP:

(1) *„kādā akadēmiskajā programmā iekļautais kurss”*;
(2) *„kurss, kas ir iekļauts kādā akadēmiskajā programmā”*;
(3) *„kurss, kas iekļauts akadēmiskajā programmā”*;
(4) *„kurss, kas ietilpst akadēmiskajā programmā”.*

Katra no šīm NP (kopumā) atbilst izstrādātās gramatikas kategorijai `C_Bar` (sk. 16. attēlu). Analīzes rezultātā, lietojot 17. un 18. attēlā dotos gramatikas likumus (un 19. attēlā doto leksikonu), tiem visiem tiek iegūts viens vienīgs, kopīgs abstraktās sintakses koks:

```
RestrictedClass Course (Restriction Link includedIn (Existential-
Quantification (ExistentialClass Some (Class AcademicProgram))))
```

```
concrete OntologyLavVar of Ontology = open OntologyLav ..
lin
  RestrictedClass c r = {
    str = case r.type.type of {
      RelCl ⇒ c.str ++ "," ++ r.str ;    -- Apzīmētāja palīgteikums
      Attr ⇒ r.str ++ c.str              -- Apzīmētājs
    } ;
    restr = {restr = Restr ; type = r.type.type}
  } ;
  Restriction lnk pred e_bar = case e_bar.restr.restr of {
    Restr ⇒ mkRelCl lnk pred e_bar ;
    NonRestr ⇒ mkRelCl lnk pred e_bar | mkAttr lnk pred e_bar ;
  } ;
  ExistentialQuantification ec_bar = {
    str = ec_bar.str ;
    restr = ec_bar.restr ;
  } ;
  ExistentialClass ex c_bar = {
    str = case c_bar.restr.restr of {
      Restr ⇒ mkExClass (Impl | Expl) ex c_bar ;
      NonRestr ⇒ mkExClass (Expl | Impl) ex c_bar
    } ;
    restr = nbar.restr
  } ;
  Class c = {
    str = c.str ;
    restr = {restr = NonRestr ; type = NoType}
  } ;
```

17. attēls. Ļoti vienkāršots fragments no domēnneatkarīgās latviešu valodas gramatikas: bez locīšanas paradigmām un teikuma locekļu sintaktiskas saskaņošanas u.c., bet ar iespējamajām alternatīvām.

```
resource OntologyLav ..
oper
  mkRelCl : Pronoun → Verb → Noun → RelClause = λpron,verb,noun → {
    str = pron.str ! Expl ++ verb.str ++ noun.str ;
    type = {restr = noun.restr.restr ; type = RelCl}
  } ;
  mkAttr : Pronoun → Verb → Noun → RelClause = λpron,verb,noun → {
    str = pron.str ! Impl ++ noun.str ++ verb.adj ;
    type = {restr = noun.restr.restr ; type = Attr}
  } ;
  mkExClass : Usage → Determiner → Noun → Str = λuse,det,noun →
    det.str ! use ++ noun.str ;
```

18. attēls. Vienkāršots fragments (bez pazīmju saskaņošanas) no latviešu valodas resursgramatikas: kopīgās funkcijas, kas var tikt izmantotas ikvienā no paralēlajām latviešu valodas gramatikām.

Domēnspecifiski, bet valodneatkarīgi jēdzieni

`C` — klases, `Pred` — īpašības (predikāti), `NPred` — ar lietvārdu izteiktas īpašības (*predicate nominatives*)

```
abstract University = Ontology ** ..
fun
  AcademicProgram, Course, MandatoryCourse, Student : C ;
  enrolls, enrolledIn, includes, includedIn, takes : Pred ;
  partOf : NPred ;
```

Domēnspecifisks latviešu valodas leksikons

```
concrete UniversityLavVar of University = OntologyLavVar ** open ResLav ..
lin
  AcademicProgram = mkMWU (mkAdjective "akadēmisks")(mkNoun "programma") ;
  Course = mkNoun "kurss" ;
  MandatoryCourse = mkMWU (mkAdjective "obligāts")(mkNoun "kurss") ;
  Student = mkNoun "students" ;
  enrolls = (mkParticiple "uzņēmis") | (mkVerb "uzņem" Object) ;
  enrolledIn = mkParticiple "uzņemts" ;
  includes = mkVerb ("iekļauj" | "ietver") Object ;
  includedIn = (mkParticiple "iekļauts") | (mkVerb "ietilpst" Place) ;
  takes = mkVerb ("apgūst" | "ņem") Object ;
  partOf = mkNoun "daļa" ;
```

Domēnspecifisks angļu valodas leksikons

```
concrete UniversityEngVar of University = OntologyEngDef ** open ResEng ..
lin
  AcademicProgram = mkMWU (mkAdjective "academic")(mkNoun "program") ;
  Course = mkNoun "course" ;
  MandatoryCourse = mkMWU (mkAdjective "mandatory")(mkNoun "course") ;
  Student = mkNoun "student" ;
  enrolls = (mkParticiple "enrolled" Object) | (mkVerb "enroll" Object) ;
  enrolledIn = (mkParticiple "enrolled" In Place) |
               (mkVerb "enroll" In Place) ;
  includes = mkVerb "include" Object ;
  includedIn = mkParticiple "included" In Place ;
  takes = mkVerb "take" "taken" Object ;
  partOf = mkNoun "part" ;
```

19. attēls. Domēnspecifiska, multilingvāla leksikona fragments: pilnnozīmes vārdi un vārdu savienojumi ar iespējamiem variantiem (gan gramatiskiem, gan leksiskiem). Tulkošanas ekvivalentu sasaiste tiek nodrošināta ar abstraktā leksikona palīdzību. Vārdu locīšanas funkcijas (piem., `mkNoun`) ir definētas atsevišķos, no lietojuma neatkarīgos gramatikas moduļos (`ResLav` un `ResEng`) līdzīgi kā 18. attēlā.

No dotajām četrām NP noklusētajai linearizācijai atbilst otrā (saskaņā ar domēnneatkarīgo gramatiku un domēnspecifisko leksikonu), kas visprecīzāk atspoguļo nozīmi[59], nezaudējot dabiskumu. Šī NP tiks ģenerēta, gan automātiski pārfrāzējot jebkuru no pārējām lietotāja ievadītajām NP, gan verbalizējot esošu ontoloģiju. Noklusēto linearizāciju lietotājs pats daļēji var mainīt: 17. un 19. attēlā dotajā gramatikā un leksikonā ir iekļautas visas iespējamās (pieļaujamās) alternatīvas, no kurām noklusētajā gramatikā un leksikonā vienmēr tiek paturēta tikai pirmā[60]. Tādējādi lietotāja (domēna eksperta) kontrolē ir piemērotāko sinonīmu izvēle (piem., īpašībai `takes` 19. attēlā dotajā latviešu valodas leksikonā), kā arī vienkāršā/saliktā laika lietošana (piem., īpašības `enrolls` gadījumā). Jāatzīmē, ka verbu, kura lietošana rada SVO struktūrai neatbilstošu teikumu (sk. iepriekšējo apakšnodaļu), nevar pārfrazēt, lietojot citu verbu vai šī paša verba citu formu, kas piesaista tiešo objektu tā, lai netiktu mainīta nozīme (piem., `includes` un `includedIn` ir divas dažādas, inversas īpašības) vai teikums nekļūtu negramatisks (sk. nākamo apakšnodaļu).

[59] .. `Course(x)` ∧ ∃`y(AcademicProgram(y)` ∧ `includedIn(x,y))`..

[60] Noklusēto gramatiku un leksikonu līdz ar to ir iespējams atvasināt automātiski.

Visa veida pārfrazēšanu, ko nodrošina ierobežotās latviešu valodas paralēlās gramatikas (un domēnu leksikoni), var iedalīt šādās grupās:

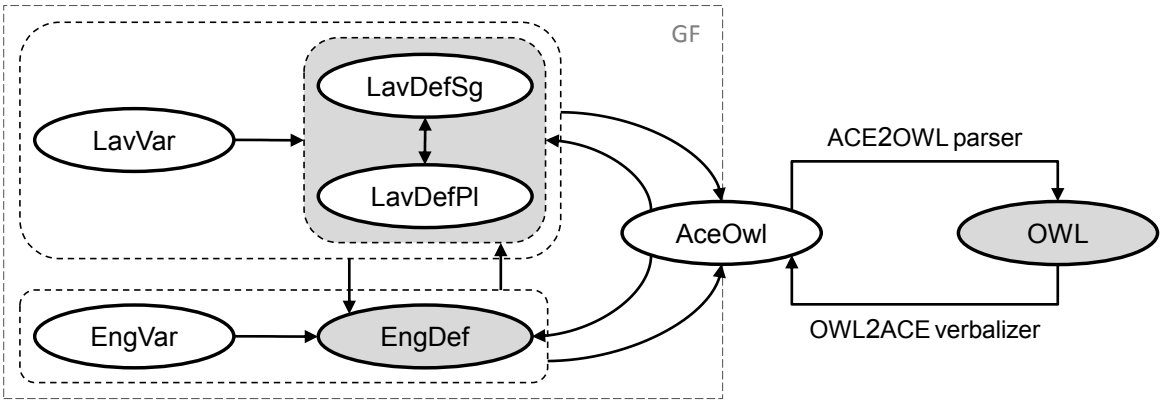
- Leksiskā pārfrazēšana, ļaujot patvaļīgi lietot sinonīmus viena un tā paša jēdziena apzīmēšanai. Tas attiecas arī uz funkcionālajiem vārdiem (piem., *„katrs"* vietā lietojot *„ikviens"*), kas paši par sevi var būt daudznozīmīgi, bet teikumā to nozīme tiek sintaktiski izšķirta (piem., *„ikviens"* var tikt interpretēts gan kā *„every"*, gan kā *„everything"*). Jāatzīmē, ka leksiskā pārfrazēšana potenciāli var tik izmantota arī 3. nodaļā aprakstītajā pusautomātiskajā nozīmju izšķiršanas procesā, daudznozīmīgos vārdus norādot kā sekundāras alternatīvas atbilstošajiem, viennozīmīgajiem vairākvārdu savienojumiem. Piemēram, vārdu *„bibliotēka"* norādot kā alternatīvu gan jēdziena `programming-library` apzīmēšanai (papildus savienojumam *„programmatūras bibliotēka"*), gan `building-library` apzīmēšanai (papildus *„bibliotēkas ēka"*). Šajā gadījumā, lietojot daudznozīmīgo vārdu *„bibliotēka"*, tiks iegūti divi analīzes koki un attiecīgi divas parafrāzes, no kurām lietotājs var pats izvēlēties īsto[61].
  - Situācijās, kad tiek dota (netieša) atsauce uz universālo klasi (`owl:Thing`), sintaktiski nav iespējams izšķirt, vai subjekts/objekts ir dzīva vai nedzīva būtne[62] un kura gramatiskā dzimte (vīriešu vai sieviešu) uz to būtu attiecināma. Šādu informāciju daļēji varētu iekļaut leksikonā (specificējot verbu saistāmību), taču (esošās) ontoloģijās šāda informācija vienalga nav pieejama. Tādēļ pārfrazējot tiek izmantotas iespējami neitrālas alternatīvas: nedzīvai būtnei un vīriešu dzimtei. Vīriešu dzimtes lietojums šādos gadījumos ir gramatiski korekts, taču pieņēmums par nedzīvu būtni (analoģiski kā valodā ACE) ir kompromiss. Tiesa, latviešu valodā nereti ir iespējams lietot norādāmo vietniekvārdu *„tas"* — ja tam seko apzīmētāja palīgteikums — padarot teikumu gan gramatiski, gan semantiski korektu (sk. 1., 5. un 6. teikumu 22. attēlā).
- Nenoteikto un norādāmo vietniekvārdu („artikulu") lietojums ir pilnībā neobligāts, taču pārfrazējot nenoteiktais vietniekvārds tiek lietots, ja attiecīgo NP nepaskaidro apzīmētāja palīgteikums. Norādāmais vietniekvārds parafrāzēs vienmēr tiek lietots.
- Atsevišķos gadījumos tiek pieļauta teikuma informācijas struktūrai neatbilstoša vārdu secība, kas pārfrazēšanas laikā tiek koriģēta (piem., *„[..], ko apgūst šis students.."* tiek pārfrazēts par *„[..], ko šis students apgūst.."*). Vēl citos specifiskos, viennozīmīgos gadījumos dabiskuma labad tiek modificēta „pareizā" informācijas struktūra (piem., *„[..] ir iekļauts kaut kur.."* tiek pārfrazēts par *„[..] ir kaut kur iekļauts.."*).
- Tiek pieļauta gan vienkāršās, gan saliktās tagadnes lietošana (lai atsauktos uz pagātnes notikumu, kam joprojām ir jūtamas sekas). Šādas alternatīvas var paredzēt pats lietotājs un pārfrazēšana ir atkarīga no lietotāja dotās alternatīvu secības domēna leksikonā.
  - Saliktās tagadnes gadījumā tiek pieļauta palīgverba *„būt"* redukcija.
- Atsevišķos gadījumos teikuma virsējā struktūrā iespējams lietot sintaktiski būtiski atšķirīgas, taču semantiski ekvivalentas konstrukcijas (piem., apzīmētāju aizstājot ar apzīmētāja palīgteikumu, kā tas tika parādīts 4.1. apakšnodaļā).
- Daudzskaitļa teikumi tiek pārfrazēti par vienskaitļa teikumiem vai otrādi (atkarībā no konfigurācijas). Turklāt T-kastes teikumos (dsk.) ir dabiski pilnībā iztikt bez tiešiem kvantoriem (piem., *„lauvas medī žirafes"*[63] → *„ikviena lauva medī kādu žirafi"*).

[61] Variantu skaitu var iepriekš mēģināt samazināt ar secinātāja palīdzību.
[62] Piemēram, *„ikviens"* (*„everyone"*) pretstatā *„jebkas"* (*„everything"*).
[63] Ja subjekts un objekts ir sieviešu dzimtē, iespējama tikai viena, neitrālā (SVO) interpretācija [105].

4.4 Divlīmeņu tulkošana uz/no OWL

Iespējamie soļi, kas tiek veikti, tulkojot latviešu valodas teikumus uz OWL (un otrādi), ir parādīti 20. attēlā. Gramatika `LavDefSg` definē noklusētos linearizācijas šablonus, lietojot vienskaitļa teikumus, `LavDefPl` ir tās analogs daudzskaitļa teikumiem[64], bet `LavVar` ir paplašināta abu noklusēto gramatiku kombinācija, kas pieļauj plašas sintaktisko konstrukciju un leksiskas variācijas (sk. iepriekšējo apakšnodaļu). `LavVar` tiek izmantota lietotāja ievadīto teikumu elastīgai, taču deterministiskai[65] parsēšanai, bet viena no noklusētajām gramatikām (atkarībā no lietotāja izvēles) — `LavVar` teikumu pārfrazēšanai un esošu ontoloģiju verbalizēšanai. Visbeidzot gramatika `AceOwl` implementē valodas ACE apakškopu (ACE-OWL), kas tiek izmantota kā starpvaloda, nodrošinot to, ka translācija starp ACE-OWL un OWL (t.sk. SWRL) ir iespējama, izmantojot esošos ACE rīkus [20, 106]. Papildus ir izstrādāts arī ierobežotas angļu valodas gramatikas prototips, kas pamatā ir balstīts uz „neierobežotas" ACE valodas konstrukcijām[66], taču ir ieviesti atsevišķi uzlabojumi:

- paplašināts saliktās tagadnes konstrukciju klāsts (piem., pieļaujot tāda veida teikumus kā *„every academic program has enrolled a student"*);
- inversu lomu verbalizēšanai virsteikumos dabiskāks ir šablons, kas tiek izmantots Sidnejas OWL sintaksē (SOS) [107] — tas tiek lietots arī `EngDef` (piem., *„for every branch its part is a leaf"* vai *„every branch has-part a leaf"* vietā lietojot *„every branch has a leaf as a part"*);
- savietojamībai ar ierobežoto latviešu valodu nepieciešams akceptēt/ģenerēt konstrukcijas, kurās tiek inversi lietotas tādas īpašības, kuru „objekts" ir vietas apstāklis (prepozicionāla konstrukcija; piem., *„every savanna is a place where a lion lives in"*).



20. attēls. Automātiskā, semantiski precīzā tulkošanas un pārfrazēšanas procesa vispārējā datu plūsma starp ierobežoto latviešu un angļu valodu, un OWL, lietojot valodu ACE-OWL kā starpvalodu. Pāreju uz/no OWL nodrošina eksistējoši rīki, bet citas pārejas nodrošina paralēlās GF gramatikas.

Lai arī angļu valodas gramatikas izstrāde un attīstīšana nav promocijas darba mērķis, paralēlā tulkošana uz angļu valodu (un otrādi) bez multilingvālā aspekta var tikt uztverta arī kā papildus pārfrazēšanas līdzeklis. Tiesa, izstrādātajā prototipā ierobežotās angļu valodas gramatika ir definēta „no nulles", taču no inženierijas viedokļa (un perspektīvā) pareizāk būtu atkalizmantot universālo GF resursgramatiku bibliotēku [108].

---

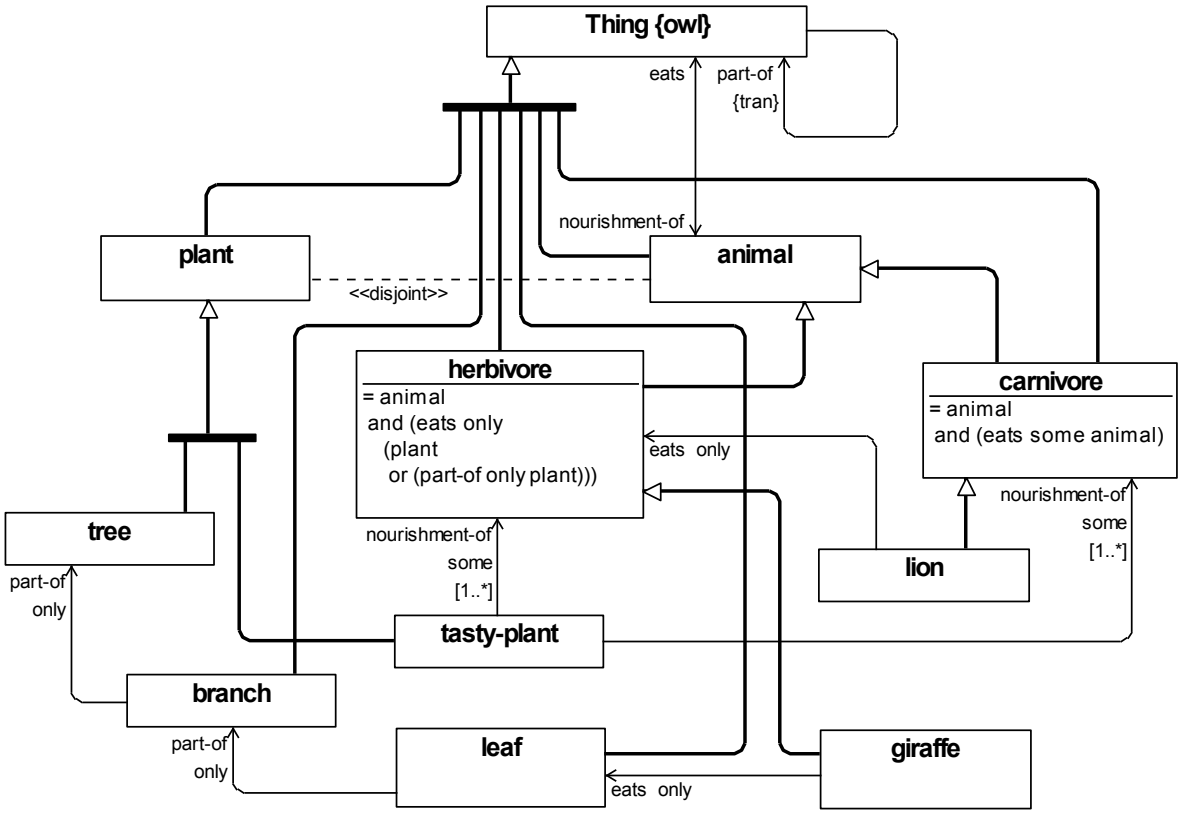
[64] Promocijas darba rakstīšanas laikā gramatika `LavDefPl` (un atbilstošās `LavVar` daļas) ir izstrādes stadijā.

[65] Dažos gadījumos var rasties sintaktiska daudznozīmība (piem., ja doti trīs vai vairāk secīgi, vienlīdzīgi apzīmētāja palīgteikumi), kuras novēršanai nepieciešams lietot vienkāršus interpretācijas likumus (analoģiski kā to dara ACE parsētājs; sk. http://goo.gl/6mmGL).

[66] ACE pilnā gramatika (pretstatā ACE-OWL) atbalsta prepozīciju, īpašības vārdu u.c. konstrukciju lietošanu.

ACE vietā par starpvalodu varētu izmantot jebkuru citu CNL, kas paredzēta OWL ontoloģiju izstrādei un verbalizācijai. Plaši zināmā un lietotā valoda ACE ir izvēlēta tās viegli pieejamo (atvērtā koda) rīku un tīmekļa servisu dēļ. Turklāt ontoloģiju verbalizāciju (virzienu no OWL uz CNL) būtu iespējams realizēt bez starpvalodas un starpniek-infrastruktūras iesaistīšanas (īpašību aksiomu verbalizācijai lietojot fiksētus šablonus), taču, realizējot CNL tulkošanu uz OWL, pretējo virzienu GF nodrošina „par brīvu".

Lai ilustrētu rezultātus un starprezultātus, kas tiek iegūti piedāvātajā divlīmeņu tulkošanas pieejā, aplūkosim ontoloģiju, kas ir vizualizēta 21. attēlā. Jāpiemin, ka šis piemērs uzskatāmi parāda arī to, ka UML un CNL ir savstarpēji papildinošas notācijas: klašu `herbivore` un `carnivore` kompleksie ierobežojumi diagrammā ir atainoti, lietojot formālo Mančestras sintaksi, taču tās vietā varētu tikt lietoti CNL teikumi.



21. attēls. Vienkāršota Āfrikas savvaļas dabas ontoloģija [109], kas attēlota UML stila grafiskā notācijā [23]. Ontoloģijas oriģinālā versija ir nebūtiski (lingvistiski motivēti [96]) modificēta, īpašību `eaten-by` pārsaucot par `nourishment-of`: lietojot lomas vārdu, nevis nekonsekventu verbālu formu.

Dotās ontoloģijas verbalizācija (neiekļaujot visas aksiomas) ir atspoguļota 22. attēlā. Tās iegūšanai paralēli ir izmantotas `LavDefSg` un `EngDef` gramatikas. Cita starpā šis piemērs uzreiz atklāj arī vairākus aspektus, kuros verbalizācija latviešu valodā būtu uzlabojama:

- T-kastes tekstos daudzskaitļa teikumi lielākoties būtu dabiskāki un intuitīvāki (piem., *„tas, ko ēd kāda žirafe, ir lapa"* pretstatā *„tas, ko ēd žirafes, ir lapas"*);
- nenoteiktais vietniekvārds *„kaut kas"* jāpārfrazē par norādāmo vietniekvārdu *„tas"* (sk. 3. teikumu), ja pirms tā dots ierobežojums *„tikai"* (`owl:allValuesFrom`); citos gadījumos tas var radīt daudznozīmību;
- noteiktās situācijās (ja seko palīgteikums) ģenitīva konstrukciju (lomu) verbalizēšanā jāizmanto analītiska forma (piem., *„kaut kā daļa"* pretstatā *„daļa no kaut kā"*; sk. 10. teikumu).

1	Tas, kas kaut ko *ēd*, ir *dzīvnieks*.	Everything that *eats* something is **an** *animal*.
2	Ikviens *plēsējs* ir *dzīvnieks*, kas *ēd* **kādu** *dzīvnieku*. Ikviens *dzīvnieks*, kas *ēd* **kādu** *dzīvnieku*, ir *plēsējs*.	Every *carnivore* is **an** *animal* that *eats* **an** *animal*. Every *animal* that *eats* **an** *animal* is a *carnivore*.
3	Ikviens *zālēdājs* ir *dzīvnieks*, kas *ēd* tikai kaut ko, kas ir *augs* vai kas ir tikai *auga daļa*.	Every *herbivore* is **an** *animal* that *eats* nothing but things that are **a** *plant* or that are **a** *part* of nothing but *plants*.
4	Ikviena *žirafe* ir *zālēdājs*.	Every *giraffe* is **a** *herbivore*.
5	Tas, ko *ēd* **kāda** *žirafe*, ir *lapa*.	Everything that is *eaten* by **a** *giraffe* is **a** *leaf*.
6	Tas, kura *daļa* ir **kāda** *lapa*, ir *zars*.	Everything that has **a** *leaf* as **a** *part* is **a** *branch*.
7	Ikviens *garšīgs augs* ir **kāda** *plēsēja barība*.	Every *tasty plant* is **a** *nourishment* of **a** *carnivore*.
8	Neviens *dzīvnieks* nav *augs*.	No *animal* is **a** *plant*.
9	Ja X-s *ēd* Y-u, tad Y-s ir X-a *barība*.	If X *eats* Y then Y is **a** *nourishment* of X.
10	Ja X-s ir kaut kā, kas ir Y-a *daļa*, *daļa*, tad X-s ir Y-a *daļa*.	If X is **a** *part* of something that is **a** *part* of Y then X is **a** *part* of Y.

22. attēls. 21. attēlā dotās ontoloģijas (daļēja) verbalizācija ierobežotā latviešu (`LavDefSg`) un angļu (`EngDef`) valodā. Slīprakstā izcelti domēnspecifiskie termini (klases un īpašības); pasvītrotas ir tās īpašības, kas izteiktas ar lietvārdu (lomu), bet semantiski tiek interpretētas kā (bināri) predikāti.

Jāpaskaidro, ka klašu ekvivalences nepieciešamie un pietiekamie nosacījumi, t.i., definīcijas valodā ACE tiek uzdotas divos atsevišķos teikumos (sk. 2. rindkopu; 3. rindkopā, taupot vietu, pretējais virziens ir izlaists). Citās CNL, piemēram, SOS definīcijas tiek verbalizētas tiešā veidā (*„..defined as..”*), taču tad lietotājam ir jāizprot šādu apgalvojumu nozīme. Savukārt, lietojot mainīgos, brīvās vārdu secības dēļ latviešu valodā tiem jāpievieno galotnes, lai varētu noteikt, kurš ir teikuma (vai teikuma daļas) subjekts, bet kurš — objekts.

1	Everything that **v:eats** something is an n:animal.	`ObjectProperty: eats Domain: animal`
2	Every n:carnivore is an n:animal that **v:eats** an n:animal. Every n:animal that **v:eats** an n:animal is a n:carnivore.	`Class: carnivore EquivalentTo: animal and (eats some animal)`
3	Every n:herbivore is an n:animal that **v:eats** nothing but things that are a n:plant or that **v:part-of** nothing but n:plant.	`Class: herbivore EquivalentTo: animal and (eats only (plant or (part-of only plant)))`
4	Every n:giraffe is a n:herbivore.	`Class: giraffe SubClassOf: herbivore`
5	Everything that is **v:eats** by a n:giraffe is a n:leaf.	`Class: inverse (eats) some giraffe SubClassOf leaf`
6	Everything that is **v:part-of** by a n:leaf is a n:branch.	`Class: inverse (part-of) some leaf SubClassOf branch`
7	Every n:tasty-plant **v:nourishment-of** a n:carnivore.	`Class: tasty-plant SubClassOf: nourishment-of some carnivore`
8	No n:animal is a n:plant.	`Class: animal DisjointWith: plant`
9	If X **v:eats** Y then Y **v:nourishment-of** X.	`ObjectProperty: eats InverseOf: nourishment-of`
10	If X **v:part-of** something that **v:part-of** Y then X **v:part-of** Y.	`ObjectProperty: part-of Characteristics: Transitive`

23. attēls. Automātiski ģenerēts ACE-OWL teksts; iegūts tulkojot 22. attēlā dotos teikumus (izmantojot paralēlo `AceOwl` gramatiku) vai verbalizējot oriģinālo OWL ontoloģiju (izmantojot ACE verbalizatoru). Skaidrības dēļ termini doti kopā ar atbilstošo vārdšķiru prefiksiem, ko ACE parsētājs akceptē. Paralēli, Mančestras notācijā, dota teikumu semantiskā interpretācija, kas iegūta ar ACE parsētāja palīdzību.

Ierobežotās latviešu un angļu valodas teikumu tulkošana (reducēšana) uz ACE-OWL ir tulkošanas procesa iekšējs solis, kura izpildes laikā visi SVO šablonam neatbilstošie teikumi tiek konvertēti par mākslīgiem SVO teikumiem (piem., *„part of something" → „part-of something"*), bet lietotie termini tiek normalizēti kā fiksētas vārdformas, kas tiešā veidā atbilst ontoloģijā lietotajiem simboliem[67] (sk. 23. attēlu). Šis starprezultāts ir tīri tehniska rakstura un no dabiskās valodas viedokļa — negramatisks[68], taču tas ir pamācošs un uzskatāmi ilustrē OWL dabu un ierobežotību. Tiesa, ne visi starpvalodas teikumi ir negramatiski (sk. 1., 2., 4. un 8. teikumu), turklāt daļa no negramatiskajiem teikumiem varētu tikt padarīti gramatiski (piem., 5. teikumā lietojot pagātnes divdabja formu), taču to nav nepieciešams darīt — galalietotājs normālā gadījumā par šo starpsoli neko pat nenojauš.

Tie paši soļi tiek veikti izveduma likumu interpretēšanai[69] un potenciāli var tikt izmantoti arī integritātes vaicājumu [110] interpretēšanai (sk. 24. attēlu). Jāatzīmē, ka ACE parsētājs nenodrošina tulkošanu uz SPARQL, taču diskursa reprezentācijas struktūru (DRS) jautājuma teikumiem tas būvē, tādējādi DRS uz/no SPARQL pāreju iespējams realizēt neatkarīgi (analoģiski kā tas ir darīts šajā pētījumā: [111]).

	SWRL	**SPARQL**
`LavDef`	Ikvienu obligāto kursu, kas ir iekļauts ***kādā*** akadēmiskajā programmā, apgūst ikviens students, ko ***šī*** akadēmiskā programma ir uzņēmusi.	Vai ir ***kāds*** students, kas apgūst kursu, kas nav iekļauts akadēmiskajā programmā, kas ***šo*** studentu ir uzņēmusi?
`EngDef`	Every mandatory course that is included in ***an*** academic program is taken by every student that is enrolled by ***the*** academic program.	Is there ***a*** student that takes ***a*** course that is not included in ***an*** academic program that has enrolled ***the*** student?
`AceOwl`	Every n:mandatory-course that v:included-in an n:academic-program is v:takes by every n:student that is v:has-enrolled by the n:academic-program.	Is there a n:student that v:takes a n:course that does not v:included-in an n:academic-program that v:has-enrolled the n:student?

Interpretācija	`mandatory-course(?x1),`
	`academic-program(?x2),`
	`student(?x3),`
	`included-in(?x1,?x2),`
	`has-enrolled(?x2,?x3)`
	`-> takes(?x3,?x1)`
		`ASK WHERE {`
		`  ?x1 rdf:type student.`
		`  ?x1 takes ?x2.`
		`  ?x2 rdf:type course.`
		`  ?x3 rdf:type academic-program.`
		`  ?x3 has-enrolled ?x1.`
		`  NOT EXISTS {?x2 included-in ?x3}}`

24. attēls. Izvedumu likumu un integritātes vaicājumu tulkošana attiecīgajā formālajā notācijā. ACE parsētājs nodrošina likumu (un aksiomu) tulkošanu OWL/FSS un OWL/XML formātā, taču piemērā uzskatāmībai ir lietota Mančestras sintakse. Tulkošanu uz SPARQL ACE parsera esošā versija neatbalsta.

Vēlreiz jāuzsver, ka CNL uzdevums ir nodrošināt tās teikumu deterministiskas interpretācijas iespējas tā, lai valodas lietotājs varētu viegli un precīzi paredzēt un uztvert nozīmi specifikācijai (ontoloģiskam tekstam), kas tiek rakstīts vai lasīts, un tā, lai cikliska tulkošana starp CNL un OWL neradītu nekādas semantiskas izmaiņas (ja vien pats lietotājs

[67] Vārdformu normalizēšanu nodrošina domēnspecifisks `AceOwl` leksikons, kuru iespējams automātiski atvasināt no `EngDef` leksikona (sk. 19. attēlu).

[68] No ACE-OWL parsētāja viedokļa šie teikumi joprojām ir gramatiski. Tas tiek panākts, parsētājam padodot automātiski ģenerētu domēna leksikonu, kurā katra termina visas formas sakrīt (sk. http://goo.gl/qQq7E).

[69] Lietotājam izvedumu likumi no aksiomām nav nekā jānošķir; lietotājs var pat nenojaust, ka daļa no teikumiem tiek interpretēti kā likumi.

pa vidu neveic kādas semantiskas izmaiņas). Līdz ar to, piemēram, tāda veida valoda kā plaši zināmā un izsmalcinātā CPL (*Computer Processable Language*) [84] mūsu gadījumā neder, jo tās interpretācija nav deterministiska (lai arī leksiskā un sintaktiskā daudznozīmība tiek risināta ar ļoti augstu precizitāti, ņemot vērā kontekstu). Turklāt CPL pieredze atklāj interesantu faktu: lai informētu lietotāju par dotā CPL teksta interpretāciju, tas tiek pārfrazēts vienkāršākā, deterministiskā CPL-*Lite* valodā; izrādās, ka lietotāji ar laiku paši pāriet uz valodas deterministiskās apakškopas lietošanu.

Šeit ir jāpiemin vēl viens aspekts: ACE verbalizators pirms teikumu ģenerēšanas veic ontoloģijas „refaktorēšanu", lielu daļu aksiomu pārrakstot vispārīgākā formā (nemainot semantiku), kā rezultātā, piemēram, predikāta subjekts, kas oriģinālajā ontoloģijā bija atomāra klase var kļūt par kompleksu klasi (klašu izteiksmi). Tas ļauj verbalizēt ontoloģiju dabiskākā angļu valodā, taču translācija atpakaļ uz OWL noved pie strukturāli modificētas ontoloģijas (salīdzinājumā ar oriģinālu). Šādas izmaiņas apgrūtina gan UML stila grafiskās notācijas, gan Mančestras sintakses paralēlu lietošanu (piem., tādā plaši lietotā redaktorā kā *Protégé*). Taču to nekādā ziņā nevar uzskatīt par CNL saskarnes trūkumu, un līdzīga situācija rodas, lietojot arī grafiskās valodas. Tādējādi ikvienai no OWL notācijām (saskarnēm) būtu jārūpējas par ontoloģiju atbilstošu refaktorēšanu (normalizēšanu).

Promocijas darbā ir piedāvāta ierobežoto valodu jomā oriģināla divlīmeņu tulkošanas pieeja, kas ļauj izstrādāt kompleksu latviešu valodas gramatiku uz salīdzinoši ierobežotās ACE-OWL bāzes. Izstrādātā ierobežotā latviešu valoda ir aprobēta uz reāliem piemēriem, definējot un verbalizējot komplicētu ontoloģiju aksiomas, izvedumu likumus un integritātes vaicājumus. Paralēli šī pieeja ir ļāvusi nodemonstrēt, ka šādi iespējams neatkarīgi un elastīgi modificēt un paplašināt arī esošas ierobežotas angļu valodas gramatiku. Protams, valoda ACE-OWL pati par sevi varētu tikt vienlīdzīgi attīstīta, bet piedāvātās pieejas priekšrocība ir tā, ka dažādus paplašinājumus un modifikācijas var veikt elastīgi un neatkarīgi, dažādu CNL konstrukcijas var tikt sapludinātas vienā gramatikā vai lietotas paralēli, kā arī starpvalodu iespējams relatīvi viegli nomainīt.

Interesanti, ka nesen veikts empīrisks pētījums OWL ontoloģiju „korpusā" parāda, ka OWL potenciāls praksē ne tuvu netiek izmantots: analizējot vairāk nekā 600 000 aksiomas no 200 ontoloģijām atklājās, ka 90% no aksiomām ir *SubClassOf*, *DisjointClasses* un *EquivalentClasses*, bet 8% ir A-kastes aksiomas [112]. Savukārt 69% lietojumu šādu T-kastes aksiomu subjekts un objekts ir atomāra klase, bet 25% lietojumu — atomārs subjekts un vienkārši ierobežots objekts: *ObjectSomeValuesFrom(ObjectProperty,Class)*. Pētījuma autori secina, ka CNL tādējādi var kalpot kā adekvāta OWL notācija. Taču promocijas darba autors šo secinājumu gribētu paplašināt: statistika ne tikai apstiprina CNL piemērotību ontoloģiju verbalizēšanai, bet, iespējams, norāda uz nepietiekamu nozaru ekspertu iesaistīšanu domēnontoloģiju izstrādē, kā arī uz formālo notāciju lietošanas grūtībām. Tas liek cerēt, ka CNL kā elastīgas, taču formālas zināšanu reprezentācijas valodas plašāka (paralēla) izmantošana, arvien vairāk iesaistot nozaru ekspertus, OWL korpusa statistikā ar laiku ieviesīs pārmaiņas.

Nobeigums

Promocijas darbā izvirzītie centrālie mērķi kopumā ir uzskatāmi par sasniegtiem un, vēlreiz apkopojot galvenos rezultātus, tie ir šādi:

- Izstrādāts oriģināls atkarību gramatikā balstīts formālās gramatikas modelis, kas ir īpaši piemērots valodu ar relatīvi brīvu vārdu secību aprakstīšanai, un nodrošina no vienas puses elastīgu, no otras — detalizētu teikuma sintaktiskās struktūras analīzi un reprezentāciju. Izstrādātais modelis ir aprobēts praksē, formalizējot nosacīti plašu latviešu valodas gramatikas apakškopu: aptverot visdažādākās sintaktiskās konstrukcijas, kas ir sastopamas vienkāršos paplašinātos teikumos.
- Ieviests mikroontoloģijas jēdziens un izveidota metodika šādu ontoloģiju sistemātiskai izstrādei un integrēšanai, parādot, ka ar to vienlaikus tiek risināta arī vārda leksisko nozīmju izšķiršanas problēma (iegūstot formālu, semantiski precīzu „skaidrojošo vārdnīcu") un ka integrētās mikroontoloģijas nodrošina vārdu nozīmju pusautomātiskas noteikšanas iespējas faktuālos, loģikā balstītas valodas tekstos.
- Izstrādāta loģikā balstīta, deterministiska latviešu valoda un tās analīzes un sintēzes (pārfrazēšanas) līdzekļi, kas nodrošina šajā valodā dabiski izteiktu komplicētu apgalvojumu un izvedumu likumu semantiski precīzu interpretāciju ontoloģiju valodā OWL un otrādi — ontoloģijas aksiomu un likumu verbalizēšanu precīzā, intuitīvā latviešu valodas apakškopā. Ierobežotās valodas teikumu tulkošanai uz/no OWL ir piedāvāts un eksperimentāli realizēts CNL jomā oriģināls divlīmeņu tulkošanas modelis, kas ļauj izmantot (pēc izvēles) kādu no esošajām CNL kā starpvalodu (t.sk. atkalizmantojot starpvalodai pieejamo infrastruktūru) un vienlaikus dod elastīgas iespējas iespējami dabiskas CNL saskarnes nodrošināšanai galalietotājam. Šī pētījumu posma rezultātus autors uzskata par nozīmīgāko promocijas darba sasniegumu.

Taču vienlaikus ir apzināti vairāki aktuāli attīstības virzieni — gan tuvākajai nākotnei, gan ilgtermiņā. Tuvākās nākotnes uzdevumi saistīti ar ierobežotās latviešu valodas attīstīšanu ontoloģiju izstrādei un verbalizēšanai, ieviešot, piemēram, daudzskaitļa konstrukcijas un šablonus kardinalitāšu uzdošanai. Tāpat interesanti būtu izpētīt deterministiskās TFA metodes pielāgojamības iespējas A-kastes teikumu analīzei. Lai turpmāko paplašinājumu realizācija un ierobežotas latviešu valodas gramatiku izstrāde citiem lietojumiem būtu iespējama vēl elastīgāka un ātrāka, universālo GF resursgramatiku bibliotēku nepieciešams papildināt ar latviešu valodas moduļiem. Savukārt ilgtermiņa uzdevumi ir saistīti ar sintaktiski anotēta latviešu valodas korpusa un korpusā balstīta statistiska parsētāja izstrādi.

Jāatzīmē, ka pētījuma rezultāti ir piemērojami arī citām sintētiskām, fleksīvām valodām, piemēram, slāvu (īpaši attiecībā uz hibrīdo sintaktiskās analīzes modeli un teikuma informācijas struktūrā balstīto zināmās un jaunās informācijas analīzi loģikā balstītā valodā). Savukārt daļa no piedāvātajām metodēm ir valodneatkarīgas, galvenokārt to attiecinot uz semantiskās analīzes līmeni un mikroontoloģiju pieeju leksiskās semantikas formalizēšanai, taču tikpat labi arī hibrīdais gramatikas modelis kā tāds ir valodneatkarīgs, un teorētiski varētu būt saistošs arī analītiskām valodām.

Lai arī neierobežotas valodas analīze visticamāk nav iespējama bez korpusā balstītu (statistisku) metožu izmantošanas, svarīgi, lai statistisko sistēmu apmācībai pieejamie modeļi iespējami precīzi un dziļi atspoguļotu valodas struktūru. Promocijas darbā izvēlētā pieeja, ievērojami ierobežojot aplūkojamo valodas apakškopu, taču pētot to padziļināti, ir devusi zināmu artavu latviešu valodas struktūras izpētē un formalizēšanā.

Atsauces

Autora publikācijas recenzētos žurnālos un starptautisku konferenču materiālos

1. Bārzdiņš G., Grūzītis N., Kudiņš R., Nešpore G., Spektors A. ***Latviešu valoda semantiskajā tīmeklī***. LZA Vēstis, A daļa, 60. sēj., 6. num., 2006, 26.–42. lpp.
2. Grūzītis N., Nešpore G., Saulīte B. ***Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas***. 11. starptautiskās zinātniskās konferences "Vārds un tā pētīšanas aspekti" rakstu krājums, 2007, 147.–159. lpp.
3. Bārzdiņš G., Grūzītis N., Kudiņš R. ***Re-engineering OntoSem Ontology Towards OWL DL Compliance***. Proceedings of the 7th Joint Conference on Knowledge-Based Software Engineering, Frontiers in Artificial Intelligence and Applications, Vol. 140, IOS Press, 2006, pp. 157–166
4. Bārzdiņš G., Grūzītis N., Nešpore G., Saulīte B. ***Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order***. Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA), 2007, pp. 13–20
5. Nešpore G., Saulīte B., Bārzdiņš G., Grūzītis N. ***Comparison of the SemTi-Kamols and Tesnière's Dependency Grammars***. Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, 2010, pp. 233–240
6. Bārzdiņš G., Grūzītis N., Nešpore G., Saulīte B., Auziņa I., Levāne-Petrova K. ***Multi-dimensional Ontologies: Integration of Frame Semantics and Ontological Semantics***. Proceedings of the 13th EURALEX International Congress, 2008, pp. 277–283
7. Bārzdiņš G., Grūzītis N., Nešpore G., Saulīte B., Auziņa I., Levāne-Petrova K. ***Ontological Word Sense Disambiguation for Discourse Representation***. Proceedings of the 3rd Baltic Conference on Human Language Technologies, 2007, pp. 33–40
8. Grūzītis N., Bārzdiņš G. ***Polysemy in Controlled Natural Language Texts***. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNCS/LNAI, Vol. 5972, Springer, 2010, pp. 102–120
9. Grūzītis N. ***Word Order Based Analysis of Given and New Information in Controlled Synthetic Languages***. Proceedings of the 1st Workshop on the Multilingual Semantic Web (at WWW 2010), CEUR, Vol. 571, 2010, pp. 29–34
10. Grūzītis N., Nešpore G., Saulīte B. ***Verbalizing Ontologies in Controlled Baltic Languages***. Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, 2010, pp. 187–194
11. Grūzītis N., Bārzdiņš G. ***Towards a More Natural Multilingual Controlled Language Interface to OWL***. Proceedings of the 9th International Conference on Computational Semantics (IWCS), 2011 (pieņemta publicēšanai)

Citas autora publikācijas starptautisku konferenču materiālos (tieši nesaistītas ar promocijas darba tēmu)

12. Milčonoka E., Grūzītis N., Spektors A. *Natural Language Processing at the Institute of Mathematics and Computer Science: Ten Years Later*. Proceedings of the 1st Baltic Conference on Human Language Technologies, 2004, pp. 6–11

13. Viļķelis M., Grundspeņķis J., Grūzītis N. *Natural Language Based Concept Map Building*. Proceedings of the 6th International Conference on e-Learning and the Knowledge Society, 2010, pp. 179–184
14. Skadiņa I., Auziņa I., Grūzītis N., Levāne-Petrova K., Nešpore G., Skadiņš R., Vasiļjevs A. *Language Resources and Technology for the Humanities in Latvia*. Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, 2010, pp. 15–22

Izmantotā literatūra

15. *OWL 2 Web Ontology Language*. W3C Recommendation, 2009 (http://www.w3.org/TR/owl2-primer/)
16. Tesnière L. *Éléments de syntaxe structurale*. Klincksieck, Paris, 1959 (Tulkojums krievu valodā: Теньер Л. *Основы структурного синтаксиса*. Ред. В.Г. Гак. Москва, Прогресс, 1988)
17. *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. W3C Member Submission, 2004 (http://www.w3.org/Submission/SWRL/)
18. *SPARQL Query Language for RDF*. W3C Recommendation, 2008 (http://www.w3.org/TR/rdf-sparql-query/)
19. Ranta A. *Grammatical Framework: A Type-Theoretical Grammar Formalism*. Journal of Functional Programming, Vol. 14, No. 2, 2004, pp. 145–189
20. Fuchs N.E., Kaljurand K., Kuhn T. *Attempto Controlled English for Knowledge Representation*. Proceedings of the 4th International Reasoning Web Summer School, LNCS, Vol. 5224, Springer, 2008, pp. 104–124
21. Džeriņš J., Džonsons K. *Harvesting National Language Text Corpora from the Web*. Proceedings of the 3rd Baltic Conference on Human Language Technologies, 2007, pp. 87–94
22. Skadiņa I., Brālītis E. *English-Latvian SMT: knowledge or data?* Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA), NEALT, Vol. 4, 2009, pp. 242–245
23. Bārzdiņš J., Bārzdiņš G., Čerāns K., Liepiņš R., Sproģis A. *OWLGrEd: a UML style graphical notation and editor for OWL 2*. Proceedings of the 7th International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 614, 2010
24. Atkins B.T.S., Rundell M. *The Oxford Guide to Practical Lexicography*. Oxford University Press, 2008
25. Zipf G.K. *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley, 1949
26. Kehoe A., Gee M. *New corpora from the web: making web text more 'text-like'*. Pahta P., Taavitsainen I., Nevalainen T., Tyrkkö J. (Eds.) Towards Multimedia in Corpus Studies, Studies in Variation, Contacts and Change in English, Vol. 2, 2007 (http://www.helsinki.fi/varieng/journal/volumes/02/kehoe_gee/)
27. Brants T. *Inter-Annotator Agreement for a German Newspaper Corpus*. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC), 2000, pp. 107–112
28. Chklovski T., Mihalcea R. *Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation*. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2003

29. Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J.D., Mercer R.L., Roossin P.S. *A statistical approach to machine translation*. Computational Linguistics, Vol. 16, No. 2, 1990, pp. 79–85
30. Callison-Burch C., Koehn P., Monz C., Schroeder J. *Findings of the 2009 Workshop on Statistical Machine Translation*. Proceedings of the 4th EACL Workshop on Statistical Machine Translation, 2009, pp. 1–28
31. Kilgarriff A. *Ontologies and terminology and how they relate to lexicography*. Course notes of the 7th Workshop in Lexicography and Lexical Computing (Lexicom), 2007
32. Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., Marsi E. *MaltParser: A language-independent system for data-driven dependency parsing*. Natural Language Engineering, Vol. 13, No. 2, 2007, pp. 95–135
33. Eisele A., Federmann C., Uszkoreit H., Saint-Amand H., Kay M., Jellinghaus M., Hunsicker S., Herrmann T., Chen Yu. *Hybrid Machine Translation Architectures within and beyond the EuroMatrix project*. Proceedings of the 12th Annual Conference of the European Association for Machine Translation, 2008, pp. 27–34
34. Burchardt A., Erk K., Frank A., Kowalski A., Padó S., Pinkal M. *The SALSA Corpus: a German Corpus Resource for Lexical Semantics*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006
35. Erk K., Padó S. *Shalmaneser — a flexible toolbox for semantic role assignment*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006
36. Fellbaum C. (Ed.) *WordNet: an Electronic Lexical Database*. MIT Press, 1998
37. Budanitsky A., Hirst G. *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*. Computational Linguistics, Vol. 32, 2006, pp. 13–47
38. Snyder B., Palmer M. *The English all-words task*. Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SensEval), 2004, pp. 41–43
39. Kilgarriff A. *How Dominant is the Commonest Sense of a Word?* Proceedings of the 7th International Conference on Text, Speech and Dialogue (TSD), LNCS, Vol. 3206, Springer, 2004, pp. 103–112
40. Chodorow M.S., Byrd R.J., Heidorn G.E. *Extracting Semantic Hierarchies from a Large On-Line Dictionary*. Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics, 1985, pp. 299–304
41. Ide N., Véronis J. *Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation*. Proceedings of the 3rd International EAMT Workshop on Machine Translation and the Lexicon, 1993, pp. 19–34
42. Hearst M.A. *Automated Discovery of WordNet Relations*. Fellbaum C. (Ed.), WordNet: an Electronic Lexical Database. The MIT Press, 1998, pp. 131–152
43. Montemagni S. *Architecture and Functioning of a System for the Acquisition of Taxonomical Information from Dictionary Definitions*. Proceedings of the 4th International Conference on Computational Lexicography, 1996, pp. 173–182
44. Dyvik H. *Translations as a semantic knowledge source*. Proceedings of the 2nd Baltic Conference on Human Language Technologies, 2005, pp. 27–38
45. Piasecki M., Szpakowicz S., Broda B. *A Wordnet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2009
46. Poesio M. *Domain modelling and NLP: Formal ontologies? Lexica? Or a bit of both?* Applied Ontology, Vol. 1, No. 1, 2005, pp. 27–33

47. Nirenburg S., Raskin V. *Ontological Semantics*. Cambridge: The MIT Press, 2004
48. Niles I., Pease A. *Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology*. Proceedings of the IEEE International Conference on Information and Knowledge Engineering, 2003, pp. 412-416
49. Nirenburg S., McShane M., Zabludowski M., Beale S., Pfeifer C. *Ontological Semantic text processing in the biomedical domain*. Working Paper #03-05, Institute for Language and Information Technologies, UMBC, 2005
50. McShane M., Beale S., Nirenburg S. *Some Meaning Procedures of Ontological Semantics*. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), 2004
51. Kilgarriff A. *I don't believe in word senses*. Computers and the Humanities, Vol. 31, No. 2, 1997, pp. 91–113
52. Navigli R., Litkowski K.C., Hargraves O. *SemEval-2007 Task 07: Coarse-Grained English All-Words Task*. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), 2007, pp. 30–35
53. Chomsky N. *Three models for the description of language*. IRE Transactions on Information Theory, Vol. 2, No. 3, 1956, pp. 113–124
54. Gaifman H. *Dependency Systems and Phrase-Structure Systems*. Information and Control, Vol. 8, No. 3, 1965, pp. 304–337
55. Nivre J. *Constraints on Non-Projective Dependency Parsing*. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006, pp. 73–80
56. Earley J. *An Efficient Context-Free Parsing Algorithm*. Communications of the ACM, Vol. 13, 1970, pp. 94–102
57. Covington M.A. *A Fundamental Algorithm for Dependency Parsing*. Proceedings of the 39th Annual ACM Southeast Conference, 2001, pp. 95–102
58. Nivre J. *Dependency Grammar and Dependency Parsing*. MSI report 05133. Växjö University: School of Mathematics and Systems Engineering, 2005
59. Johansson R., Nugues P. *Extended Constituent-to-Dependency Conversion for English*. Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA), 2007, pp. 105–112
60. Stevenson M., Greenwood M. *Comparing Information Extraction Pattern Models*. Proceedings of the Workshop on Information Extraction Beyond the Document (at COLING/ACL), 2006, pp. 12–19
61. Johansson R., Nugues P. *LTH: Semantic Structure Extraction using Nonprojective Dependency Trees*. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval), 2007, pp. 227–230
62. Johansson R., Nugues P. *Comparing Dependency and Constituent Syntax for Frame-semantic Analysis*. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), 2008
63. Eisner J.M. *Bilexical grammars and their cubic-time parsing algorithms*. Bunt H., Nijholt A. (Eds.), Advances in Probabilistic and Other Parsing Technologies, Kluwer, 2000, pp. 29–62
64. Pollard C., Sag I.A. *Head-driven phrase structure grammar*. University of Chicago Press, 1994

65. Brants S., Hansen S. *Developments in the TIGER Annotation Scheme and their Realization in the Corpus*. Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC), 2002, pp. 1643–1649
66. Järvinen T., Tapanainen P. *Towards an implementable dependency grammar*. Proceedings of the Workshop on Processing of Dependency-Based Grammars, 1998, pp. 1–10
67. Podiņš K. *Ātrs atkarību gramatiku sintaktiskais analizators*. Maģistra darbs. Latvijas Universitātes Datorikas fakultāte, 2008
68. Lokmane I. *Sekundāri predikatīvs komponents kā teikuma loceklis*. Baltu filoloģija, 11. sēj., 1. num., Rīga: LU, 2002, 47.–64. lpp.
69. Sangati F., Mazza C. *An English Dependency Treebank à la Tesnière*. Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories, 2009, pp. 173–184
70. Marcus M.P., Santorini B., Marcinkiewicz M.A. *Building a large annotated corpus of English: The Penn treebank*. Computational Linguistics, Vol. 19, No. 2, 1993, pp. 313–330
71. Sgall P., Hajičová E., Panevová J. *The Meaning of the Sentence in Its Pragmatic Aspects*. Reidel, 1986
72. Pretkalniņa L., *Integrēta sistēma sintaktiski anotēta latviešu valodas tekstu korpusa izveidei*. Maģistra darbs. Latvijas Universitātes Datorikas fakultāte, 2011 (melnraksts)
73. Petr P., Stepanek J. *Recent Advances in a Feature-Rich Framework for Treebank Annotation*. Proceedings of the 22nd International Conference on Computational Linguistics (COLING), 2008, pp. 673–680
74. Fillmore C.J., Johnson C.R., Petruck M.R.L. *Background to FrameNet*. International Journal of Lexicography, Vol. 16, 2003, pp. 235–250
75. Baker C., Ellsworth M., Erk K. *SemEval'07 Task 19: Frame Semantic Structure Extraction*. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), 2007, pp. 99–104
76. Tonelli S., Pianta E. *A novel approach to mapping FrameNet lexical units to WordNet synsets*. Proceedings of the 8th International Conference on Computational Semantics (IWCS), 2009, pp. 342–345
77. Scheffczyk J., Baker C.F., Narayanan S. *Ontology-based reasoning over lexical resources by means of ontologies*. Proceedings of the Workshop on Ontologies and Lexical Resources (OntoLex), 2006, pp. 1–8
78. Scheffczyk J., Pease A., Ellsworth M. *Linking FrameNet to the Suggested Upper Merged Ontology*. Proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS), Frontiers in Artificial Intelligence and Applications, Vol. 150, IOS Press, 2006, pp. 289–300
79. Halpin H., Hayes P.J. *When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web*. Proceedings of the WWW2010 Workshop on Linked Data on the Web (LDOW), CEUR, Vol. 628, 2010
80. Hanks P., Pustejovsky J. *A Pattern Dictionary for Natural Language Processing*. Revue Francaise de linguistique appliquée, Vol. 10, No. 2, 2005
81. Pustejovsky J., Havasi C., Littman J., Rumshisky A., Verhagen M. *Towards a Generative Lexical Resource: The Brandeis Semantic Ontology*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006
82. Meņšikovs K. *Teksta nozīmes vizualizēšana*. Maģistra darbs. Latvijas Universitātes Datorikas fakultāte, 2009

83. Wyner A., Angelov K., Barzdins G., Damljanovic D., Davis B., Fuchs N., Hoefler S., Jones K., Kaljurand K., Kuhn T., et al. *On Controlled Natural Languages: Properties and Prospects*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNAI, Vol. 5972, Springer, 2010, pp. 281–289
84. Clark P., Murray W.R., Harrison P., Thompson J. *Naturalness vs. Predictability: A Key Debate in Controlled Languages*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNAI, Vol. 5972, Springer, 2010, pp. 65–81
85. Schwitter R., Kaljurand K., Cregan A., Dolbear C., Hart G. *A Comparison of three Controlled Natural Languages for OWL 1.1*. Proceedings of the 4th International Workshop on OWL Experiences and Directions (OWLED), CEUR, Vol. 496, 2008
86. Blackburn P., Bos J. *Working with Discourse Representation Structures*. Representation and Inference for Natural Language: A First Course in Computational Linguistics, Vol. 2, 1999
87. Kuhn T. *Codeco: A Grammar Notation for Controlled Natural Language in Predictive Editors*. Pre-Proceedings of the 2nd Workshop on Controlled Natural Languages (CNL 2010), CEUR, Vol. 622, 2010
88. Leary D.E. (Ed.) *Metaphors in the history of psychology*. Cambridge: Cambridge University Press, 1994
89. Ravin Y., Leacock C. *Polysemy*. Oxford: Oxford University Press, 2000
90. Magnini B., Strapparava C., Pezzulo G., Gliozzo A. *The role of domain information in word sense disambiguation*. Natural Language Engineering, Vol. 8, No. 4, 2002, pp. 359–373
91. Rinaldi F., Dowdall J., Hess M., Molla D., Schwitter R., Kaljurand K. *Knowledge-Based Question Answering*. Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES), LNCS/LNAI, Vol. 2773, Springer, 2003, pp. 785–792
92. Wilks Y., Barnden J., Wang J. *Your metaphor or mine: belief ascription and metaphor interpretation*. Proceedings of the 12th International Joint Conference on Artificial Intelligence, 1991, pp. 945–950
93. Lenat D. *Cyc: A Large-Scale Investment in Knowledge Infrastructure*. Communications of the ACM, Vol. 38, No. 11, 1995, pp. 33–38
94. Euzenat J., Shvaiko P. *Ontology Matching*. New York: Springer, 2007
95. Banek M., Vrdoljak B., Tjoa A.M. *Word Sense Disambiguation as the Primary Step of Ontology Integration*. Proceedings of the 19th International Conference on Database and Expert Systems Applications, LNCS, Vol. 5181, Springer, 2008, pp. 65–72
96. Schwitter R. *Creating and Querying Linguistically Motivated Ontologies*. Proceedings of the Knowledge Representation Ontology Workshop, Conference in Research and Practice in Information Technology, Vol. 90, 2008, pp. 71–80
97. Kuhn T., Schwitter R. *Writing Support for Controlled Natural Languages*. Proceedings of the Australasian Language Technology Association Workshop, 2008, pp. 46–54
98. Horridge M., Drummond N., Goodwin J., Rector A., Stevens R., Wang H. *The Manchester OWL syntax*. Proceedings of the 2nd International Workshop on OWL: Experiences and Directions (OWLED), 2006
99. Dimitrova V., Denaux R., Hart G., Dolbear C., Holt I., Cohn A.G. *Involving domain experts in authoring OWL ontologies*. Proceedings of the 7th International Conference on the Semantic Web (ISWC), LNCS, Vol. 5318, Springer, 2008, pp. 1–16

100. Bārzdiņš G., Liepiņš E., Veilande M., Zviedris M. *Semantic Latvia Approach in the Medical Domain*. Proceedings of the 8th International Baltic Conference on Databases and Information Systems, 2008, pp. 89–102
101. Angelov K., Ranta A. *Implementing controlled languages in GF*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNAI, Vol. 5972, Springer, 2010, pp. 82–101
102. Hajičová E. *Issues of Sentence Structure and Discourse Patterns*. Prague: Charles University, 1993
103. Saulīte B. *Linguistic Markers of Information Structure in Latvian*. Proceedings of the 18th International Congress of Linguists, 2008, pp. 3067–3076
104. Ranta A. *Type Theoretical Grammar*. Oxford: Oxford University Press, 1994
105. Lokmane I. *Vārdu secības funkcijas latviešu valodā*. Latvistika un somugristika Latvijas Universitātē, 2010, 59.–68. lpp.
106. Kaljurand K., Fuchs N.E. *Verbalizing OWL in Attempto Controlled English*. Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 258, 2007
107. Cregan A., Schwitter R., Meyer T. *Sydney OWL Syntax — towards a Controlled Natural Language Syntax for OWL 1.1*. Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 258, 2007
108. Ranta A. *Grammars as Software Libraries*. Bertot Y., Huet G., Lévy J-J., Plotkin G. (Eds.), From Semantics to Computer Science, Cambridge University Press, 2009, pp. 281–308
109. Antoniou G., van Harmelen F. *A Semantic Web Primer*. Cambridge: MIT Press, 2003
110. Sirin E., Tao J. *Towards Integrity Constraints in OWL*. Proceedings of the 6th International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 529, 2009
111. Bernstein A., Kaufmann E., Göhring A., Kiefer C. *Querying ontologies: A controlled English interface for end-users*. Proceedings of the 4th International Conference on the Semantic Web (ISWC), 2005, pp. 112–126
112. Power R., Third A. *Expressing OWL axioms by English sentences: dubious in theory, feasible in practice*. Posters of the 23rd International Conference on Computational Linguistics (COLING), 2010, pp. 1006–1013

Pielikums

Autora referāti par darba rezultātiem starptautiskās zinātniskās konferencēs un semināros

1. **Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas** — *11. starptautiskā zinātniskā konference "Vārds un tā pētīšanas aspekti"*, Liepājā, 2006. gada decembrī
2. **Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order** — *16th Nordic Conference of Computational Linguistics* (NODALIDA), Tartu (Igaunija), 2007. gada maijā
3. **Ontological Word Sense Disambiguation for Discourse Representation** — *3rd Baltic Conference on Human Language Technologies*, Kauņā (Lietuva), 2007. gada oktobrī
4. **Polysemy in Controlled Natural Language Texts** — *1st Workshop on Controlled Natural Language*, Maretimo salā (Itālija), 2009. gada jūnijā
5. **Word Order Based Analysis of Given and New Information in Controlled Synthetic Languages** — *1st Workshop on the Multilingual Semantic Web* (konferences WWW 2010 ietvaros), Roli (Ziemeļkarolīna, ASV), 2010. gada aprīlī
6. **Verbalizing Ontologies in Controlled Baltic Languages** — *4th International Conference on Human Language Technologies — the Baltic Perspective*, Rīgā, 2010. gada oktobrī

Promocijas darbā iekļautās publikācijas un promocijas darba autora personiskais ieguldījums

Autori	Publikācija	Ieguldījums (%)	Piezīmes
G. Bārzdiņš N. Grūzītis R. Kudiņš G. Nešpore A. Spektors	*Latviešu valoda semantiskajā tīmeklī*	70	• Piedalīšanās koncepcijas izstrādē. • Pamatprincipu izstrāde esošu ontoloģiju izmantošanai leksiskās daudznozīmības risināšanā latviešu valodā. • Pamatprincipu izstrāde un eksperimentāla pārbaude latviešu valodas leksikona pusautomātiskai sasaistei ar ontoloģijas jēdzieniem, saglabājot nozīmju nianses.
N. Grūzītis G. Nešpore B. Saulīte	*Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas*	90	• Idejas izstrāde. • Semantisko attieksmju leksiski sintaktisko šablonu definēšana. • Eksperimentāla leksiskās taksonomijas automātiskas izguves rīka izveide.
G. Bārzdiņš N. Grūzītis R. Kudiņš	*Re-engineering OntoSem Ontology Towards OWL DL Compliance*	40	• *OntoSem* pieejas analīze. • Piedalīšanās ontoloģijas konvertēšanas un „atkļūdošanas" principu izstrādē. • Pamatprincipu izstrāde ontoloģijas testpiemēru ģenerēšanai no skaidrojošo vārdnīcu nozīmju definīcijām.

G. Bārzdiņš N. Grūzītis G. Nešpore B. Saulīte	*Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order*	70	• Piedalīšanās idejas izstrādē. • Sākotnējā *SemTi-Kamols* atkarību gramatikas modeļa attīstīšana un piemērošana dažādu sintaktisko konstrukciju aprakstīšanai. • Latviešu valodas gramatikas apakškopas formalizēšana, aptverot vienkāršus paplašinātus teikumus (t.sk. komplicētas analītiskās konstrukcijas).
G. Nešpore B. Saulīte G. Bārzdiņš N. Grūzītis	*Comparison of the SemTi-Kamols and Tesnière's Dependency Grammars*	30	• *SemTi-Kamols* atkarību gramatikas modeļa salīdzinājums ar citām pieejām; tā pamatīpašību formulēšana oriģinālās Tenjēra gramatikas kontekstā.
G. Bārzdiņš N. Grūzītis G. Nešpore B. Saulīte I. Auziņa K. Levāne-Petrova	*Multidimensional Ontologies: Integration of Frame Semantics and Ontological Semantics*	50	• Piedalīšanās ideju izstrādē, daudzdimensiju ontoloģijas jēdziena definēšanā un šīs pieejas attīstīšanā. • *FrameNet* analīze un tā līdzekļu attēlošana daudzdimensiju ontoloģijas modelī; *FrameNet* konvertēšanas iespēju izpēte.
G. Bārzdiņš N. Grūzītis G. Nešpore B. Saulīte I. Auziņa K. Levāne-Petrova	*Ontological Word Sense Disambiguation for Discourse Representation*	60	• Piedalīšanās ideju izstrādē. • Līdzdalība *Protégé* spraudņa izstrādē OWL ontoloģiju konvertēšanai un paplašināšanai uz FOL, minimālā modeļa būvēšanai (izmantojot *Mace4*) un iegūtā modeļa vizualizācijai.
N. Grūzītis G. Bārzdiņš	*Polysemy in Controlled Natural Language Texts*	50	• Piedalīšanās ideju izstrādē. • Mikroontoloģiju pieejas formulēšana un metodikas izveide šādu ontoloģiju sistemātiskai izstrādei un integrēšanai, nodrošinot daudznozīmīgo klašu automatizētu izšķiršanu.
N. Grūzītis	*Word Order Based Analysis of Given and New Information in Controlled Synthetic Languages*	100	• Idejas izstrāde. • Ierobežotās, aprakstošajā loģikā balstītās latviešu valodas pamatversijas projektēšana un realizācija *Grammatical Framework* (GF) vidē, parādot, ka teikuma informatīvās struktūras analīze ir pietiekams un precīzs līdzeklis viennozīmīgai kvantoru un koreferenču noteikšanai ierobežotas, izteikti sintētiskas un fleksīvas valodas formā dotās OWL terminoloģiskajās aksiomās, SWRL izvedumu likumos un SPARQL integritātes vaicājumos.
N. Grūzītis G. Nešpore B. Saulīte	*Verbalizing Ontologies in Controlled Baltic Languages*	90	• Lietotāju aptaujas izstrāde, rezultātu apkopošana un analīze. • Būtiski paplašinātas ierobežotās latviešu valodas projektēšana,

			- nodrošinot plašas alternatīvas vienu un to pašu abstraktās sintakses konstrukciju realizācijai (linearizācijai), - pieļaujot izteikt predikātus ar lietvārdu (lomu) palīdzību, bet objektus — kā vietas apstākļus (izteiktus ar lietvārdiem). • Paralēlu gramatiku izstrāde ierobežotās latviešu valodas elastīgai, taču viennozīmīgai analīzei un iespējami dabiskai, taču precīzai ģenerēšanai (t.sk. pārfrazēšanai); šo gramatiku efektīva realizācija GF vidē.
N. Grūzītis G. Bārzdiņš	*Towards a More Natural Multilingual Controlled Language Interface to OWL*	90	• Idejas izstrāde. • Prototipa realizācija, kas, izmantojot piedāvāto, oriģinālo divlīmeņu tulkošanas pieeju, demonstrē ierobežotās latviešu valodas teikumu translēšanu uz OWL (un otrādi), lietojot valodas *Attempto Controlled English* (ACE) OWL apakškopu kā starpvalodu (un atkalizmantojot tās brīvi pieejamo infrastruktūru). • Eksperimentālas, paralēlas gramatikas izstrāde, parādot, ka šī pieeja dod iespēju uz ACE-OWL bāzes elastīgi un neatkarīgi izveidot arī dabiskāku, viegli pielāgojamu un paplašināmu ierobežotu angļu valodu.