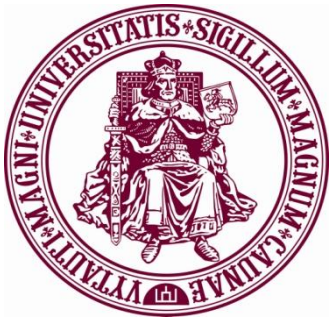


VYTAUTO DIDŽIOJO UNIVERSITETAS
HUMANITARINIŲ MOKSLŲ FAKULTETAS
LIETUVIŲ KALBOS KATEDRA

Vilmantas Ramonas

NATŪRALIOS KALBOS APDOROJIMO TERMINŲ ONTOLOGIJA: KŪRIMO PROBLEMOS IR JŲ SPRENDIMO BŪDAI

Magistro baigiamasis darbas

Skaitmeninės lingvistikos studijų programa, valstybinis kodas 62604H107
Filologijos studijų kryptis

Vadovas dr. Loiz Boizou _____ _____
(parašas) (data)

Apginta_____ _____ _____
(Fakulteto dekanas) (parašas) (data)

Kaunas, 2010

TURINYS

Natūralios kalbos apdorojimo terminų ontologija: kūrimo problemos ir jų sprendimo būdai

Santrauka

Šiame darbe aptariamas natūralios kalbos apdorojimo terminų ontologijos kūrimas, kūrimo problemos ir jų sprendimo būdai. Tam, iš skirtingų šaltinių surinkta 217 NLP terminų. Terminai išversti į lietuvių kalbą. Trumpai aptartos problemos verčiant. Aprašytos tiek kompiuterinės, tiek filosofinės ontologijos, paminėti jų panašumai ir skirtumai. Išsamiau aptartas filosofinis požiūris į sąvokų ir daiktų panašumą, ką reikia žinoti, siekiant kiek galima geriau suprasti kompiuterinių ontologijų sudarymo principus. Išnagrinėtas pats NLP terminas, kas sudaro NLP, kokios natūralios kalbos apdorojimo technologijos jau sukurtos, kokios dar kuriamos.

NLP terminų ontologijos sudarymui pasirinkus Teminių žemėlapių ontologijos struktūrą ir principus, plačiai aprašyti Teminių žemėlapių (TM) sudarymo principai, pagrindinės TM sudedamosios dalys: *temos, temų vardai, asociacijos, vaidmenys asociacijose* ir kiti.

Vėliau, iš turimų terminų, paliekant tokią struktūrą, kokia rasta šaltinyje, nubraižytas medis. Prieita išvados, jog terminų skaičių reikia mažinti ir atsisakyti pirminės iš šaltinių atsineštos struktūros. Tad palikti tik 69 terminai, darant prielaidą, jog šie svarbiausi. Šiems terminams priskirta keliolika *tipų*, taip juos suskirstant į grupes.

Ieškant dar geresnio skirstymo būdo, kiekvienam iš terminų priskirtas vienas ar keli jį geriausiai nusakantys meta aprašymai, pvz.: *mašininis vertimas – vertimas, aukštas automatizavimo lygis*. Visi meta aprašymai suskirstyti į 7 stambiausias grupes, tarp meta aprašymų ir pačių grupių nustatytos asociacijos ir vaidmenys jose. Šitaip įrodžius, jog ontologijos modelį galima sukurti ir iš metą aprašymų. Aptartas galimas konkrečių atvejų modelis NLP ontologijoje.

Šį tyrinėjimų sritis dar nauja, ir nėra vieno kelio ar atsakymo, kaip sukurti ontologiją. Šiame darbe bandyta į šį procesą pažiūrėti nuo pat pradžių. Aptartos problemos su kuriomis susidurta, pasiūlyti jų sprendimai. Akivaizdu tai, jog žmogaus mąstyme savaime suprantamas asociacijas sunku aprašyti, tačiau įmanoma.

Ontology of Natural Language Processing Terms: Development Issues and Their Solutions

Summary

In this work it is discussed the development of ontology of natural language processing terms, developmental problems and their solutions. In order to reveal the topic of this work was gathered a collection of 217 NLP terms from different sources. The terms were translated into Lithuanian language. Briefly were revealed the problems of translation. There were described both the computer and philosophical ontology, mentioned their similarities and differences. There was discussed in detail the philosophical approach to the similarity of concepts and objects which is needed to know seeking to understand the ontology of computer principles as much as possible. There was examined the term of NLP, what is the NLP, which natural language processing technologies have already been developed, which are still being developed.

For the composition of ontology of NLP terms were chosen the structure and principles of the Topic Maps in order to describe in broad the principles of composition of Topic Maps (TM), the main components of TM: *theme, topic names, associations, role in association and others.*

Later from the got terms there was drawn the tree leaving the structure which was found in the source. It was found that the number of terms should be reduced and it is needed to refuse the primary structure taken from the sources. So, there were left only 69 terms, assuming that they are the most important. There were assigned several types for these terms dividing them into the groups.

Finding better way to improve the distribution for each of the terms were assigned one or more meta descriptions defining it the best, for example: *machine translation - the translation, high level of automation*. All meta descriptions were divided into the seven largest groups, associations and roles within them were set between meta description and the groups. This showed that the ontology model could be created from the meta descriptions. It was discussed a possible model of concrete cases of NLP ontology.

This research area is still new, and there is no one way or the answer how to create the ontology. In this work it was attempted to look at this process from the beginning. There were discussed the problems and offered the decisions. Obviously that it is difficult to describe the associations of human thinking, but it is possible.

1. ĮVADAS

Informacinės technologijos keičia mūsų supratimą apie pasaulį ir tuo pačiu pačia visuomenę. Gutenbergui XV amžiuje išradus pirmąją spausdinimo mašiną, kurios pagrindinė paskirtis buvo pagreitinti Biblijos leidimą, o kartu ir paplitimą, įsibėgėjo informacijos sklaida. Nei jis pats, nei kiti to meto žmonės, tuomet dar nė nenumanė, kad knygos (spausdintas žodis) taip paplis, jog jos taps nauja žinių kontroliavimo ir talpinimo forma, kurią skaitant – mokomasi, sužinoma, jose esančia informacija galima manipuliuoti kitais, pvz.: spauda (J. Davies 2006).

Nepaisant to, kad ir kaip spausdintas žodis paplitęs, jį skaityti, gabenti iš vienos vietos į kitą (laikraščiai, knygos), reikia daug laiko. Šią problemą išsprendė *Internetas*. Be to, kad sutrumpėjo ir išnyko daugelis barjerų, iki tol ribojusių informacijos sklaidą, bei patį jos apdorojimą (kiekvieną tekstą žmogus turi perskaityti, susisteminti ir priskirti jį kažkokiai sričiai, stiliui ar žanrui ir t.t.), *Internetas*, pats žinių vertimas skaitmeninėmis, suteikė galimybę informaciją apdoroti ne tik žmogui, bet ir mašinai. Internetas „artėja“ prie Web 3.0, kitaip dar vadinamo *Semantiniu internetu*. Šis leidžia ieškoti informacijos susijusios *pagal prasmę*, o ne vien pagal *reikšminius žodžius*, kur paieškos frazės žodžiai turi būti ir rastame dokumente. Jei, *Internete* ieškome informacijos apie mašininio vertimo įrankius, mus gali dominti ir įrankį sukūrę žmonės, bei vartotojų atsiliepimai. Visą šią informaciją norime matyti viename lange jau apdorotą ir surinktą iš skirtingų šaltinių. Tokiai paieškai jau nebeužtenka vien reikšminių žodžių, reikalingi ir ryšiai tarp įvairių šaltinių, bei gebėjimas iš jų išrinkti reikiamas žinias.

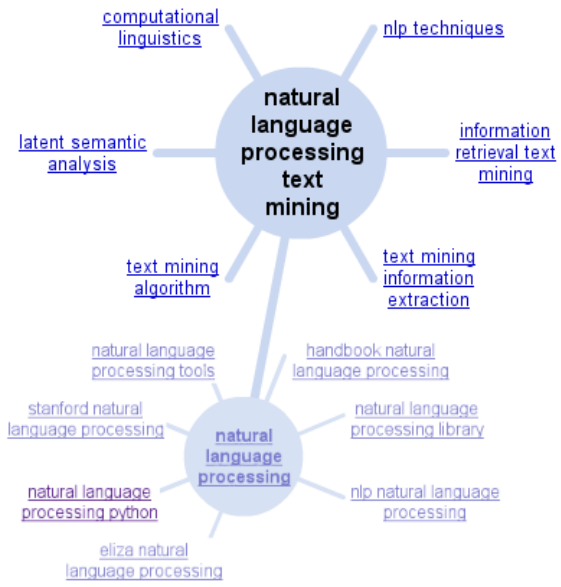
Semantinio interneto „širdimi“ galima vadinti *ontologijas*. Nors filosofijoje ir kompiuterijoje (dirbtinio intelekto kūrime), ontologijos apibrėžimas turi bendrų bruožų, tačiau kartu ir skiriasi. Filosofijoje **ontologija** galima nusakyti šiais žodžiais – *ieškojimas to, kas yra*. Siekiama atsakyti kas yra būtis. Svarbiausia čia **egzistencija** ir **būties kaip tokios** suvokimas. Pagrindinis ontologijos klausimas – *kas egzistuoja?*

Tuo tarpu kompiuterijoje – šio termino daugiskaitine forma **ontologijos** (skirtingai nuo filosofijoje – ontologija) vadinamas tam tikros srities sąvokų visumos specifikavimas išreikštas loginiais ar matematiniais ryšiais tarp jų, į ontologiją galima įtraukti viską, kas yra tikra, tai, kas aptinkama realybėje (P. Valore 2006: 11-13).

Pačią kompiuterinių *ontologijų* kūrimo idėją, netgi galima pavadinti savotišku „kalbos žaidimu“. Žmogui kalbant, mąstant ar rašant, žodžiai dėliojami vienas prie kito,

pasirenkami priklausomai nuo kalbinio konteksto, nuo to, ką jais norima pasakyti. Žodžiais perteikiamas ir suprantamas pasaulis, ir netgi yra teorija (dar kitaip tai galima pavadinti *siekiamybe*), jog egzistuoja tam tikras „to žaidimo“ derinys (kuomet aprašytos visos žmogaus gyvenimo sritys (ontologijų principu), nusakyti ryšiai tarp sričių ir jų viduje), kuris aprašo visą tikrovę (H. Putman 2004). Lipdant žodžius, jų ryšius, kažkada galima pasiekti tašką, kai bus aprašyta visa realybė ir sukurtas dirbtinis intelektas. Bent jau tokios teorinės prielaidos, bet kol kas niekas to nesukūrė. Realu manyti, kad kažkada visa *Internete* esanti informacija virs dirbtiniu intelektu, jei taip nutiktų, iškart kažkas pasikeistų ir tai pajustume. Jeigu toks „kalbos žaidimo“ derinys ir egzistuoja, vargu ar jis statiškas, nekintantis. Atsiradus galingoms skaičiavimo mašinoms, mes pagaliau bandome tai padaryti praktiškai, apie ką ir yra šis darbas.

Naudojantis kompiuterinėmis *ontologijomis*, galima ne tik bandyti modeliuoti žmogaus suvokimą, ar bent jau šiek tiek jį aprašyti, bet ir pagreitinti bei patikslinti paiešką, naudoti naujus informacijos atvaizdavimo būdus (paieškos sistemose pateikiami rezultatai nuo geriausiai atitinkančio, iki prasčiausio, nuo viršaus – apačion), tuo tarpu naudojant *ontologijas* (jos suteikia galimybę informaciją filtruoti įvairias būdais, pažiūrėti į ją iš įvairių žiūros taškų), informacija gali būti pateikiama taip (žr. į 1 paveikslą Google.com informacijos vizualizavimas). Lėtai, bet tai ateina į *Internetą* ir jau galima tuo naudotis.



1 paveikslas. **Pirmieji *Semantinio interneto* pritaikymo pavyzdžiai masiniam vartojimui**

Šiame darbe bandoma atlikti tai, kas dar visai nauja ir bent jau nepavyko rasti, kad Lietuvoje būtų daryta – sukurti Natūralios kalbos apdorojimo (nuo šiol vartojamas trumpinys – NLP) ontologiją.

Darbo objektas – surinkti 217 NLP terminai, iš kurių po to palikti 69 kaip svarbiausi. Šio **darbo tikslas** yra aptarti *Teminių žemėlapių* ontologijos sudarymo principus, aprašyti pagrindines problemas kuriant ontologiją ir pateikti pasiūlymų.

Darbo uždaviniai:

1. Aptarti *Teminių žemėlapių* ontologijos sudarymo principus.
2. Surinkti NLP terminus, juos išversti į lietuvių kalbą.
3. Pabandyti *Teminių žemėlapių* pagrindu sukurti NLP terminų ontologiją, aprašyti jos struktūrą ir sudarymą, iškilusias problemas ir pateikti sprendimus.
4. Suvesti NLP terminus, apsirašytas asociacijas, vaidmenis jose į Ontopia Omnigator programą, aptarti problemas ir pasiūlymus.

Darbe bandoma patikrinti šias hipotezes:

1. Surinkus terminus iš skirtingų šaltinių su jau nusakyta jų struktūra, ontologiją sukurti yra lengviau, nei pirma kurti pačią struktūrą, o tik po to parinkti jai terminus.
2. Terminų išsidėstymo struktūrą ontologijoje lengviau nustatyti pagal terminų meta aprašymus ir šių grupavimą.

Naudojami metodai:

1. Ontologijos kūrimas pagal *Teminių žemėlapių* sudarymo standartą.
2. NLP terminijos kūrimas.

Verta aptarti ir darbo aktualumą. Kompiuterinės ontologijos sąvoką gyvuoja tik apie 20 metų. Augant informacijos kiekiui žmogus suprato, kad pats visko apdoroti nebegali, todėl žmogiško mąstymo principus paremtus logika, semantika, bandoma perkelti į kompiuterines sistemas, kad šios perimtų dalį žmogaus darbų. Tai glaudžiai siejasi du dirbtinio, visą apimančio proto kūrimu. Darbe nagrinėjama NLP terminų sritis, o juk NLP apimančių mokslo sričių tikslas ir yra „protingų“ sistemų sukūrimas.

Ontologijos, jų pagrindų grindžiamas semantinis *Internetas* dar tik įgauna pagreitį ir kiekvienas bandymas aptarti šią sritį ir iškelti ją viešumon, tik priartina „sintetinį“ protą ir visai kitas informacijos apdorojimo, įsisavinimo ir naudojimo galimybes prie mūsų.

2. NLP TERMINAS IR TŽ TERMINŲ VERTIMAS

Norint sudaryti Natūralios kalbos apdorojimo terminų ontologiją, reikia apsibrėžti, kokie terminai šią terminų sritį sudaro. Ne ką mažiau svarbu aptarti NLP apimančias mokslo šakas ir uždavinius, bei tikslus, kuriuos įgyvendinti Natūralios kalbos apdorojimu siekiama.

Toliau šiame skyriuje kalbama apie NLP ontologijos kūrimui pasirinkto ontologijų tipo – *Teminiai žemėlapiai* (jiems skirtas 4 skyrius) sudedamųjų dalių terminų vertimą į lietuvių kalbą, ką reikia atlikti iš anksto, dar prieš pradedant kalbėti apie pačius *Teminius žemėlapius* (nuo šiol TŽ).

2.1 Natūralios kalbos apdorojimas (NLP)

NLP, tai eilė teoriškai pagrįstų ir praktiškai taikomų kompiuterinių technologijų, kurių tikslas analizuoti tekstus (šie gali būti tiek užrašyti, tiek ištarti ir įrašyti žmonėms natūraliai bendraujant, o tada transkribuoti) lingvistiniais metodais, tam, kad būtų galima pritaikyti žmogaus kalbą įvairioms užduotims ar kompiuterinėms programoms.

NLP galutiniu tikslu galima vadinti visišką natūralios kalbos supratimą, tai įmanoma tada, kai kompiuterinės sistemos pajėgs susidoroti su šiomis užduotimis:

1. Įvedamo (tiek kalbant, tiek rašant) teksto parafrazavimu, atrenkant tik tai, kas svarbiausia, tai ir automatinis santraukų darymas, ir konkrečios informacijos išryškinimas tekste.
2. Vertimu iš/į kitas kalbas. Galima teigti, jog tai plačiausiai šiuo metu *Internete* paplitusi NLP technologija.
3. Kompiuterinės sistemos atsakinėjimu į žmogaus užduotą klausimą apie ką yra tekstas, ar daug konkretesnius klausimus (kas su kuo susitiko, kokie miestai minimi tekste ir panašiai).
4. Automatiniu išvadų apie tekstą generavimu – išrinkti tekstus su vertinga informacija, atmetant galimai nereikšmingą, ištaisyti ir pateikti įvertinimą.

Tai tikrai dar ne visos NLP siekiamybės, kadangi Natūralios kalbos apdorojimas yra viena iš dirbtinio intelekto sudedamųjų dalių, kas verčia kurti tokias sistemas, kurios kiek įmanoma geriau imituotų žmogaus mąstymą ir palengvintų informacijos apdorojimą. Manoma, jog geriausiai žmogaus mąstymą galima imituoti remiantis kalba.

Dabar plačiau apie mokslus, kurie sudaro NLP technologijų pagrindą:

Lingvistika, kuri remiasi formaliais ir struktūruotais kalbos modeliais, bei ieško kalboje universalijų, to, kas būdinga konkrečiai ar visoms kalboms (apima sintaksę, morfologiją, leksikografija ir t.t.). Be lingvistikos, NLP neatsiejama nuo **informatikos**, nes būtent kompiuterių atsiradimas suteikė galimybę į kalbą pažiūrėti plačiau, ją apdoroti ir modeliuoti. Trečiasis mokslas besisiejantis su Natūralios kalbos apdorojimu – **kognityvinė (pažinimo) psichologija**, kuri akcentuoja mąstymą ir kognityvinių procesų įtaką elgesiui. Svarbi ir kalba, į kurią žiūrima kaip į bendravimo priemonę ir būdą geriau suprasti žmonių bendravimo ypatybes, o formalizuojant kalbą, netgi suprasti, kodėl žmonės elgiasi taip, o ne kitaip.

Kaip jau ankščiau paminėta – NLP skirta analizuoti tiek rašytinę, tiek sakytinę kalbą, tačiau šį teiginį būtina papildyti ir tuo, jog išskiriamos dvi pagrindinės NLP šakos, kur skiriamos NLP technologijos, jų paskirtis ir mėginimas atkartoti žmogaus galimybes, tai: 1) **kalbos apdorojimas** ir 2) **kalbos generavimas**.

Apdorojant kalbą, kompiuterinė sistema atlieka *skaitytojo*, arba *klausytojo* vaidmenį. *Skaitytojo*, nes priima tekstinę informaciją ir ją apdoroja, interpretuoja. *Klausytojo*, kai apdorojamas sakytinis tekstas ar natūrali šneka realiu laiku (čia kalbos atpažinimo sistemos, automatinio diktavimo ir valdymo balsu programos.

Be kalbos apdorojimo, svarbus ir jos generavimas – sintezavimas. Generuodama informaciją, sistema atlieka *rašytojo*, arba *kalbėtojo* vaidmenį. *Rašytojo*, kuomet sistemai uždavus klausimą, ši pateikia apdorotą informaciją į ekraną, ar kitą įrenginį, galop kitai programai – tolesniam informacijos apdorojimui. *Kalbėtojo*, kai tekstas sintezuojamas į kalbą. Kompiuterinei sistemai atpažįstant kalbą, pirmiausia apdorojami garsiniai signalai, tačiau jie vis tiek tolesniuose žingsniuose surišami su tekstu, konkrečiau juo užrašytais garsais. Taigi, visi šie 4 paminėti vaidmenys imituoja žmogaus sugebėjimus suprasti, apdoroti ir keistis informacija su kitais.

Aptarus pačią NLP esmę, šios mokslo srities tikslus, būtina užsiminti ir nuo ko visa tai prasidėjo. Pačia NLP pradžia, galima laikyti ***mašininio vertimo*** atsiradimą dar praeito amžiaus 5-tame dešimtmetyje Amerikoje, siekiant automatiškai išversti priešingoje pusėje kariavusių ((rusų) tuomet vyko Antrasis Pasaulinis karas) žinutes. Pats vertimas buvo primityvus – pažodinis, o išversti žodžiai sudėliojami pagal kalbos, į kurią verčiama, struktūros taisykles. Kūrėjai suprato, kad ne viskas taip paprasta, kaip atrodė, ir taip tikėtas visiško mašininio vertimo sukūrimas per keletą metų nedavė rezultatų. Garsusis ALPAC

pranešimas 1966 metais Amerikoje, kritikavęs visiško mašininio vertimo (angl. *full machine translation*) sukūrimo galimybę, ilgam sustabdo šios technologijos tobulinimą. Negalima teigti, kad kalbos kompiuterizavimas sustojo, buvo kuriami prototipiniai kalbos generavimo, dialoginių sistemų projektai, tačiau atsigavimas pastebimas tik po 80-tųjų metų. Imami plačiai taikyti statistiniai metodai, daugėja skaitmenizuoto teksto, spartėja kompiuteriai, interneto atsiradimas (E. D. Liddy 2001).

Kalbant apie mašininį vertimą dabar, tobulėjant technikai ir atsiradus dideliems tekstynams, kuriuos galima vadinti „dideliu konkrečios kalbos pavyzdžiu“, buvo pritaikyti statistiniai metodai, kurie kartu su taisyklių rinkiniais, dar iki šiol tobulinami, nes kalbos pačios nuolat kinta, pasipildo nauja leksika, kai dalis jos palaipsniui užmirštama. Akivaizdu tai, kad netgi rašant šį darbą, naudota ne viena mašininio vertimo technologija (Tildės biuro vertimo vedlys, VDU vertimas (2010), Google vertėjas (2010)), kas rodo, jog technologija gyva ir plačiai taikoma. Ji nėra tobula, tačiau palengvina supratimą, o verčiant paprastus sakinius, gaunami neblogi rezultatai.

Trumpai aptarus istoriją, toliau reikia paminėti NLP lygius. Kuo NLP sistema tobulesnė, tuo šių lygių ji apims daugiau, kartu priartėdama prie žmogiško supratimo.

1. **Fonetika**. Šis lygmuo siekia teisingai interpretuoti kalbos garsus. Bandoma išspręsti tokias problemas, kaip žodžių atpažinimas šnekamojoje kalboje, kur kalbama greitai ir žodžiai persidengia vieni su kitais, garsai tariami nevienodo ilgumo, be to, balsu dar ir perteikiama nuotaika, intonacija.
2. **Morfologija**. Morfemų – mažiausių reikšmės vienetų, problematikos sprendimai. Ypač aktualu kaitomose kalbose, kur net ir galūnės pasikeitimas, keičia žodžio prasmę (*tu – bėg**i**, jis – bėg**a***), arba norint surinkti informaciją apie žmogų, turint jo pavardę, neužtenka vieno linksnio, o reikia visų jo variantų. Šioje vietoje verta užsiminti apie Ramono V. (2009) darbą, kuriame nagrinėtas asmenvardžių atpažinimas. Surinkus kiek įmanoma daugiau vardų ir apsirašius moteriškų pavardžių galūnes su priesagomis, bei remiantis tuo, jog vardas ir pavardė eis šalia ir prasidės iš didžiųjų raidžių, pavyko iš didelio tekstų rinkinio ištraukti moteriškus asmenvardžius (nes moterų pavardžių sudarymo sistemą griežtesnė, nei vyrų), paklaidos neišvengta, tačiau tai puikus pavyzdys, kaip galima pritaikyti NLP.
3. **Leksika**. Žodžių reikšmės supratimas. Vienas iš paprasčiausių būdų suprasti reikšmę – priskirti žodžiui kalbos dalies žymę, tačiau šie gali priklausyti ir kelioms kalbos dalims, arba būti daugiareikšmiai, priklausomai nuo juos supančio konteksto ar kalbinės

situacijos. Šios problemos sprendžiamos ir statistiniais metodais, kai skaičiuojama, kokia tikimybė, kad žodis eis kartu su vienu ar kitu žodžiu.

4. **Sintaksė**. Aprašomos gramatinės sakinių struktūros, kokia kalbos dalis eina po kurios. Čia bene didžiausia problema, kai sakinių gramatinė struktūra identiška, tačiau reikšmes visai kitokios, pvz.: *Jonas eina pas Petrą į svečius ir Petras eina pas Joną į svečius* (pavyzdžiai darbo autoriaus) – kalbos dalių išsidėstymas identiškas, tačiau keičiasi reikšmė kas pas ką eina.
5. **Semantika**. Norint nusakyti žodžio ar sakinio prasmę, neužtenka pažiūrėti į žodyną ir pagal užrašytas reikšmes pasakyti, jog sakinys ar žodis apie tai ir tik tai. Kaip ir leksikoje, norint įvardinti prasmę, būtinas kontekstas, ryšių tarp sąvokų nusakymas ir įvardijimas. Su semantikos problemomis kovojama kuriant ontologijas, kas padeda kompiuterinėms sistemoms suprantama kalba pateikti pasaulio sąvokas, ryšius tarp jų.
6. **Diskursas.** Nagrinėjant diskursą, žiūrima į didesnius teksto blokus, nei pavieniai sakiniai, siekiant nusakyti jų prasmę. Didžiausios problemos – teisingas įvardžių interpretavimas, vertimas, arba kai jie tik nuspėjami, tačiau be jų negalima suprasti prasmės. Pvz.: *Jis nuėjo į mišką. Čia Petras kirs medžius* – jei sistema nesusies *jis* ir *Petras* ar *mišką* ir *čia*, sakiniai nesisies ir uždavus sistemai klausimą: *Kur Petras kirs medžius*?, jokio atsakymo nebus. Diskursas apima ir teksto dalių nustatymą, pvz.: kur teksto dalis svarbiausia, o kur ne tokia svarbi informacija. Bent jau internetinėje spaudoje galima interpretuoti, jog svarbiausia informacija pačioje straipsnio pradžioje. Šių problemų sprendimų ieško ir paieškos sistemos, parenkančios, kuri informacija svarbiausia pagal žodžių svorius, statistiką ir pan.
7. **Pragmatika.** Pragmatika siekiama atsakyti į klausimą – kas ištikto norėta tuo pasakyti, ar kas slepiasi po vienu ar kitu pasakymu, čia gali būti iš ironija, pašaipa, pamokymas, nutylėjimas. Žiūrima į intonaciją, ištarimo būdą, anekdotai, kalambūrai, kuriems suprasti dažnai reikalingos foninės žinios. Ne visada žmogus supranta, kad su juo kalbama ironiškai ar nutylint, ne visada visi vienodai reaguoja į anekdotus, tad kompiuterizuoti pragmatiką bene sudėtingiausia.

Aptarus visus 7 NLP lygius, kuriuos visus pritaikius kompiuterinėms sistemoms, būtų priartėta tiek prie natūralios kalbos supratimo, tiek prie dirbtinio intelekto sukūrimo, akivaizdu tai, jog paplitus internetui, atsirado ir tikslios paieškos, informacijos grupavimo, jos susiejimo su besisiejančia informacija strategijų poreikis. Žmogus įvedęs kelias raides į paieškos lauką (šie dabar visur, nuo paieškos sistemų, operacinių sistemų, ar socialinių tinklų), jau tikisi rasti tai, ko jam reikia. Informacijos paieškos rezultatus siekiama pateikti

viename lange. Pasiūlymas išversti tinklalapį į užsienio ar gimtąją kalbą jau nebestebina, tad kiekvienais metais NLP technologijų svarba tampa vis aktualesne.

2.2 Teminių žemėlapių (TŽ) terminai

Kadangi darbe susiduriama su specifiniais (*Teminių žemėlapių* sandaros elementai), tiek ir Natūralios kalbos apdorojimo terminais, ir galiausiai terminai verčiami iš anglų kalbos į lietuvių kalbą, būtina apsibrėžti, jog ***terminas*** – žodis ar žodžių junginys, įvardijantis specialią mokslo, technikos, meno ar kitos visuomeninės gyvenimo srities sąvoką (Keinys. S. 2005).

Patys terminai privalo pasižymėti (pasirinkti svarbiausi): 1) **Vienareikšmiškumu** – jokių šalutinių reikšmių, 2) **Tikslumu** – sietinas su jų vienareikšmiškumu, vengiant bet kokio daugiaprasmiškumo 3) **Sistemiškumu** – sistematiškai sietis su kita tos srities terminija, 4) **Taisyklingumu** – atitikti kalbos normas, gramatines, rašybos, kirčiavimo, 5) **Produktyvumu** – jo pagrindu galima sudaryti kitus terminus), 6) **Trumpumu** – kuo trumpesnis ir konkretesnis, kiek įmanoma keliažodžius terminus trumpinant, pvz.: jei kontekstas leidžia, gimininį terminą galima pakeisti rūšiniu, kad ir *dantų šepetėlis* į *šepetėlis*, 7) **Stilistiniu neutralumu** – atsisakant išraiškingumo ir vaizdingumo. (Keinys. S. 2005: 36-50)

Apsibrėžus patį terminą ir prieš pradedant gilintis į *Teminių Žemėlapių* sudarymo elementus, į lietuvių kalbą išsiversti pagrindiniai darbe vartojami TŽ terminai. Verčiant laikytasi reikalavimų, susijusių su terminų sudarymo logikos dėsniais: a) *tikslumo,* b) *trumpumo*. Aišku viena, sunkiausia rasti tikslų ir gerai kalbą atitinkantį termino atitikmenį, kad ir šis pavyzdys: angl. *Occurrence* – vienažodis terminas, o išvertus jį tokiu pat vienažodžiu terminu į lietuvių kalbą – *atvejis*, iškarto pažeidžiamos neutralumo, bei vienareikšmiškumo taisyklės, kuomet terminas turi būti neutralus ir neturėti kitų reikšmių ir asociacijų.

Pats *Teminių žemėlapių* (angl. *Topic Maps*) terminas gali būti verčiamas dvejopai (žiūrėti į 1 lentelę) – *temų žemėlapiai* ir *teminiai žemėlapiai*, pasirinktas pastarasis variantas, nes kalbant apie žemėlapius, lietuvių kalboje paprastai vartojama *–in* priesaga su galūne *–is*, pvz.: *kartografinis, panoraminis* (remiamasi darbo autoriaus, kaip gimtakalbio nuomone).

1 lentelė. **Pagrindiniai Teminių žemėlapių terminai**

Angliškas terminas	**Lietuviškas terminas**
Association role	Galimi keli vertimo variantai: *asociacijos rolė, asociacijos vaidmuo,* tačiau pagal savo prasmę teisingiausia šį terminą versti kaip: ***vaidmuo asociacijoje***
Association type	*Asociacijos tipas* arba tiesiog *asociacija*
Facet	*Aspektas*
Name type	*Vardo tipas* arba tiesiog *vardas*
Occurrence	*Konkretus atvejis* arba tiesiog *atvejis*
Scope	*Nagrinėjimo sfera*
Topic Maps	*Temų žemėlapiai*, kitas galimas variantas – ***Teminiai žemėlapiai***
Topic type	*Temos tipas*
Topic	*Tema*
Theme	*Tematika*

Aptarus terminus, jų sudarymo reikalavimus ir tai, kas yra NLP, kitame skyriuje aprašomas sąvokų panašumo klausimas, aptariamos kompiuterinių ir filosofinių ontologijų panašumai ir skirtumai.

3. SĄVOKŲ PANAŠUMAS IR ŠIOS JŲ SAVYBĖS TAIKYMAS ONTOLOGIJOSE

Norint pradėti kalbėti apie kompiuterines ontologijas ir jų kūrimą, būtina kiek plačiau aptarti ir filosofinį to pagrindą, pačias pagrindines idėjas, kurių dėka imtos kurti kompiuterinės ontologijos.

Taigi, daiktai ar mus supančios sąvokos turi tam tikras savybes, kuriomis išsiskiria, arba tiesiog jas turi. Pagal šias skirtybes, mes skirstome daiktus ar sąvokas į kategorijas. Kad ir kėdės, jos turi tai, ką galime pavadinti „kėdiškumu“ ir kitame daikte atpažinę „kėdiškumą“ – priskirsime jį kėdėms. Šias savybes galima vadinti universalijomis arba formomis (Platonas manė, kad yra idėjų, sąvokų (plačiau galima jas tiesiog vadinti universalijomis) pasaulis, ir realybėje mes tik bandome atsiminti tai, ką ten žinojome). Dažnai yra sunku nusakyti, pagal ką mes skirstome daiktus, viena logiškiausių idėjų – pagal panašumą, tačiau kas gali nulemti panašumą ir ar jis iš viso svarbus? Į tai, ir kas gi tos universalijos, bando atsakyti trys filosofinės kryptys:

1. *Realizmas* – tarp konkrečių daiktų egzistuoja panašumas ir tai nepaneigiamas faktas. Kalbant apie realizmą ir universalijas, verta paminėti ir tai, jog tam tikros sąvokos (žodžiai), turi savo etalonus (labiausiai panašius į sąvoką), pavyzdžiui pagalvojus apie medį – pirma mintis: ąžuolas, pelę: graužikas, saulę: geltona. Aišku, šie etalonai tarp kultūrų gali skirtis (vien jau pažiūrėjus į patarles, kaip jos skirtingai sudaromos priklausomai nuo kalbos), tačiau tai akivaizdžiai parodo ryšius ir panašumų ieškojimą, iš šalies žiūrint tarp visiškai nepanašių dalykų (kad ir metafora – *vėjo pamušalas*, kur vėjas, remiantis logika niekaip nesisieja su pamušalu).
2. Nominalizmas – panašumai tarp konkrečių daiktų egzistuoja, tačiau panašumai patys savaime nėra kažkas tikro.
3. Antirealizmas – paneigia bet kokias panašumo egzistavimo galimybes.

Dabar plačiau apie nominalizmą ir antirealizmą. Taigi, plačiau kalbant apie nominalizmą – ši sritis paneigia bet kokias Platono formų ar universalų egzistavimo idėjas ir tiesiog pripažįsta, jog egzistuoja panašumas ir daugiau į tai gilintis neverta. Ką jaučiu savo pojūčiais – to gana. Koncentruojamasi į tai, kas žmogaus viduje, kone paneigiant, kad mintys ir vaizdai turi ryšį su realybe. Tačiau šis požiūris susiduria su problema jau iš karto po to, kai keli žmonės kalba apie ta patį, žiūrėdami ir matydami eilę kėdžių, šie priskiria jas kėdėms, tad tai prieštarauja minčiai, jog visa tai, ką matau, yra tik mano galvoje (kuriant bet kokias

kompiuterines ontologijas būtinas bendras suvokimas ir susitarimas, kaip vieną ar kitą ryšį arba panašumą tarp daiktų traktuoti, jau nekalbant apie terminų ontologijas, kur visus terminus kiekvienas individas turi suprasti vienodai, tad čia kyla universalios kalbos idėja, kas vargu ar įmanoma).

Antirealizmas – kuomet kalbama apie formas ar universalijas – tiesiog kalbama žodžiais ir apie žodžius. Kiekvienas bandymas kalbėt apie šias bendras viskam savybes, tėra žmogaus sukurto mąstymo projektavimas į realybę, pasitelkiant kalbą. Reikštų, jog mūsų kalbos struktūra, formuoja mūsų suvokimą ir tai, kaip elgiamės pasaulyje, kaip atpažįstame daiktus ir tai, iš kur žinome, kaip su jais elgtis (remiantis šiuo požiūriu, bet koks klaidingos ontologijos sukūrimas ir pritaikymas masėms, gali iškreipti bendrą suvokimą, iš esmės logiška, nes ko nėra kalboje, to nėra ir pasaulyje, jei paieškos variklis paremtas ontologija neaprėps viso pasaulio, reiškias ši pasaulio dalis nebus matoma).

Aptarus požiūrį į sąvokas, panašumo tarp jų ir universalijų paieškas iš filosofinio žiūros kampo, būtina nepamiršti, jog kompiuterinė ontologija – tai loginis pasaulio sąvokų atvaizdavimo būdas. Tai kartu ir teorija, kad realybę galima kategorizuoti nusakant ryšius tarp sąvokų. Taigi, kiekvienas iš šių požiūrių gali būti naudingas kuriant Teminių žemėlapių kompiuterinę ontologiją. Realizmas siūlo ontologiją struktūruoti taip, jog patys panašiausi terminai būtų kuo arčiau pagrindinio termino (netgi nesilaikant abėcėlinio išrūšiavimo), vėlgi, panašumas gali būti apsprendžiamas keliais būdais:

1. Skaičiuojant kaip dažnai terminai minimi šalia kitų (tam reikalingas didelis konkrečios srities tekstynas). Taip galima gauti *Teminius žemėlapius*, kuriuose greičiau aptinkama informacija. Tokią galimybę turi paieškos sistemos, aprėpiančios didelį kiekį informacijos, kaip realus to taikymas (žr. į 1 paveikslą) egzistuoja paieškos vizualizavime.
2. Renkant jau struktūruotus tekstus ir žiūrint į teksto turinio struktūrą. Turint galimybę pasitelkti internetą, naudoti reitingavimą, kuomet pagal vartotojų užklausų skaičių ar suteikiama galimybę balsuot, kuri informacija svarbesnė – taip svarbiausi terminai atsiduria viršuje, mažiau svarbūs – apačioje.

Nominalizmas siūlo pasitelkti savo nuomonę ieškant panašumo, na o antirealizmas – dar labiau plėtoti kalbinę analizę kaip svarbiausia, mažiau atsižvelgiant į žmogaus siūlomą struktūravimą, o remiantis didelių kalbos išteklių duomenimis. Vėlgi, čia tikroji tiesa, jei ji yra, niekad nebus pasiekta, kol be tekstų apdorojimo, nepasieksime ir gero, bei greito spontaninės šnekamosios kalbos apdorojimo. Kalba tekstu užrašoma pagal tam tikras

taisykles, logiškai dėstant mintis, tuo tarpu šnekamoji kalba nepasižymi nuoseklumu ir tų taisyklių laikymusi (T. Horgan 2004). Filosofinė ontologijos koncepcija teigia sąryšių buvimą, tuo tarpu kompiuterinės ontologijos skirtos tuos ryšius perteikti loginiais principais, struktūruojant sąvokas ir asocijuojant jas ryšiais, kai kuriuos ryšius apribojant ir pan., bei pritaikant jas realiame pasaulyje.

4. ONTOLOGIJŲ UŽRAŠYMO KALBOS

Aptarus filosofinės ontologijos principus, bei šių idėjų galimą pritaikymą kompiuterinėms ontologijoms, toliau darbe aprašomos kompiuterinių ontologijų užrašymui naudojamos kalbos. Kalbame natūralia kalba, tačiau norint aprašyti ontologijų ryšius, logines ar semantines sąsajas, būtinos formalios kalbos, turinčios griežtą specifikaciją, priimtos kaip visuotinis standartas. Neturint griežtai apibrėžtos kalbos, norintys kurti ontologiją, susikurtų naujas užrašymo kalbas, arba laisvai modifikuotų jau esamas, o kuo daugiau tokių variantų, tuo mažesnė susikalbėjimo ir technologijos paplitimo galimybė. Tad apie pagrindines kalbas:

4.1 XTM

XTM kalba yra pagrįsta XML (angl. *Extensible Markup Language*) ir yra kaip *XML Topic Maps* akronimas. Oficiali XTM specifikacijos svetainė (XML Topic Maps, 2010) pateikia štai tokius šiai specifikacijai keliamus reikalavimus:

1. XTM turi būti lengvai panaudojamas *Internete*.
2. XTM turi palaikyti įvairiausius taikymo variantus.
3. XTM turi būti suderinamas su XML, XLink ir ISO 13250.
4. XTM turi turėti kuo mažiau papildomų savybių (angl. *optional features*).
5. XTM užrašyti dokumentai turėtų būti pakankamai lengvai ir aiškiai skaitomi.
6. XTM žymos turi būti trumpos ir aiškios.

Įdomu tai, jog nors XTM žymų ir nėra daug, jomis galima aprašyti įvairiausius santykius tarp duomenų ar sąvokų (leidžia kartu naudoti ir toliau aptariamas RDF ir OWL kalbas kartu su TŽ specifikacija). Būtent ši kalba ir naudojama TŽ aprašymui, aišku norint kurti Teminių žemėlapių ontologijas, nėra būtina išmanyti pačios XTM sintaksės ir sudarymo

principų, nes yra programų (pvz.: *Ontopia Omnigator*), kurios TŽ kūrimui naudoja grafinę sąsają.

4.2 RDF

RDF (angl. *Resource Description Framework*) – standartas buvo sukurtas tam, kad aprašyti praktiškai viską ir kelius, bei ryšius vedančius iš viso to, struktūriškai modeliuoti informaciją, paaiškinti ją ir praplėsti jau turimą – nauja. Pačią RDF schemą galima vadinti trilype, nes ją paprastai sudaro: subjektas, predikatas ir objektas, pvz.: *Jonas gyvena Kaune*, *Jonas* – subjektas, *gyvena* – veiksmažodis (predikatas), *Kaune* – objektas.

Teigiamos pusės: plačiai pritaikoma tiek internete, tiek kuriant duomenų bazes, aprašant loginius ryšius, ar kaip įvairių internetinių paslaugų dalis.

Trūkumai: nors ši ryšių aprašymo kalba ir laikoma vienu iš standartu, tačiau ir čia neišvengiama problemų. Kiekvienas RDF fragmentas (panašiai kaip TŽ specifikacijoje – *Tema*) gali turėti URI nuorodą (vėlgi TŽ – konkrečių atvejų atitikmuo), nukreipiančią į su fragmentu susijusią informaciją ar aprašą (siekiant patikslinti informaciją), bet kai net tame nukreipiančiame šaltinyje nėra iki galo aišku apie ką fragmentas (ypač, kuomet keli fragmentai identiški ir tik URI kiek skiriasi), sunku nustatyti fragmento reikšmę (B. Passin 2004: 155-60). Kadangi internete informacija nuolat atnaujinama ir duomenys gaunami iš skirtingų šaltinių, šie gali konfliktuoti tarpusavyje, atsirasti identiškų ar klaidingai RDF kalba aprašytų loginių ryšių.

4.3 OWL

Trečia plačiai paplitusi ontologijų užrašymo kalba – OWL (angl. *Web Ontology Language*). Kaip jau ir prieš tai minėta RDF kalba, OWL naudojama ontologijų kūrimui ir iš esmės remiasi RDF schema, tačiau gerokai ją papildydama. OWL galima aprašyti griežtas ribas – pvz.: automobilis negali turėti daugiau nei keturis ratus, papildomas daiktų galimybes – *namas gali ir neturėti langų*, sujungti keletą klasių – *žmonės iš skirtingų miestų, tačiau lanko tą pačia mokyklą*. Šiuo metu tai pagrindinė ontologijų kūrimo kalba naudojama internete (B. Passin 2004: 161-63).

5. TEMINIAI ŽEMĖLAPIAI

Jau aptarus NLP terminą, ontologijų tipus, toliau bus kalbama apie vieną iš ontologijų modelių, kuris buvo pasirinktas, kaip geriausiai tinkantis Natūralios kalbos apdorojimo terminų ontologijos sudarymui – *Teminius Žemėlapius*. O Teminiai žemėlapiai pasirinkti todėl, kad jų sudarymo principas nėra per daug sudėtingas, turi aiškiai aprašytus sudarymo principus.

5.1 Indeksai (rodyklės)

Tam, kad geriau suprasti *Teminių žemėlapių* sudarymą, neišvengiamai reikia aptarti ir pagrindinę jų paskirtį (iš informacijos automatiškai generuoti indeksus (rodykles)). Paprastai *ontologijų* sudarymo principas organizuojamas taip (šitai padės geriau suvokti jų sudarymo esmę):

1. Konceptų (temų, vardų) rinkinys.
2. Ryšių, siejančių tuos konceptus rinkinys.
3. Ir dar kiti ryšiai, nukreipti į konkrečia informaciją (iš konceptų ar ryšių). (Teminių žemėlapių sudarymo aprašymas 2009)

Pačios ontologijos palaipsniui įgauna vis didesnę reikšmę, palaikant keitimosi informacija procesus, nes jos gali perteikti daugumai priimtą ir bendrai naudojamą supratimą apie taikomąsias sritis. Šiuo metu, kuomet informacijos kiekis nuolat auga, ji atnaujinama, papildoma, ir per dieną gali pasikeisti bent kelis kartus, reikalingi būdas, kaip visame tame rasti dominančią informaciją. Norint ją rasti kuo greičiau, reikalingas indeksavimas (duomenų sutraukimas į rodyklę, kurios pagalba vėliau po tuos duomenis būtų galima „keliauti“), kurio pagalba randamas trumpiausias atstumas (trunkantis laikas), kaip iš taško A pasiekti tašką B.

Vien jau atsivertę ir peržvelgę knygos turinį ar rodyklę, įvertiname apie ką bus knyga ir galime pasirinkti, kurią jos dalį atsiversti, kad rastume norimą informaciją. Turinio – knygose, abėcėlinės rodyklės – žodynuose, bei knygose, dokumentuose, pagalba išvengiama blaškymosi po visą tekstą.

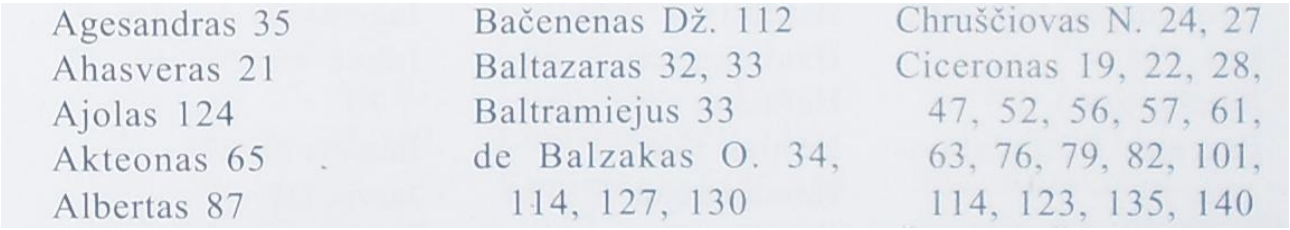
Tekstiniuose dokumentuose sužymimi reikšminiai žodžiai ar frazės, ir po to iš jų sukuriami indeksai, tačiau ši rodyklė tinka tik konkrečiam dokumentui. Taigi, tekstą žmogus suindeksuoja savo nuožiūra, visa tai paversdamas rodykle, tačiau kiekvienas žmogus tą patį

teksto gabalą (pastraipą ar dalį) pavadintų skirtingai, reikšminiai žodžiai irgi varijuoja priklausomai nuo žmogaus pasirinkimo.

Paplitus *Internetui*, dokumento autentiškumas prarandą prasmę, jei randame mus dominantį dokumentą su ilgai ieškota informacija, dažnai kyla poreikis rasti ir daugiau tokių dokumentų su panašia informacija (galima pateikt pavyzdį ir su muzika: jei jums patinka konkrečios dainos skambesys, norite rasti panašių į šį skambesį dainų), tad nebeužtenka senųjų indeksavimo galimybių, kuomet žmogus savo rankomis išrenka reikšminius žodžius, sakinius, geriausiai apibūdinančius tekstą. Nebeužtenka ir dėl greičio stokos tai atliekant, ir galimų skirtumų tarp skirtingų žmonių nuostatų.

Vienas galimų būdų – automatiškai suindeksuoti visą tekstą. Pagrindinė šio principo problema ta, jog imami visi tekste esantys žodžiai, dažniausiai besikartojantys išmetami, o likę, priskiriami prie dokumento **reikšminių žodžių** (geriausiai jį apibūdinančių). Čia iš karto iškyla **homonimų, sinonimų** problema, nes iš teksto ištraukti žodžiai dažnai turi ne vieną reikšmę. Šiuo principu veikia interneto paieškos sistemos.

Norėdamas sukurti indeksą, kad ir nedidelei knygai, turėtumėte imti kiekvieną reikšminį žodį, ieškoti jo pavartojimo atvejų ir žymėti kur jis pasitaikė tekste.



2 paveikslas. **Indekso rodyklės pavyzdys (A. Butkus 2009)**

Dabar detaliau apie pačius indeksus (rodykles nukreipiančias į konkrečią teksto dalį). Pavyzdžiui, knygų gale sąrašas pavardžių ar svarbiausių faktų paminėtų knygoje ir nurodytas puslapis, kuriame apie juos rašoma (žr. į 2 paveikslą).

Būtina pakalbėt ir apie skirtingą žymėjimą, norint atskirt skirtingus dalykus:

Žalakevičius V. 98, 136

Puccini, Giacomo, 69-71

La Bohème, 10, 70, **197-198**, 326 (pavyzdžiai darbo autoriaus)

1. Skirtingi lygiai išskiriami skirtingai atspausdinant (pvz.: autoriaus kūriniai išskiriami kursyvu ar atitraukimu).

2. Nuoroda į aprašymą (šiuo atveju kūrinio), žymima pajuodinimu.

Ir įvairūs kiti, priklausomai, nuo leidėjų ir nuo rodyklės tipo. Knygose galima aptikti skirtingas rodykles vardams, vietovėms, lotyniškiems apibrėžimams ir t.t. Homonimai gali būti atskirti ir paaiškinti atskirai, nurodant puslapį knygoje (pvz.: Nemunas (vardas), Nemunas (upė)). Galimi ir įvairūs nurodymai rodyklėje, kurioje puslapio dalyje konkretų atveji galima rasti. Dar viena indeksavimo alternatyvų – Biblijos teksto padalinimas į eilutes (skyrius, eilutė).

Galima manyti ir taip, jog žodžiai patenkantys į antraštes, turi didesnius svorius, geriau apibūdina apie ką tekstas, nusako jį, tad NLP sritis susijusi su automatizuotu teksto apdorojimu.

5.2 Terminų žodynai, Tezaurai

Aptarus indeksus, kita svarbi dalis (taip pat įeinanti į *Teminių žemėlapio* specifiką) – Terminų žodynas – terminų sąrašas su jų definicijomis. Galimą jį suprasti ir kaip tam tikrą indeksą, kuriame svarbus tik tas pavartojimo atvejis, kuriam priskiriama konkreti definicija (žodžio reikšmė) nenurodant į kažkokią teksto dalį.

Be terminų žodyno, svarbus ir *Tezauras* – žodynas, pateikiantis žodžio sinonimus, giminingus žodžius, ryšius tarp žodžių. Tezaurą sudaro skyreliai, kuriuose surašyti žodžiai, susiję su antraštiniu skyrelio žodžiu, ir abėcėlinis žodžių, bei ženklų, žyminčių skyrelius, kuriuose tie žodžiai minimi, sąrašas. Ne visi tezaurai turi visas čia minėtas savybes. Kartais tezauru vadinamas vien sinonimų žodynas.

Tad nors terminų žodyno definicijose ir būna paaiškinta iš kurios kalbos žodis kilęs, ar užrašytas tarimas, tačiau svarbiausia vis tiek lieka pats terminas ir jo reikšmė, tuo tarpu tezauras jau nusako ryšius tarp žodžių, skirsto juos į klases (kas panašu į ontologiją ir svarbu tai suprasti, norint geriau suvokti ontologijos sudarymo principus).

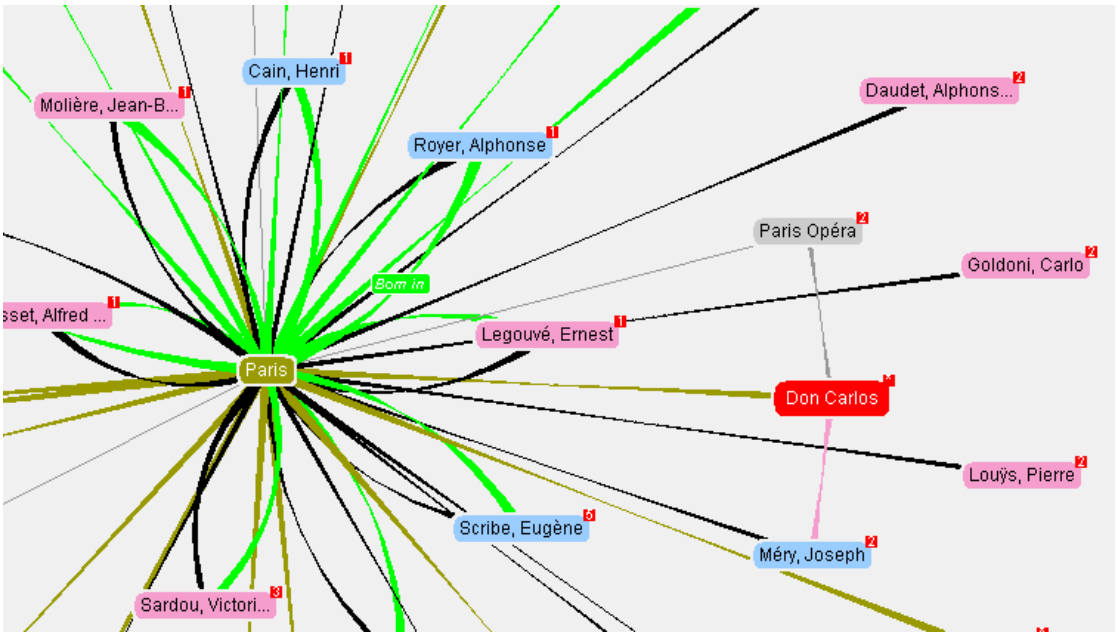
Kalbant apie ryšius tarp terminų tezaure, lyginant su įprastiniu indeksu ir terminų žodynu, svarbu yra tai, kaip šie ryšiai **užrašomi**. Šitai svarbu, nes galima matyti ne tik, kad terminai susiję tarpusavyje, bet ir kaip ir kodėl jie susiję (pagal tai, kokiai klasei žodis priskirtas, pvz.: platesnė reikšmė, siauresnė, besisiejantys terminai – angl. *broader term, narrower term, related term* ir pan.).

Minėti žmogiškų žinių struktūravimo ir žymėjimo būdai, (terminų žodynas, tezauras, indeksas (rodyklė)) be vargo suvokiami žmonių, tačiau, kokiais būdais šią sampratą (žinias ir reikšmes) užrašyti taip, kad ji būtų suprantama ir žmogui ir mašinai, kaip suvokimą transformuoti į mašinoms „suvokiamą“ kalbą. Kaip pasinaudojus geriausiomis indeksavimo, terminų žodynų ir tezaurų savybėmis tai užrašyt kalba, suprantama tiek žmogui, kuris ją rašys, tiek mašinai, kuri ją naudodama generuos indeksus iš didelių duomenų bazių. Vienas iš paplitusių žinių formalizavimo būdų yra sąvokiniai grafikai (angl. *conceptual graphs*), kurie „lipdomi“ iš sąvokų ir ryšių tarp jų. Arba dar kitaip – semantiniai tinklai. Pavyzdys:

[žmogus] <- (agentas) <- [valgo] -> (objektas) -> [sumuštinį] – tai jau gana panašu į ontologijų sudarymo principą (mašinai suprantamą formalizuotą kalbą). Ryšiai tarp žodžių, sąvokų užrašomi įvairiai, pasitelkiama predikatų logika, matematinės išraiškos. Neliečiant šių būdų, toliau darbe kalbama konkrečiai apie *Teminių žemėlapių* ontologijos tipo sandarą.

6. TEMINIŲ ŽEMĖLAPIŲ SANDARA

Prieš tai buvusiuose skyriuose aptarta, kas yra ontologija, kalbėta apie indeksus, tezaurus ir terminų žodynus. Kadangi TŽ principais šiuos informacijos rūšiavimo būdus galima aprašyti mašinoms suprantama kalba, šiame skyriuje bus kalbama apie pačius *Teminių žemėlapių* tipo ontologijos sudarymo principus. TŽ sandara aprašoma remiantis Teminių žemėlapių sudarymo aprašymu (2009).

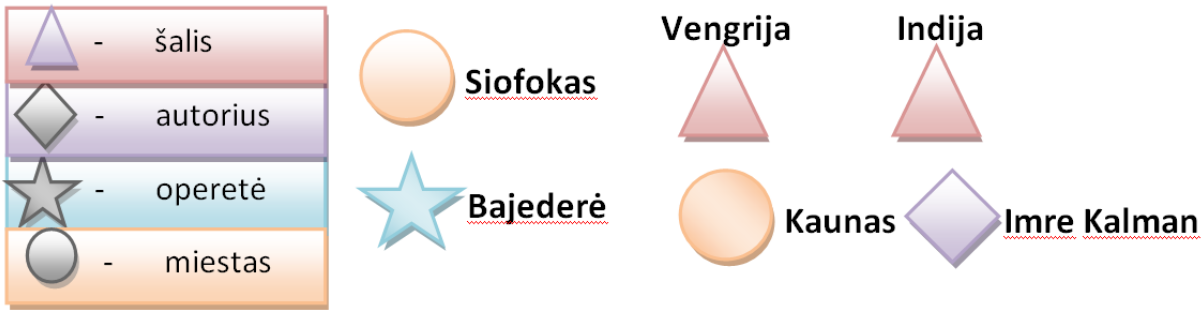


3 paveikslas. ***Temininio žemėlapio*** **grafinis vaizdavimas**

Tad nuo pradžių: *Tema* – plačiąja savo prasme gali būti bet kas. Kitaip sakant šiuo žodžiu *Teminių žemėlapių* standartas įvardija subjektą.

Kitaip tariant, terminas *Topic* (tema) veda į objektą arba mazgą. Kaip matyti 3 paveiksle, šitaip atrodo TŽ modelis su nurodytomis sąsajomis (operos ontologija pateikiama kūrėjų tinklapyje). Pasirinkus konkrečią Temą, ir paspaudus ant jos, galima pamatyti kaip ši išsišakoja ir kur šakos veda, savo ruožtu šios vėl rodo į kitas Temas.

6.1 Temų tipai (angl. *Topic Types*)



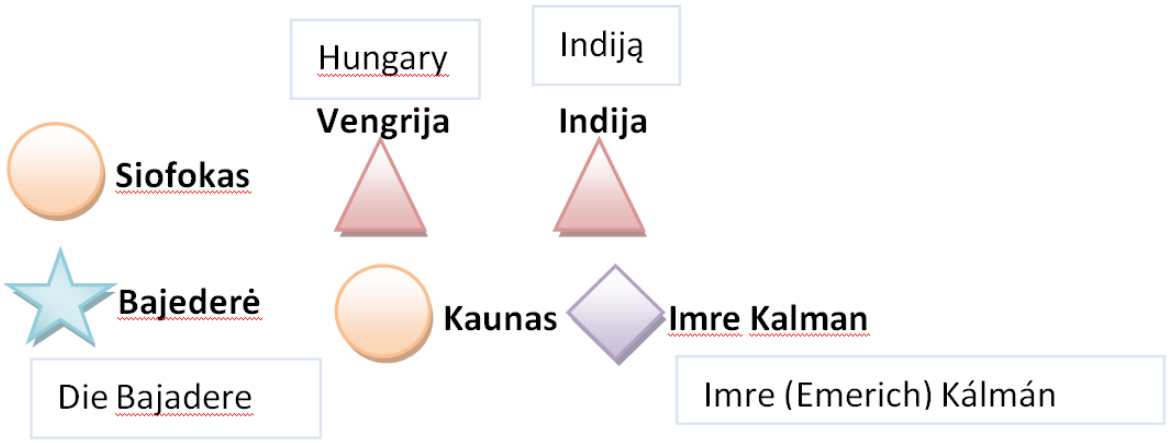
4 paveikslas. ***Temos* ir joms priskirti *Tipai* (kairėje paveikslo pusėje)**

Topic (nuo šiol bus vadinama – *Temos*) gali būti **skirstomos** priklausomai nuo to, kokiam **skiriamajam požymiui** (šį vadinsim *Tipu*) priklauso (žr. į 4 paveikslą). Kiekvienai tokiai *Temai* gali būti, arba nebūti priskirtas konkretus *Tipas* (pvz.: *V. Kernagis* (kaip *Tema*, - aktorius, dainininkas (kaip *Tipas*).

Tarkime tezaure *Tema* nurodys terminą, jo reikšmę ir sritį, kuriai jis priklauso, o ontologija, kaip toliau rašoma darbe, apima ir ryšius su kitomis temomis.

Temai galima ir nepriskirti *Tipo*, tačiau jį priskrus ir turint didelę ontologiją, iš jos bus galima išgauti daug daugiau informacijos (tarkim, mus domina kažkoks muzikos stilius (kaip *Tipas*), o *Temos* – to stiliaus atlikėjai, tad pasirinkę informaciją filtruoti pagal *Tipą* (Pop), rasime visus šio stiliaus atlikėjus).

6.2 Temų vardai (angl. *Topic Names*)

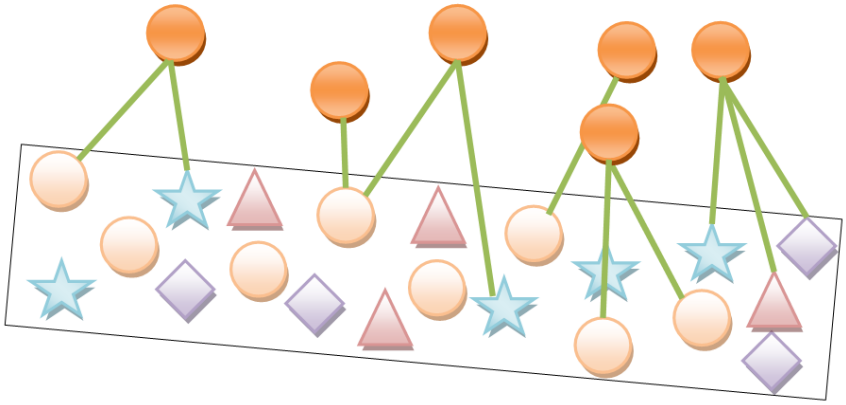


5 paveikslas. **Temos gali turėti keletą vardų variantų**

Kiekviena *Tema* turi savo vardą (nėra griežtų taisyklių, koks jis turi būti, tai ir slapyvardžiai, gyvūnų vardai, prisijungimo vardai ir t.t.). Yra galimybė nusirodyti įvairius vardo variantus (priklausomai nuo laikotarpio, konteksto (kuriam skiriantis, vardas gali reikšti visai ką kita) tarimo, lietuvių kalbos atvejų ir linksniavimo) (žr. į 5 paveikslą). Kiekvienas *Temos* vardas savo ruožtu gali būti priskirtas kitam kontekstui.

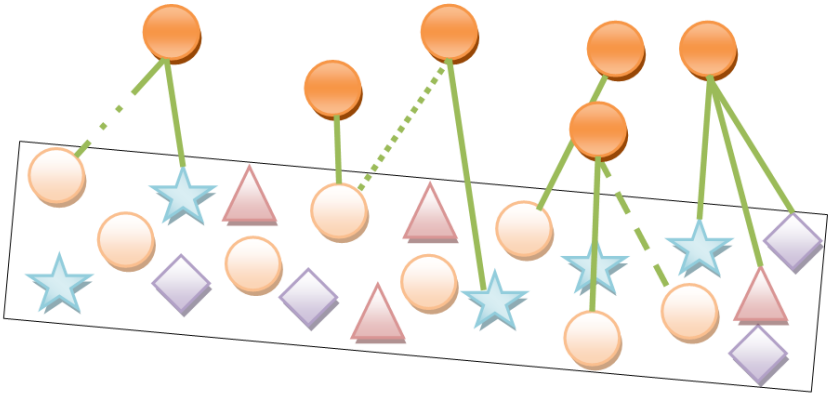
6.3 Konkretūs atvejai (angl. *Occurrences*)

Tema gali būti susieta su vienu ar daugiau informacijos resursų, kurie, kaip spėjama susiję su ta *Tema* (tai gali būti paveiksliukas, citata, garso įrašas, ar bet kokia kita besisiejanti informacijos forma). Ši *Teminių žemėlapių* savybė gali būti naudinga sužymint tekste tam tiktus žodžius (priklausomai nuo ontologijos tipo, jei tai *Natūralios kalbos apdorojimo* terminų ontologija, tekste aptikus vieną iš šių terminų, jis pažymimas, o užvedus pelę ant jo, gaunama papildoma informacija, paveiksliukas ir panašiai) (žr. į 6 paveikslą (oranžiniais apskritimais žymima *Tema*, iš jus išeina sąsajos į susijusią informaciją).



6 paveikslas. **Iš *Temų* eina nuorodos į konkretų su jomis besisiejantį turinį**

6.4 Konkrečių atvejų vaidmenys (angl. *Occurrence roles*)



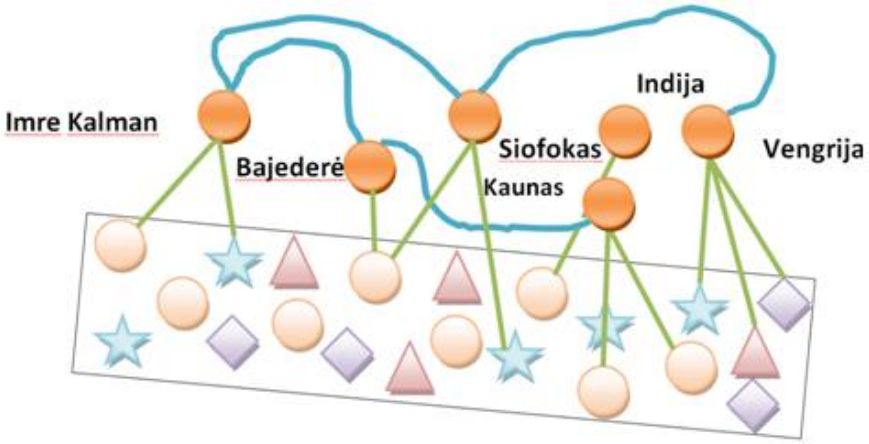
7 paveikslas. **Iš *Temų* eina nuorodos į konkretų su jomis besisiejantį turinį**

Būtina apsibrėžti ir *Konkrečių atvejų* vaidmenis. Jei iš *Temos* išeina kelios nuorodos į su ja besisiejančią informaciją, kiekvieną šį ryšį galima apsibrėžt nurodant jo *Tipą* (ar tiesiog klasę kitaip tariant) (pvz.: Jei iš *Temos Bajaderė* išeina nuorodos į **nuotraukas** ir **vaizdo įrašą** iš operetės), tai šioms šakoms priskiriama „vertė“ – pvz.: *nuotraukos* ir *vaizdas* (žr. į 7 paveikslą – nuorodos iš *Temų* pažymėtos skirtingomis rodyklėmis).

Kiekvienos toks *Temos* „išsišakojimas“ su priskirtu *Tipu*, įgalina ontologijos pagalba atrinkti ar ieškoti vis įvairesnę informaciją.

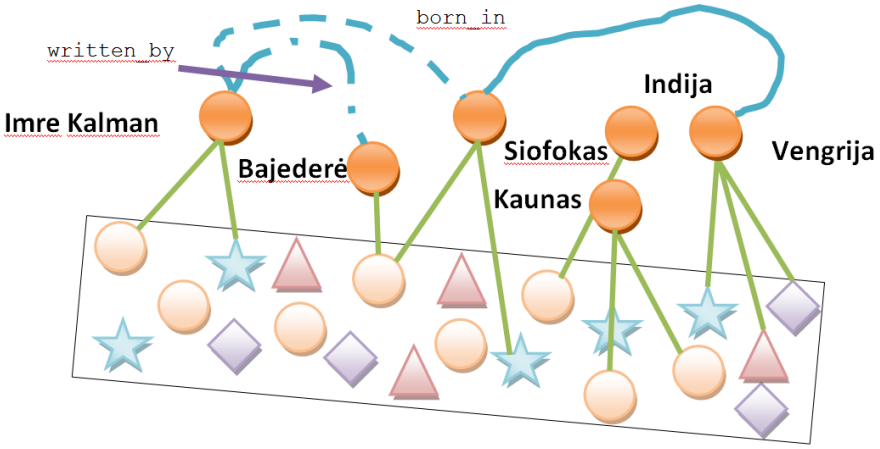
6.5 Asociacijos (ryšiai tarp temų) (angl. *Association types*)

Asociacijos terminu apibūdinamas ryšys tarp dviejų ar daugiau *Temų* (angl. *topic*). Iš ankščiau aptartų sudedamųjų dalių, jau galima sudaryti esminius informacijos organizavimo principus. Galime organizuoti turimą informaciją pagal subjektus ir kurti paprastus indeksus. Kaip matyti iš 8 paveikslo, nusirodžius ryšius tarp atskirų *Temų*, galima apsirašyti kad ir tokius sakinius: Inre Kalman Bajaderės autorius, Bajaderė pastatyta Kaune ir panašius (pavyzdžiai darbo autoriaus).



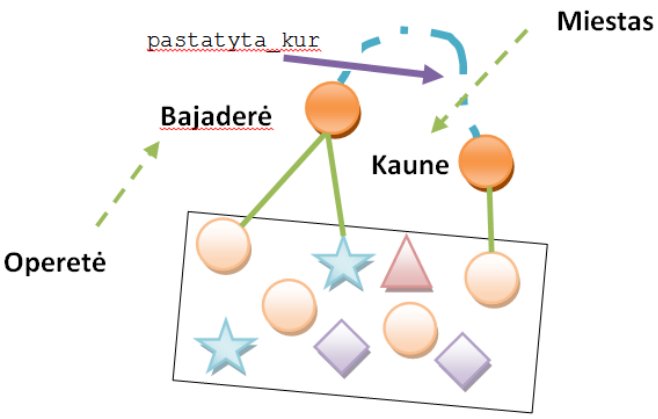
8 paveikslas. **Temos susiejamos ryšiais**

Žmogui toks susiejimas aiškus, tačiau mašina pati ryšiams kategorijų nepriskirs. Žmogus matydamas nubrėžtą ryšį tarp Bajaderė – Kaunas, ir turėdamas foninių žinių, žinos, jog operetė pastatyta Kaune, mašinai to padaryti nepavys, tad ryšiams reikia priskirt kategorijas, pvz.: angl. *written_by, born_in* (parašytas_ko, gimęs_kur) (žr. į 9 paveikslą).



9 paveikslas. **Kiekvienam ryšiui priskiriama klasė**

Šitaip jau galima informaciją filtruoti pagal pvz.: mus domina kur gimė autorius ir gausime atsakymą: vietovę.



10 paveikslas. **Kiekvienai asociacijai priskirtai Temai suteikiamas vaidmuo**

Taigi, *Temas* galima jungti *asociacijomis*, o kiekviena *Tema*, kuri sujungta tąja asociacija, turi savo vaidmenį toje asociacijoje (žr. į 10 paveikslą). Pvz.: Bajaderė pastatyta Kaune, tad Bajaderė ir Kaune yra *Temos*, jas sieja *asociacija*, kurią pavadinsime – *pastatyta_kur*, šioje *pastatyta_kur* asociacijoje *Temos* Bajaderė ir *Kaune* įgyja vaidmenis (angl. *Association roles*) Bajaderė: operetė, o Kaune: miestas.

Šitaip nusakoma ne tik asociacija, tačiau ir asociacijos dalyviams priskiriamos konkrečios reikmės, taip padedančios dar tiksliau nusakyti sakinio esmę ir užrašyti jį kaip ontologiją. Kitaip tariant, *temų* vaidmenys asociacijoje skirtos tam, kad nusakyti semantiniams ryšiams, kodėl viena ar kita tema priskiriama konkrečiai asociacijai.

6.6 Nagrinėjimo sferos (angl. *Scopes*)

Kalbant apie nagrinėjimo sferas, remtasi Teminių žemėlapių konteksto sprendimas aprašymu (2010). Nors jau žinome, kaip apsibrėžt *Temas*, jų vardus, nusakyti ryšius ir nuorodas į besisiejančią informaciją, tačiau viso to negana. Norint teisingai suprasti informaciją ir ją interpretuoti, susiduriame su konteksto problema.

Tam, kad apriboti kontekstui, kiekviena *Tema* gali turėti *Nagrinėjimo sferas*, kurios skirstomos pagal *Tematikas* (angl. *Theme*).

Nagrinėjimo sferų paskirtis – leisti *Teminio žemėlapio* autoriui aprašyti ribas, kurios būtų teisingos *konkretiems atvejams*. Pvz.: „Vilnius“ tinkamas tik tam tikruose kontekstuose (nuo 1941 ir po 1991 metų), bet ne kituose.

Pagrindinė Nagrinėjimo sferų mintis, kaip apriboti *Temą* nuo ryšių su netinkamais kontekstais (pvz.: mus domina pelė tik kompiuterijos srities kontekste, bet ne gyvūnų pasaulio).

6.7 Aspektai (angl. *Facets*)

11 paveikslas. **TŽ konkretūs atvejai surūšiuojami pagal pasirinktą tipą (pvz.: kalba), pritaikius filtravimą, galima palikti tik vienas ar kitas nuorodas (pagal pasirinktą aspektą)**

Kartais svarbu įvertinti į kokią informaciją *Temos* veda, neatsižvelgiant į patį *Teminį žemėlapį*, mums svarbi pati informacija, ne tai, kur ji veda ar su kuo siejasi. Šiuos šaltinius galima įvairiai rūšiuoti, pvz.: pagal kalbą, pagal tai, ar jie patalpinti Internete, ar vartotojo kompiuteryje, bylos plėtinį ir pan.). Tarkim, nuo apačios į viršų: vaizdo įrašas yra **lietuvių kalba** (bus *Aspektas*) – *konkretaus atvejo tipas* (vaizdo įrašas), šis veda į *Temą* (Bajaderė) (žr. į 11 paveikslą).

7. ONTOLOGIJOS KŪRIMAS

Ankstesniuose skyriuose aptarus pačias kompiuterines ontologijas, jų skirtumus nuo filosofinės, Teminių žemėlapių sudarymo principus, šioje dalyje aprašomas NLP terminų ontologijos kūrimas, remiantis TŽ specifikacija, visi jo etapai ir galimi problemų sprendimai.

Kaip jau darbe anksčiau minėta, ontologija bandoma nusakyti visus ryšius, asociacijas tarp sąvokų, sudėlioti sąvokinį tinklą. Priklausomai nuo ontologijos srities, šią gali sudaryti nuo kelių, iki kelių šimtų ar net daugiau sąvokų rinkinys. Atrodo, jog nusakyti sąvokų ryšius iš to, kas supa žmogų – yra nesudėtinga. Kaip pavyzdį galime paimti darbinio stalo ir ant jo esančių daiktų ontologiją. Gali pasirodyti nieko nėra lengviau, tačiau: ant stalo guli popieriaus lapai, rašikliai, pieštukai, įvairių formų ir spalvų spalvoti lapukai, susegimo priemonės, trintukai. Tad trintukas siejasi su lapu, bet tik tuo lapu, ant kurio parašyta pieštuku, lapas prirašytas šratinuku jau kaip išimtis lapo – trintuko asociacijai. Tuomet vėlgi kokius vaidmenis asociacijoje priskirti lapui ir trintukui, kaip pavadinti asociaciją? *Tas kuris* trina ir *Kuriame trinama*, o asociacija Lapo trynimas? Aiškėja, kad vien jau taip paprastai iš šalies atrodanti asociacija, sukelia abejonių, o jau ką kalbėti apie viso to, kas ant stalo susiejimą.

Išdėsčius mintis apie ontologijos sąvokų pasirinkimą ir priskyrimą konkrečiai kategorijai, verta paminėti ir ontologijų grafinį vizualizavimą. Jei ontologija didžiulė ir ją bandoma taikyti paieškai naudojant vizualizavimą (kuomet iš *Temos* per ryšius einama link kitų *Temų*, ar įvairiai ją filtruojant, labai lengva pasiklysti. Tad būtini tam tikri „kelio ženklai“, kurie rodytu, kurioje ontologijos vietoje esi, reikalinga gera navigacija (galimybė grįžti atgal (angl. *undo*)), nes vien jau „pakeliavus“ po pavyzdinę operos ontologiją (žr. į 3 paveikslą), greitai pasimetama kur esama ir nėra galimybės grįžti.

Dar viena problema, jei taip galima tai vadinti – kalba. Plačiai paplitusios kalbos kone visuomet pirmauja. Lietuvių kalbai pritaikytų veikiančių ontologijų ar jomis pagristų tarnybų nepavyko rasti, tačiau anglakalbiai tokių turi, tai nėra NLP terminų ontologijos, tačiau jas verta apžvelgti, dar prieš pradedant TŽ ontologijos kūrimo aprašymą. Pirmiausia **Cyc** (2010), ontologijos principu veikianti duomenų bazė (pvz.: paspaudus ant *Bird* – pateikiamas paukščių sąrašas, jų ryšiai su kitais) – ši duomenų bazė gali būt naudinga nusakyt ryšiam tarp konkrečios informacijos.

Būtent ryšių nustatymas leidžia generuoti tinklapius, kurių turinį sudaro iš įvairių šaltinių realiu laiku „surenkama“ informacija. Puikus pavyzdys kaip būtų galima kurti kažką panašaus ir lietuvių kalbai.

Galimas pritaikymo pavyzdys: esate skaitmenine lingvistika besidomintis žmogus ir jus domina visos programos, kurios skirtos konkordansų kūrimui, bei informacija apie jas. Jeigu jos visos sužymėtos ontologijoje ir priskirtos konkordavimo programų kategorijai – gausite sugeneruotą visų tokių programų sąrašą ir visai nebūtina, kad viskas būtų viename tinklalapyje.

Antras projektas: **Calais** (2010), jis pritaikytas tik anglų ir prancūzų kalboms (**Cyc** – anglų). **Calais** pateiktame tekste sužymi miestus (pvz.: šalis – miestas, valstija), užimamas pareigybės, konkrečius asmenis, citatas, organizacijas, tam tikrus ryšius, priskirią tekstą konkrečiai kategorijai (įkėlus tekstą iš tinklapio rašančio apie NLP (http://nlpers.blogspot.com/ žiūrėta 2010-04-30), *mašininio vertimo* terminas priskirtas technologijoms, o tai jau puikus būdas automatiškai filtruot turinį, kad ir interneto portaluose nauji straipsniai gali būti automatiškai patalpinami tam tikroje portalo dalyje, pvz.: Lingvistika. Išbandžius abu ontologijų principu veikiančius įrankius, susidarė įspūdis, jog tokiais projektais, kaip **Calais** ir **Cyc**, pirmiausiai siekiama automatizuoti informacijos skirstymą ir suteikti jai reikšmes jau be žmogaus įsikišimo. Tad galima spėti, jog nyks atskiri tinklapiai ir juos „maitins“ specialiai tam ontologijų pagrindu sukurtos ir nuolat pildomos duomenų bazės (o iš šių duomenų automatiškai bus generuojami duomenys ir skirstomos į kategorijas).

Taigi, akivaizdu, kad anglų kalbai jau yra veikiančių ontologijų pagrindu sudarytų sistemų, na o šio darbo tikslas yra aptarti, kaip kažką tokio panašaus galima pritaikyti ir lietuvių kalbai.

Viena yra kurti ontologiją iš mus nuolat supančių dalykų, kita iš konkrečios terminų šakos, šio darbo atveju – Natūralios kalbos apdorojimo terminų. Kaip jau darbe minėta, terminu siekiama kuo konkrečiau apibrėžti sąvoką, veiksmą ar kažkokį dalyką. Čia viskas daug tiksliau ir glausčiau. Tad vėl mintis, jei jau viskas taip tikslu, tuomet terminų ontologija sukurti dar paprasčiau. Tokios pirmos mintys ir aišku vis dar neatsakyta į darbo pradžioje išsikeltą hipotezę: ar surinkus terminus iš skirtingų šaltinių su jau nusakyta jų hierarchija, ontologiją sukurti yra lengviau, nei pirma kurti pačią hierarchiją, o tik po to parinkti jai terminus. Taigi, nuo to ir prasideda NLP terminų ontologijos kūrimo aprašymas. Pats aprašymas skiriamas į du bandymus, taigi, toliau apie juos.

8. PIRMAS BANDYMAS

Kaip užsiminta praeitame skyriuje, ontologijos kūrimą šiame darbe skiriamas į du bandymus. Pirmasis – mėginimas struktūruoti terminus, jų struktūra vaizduoti medžiu ir sukurti tipus, kurie terminus skirstytų į grupes pagal reikšmę.

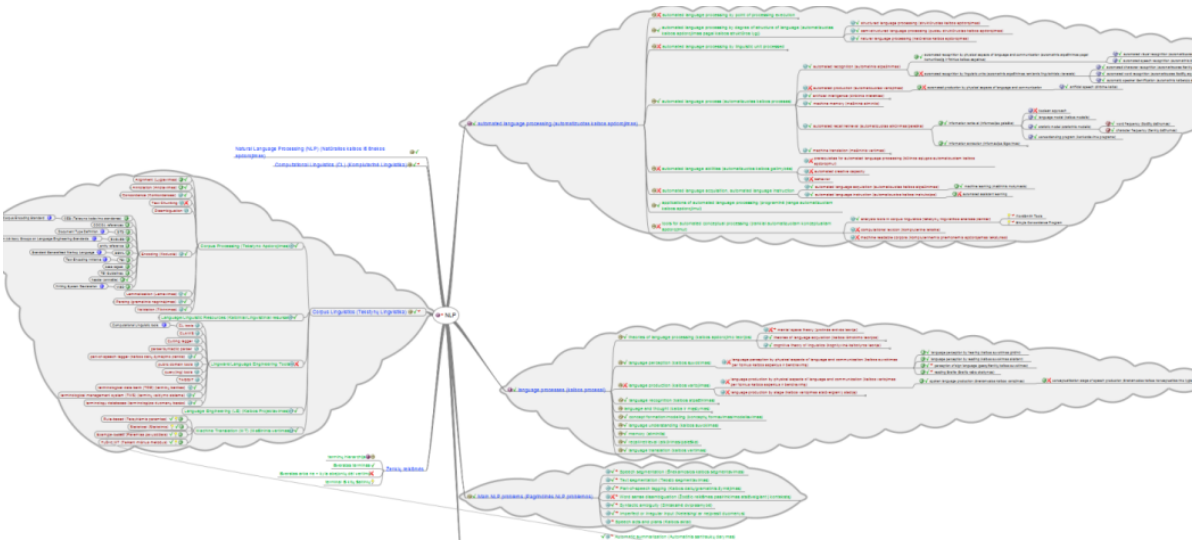
8.1 1 etapas: Terminų surinkimas, išvertimas, minčių medis

Prieš kuriant NLP terminų ontologiją, pačios pirmos idėjos buvo tokios: kadangi ontologija sudarinėjama iš terminų ir žinant, jog terminų žodynuose šie paprastai struktūruojami – nusakoma konkreti kategorija, tuomet eina rūšinis terminas ir iš jo seka gimininiai, tad struktūra jau aprašyta, todėl pirmiausiai nuspręsta pasiieškoti tokių terminų šaltinių, kuriuose terminai nėra vien išvardinti, o jau su priskirta struktūra.

Siekiant įsitikinti darbo pradžioje išsikeltos hipotezės pagrįstumu, iš 3 skirtingų šaltinių – 1) Sisteminis tekstynų lingvistikos žodynas (angl. *Systematic Dictionary of Corpus Linguistics* (2010)), 2) Lingvistikos tezauras (angl. *Linguistics Thesaurus* (2010), 3) Natūralios kalbos apdorojimas (angl. *Natural language proccesing* (2010)), surinkti 5 NLP **terminų sąrašai** išlaikant šaltiniuose pateiktą terminų hierarchiją. Surinkti manant, jog išankstinis struktūros buvimas išspręs jos kūrimo nuo pat pradžių problemą. Terminų sąrašai įvardinti kaip: NLP1 (Tekstynų lingvistika, 90 terminų (angl. *Corpus Linguistics*)), kaip terminai pridėti ir dviejų šioje srityje naudojamų programų pavadinimai, NLP2 (Automatizuotas kalbos apdorojimas, 59 terminai (angl. *Automated language processing*)), NLP3, 41 terminas (Kalbos procesai (angl. *Language processes*)), NLP4 (Pagrindinės NLP problemos, 7 terminai (angl. *Main NLP problems*)) ir NLP5 (Pagrindiniai NLP uždaviniai, 20 terminų (angl. *Major tasks in NLP*)). Viso 217 terminų. Būtina užsiminti, jog šie terminai anglų kalbą.

Atlikus pirmą žingsnį ir priskyrus **terminų sąrašams** pavadinimus, kitas žingsnis – naudojantis *minčių medžių* sudarymo programinę įrangą FreeMind (2010), kuri skirta vizualiai aprašyti ir struktūruoti žmogaus mąstymo procesus, struktūriškai pavaizduotos terminų sąsajos ir išsidėstymas atkartojant išsidėstymą, kuris sutiktas šaltiniuose. *Minčių medžių* atvaizdavimo būdas pasirinktas, nes taip geriau suvokiamas terminų pasiskirstymas, galima matyti visą medį iš karto ir prireikus šalinti terminus ar papildyti šakas naujais. Ne ką

mažiau svarbios programos galimybės įvairiai ženklinti terminų šakas (piktogramomis, formatavimu), bei brėžti nuorodas tarp bet kurių šakų.

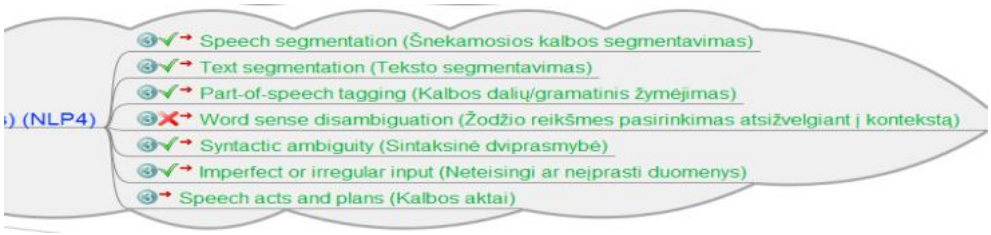


12 paveikslas. **NLP terminų hierarchijos vizualizavimas naudojant minčių medžių programinę įrangą**

Minčių medžiui sudaryti panaudojus tas pačias terminų išdėstymo struktūras kaip ir šaltiniuose ir kaip motininę šaką pasirinkus NLP terminą, trečia kas atlikta: kaip jau minėta, terminai angliški, todėl prie angliško termino skliausteliuose įrašytas ir lietuviškas vertimas (žr. į 12 paveikslą).

Trumpai apie patį vertimą. Sudarytame minčių medyje, žalios varnelės ženklu sužymėti terminai (žr. į 13 paveikslą), kuriuos verčiant nekyla abejonių, raudonu x ženklų – neišversti arba, tie, kurių vertimas į lietuvių kalbą kelia abejonių, šitai daryta vėlgi tam, kad būtų supaprastintas tolesnis terminų skirstymas, raudonomis rodyklėmis priskirti *konkretūs atvejai* – nuorodos į tinklalapius, aprašančius terminą (plačiau apie tai 8.3 poskyryje, 56 psl.).

Sudarius medį, iš karto matyti aiški atskirtis tarp terminų iš skirtingų šaltinių, beje, ne visuose iš 5 sąrašų terminai suskirstyti aiškiai, pvz.: angl. *language perception by physical aspects of language and communication* (kalbos suvokimas per fizinius kalbos aspektus ir bendravimą) itin plati sąvoka, kas visiškai netinka ontologijos sudarinėjimui.



13 paveikslas. **Terminų žymėjimas piktogramomis, palengvinant visos struktūros suvokimą**

Sudarius *minčių medį*, išvertus terminus iš karto pastebėtos kelios problemos:

1. Keliant terminus iš skirtingų šaltinių, dalis jų kartojasi, bei yra priskirti skirtingai struktūrai (pvz.: *mašininis vertimas* NLP5 priskiriamas prie pagrindinių NLP problemų, o jau NLP1 – kaip tekstynų lingvistikos šaka, *informacijos išgavimas* – NLP2 kaip automatizuoto atkūrimo/paieškos šaka, o NLP5 kaip NLP uždavinys.
2. Kiekvienoje iš 5 NLP terminų kategorijų ne tik kartojasi terminai, tačiau trūksta bendros sistemos, ar bent jau junglumo taškų, kurie būtų kaip lizdai prijungti šalutiniams terminams, todėl būtinas dabartinės NLP terminų struktūros pertvarkymas.
3. Kai kurie terminai (pvz.: *kalbos suvokimas per fizinius kalbos aspektus ir bendravimą* (angl. *language perception by physical aspects of language and communication*)), ne tik kad neatitinka vienos iš esminių terminų sudarymo taisyklių – trumpumo, tačiau yra ir pernelyg platūs savo reikšme, ko būtina vengti sudarinėjant ontologiją.
4. Akivaizdu, kad kol kas net žmogui sunku suprasti struktūrą.

Pirmo etapo išvada: kuo daugiau terminų šaltinių, tuo įvairesnė šiuose pateikiama jų išdėstymo struktūra, tai priklauso nuo jos sudarytojų, tad pirmiausia sudarinėjant terminų ontologiją reikia vengti daug ir iš įvairių šaltinių terminų pridėjimo. Ima atrodyti, jog darbe išsikeltą hipotezę galima paneigti. Prieš rašant darbą atrodė, kad įsikėlus terminus jau su iš anksto nustatyta struktūra, ontologiją sudaryti bus kur kas lengviau, tačiau yra priešingai ir toji išankstinė struktūra tik trukdo. Paaiškėjo, jog nueita klaidingu keliu ir pirma būtina apsirašyti struktūrą, o tik tada rinkti terminus, o ne turint terminus, ieškoti juose struktūros.

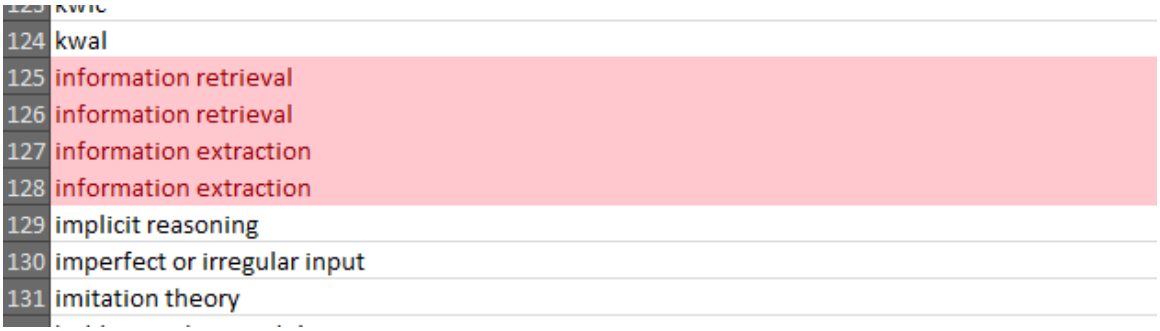
8.2 2 etapas: Terminų šalinimas, struktūros keitimas, *temų tipų* priskyrimas

Aprašius 1 etapą ir aptarus problemomis su kuriomis susidurta, kitas žingsnis – išspręsti jas, tam šiame etape atlikti šie darbai:

1. Dalis terminų pašalinta, nustatyti šalinimo kriterijai.
2. Pakeista NLP terminų išdėstymo struktūra.

Terminų šalinimas. Kadangi praeitame poskyryje nuspręsta, jog kai kurie iš turimų NLP terminų yra netinkami ontologijos sudarymui dėl savo reikšmės platumo, neatitinka trumpumo kriterijaus ar sutinkami ne vieną kartą, nuspręsta dalį jų pašalinti. Pirminė prielaida, jog tai padės sudaryti aiškesnę ir tikslesnę struktūrą.

Visi 5 sąrašai su 217 terminų (be vertimų į lietuvių kalbą), perkelti į vieną Microsoft Excel programos stulpelį. Kiekvienas terminas užrašytas mažosiomis raidėmis ir pašalinti šalimais esantys termino trumpiniai (jei yra), pvz.: angl. *terminological data bank **(TDB)** (terminų bankas)*, taip išvengiant skirtingų termino užrašymo variantų norint aptikti identiškus. Terminai išrūšiuoti pagal abėcėlę ir automatiškai paryškinti besikartojantys (žr. į 14 paveikslą). Rasti 7: *part-of-speech tagging, optical character recognition, natural language processing, machine translation, information retrieval, information extraction, artificial intelligence.*



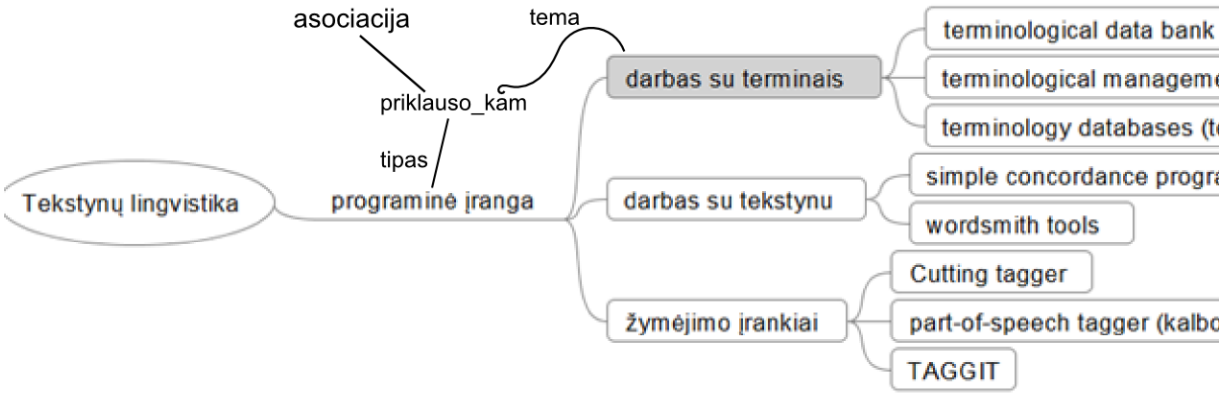
14 paveikslas. **Automatiškai atrinkti pasikartojantys terminai**

Pašalinus besikartojančius terminus Excel programa, bei išmetus juos iš jau minėto NLP terminų *minčių medžio*, atsisakyta ir tų, kurie pasirodė per platūs savo reikme, besikartojantys su kitais savo reikšmės panašumu. Palikti tik patys bendriausi 69 terminai, nes ankstesnės klaidos parodė, jog reikia laikytis principo: **nuo mažai link daugiau.**

Norint terminų struktūrą paversti aiškesne, kitas žingsnis, kurį reikia atlikti po to, kai dalis terminų pašalinta – sugalvoti *temų tipus*, atsisakant ankstesnės terminų išsidėstymo struktūros ir sukurti naują, kiek įmanoma labiau tipizuojant turimus terminus.

Kadangi struktūra kuriama iš naujo, pradėta nuo Tekstynų lingvistikos (NLP1) terminų struktūros pertvarkymo. Idėja tokia: panašius savo reikšme terminus (*temas*), galima grupuoti priskiriant juos geriausiai apibūdinančiam *tipui*. Šitaip skaidant informaciją į kiek galima daugiau dalių. Toks skaidymas palengvina paiešką ir pačią navigaciją ontologijoje (jei šie duomenys vizualizuojami grafiškai), nes nebereikia skaityti ilgų terminų sąrašų ieškant reikalingo. Žinant ko ieškai, kiekvienas teisingas *temos tipo* pasirinkimas priartina prie paieškos rezultato. Priskyrus tipą, tema su tipu susiejama asociacija, bei priskiriami vaidmenys toje asociacijoje. Pradėta nuo *programinė įranga* (kaip *tipas*), šiam *tipui* priskirtos temos: *darbas su terminais*, *darbas su tekstynu*, *žymėjimo įrankiai,* kiekvienam iš šių priskiriami geriausiai pagal prasmę tinkantys terminai.

Šiame nedideliame ontologijos sudarymo pavyzdyje (žr. į 15 paveikslą) nusistačius *tipus* ir priskyrus jiems *temas*, priskirtos asociacijos, kaip *asociacija* pasirinkta: *priklauso_kam*, o vaidmenys asociacijose kol kas nustatyti kaip *tipas* ir *tema*. Pagal tai, kompiuterinė sistema jau gali skirti, kurie terminai susieti vieni su kitais ir kuris terminas, kokį vaidmenį atlieka asociacijoje.

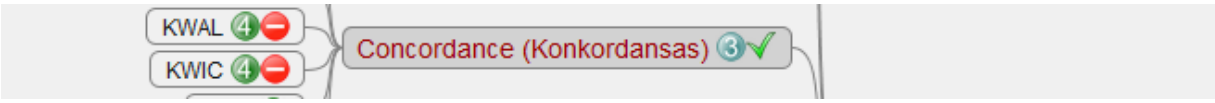


15 paveikslas. **Ontologijos hierarchijos keitimas**

Antro etapo išvados: terminų šalinimas tik parodė, kad terminai gali atlikti ir „šiukšlių“ vaidmenį, kuomet siekiant optimalios ontologijos struktūros, šie niekur netinka. Pirminė idėja, jog turint praktiškai visus NLP terminus ir jų struktūrą bus lengviausia sukurti ontologiją – visiškai nepasitvirtino, tad nuo šiol laikomasi principo: **nuo mažiau link**

daugiau. Taigi, toliau darbe aprašinėjama kokie **temų tipai** pasirinkti ir, kokie terminai jungti prie jų.

8.3 3 etapas: Sudarinėjimo aprašymas



16 paveikslas. **MM1 perkėlimas į MM2**

3 etape aptariamas struktūros keitimas ir įvairių *tipų* priskyrimas NLP terminams. Pirmiausia atsidarius pirmąjį jau prieš tai minėtą minčių medį (nuo šiol jis vadinamas MM1), kuriame buvo suvesti terminai, o vėliau dalis jų pašalinta, sukurtas kitas minčių medis (MM2).

Iš MM1 terminai kopijuoti į MM2. MM2 medyje panaudotas terminas MM1 pažymėtas raudona apvalia piktograma su baltu brūkšniu viduryje (žr. į 16 paveikslą) šitaip pasilengvinant darbą ir iš karto matant, kas jau panaudota, kas ne.

Kaip ir MM1, MM2 pagrindiniu terminu pasirinktas NLP. Iš jo išeina 4 šakos (čia terminų pavadinimai vardijami lietuvių klaba): 1) Tekstynų lingvistika, 2) Pagrindiniai NLP uždaviniai, 3) Kalbos atpažinimas, 4) Pagrindinės NLP problemos. Kaip pagrindinės šios šakos pasirinktos todėl, kad jos bendriausios ir iš jų gali sekti tolimesni ryšiai. Taigi, prieš keliant terminus iš MM1 į MM2, apsirašytas naujos MM2 medžio struktūros modelis. Sudarymo hierarchija žymima taip:

1), **2)** ir t.t – aukščiausia vieta hierarchijoje

I), **II)** ir t.t – pirmos klasės šalutinė atšaka (antra klasė)

i), **ii)** ir taip toliau – antros klasės šalutinė atšaka (trečia klasė)

a), **b)** ir t.t – trečios klasės šalutinė atšaka (ketvirta klasė)

aa), **bb)** ir t.t – ketvirtos klasės šalutinė atšaka (penkta klasė)

tolesnės klasės – pridedant papildomą raidę pvz.: **aaa)** (šešta klasė) ir t.t

Pirmoji šaka ***TEKSTYNŲ LINGVISTIKA*** perskirta į:

I) ***Programinė įranga*** suskirstyta į:

i) **Darbas su terminais** (šakojasi į: a) terminų bankas, b) terminų valdymo sistema, c) terminologijos duomenų bazės).

ii) **Darbas su tekstynu** (šakojasi į: a) simple concordance program, b) wordsmitht tools (programų pavadinimai neverčiami į lietuvių kalbą)).

iii) **Žymėjimo įrankiai** (šakojasi į: a) Cutting tagger, b) kalbos dalių žymėjimo įrankis, c) TAGGIT.

iv) **Vertimo programinė įranga** (šakojasi į: a) vertimo atminties programinė įranga.

II) **Kalbos vertimas**. Suskirstyta į:

i) **Mašininis vertimas** (šis į: a) **mašininio vertimo metodai** (šie šakojasi į: aa) Taisyklėmis paremtas, bb) Statistinis, cc) Paremtas pavyzdžiais, dd) Taikant mišrius metodus.

ii) **Vertimas kompiuterio pagalba**

III) **Tekstynai**. Suskirstyta į:

i) **Tekstyno Apdorojimas** (šis į: a) Apdorojimo būdai, o šie į: aa) Lygiavimas, bb) Anotavimas (šis į aaa) anotavimo tipai, šie į: aaaa) anaforų anotavimas, bbbb) kalbos dalių anotavimas, cccc) fonetinė transkripcija, dddd) semantinis anotavimas, eeee) prozodijos anotavimas, ffff) diskurso anotavimas), cc) Konkordansas (šis į: aaa) konkardanso eilučių išdėstymo būdai, o šie savo ruožtu į: aaaa) KWAL, bbbb) KWIC, dd) Lemavimas, ee) Gramatinis nagrinėjimas, ff) Tikrinimas (šis į: aaa) Klaidų taisymas, bbb) pagrindinis klaidų tekste taisymo įrankis, ccc) klaidų taisymo įrankis, ddd) stiliaus klaidų taisymas), dd) Konkordansą sudaro (šis į: aaa) Kolokatas, bbb) kolokacija), ccc) Konkordanso eilutė).

ii) **Tekstyno tipas** (šis į; a) pagal kalbų skaičių (skirstoma į: aa) vienakalbis tekstynas, bb) daugiakalbis tekstynas), b) subalansuotas tekstynas, c) palyginamasis tekstynas, d) lygiagretus tekstynas, e) oportunistinis tekstynas, e) pagal sužymėjimą (skiriamas į aa) nesužymėtas tekstynas, bb) anotuotas tekstynas.

Antroji šaka ***PAGRINDINIAI NLP UŽDAVINIAI*** perskirta į:

I) **Užsienio kalba** (šis į: i) Pagalbinės priemonės užsienio kalba užrašytam tekstui skaityti, ii) Pagalbinės priemonės užsienio kalba užrašytam tekstui skaityti, iii) Mašininis vertimas).

II) **Šnekamoji kalba** (šis į: i) Teksto vertimas balsu/Sintezavimas, ii) Šnekamosios kalbos atpažinimas).

III) **Paieškos sistemos** (šis į: i) Informacijos paieška, ii) Informacijos išgavimas, iii) paieška naudojant natūralią kalbą).

IV) **Klaidų taisymas**.

V) **Tikrinių vardų atpažinimas**.

VI) **Natūralios kalbos generavimas**.

VII) **Kalbos konkretinimas** (šis į: i) Automatinis santraukų darymas, ii) Teksto supaprastinimas neprarandant esminės informacijos).

Trečioji šaka ***KALBOS ATPAŽINIMAS*** perskirta į:

Du terminai: *vizualus ženklų atpažinimas* ir *vizualus žodžių atpažinimas* sujungti į vieną naują terminą – *optinis atpažinimas* kaip temos tipą.

I) **Optinis atpažinimas** (išskirtas į dvi šakas, vėlgi kiek pakeitus terminus, nes taip jie geriau nusako savo paskirtį: a) optinis simbolių/ženklų atpažinimas, b) optinis žodžių atpažinimas.

II) **Šnekamosios kalbos atpažinimas.**

Ketvirtoji šaka ***PAGRINDINĖS NLP PROBLEMOS*** skiriama į:

I) **Segmentavimas** (skiriamas į: i) Šnekamosios kalbos segmentavimas, ii) Teksto segmentavimas. II) Kalbos dalių/gramatinis žymėjimas, III) Žodžio reikšmes pasirinkimas atsižvelgiant į kontekstą, IV) Sintaksinė dviprasmybė, V) Neteisingi ar neįprasti duomenys, IV) Kalbos aktai.

Atlikus štai tokį struktūros pakeitimą, išryškėjo šie dalykai:

1. Pritaikius tokius **temų tipus** (13), kaip: *Programinė įranga, Darbas su terminais, Darbas su tekstynu, Žymėjimo įrankiai, Vertimo programos, Mašininio vertimo metodai, Paieškos sistemos, Tekstyno tipas, Segmentavimas, Užsienio kalba, Šnekamoji kalba,*

Kalbos konkretinimas, *Optinis atpažinimas*, pati terminų išsidėstymo struktūra tapo aiškesne.

2. Paaiškėjo ir tai, jog pirmame šio bandymo etape atliktas besikartojančių terminų pašalinimas nėra visiškai tikslingas. Terminas nebūtinai turi priklausyti tik vienai kategorijai. Pvz.: *kalbos atpažinimas* gali būti priskirtas prie kalbos atpažinimo, ir būti laikomas kaip vienu iš NLP uždavinių.

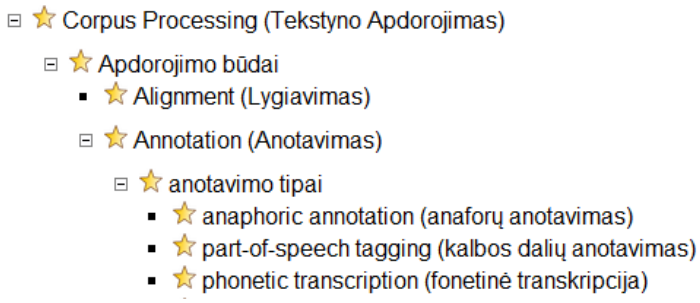
Pirmo bandymo išvados. Kuriant ontologiją ir remiantis iš kelių šaltinių surinktais terminais, nevertėtų pasikliauti iš šaltinių atsinešta struktūra (tai taikytina nebent turint itin išsamų ir patikimą terminų šaltinį (NLP atveju būtų šios srities terminų žodynas). Pirmiausia reikia apsibrėžti struktūrą, o tik tada pradėti rinkti terminus, darbe buvo pasirinktas priešingas kelias. Visgi, sumažinus terminų kiekį ir priskyrus jiems kiek galima daugiau *tipų* ir taip sukūrus naują ir aiškesnę struktūrą, pasiektas geresnis rezultatas. Siekiant dar geresnio rezultato, atliktas antras bandymas.

9. ANTRAS BANDYMAS

Kaip parodė pirmo bandymo rezultatai, pašalinus didžiąją dalį terminų, o likusiems priskyrus tipus, pati terminų struktūra pasidarė aiškesnė, tačiau liko nepatvirtinta arba paneigta antroji hipotezė, jog terminus aprašius geriausiai šiuos nusakančiais žodžiais ir juos sugrupavus, jungtys tarp NLP terminų taps aiškesnėmis nei pirmame bandyme.

9.1 1 etapas: Temų tipų konkretinimas ir priskyrimas NLP terminams

Šiame etape nuspręsta terminus dar labiau kategorizuoti, priskiriant kiekvienam iš jų vieną ar kelis geriausiai jį apibūdinančius **meta aprašymus** (toks apibūdinimas pasirinktas dėl to, jog meta parašymai atlieka specialią paskirtį – apie kalbą kalba kita kalba, šiuos atveju patys terminai aprašomi kitais žodžiais). Pats aprašymas vyko taip: kaip jau minėta anksčiau (žiūrėti į 1 bandymą), iš susidarytos naujos struktūros, sukurtas MM2 minčių medis, šis, panaudojus programos FreeMind galimybę eksportuotas į HTML failą (žiūrėti į 17 paveikslą), atspausdintas ir prie kiekvieno termino prirašyta po vieną ar daugiau jį geriausiai apibūdinančių žodžių (terminai kartu su meta aprašymais 1 priede).



17 paveikslas. **MM2 eksportuotas į HTML**

Štai, kaip atrodo patys terminai (čia pateikiami tik lietuviški terminų vertimai, be angliškų atitikmenų, programų pavadinimai neverčiami į lietuvių kalbą) su jau priskirtais naujais **meta aprašymais** (tai netaikyta pirmame bandyme NLP terminams jau priskirtiems tipams, imti tik terminai, o pagal *Teminių žemėlapių* specifikaciją – temos, atsisakyta priskirti *meta aprašymus* ir pagrindinėms MM2 medžio šakoms priskirtiems terminams (*Tekstynų lingvistika* ir pan.), nes šie jau ir taip yra tipai):

Terminų bankas – *terminai*, **Terminų valdymo sistema** – *terminai*, **Terminologijos duomenų bazės** – *terminai*, **Simple concordance program** – *programa*, *tekstynas*,

konkordansas, **Wordsmith tools** – *programa*, *tekstynas*, *konkordansas*, **Cutting tagger** – *žymėjimas*, *programa*, **Kalbos dalių žymėjimo įrankis** – *programa*, *žymėjimas*, *kalbos dalys*, **TAGGIT** – *žymėjimas*, *programa*, **Vertimo atminties programinė įranga** – *vertimas*, *programa*, *vidutinis automatizavimo lygis*, **Kalbos vertimas** – *vertimas*, **Mašininis vertimas** – *vertimas*, *aukštas automatizavimo lygis*, **Taisyklėmis paremtas** – *vertimas*, *taisyklės*, **Statistinis** – *vertimas*, *statistika*, **Paremtas pavyzdžiais** – *vertimas*, *pavyzdžiai*, **Taikant mišrius metodus** – *vertimas*, *mišrūs metodai*, **Vertimas kompiuterio pagalba** – *vertimas*, *vidutinis automatizavimo lygis*, **Tekstynai** – *tekstas*, *tekstynas*, **Tekstyno Apdorojimas** – *tekstas*, *tekstynas*, **Lygiavimas** – *tekstas*, **Anotavimas** – *tekstas*, *žymėjimas*, *anotavimas*, **Anaforų anotavimas** – *žymėjimas*, *anafora*, *anotavimas*, **Kalbos dalių anotavimas** – *žymėjimas*, *kalbos dalys*, *anotavimas*, **Fonetinė transkripcija** – *fonetika*, *transkribavimas*, **Semantinis anotavimas** – *žymėjimas*, *semantika*, *anotavimas*, **Prozodijos anotavimas** – *prozodija*, *žymėjimas*, *anotavimas*, **Diskurso anotavimas** – *diskursas*, *žymėjimas*, *anotavimas*, **Konkordansas** – *tekstas*, *eilutė*, *reikšminis žodis*, *kontekstas*, **KWAL** – *reikšminis žodis*, *eilutė*, **KWIC** – *reikšminis žodis*, *kontekstas*, **Kolokatas** – *žodžių junginys*, **Kolokacija** – *žodžių junginys*, **Konkordanso eilutė** – *tekstas*, *konkordansas*, *reikšminis žodis*, *eilutė*, *kontekstas*, **Lemavimas** – *morfologija*, **Gramatinis nagrinėjimas** – *gramatika*, **Tikrinimas** – *klaida*, **Klaidų taisymas** – *klaida*, *taisymas*, **Pagrindinis klaidų tekste taisymo įrankis** – *klaida*, *programa*, *taisymas*, **Klaidų taisymo įrankis** – *klaida*, *programa*, *taisymas*, **Stiliaus taisymo įrankis** – *stilius*, *klaida*, *programa*, **Vienakalbis tekstynas** – *tekstynas*, *tekstas*, **Daugiakalbis tekstynas** – *tekstynas*, *tekstas*, **Subalansuotas tekstynas** – *tekstynas*, **Palyginamasis tekstynas** – *vertimas*, *tekstynas*, **Lygiagretus tekstynas** – *vertimas*, *tekstynas*, **Oportunistinis tekstynas** – *tekstynas*, **Nesužymėtas tekstynas** – *tekstynas*, **Anotuotas tekstynas** – *žymėjimas*, *tekstynas*, *anotavimas*, **Pagalbinės priemonės užsienio kalba užrašytam tekstui skaityti** – *tekstas*, *skaitymas*, *pagalbinė priemonė*, *programa*, **Pagalbinės priemonės rašyti užsienio kalba** – *pagalbinė priemonė*, *programa*, *rašymas*, **Šnekamoji kalba** – *šneka*, **Teksto vertimas balsu/Sintezavimas** – *tekstas į balsą*, **Šnekamosios kalbos atpažinimas** – *šneka*, *atpažinimas*, **Informacijos paieška** – *paieška*, **Informacijos išgavimas** – *išgavimas*, **Paieška naudojant natūralią kalbą** – *paieška*, *natūrali kalba*, **Tikrinių vardų atpažinimas** – *tikriniai vardai*, *atpažinimas*, **Natūralios kalbos generavimas** – *generavimas*, *natūrali kalba*, **Automatinis santraukų darymas** – *aukštas automatizavimo lygis*, *santrauka*, *konkretinimas*, **Teksto supaprastinimas neprarandant esminės informacijos** – *tekstas*, *supaprastinimas*, **Kalbos atpažinimas** – *atpažinimas*, **Šnekamosios kalbos segmentavimas** – *šneka*, *segmentavimas*, **Teksto segmentavimas** – *tekstas*, *segmentavimas*, **Kalbos dalių/gramatinis žymėjimas** – *kalbos*

dalys, gramatika, žymėjimas, anotavimas, **Žodžio reikšmes pasirinkimas atsižvelgiant į kontekstą** – *kontekstas, reikšmė,* **Sintaksinė dviprasmybė** – *sintaksė, dviprasmybė,* **Neteisingi ar neįprasti duomenys** – *klaida,* **Kalbos aktai** – *verbalinė informacija,* **Optinis simbolių/ženklų atpažinimas** – *atpažinimas, optinis, ženklai, simboliai,* **Optinis žodžių atpažinimas** – *atpažinimas, optinis, žodžiai.*

Kiekvienam ir 69 terminų priskyrus *meta aprašymus* (nuo šiol MA), šios surūšiavus ir atmetus pasikartojančius, liko 51.

Turint terminų MA ir pvz.: ontologijos paieškoje suvedus žodį **tekstas**, pateikti rezultatai bus NLP terminai, kurie asociacijomis sujungti su šiuo *meta aprašymu* ir lygiai taip pat bus galima matyti kitas termino asociacijas, vertimą į kitą kalbą ar nuorodą į konkretų šaltinį su termino apibrėžtimi.

9.2 2 etapas. 7 meta aprašymų grupės

Kadangi turimiems NLP terminams priskirti *meta aprašymai*, kitas žingsnis – juos suskirstyti į stambesnes grupes. Būtina pabrėžti, kad skirstant *meta aprašymus* į grupes, visiškai nežiūrėta, kokius NLP terminus jie aprašo, darant prielaidą, jog pirma suskirsčius MA, o tada prie šių *meta aprašų* gražinus jiems priklausančius terminus, išryškės logiška struktūra.

Pirmame bandyme parodyta, kaip kuriama ontologiją, nusakant pagrindines šakas su terminais, prie šių jungiant šalutines ir t.t. Naujas NLP terminų ontologijos modelis sudaromas pradedant ne nuo pačių terminų, o nuo jų *meta aprašymų*, šių kategorizavimo ir galiausiai, prie naujai sukurtų kategorijų, prijungiant pačius terminus. MA pagalba, termino reikšmė išskaidoma ir kartu plačiau nusakoma. Panašūs principai taikomi interneto tinklaraščių naujiems įrašams sužymėti, kuomet temai priskiriamos žymės geriausiai nusakančios įrašą, vėliau jos panaudojamos paieškai.

Pirmiausia MA suskirstyti į stambesnes grupes, o tose grupėse ieškota asociacijų tarp *meta aprašymų*. Priskirtos asociacijos yra binarės, tai yra – apima du *meta aprašymus*, kiekvienam MA asociacijoje priskirta po vaidmenį (remiantis *Teminių žemėlapių* sudarymo principu). Asociacijos priskirtos tokiu principu – privalo nusakyti asociacijos prasmę ir būti artimos natūraliai kalbai, neformalizuojant jų. Vaidmenys asociacijose – geriausiai įvardijantys *meta aprašymo* vaidmenį toje asociacijoje. Taigi, dabar apie visa tai pažingsniui.

Pirmiausia *meta aprašymai* atsispausdinti, iškirpti ir dėliojant vienus šalia kitų, ieškota tinkamiausios kombinacijos. Iškarto išryškėjo keletas dalykų:

Bandant šios *meta aprašymus* suskirstyti į grupes, akivaizdus kiekvieno žmogaus individualus MA priskyrimas kategorijai. Nuomonės tarp skirtingų žmonių skiriasi, todėl arba būtina iš anksto nustatyta ir gerai aptarta struktūra, arba bendras sudarinėtojų sutarimas. Ankščiau darbe minėta idėja (žr. į 6 psl.), jog galimas modelis, kuomet ontologijų principu aprašomos visos įmanomos egzistuojančios sąvokos, kelia abejonių, nes tas modelis, kiekvienam individui individualus.

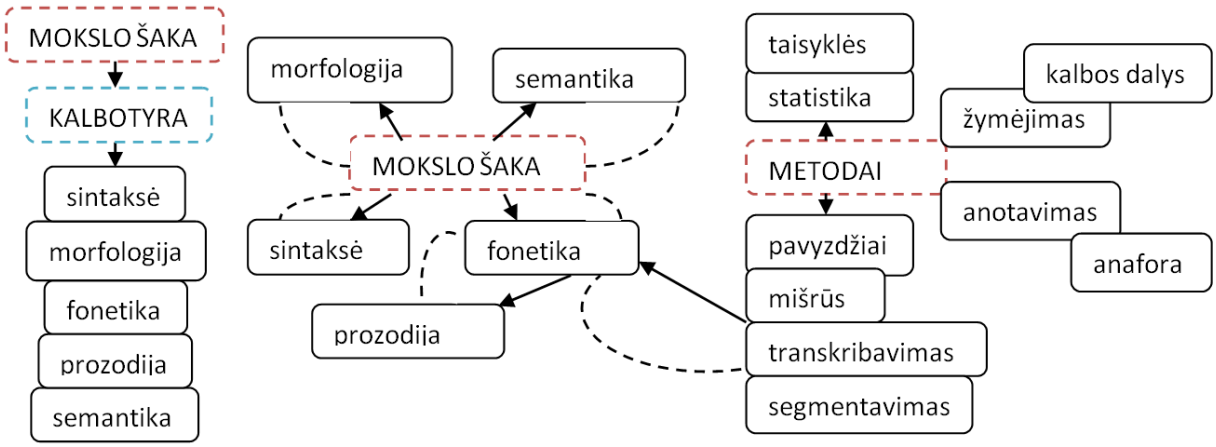
Dėliojant MA, pagal savo prasmės ir logines sąsajas, išryškėjo 7 pagrindinės grupės:

1) Mokslo šaka
2) Metodai
3) Automatizavimo lygis
4) Informacijos išgavimas
5) Informacijos pateikimas
6) Kalbos vienetas
7) Taisymas

Išskyrus šias grupes ir darant prielaidą, jog jos geriausia apibūdina turimus *meta aprašymus*, kitas veiksmas – aprašyti, kokie MA priskirti šioms grupėms, kodėl ir kokios galimos asociacijos tarp grupės pavadinimo ir *meta aprašymų*, bei pačių MA santykiai toje grupėje.

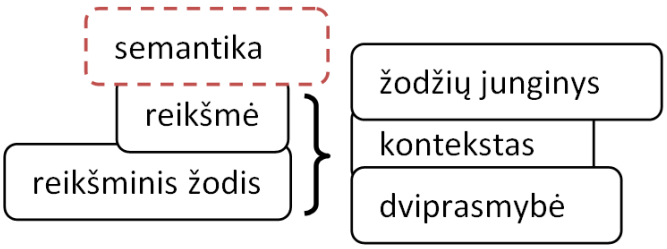
Pirma grupė: ***MOKSLO ŠAKA***, šiai grupei priskirti šie *meta aprašymai* – *morfologija, fonetika, sintaksė, semantika*. Priskirti, remiantis tuo, kad yra kalbotyros mokslo šakos, šioje vietoje nederėtų atmesti ir sudarinėtojo foninių žinių, kuomet sudarinėjant modelį ir daugiau ar mažiau išmanant sritį, MA skirstymas yra šališkas. Sudarinėtojui be šių foninių žinių, skirstymas gali atrodyti visai kitoks.

Kaip matyti iš 18 paveikslo (44 psl.) (punktyrinėmis juodomis linijomis žymimos asociacijos, raudonomis – MA grupės), iš MA *fonetika* išeina ryšys į *transkribavimas*, tad *transkribavimas* priklauso ir *fonetika*, ir tuo pačiu siejasi su kita MA grupe – *METODAI*. Žmogus gali suprasti šį ryšį, tačiau norint ryšius pritaikyti kompiuterinėms sistemoms, reikia apsibrėžti asociacijas ir vaidmenis asociacijose, kas toliau ir atliekama:



18 paveikslas. **Ryšiai tarp MA grupės *MOKSLO ŠAKA***

1. ryšio **mokslo šaka** – **sintaksė** asociacija: **tyrinėja_ką**, vaidmenys asociacijoje **mokslo šaka – mokslas**, **sintaksė** – **ryšius tarp žodžių ir sakinių**.
2. ryšio **mokslo šaka** – **morfologija** asociacija: **tyrinėja_ką**, vaidmenys asociacijoje: **mokslo šaka** – **mokslas**, **morfologija** – **žodžių semantinę gramatinę morfeminę sudėtį**. Vaidmenis norėtųsi apsibrėžti kiek galima trumpesnius, tačiau šiuo atveju norėta trumpai nusakyti MA esmę, kad dėliojant meta aprašymus vienus prie kitų, būtų aiškiau kas su kuo siejasi.
3. ryšio **mokslo šaka** – **semantika** asociacija: **tyrinėja_ką**, vaidmenys asociacijoje: **mokslo šaka** – **mokslas**, **semantika** – **reikšmę prasmę**.
4. ryšio **mokslo šaka** – **fonetika** asociacija: **tyrinėja_ką**, vaidmenys asociacijoje: **mokslo šaka** – **mokslas**, **fonetika** – **garsus**. Galima pridėti ir platesnį apibrėžimą: **šnekamąją kalbą** ir pan.
5. MA **prozodija** pasirinkta priskirti prie **fonetika** dėl jos sąsajų su garsais, pasirinkta asociacija: **tyrinėja_ką**, o vaidmenis asociacijoje: **fonetika** – **kalbotyros šaka**, **prozodija** – **kirtis intonacija**. Iš čia kilo klausimas, kaip teisingai pasirinkti asociaciją, bei ar turimas MA grupavimas teisingas, nes *prozodiją* galima laikyti atskira kalbotyros šaka, todėl struktūra šiek tiek patikslinta, MA grupė *MOKSLO ŠAKA* įgauna smulkesnį skirstymo vienetą: *KALBOTYRA* (žr. į 18 paveikslo kairę pusę).



19 paveikslas. ***Semantika* ir iš šio MA sekantys meta aprašymai**

Aptarus kaip MA patekusius į grupę ***MOKSLO ŠAKA***, išryškėjo tai, kad keletą *meta aprašymų* pagal jų reikšmę, geriausia priskirti prie – *semantika* (kaip Kalbotyros šaka, o pačią semantiką laikant lingvistine). Iš *semantika* seka šie meta aprašymai (žr. į 19 paveikslą): *reikšmė, reikšminis žodis, žodžių junginys, kontekstas, dviprasmybė*. Šie MA priskirti prie *semantika,* nes *reikšmė* su *reikšminis žodis* nusako pačią semantikos esmę – reikšmių tyrimą, iš šių dviejų meta aprašymų, savo ruožtu seka – *žodžių junginys*, savaime savyje talpinantis vienokią ar kitokią reikšmę, *kontekstas* – būtent jis dažnai patikslina ar apsprendžia reikšmę, o *dviprasmybė* – savaime talpina kelias reikšmes. Paaiškinus, kodėl šie MA priskirti prie *semantika*, būtina apsibrėžti ir asociacijas:

1. ryšio **semantika** – **reikšmė** asociacija: **siejasi_su**, vaidmenys asociacijoje: **semantika** – **semantika**, **reikšmė** – **reikšme**.
2. ryšio **semantika** – **reikšminis žodis** asociacija: **siejasi_su**, vaidmenys asociacijoje: **semantika – semantika**, **reikšminis žodis – reikšminiu žodžiu**.
3. ryšio **reikšmė** – **kontekstas** asociacija: **išryškėja_iš**, vaidmenys asociacijoje: **reikšmė** – **reikšmė**, **kontekstas** – **konteksto**. Lygiai tą pačią asociaciją ir vaidmenį asociacijoje galima taikyti ir junginiui: **reikšminis žodis** – **kontekstas**.
4. ryšio **reikšmė** – **žodžių junginys** asociacija: **išryškėja_iš**, vaidmenys asociacijoje: **reikšmė** – **reikšmė**, **žodžių junginys – žodžių junginio**. Vėlgi tas pats su junginiu **reikšminis žodis** – **žodžių junginys**.
5. ryšio **reikšmė** – **dviprasmybė** asociacija **reiškia_mažiau_už**, vaidmenys asociacijoje: **reikšmė** – **reikšmė**, **dviprasmybė** – **dviprasmybę**. Junginiui **reikšminis žodis** – **dviprasmybė**, galioja ta pati asociacija ir vaidmenys joje. Norint įvardinti šią asociaciją, reikia apsibrėžti, jog **reikšmė** – tai, kas kažką reiškia, *dviprasmybė* – kas reiškia daugiau, nei reikšmė ir savyje talpina dvi ar daugiau reikšmių. Remiantis reikšmių kiekiu, asociacija įvardinta: **reiškia_mažiau_už**, svarstytina, tačiau galima asociacija.

Antra grupė: ***METODAI***

Šiai grupei priskirti šie MA (žiūrėti į 18 paveikslą, 44 psl.): *taisyklės*, *statistika*, *pavyzdžiai*, *mišrūs, transkribavimas*, *segmentavimas*, *anotavimas*, *žymėjimas* – šie pridėti prie *METODŲ,* nes siejasi ryšių*:* **metodas paremtas kuo**. Aišku tai, jog sudarinėtojas neišmanydamas terminų srities, gali ir nepriskirti šių MA prie *METODŲ*, tačiau šiuo atveju pasirinktas toks skirstymas. Aptarus, kodėl šie meta aprašymai priskirti antrai grupei, toliau išvardintos asociacijos ir vaidmenys jose:

1. ryšio **metodai** – **taisyklės** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **taisyklės** – **taisyklėmis**.
2. ryšio **metodai** – **statistika** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **statistika** – **statistika**.
3. ryšio **metodai** – **pavyzdžiai** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **pavyzdžiai** – **pavyzdžiais**.
4. ryšio **metodai** – **mišrūs** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **mišrūs** – **mišriais metodais**.
5. ryšio **metodai** – **transkribavimas** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **transkribavimas** – **transkribavimu**.
6. ryšio **metodai – segmentavimas** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **segmentavimas** – **segmentavimu**.
7. ryšio **metodai** – **žymėjimas** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **žymėjimas** – **žymėjimu**.
8. ryšio **metodai** – **anotavimas** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **anotavimas** – **anotavimu**.

Iš MA *žymėjimas* ir *anotavimas* pagal prasmę seka šie *meta aprašymai*:

9. ryšio **žymėjimas – kalbos dalys** asociacija: **žymi_ką**, vaidmenys asociacijoje: **žymėjimas – žymėjimas**, **kalbos dalys** – **kalbos dalis**. Tokios pat asociacijos ir vaidmenys jose tinka šiems junginiams: **žymėjimas** – **anafora**, **anotavimas** – **kalbos dalys**, **anotavimas** – **anafora** (čia žymėjimas keičiamas į anotavimą).

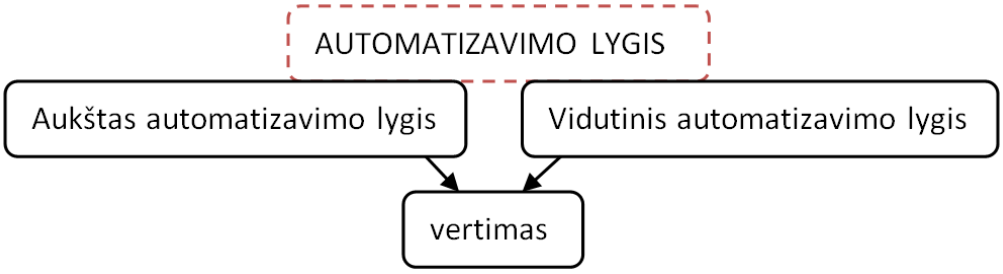
Aprašius dvi MA grupes, verta plačiau pakalbėti apie aukščiau jau užsimintą (žr. į 43 psl.) tarp MA asociacijų atsiradusį sudėtingesnį atvejį: nusakyti kylančią sąsają tarp **fonetika**, **transkribavimas** ir akivaizdaus ryšio tarp **metodai**. **Transkribavimas**, tai šnekamosios kalbos užrašymas tekstu, o **fonetika** – mokslas nagrinėjantis kalbą, jos garsus, tad bendrą šių

MA asociacija galima laikyti trinare ir perteikti ją tokiu loginį ryšį nusakančiu sakiniu: *fonetika yra mokslas, šis mokslas turi metodą, o tas metodas yra transkribavimas*. Asociacija šalima užrašyti ir taip: **mokslo_metodas_yra**, vaidmenys: fonetika – mokslas, metodas – metodas, transkribavimas – transkribavimas.

Trečia grupė: ***AUTOMATIZAVIMO LYGIS*** (žr. į 20 paveikslą)

Ši MA grupė nusako programinės įrangos ar metodo automatizavimo lygį, kitaip tariant, kuo šis lygis aukštesnis, tuo mažiau žmogaus įsikišimo reikia.

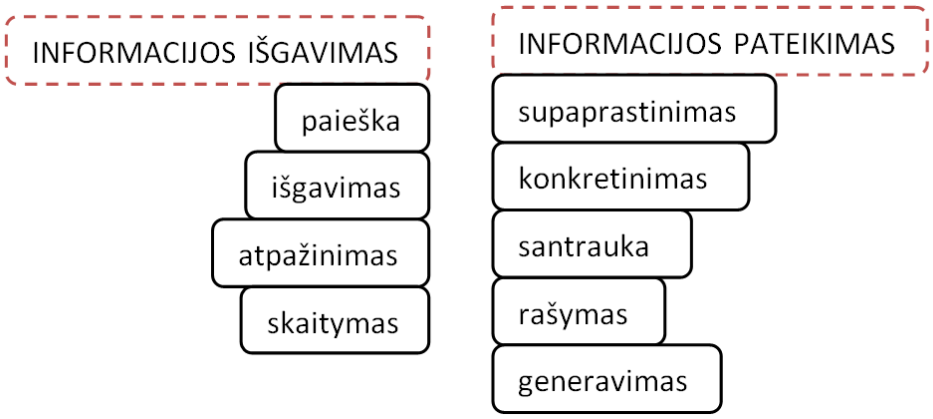
1. ryšio **automatizavimo lygis** – **aukštas automatizavimo lygis** asociacija: **automatizavimo_lygmuo**, vaidmenys asociacijoje: **automatizavimo lygis** – **automatizavimas**, **aukštas automatizavimo lygis** – **aukštas**.
2. ryšio **automatizavimo lygis** – **vidutinis automatizavimo lygis** asociacija: **automatizavimo_lygmuo**, vaidmenys asociacijoje: **automatizavimo lygis** – **automatizavimas**, **vidutinis automatizavimo lygis** – **vidutinis**.



20 paveikslas. MA grupė ***AUTOMATIZAVIMO LYGIS***

Iš visų turimų MA, logiškiausiai ***AUTOMATIZAVIMO LYGIS*** siejasi su **vertimas** (tiek su aukštu, tiek su vidutiniu). Akivaizdu ir tai, jog jei tarp įtrauktų terminų priskyrus apibrėžiantį sąvoką – **verčia žmogus**, nesinaudodamas jokia automatizavimo įranga, tuomet prisidėtų dar vienas MA – žemas automatizavimo lygis. Kalbant apie **vertimas** ir **automatizavimo lygis** asociaciją, ši priskirta tokia: **vertimo_automatizavimo_lygmuo**, vaidmenys asociacijoje: **aukštas automatizavimo lygis** – **aukštas lygmuo**, **vertimas** – **vertimo**, **vidutinis automatizavimo lygis** – **vidutinis lygmuo**, **vertimas** – **vertimo**.

Ketvirta grupė: ***INFORMACIJOS IŠGAVIMAS***



21 paveikslas. **MA grupės: *INFORMACIJOS IŠGAVIMAS* ir *INFORMACIJOS PATEIKIMAS***

INFORMACIJOS IŠGAVIMAS grupei priskirti šie MA: *paieška*, *išgavimas*, *atpažinimas*, *skaitymas* (žr. į 21 paveikslą), pagal bendrą savybę – ***kažkas* iš *kažko išgaunama***: *paieška* – skirta gauti dominančią informaciją; *išgavimas* – ją išgauti, *atpažinimas* – atpažinti iš neatpažinto; o *skaitymas* – priimti užrašytą informaciją ir šią apdoroti. Štai, kokios asociacijos priskirtos šiems MA grupėje:

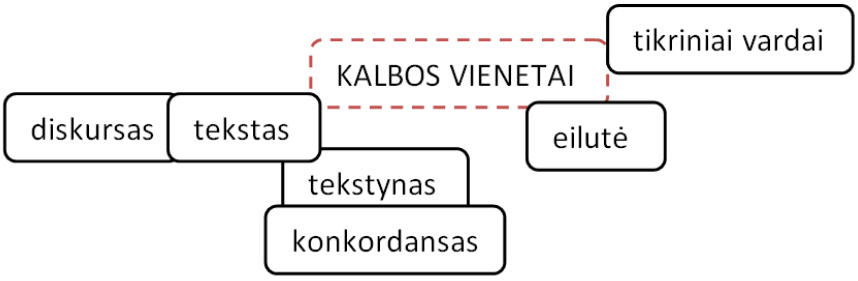
1. ryšio **informacijos išgavimas** – **paieška** asociacija: **išgauta_kaip**, vaidmenys asociacijoje: **informacijos išgavimas** – **informacija**, **paieška** – **ieškant**.
2. ryšio **informacijos išgavimas** – **išgavimas** asociacija: **išgauta_kaip**, vaidmenys asociacijoje: **informacijos išgavimas** – **informacija**, **išgavimas** – **išgaunant**.
3. ryšio **informacijos išgavimas** – **atpažinimas** asociacija: **išgauta_kaip**, vaidmenys asociacijoje: **informacijos išgavimas** – **informacija**, **atpažinimas** – **atpažįstant**.
4. ryšio **informacijos atpažinimas** – **skaitymas** asociacija: **išgauta_kaip**, vaidmenys asociacijoje: **informacijos atpažinimas** – **informacija**, **skaitymas** – **skaitant**.

Penkta grupė: ***INFORMACIJOS PATEIKIMAS*** (žr. į 21 paveikslą)

INFORMACIJOS PATEIKIMAS grupei priskirti šie MA: *supaprastinimas*, *konkretinimas*, *santrauka*, *rašymas*, *generavimas*, nes turi bendrą savybę – ***versti iš kažko į kažką***. *Supaprastinimas* – iš didelio į mažą; *konkretinimas* – iš daug į mažiau; *santrauka* – vėlgi, *daug-mažai*; *rašymas* – žodžiai, mintys, ar šneka į tekstą, *generavimas* – iš tam tikros informacijos gaunama kita. Nusakius pagrindinį visų šiai grupei priklausančių MA reikšmių panašumą, dabar pačios asociacijos ir vaidmenys jose:

1. ryšio **informacijos pateikimas** – **supaprastinimas** asociacija: **pateikiama_kaip**, vaidmenys asociacijoje: **informacijos pateikimas** – **informacija**, **supaprastinimas** – **supaprastinant**. Iš esmės, toks vaidmuo atkartoja patį *meta aprašymą*, tačiau platesnis apsirašymas padeda geriau suprasti, o ir vėliau sužymint asociacijas – nepasimesti šių gausoje.
2. ryšio **informacijos pateikimas** – **konkretinimas** asociacija: **pateikiama_kaip**, vaidmenys asociacijoje: **informacijos pateikimas** – **informacija**, **konkretinimas** – **sukonkretinant**.
3. ryšio **informacijos pateikimas** – **santrauka** asociacija: **pateikiama_kaip**, vaidmenys asociacijoje: **informacijos pateikimas** – **informacija**, **santrauka** – **sutraukiant**.
4. ryšio **informacijos pateikimas** – **rašymas** asociacija: **pateikiama_kaip**, vaidmenys asociacijoje: **informacijos pateikimas** – **informacija**, **konkretinimas** – **rašant**.
5. ryšio **informacijos pateikimas** – **generavimas** asociacija: **pateikiama_kaip**, vaidmenys asociacijoje: **informacijos pateikimas** – **informacija**, **generuojant** – **generuojant**. Pasivertus į sakinį, šios asociacijos su vaidmenimis skamba taip: informacija pateikiama kaip? – generuojant.

Šešta grupė: ***KALBOS VIENETAI*** (žr. į 22 paveikslą)



22 paveikslas. **MA grupė *KALBOS VIENETAI***

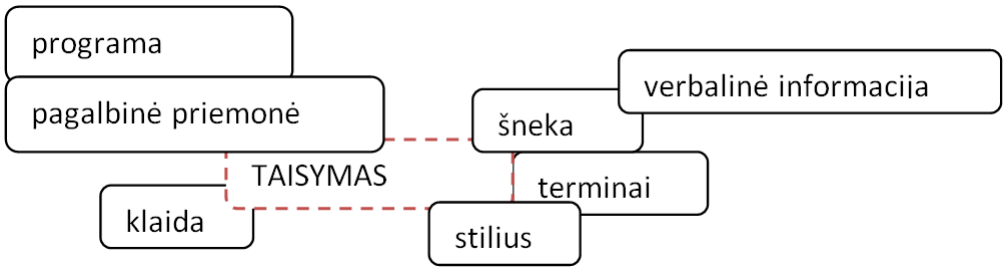
Grupės pavadinimas ***KALBOS VIENETAI*** pasirinktas todėl, kad šitaip bandyta paaiškinti trijų besisiejančių MA: *eilutė*, *tikriniai vardai* ir *tekstas* reikšmę. Šiuos MA savaime galima priskirti vienetams: *eilutė* – kaip teksto dalis, vienetas; *tikriniai vardai* – žodžiai iš daugumos; o *tekstas* – kaip vienas iš daugumos tekstų. Be to, juos visus galima priskirti kalbai, nes jie – jos sudedamosios dalys. Taigi, kokios asociacijos ir vaidmenys nustatyti šioje grupėje:

1. ryšio **kalbos vienetai** – **tikriniai vardai** asociacija: **priklauso_kam**, vaidmenys asociacijoje: **kalbos vienetai** – **kalbos vienetams**, **tikriniai vardai** – **tikriniai vardai**.

Tokios pat asociacijos ir vaidmenys jose, tinka ir junginiams: **kalbos vienetai** – **eilutė**, **kalbos vienetai** – **tekstas**. Šitokias asociacijas galima įvardinti ir hierarchiniu principu, kaip asociaciją nurodžius **priklauso_kam**, o vaidmenis įvardinant: **superklasė** (kaip pagrindinis MA) ir **subklasė** (kaip iš pagrindinio į šalutinį einantis *meta aprašymas*)

2. ryšio **tekstas** – **tekstynas** asociacija: **rinkinys_sudaro**, vaidmenys asociacijoje: **tekstas** – **tekstų**, **tekstynas** – **tekstyną**.
3. ryšio **tekstynas** – **konkordansas** asociacija: **pateikimas_kaip**, vaidmenys asociacijoje: **tekstynas** – **tekstų rinkinys**, **konkordansas** – **eilutėmis**. Asociacija sudaryta remiantis sąsaja tarp MA – **iš** *tekstyno* **į** *konkordansą*, o *konkordansas* – vienas iš tekstyno pateikimo būdų.

Septinta grupė: ***TAISYMAS*** (žr. į 23 paveikslą)



23 paveikslas. **MA grupė *TAISYMAS***

Septinta grupė pavadinta ***TAISYMAS***. pagal šių MA reikšmę ir ryšį tarp jų. Grupę sudaro šie *meta aprašymai*: *klaida* – priežastis, kuri reikalauja taisymo; *stilius*, *terminai* – tai, ką galima taisyti; *pagalbinė priemonė* – priemonė, kuri naudojama taisymui (keliant prielaidą, jog šis MA siejasi būtent su taisymu), *programa* – skirta taisymui; *šneka* – taip pat gali būti taisoma, o prie šio MA prijungta *verbalinė informacija* – kaip informacija gaunama iš šnekos. Apibendrinus grupės pavadinimo pasirinkimą, kitas žingsnis – išvardinti asociacijas ir vaidmenis jose:

1. ryšio **taisymas** – **klaida** asociacija: **įtakoja_kas**, vaidmenys asociacijoje: **taisymas** – **veiksmas**, **klaida** – **klaida**. Asociacija pasirinkta štai tokiu principu: taisymas reikalingas tik tada, kuomet randama klaida, taisymas laikytinas veiksmų, kurį įtakoja klaidos atsiradimas.

2. ryšio **taisymas** – **šneka** asociacija: **taisoma_kas**, vaidmenys asociacijoje: **taisymas** – **veiksmas**, **šneka** – **šneka**. Verčiant į paprastą sakinį – taisymas tai veiksmas, kuriuo taisoma kas – šneka.
3. ryšio **taisymas** – **terminai** asociacija: **taisoma_kas**, vaidmenys asociacijoje: **taisymas** – **veiksmas**, **terminai** – **terminai**. Vietoje to, kad kaip vaidmenį užrašinėti kas taisoma, kaip šiuo atveju – terminai, galima šį vaidmenį apsirašyti ir kaip – **tikslas** (angl. *target*), nes taisymas, yra kažkuria linkme nukreiptas tikslas.
4. ryšio **taisymas** – **stilius** asociacija: **taisoma_kas**, vaidmenys asociacijoje: **taisymas** – **veiksmas**, **stilius** – **stilius**.
5. ryšio **taisymas** – **programa** asociacija: **taisoma_kuo**, vaidmenys asociacijoje: **taisymas** – **veiksmas**, **programa** – **taisymo priemonė.**
6. ryšio **taisymas** – **pagalbinė priemonė** asociacija: **taisoma_kuo**, vaidmenys asociacijoje: **taisymas** – **veiksmas**, **pagalbinė priemonė** – **taisymo priemonė**.
7. ryšio **šneka** – **verbalinė informacija**: **priklauso_kam**, vaidmenys asociacijoje: **šneka** – **šneka**, **verbalinė informacija** – **šnekai**. Šiuo atveju vėlgi, šneką galima apsibrėžti kaip šaltinis (angl. *source*), o verbalinę informacija – iš šaltinio išeinančią informaciją (angl. *output*)

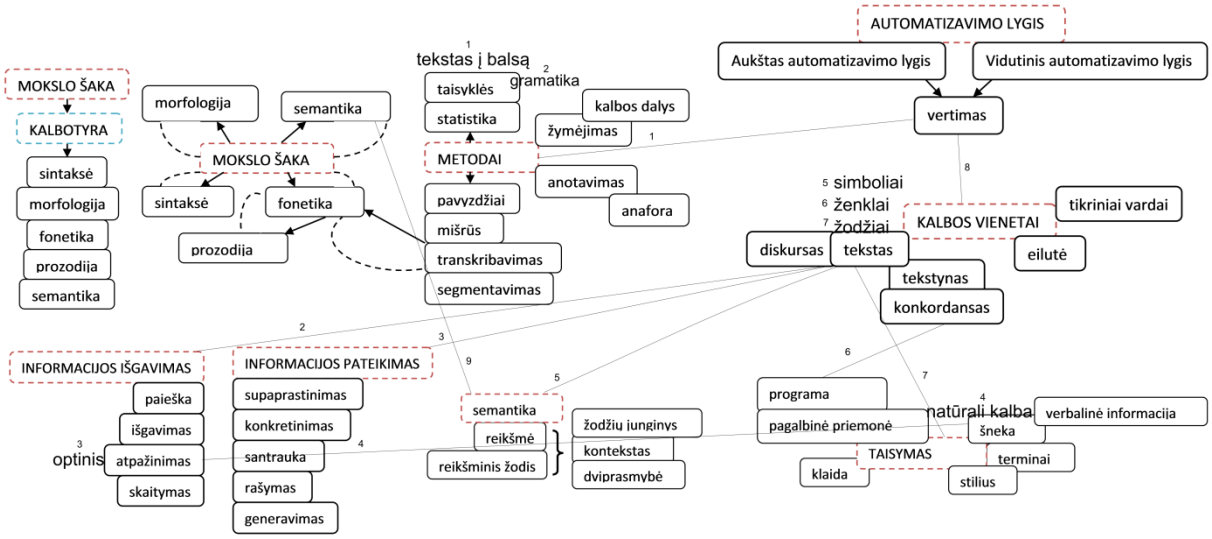
Aptarus visas 7 grupes išaiškėjo šie dalykai: dažnai sudėtinga priskirti teisingą asociaciją, bei vaidmenis jose, net ir suskaldžius *meta aprašymus* į kelias grupes. Nurodžius asociacija, kartais geriausia kaip vaidmenis nurodyti: šaltinį, jungties tašką ar išeities tašką, ką geriausiai perteikti angliškais terminais – angl. *source*, *output*, *input*, dėl jų reikšmės apibrėžtumo, ką bandant nusakyti lietuvišku terminu sunku, nes trūksta taiklaus posakio.

Vaidmenis asociacijose galima aprašyti ir hierarchiniu principu, naudojant **superklasės** ir **subklasės** apibrėžtis. **Superklasė** – visada bus pagrindinis terminas ar meta aprašymas, tas savo ruožtu gali būti ir kito termino **subklase**, taip nusakant terminų išsidėstymą.

Akivaizdu ir tai, jog kol grupės nesudėtos šalia viena kitos, sąryšiai grupių viduje atrodo tvarkingai, tačiau visa tai pasikeičia, kai šios MA grupės sujungiamos – atsiranda naujos loginės asociacijos.

Verta paminėti ir tai, jog suskirsčius MA į 7 grupes, liko *meta aprašymų*, nepriskirtų jokiai iš grupių, štai jie: *natūrali kalba*, *gramatika*, *tekstas į balsą*, *optinis, simboliai, ženklai, žodžiai*.

Taigi, kokie kiti žingsniai siekiant kaip galima geriau aprašyti NLP ontologijos struktūrą: sudėti grupes viena šalia kitos, nusakyti pagrindines naujai atsiradusias asociacijas tarp jų, bei turint galutinę MA struktūrą – sudėti šalia *meta aprašymų*, terminus, kuriems MA buvo priskirti ir žiūrėti, koks rezultatas gavosi, o tada aptarti problemas.



24 paveikslas. **Visos 7 MA grupės ir naujos pastebėtos asociacijos**

Sudėjus 7 MA grupes viena šalia kitos, kaip iš anksto ir manyta, ryškėja naujos asociacijos (žr. į 24 paveikslą), 8 akivaizdžiausias verta aptari, devintoji tik kaip *semantika* netilpusio paveikslo pratęsimas. Tad 8 naujos asociacijos:

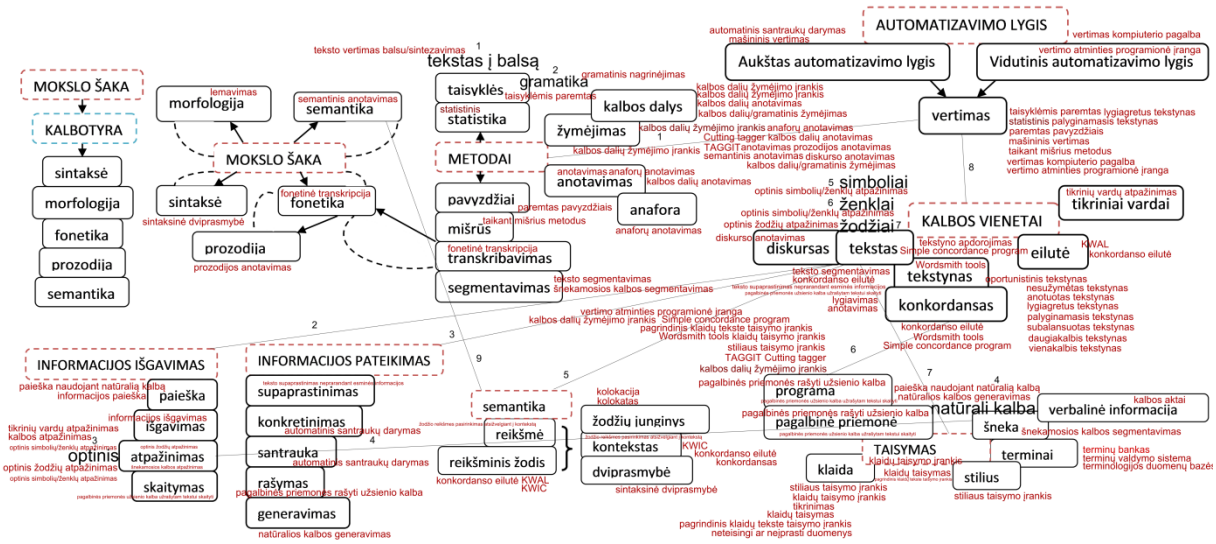
1. ryšio **metodai** – **vertimas** asociacija: **skirta_kam**, vaidmenys asociacijoje: **metodai** – **metodas**, **vertimas** – **vertimui**. Galima teigti, jog visi MA esantys ***METODAI*** grupėje susiję su *vertimas*, transkribuotas tekstas irgi gali būti išverstas.
2. ryšio **informacijos išgavimas** – **tekstas** asociacija: **išgaunama_iš**, vaidmenys asociacijoje: **informacijos išgavimas** – **informacija**, **tekstas** – **tekstas**. Vėlgi, kiekvienas iš ***INFORMACIJOS IŠGAVIMAS*** grupėje esančių MA sietinas su tekstu.
3. ryšio **informacijos pateikimas** – **tekstas** asociacija: **pateikiama_kuo**, vaidmenys asociacijoje: **informacijos pateikimas** – **informacija**, **tekstas** – **tekstu**. Kaip ir ankstesnėse dviejose asociacijose grupės MA siejasi su tekstu, nes paprastai santraukos ar konkretinimas daromas iš teksto ir kaip rezultatas taip pat pateikiama tekstu.
4. ryšio **atpažinimas** – **šneka** asociacija: **atpažįstama_iš**, vaidmenys asociacijoje: **atpažinimas** – **atpažįstama**, **šneka** – **šnekos**.
5. ryšio **semantika** – **tekstas** asociacija: **sutinkama_kur**, vaidmenys asociacijoje: **semantika** – **reikšmė**, **tekstas** – **tekste**. Tokia asociacija priskirta, nes tekstas kaip reikšmių šaltinis.

6. ryšio **programa** – **konkordansas** asociacija: **skirta_sukurti**, vaidmenys asociacijoje: **programa** – **programa**, **konkordansas** – **konkordansą**.
7. ryšio **taisymas – tekstas** asociacija: **taisyti_ką**, vaidmenys asociacijoje: **taisymas** – **taisyti**, **tekstas** – **tekstą**. Asociacija priskirta, nes taisymas paprastai sietinas su teksto, klaidų jame paieška ir taisymu.
8. ryšio **kalbos vienetai** – **vertimas** asociacija: **verčiama_kas**, vaidmenys asociacijoje: **kalbos vienetai – kas verčiama**, **vertimas** – **vertimas**.

Pridėti ir 7 prieš tai nei vienai grupei nepriskirti ar pasimetę *meta aprašai*: *gramatika*, *tekstas į balsą*, *optinis*, *natūrali kalba, simboliai, ženklai, žodžiai* (žr. į 24 paveikslą, kur jie sužymėti skaičiais eilės tvarka, kuria įtraukti į paveikslą), šių asociacijos:

1. ryšio **taisyklės** – **gramatika** asociacija: **paremtos_kuo**, vaidmenys asociacijoje: **taisyklės** – **taisyklės**, **gramatika** – **gramatika**.
2. ryšio **metodai** – **tekstas į balsą** asociacija: **paremtas_kuo**, vaidmenys asociacijoje: **metodai** – **metodas**, **tekstas į balsą** – **tekstu į balsą**.
3. ryšio **atpažinimas** – **optinis** asociacija: **atpažįstama_kaip**, vaidmenys asociacijoje: **atpažinimas** – **atpažinimas**, **optinis** – **būdas**.
4. ryšio **šneka** – **natūrali kalba** asociacija: **priklauso_kam**, vaidmenys asociacijoje: **šneka** – **šneka**, **natūrali kalba** – **natūraliai kalbai**.
5. ryšio **tekstas** – **simboliai** asociacija: **sudaro_kas**, vaidmenys asociacijoje: **tekstas** – **tekstą**, **simboliai** – **simboliai**. Asociacijose **tekstas** – **ženklai**, **tekstas** – **žodžiai** galioja tos pačios asociacijos ir vaidmenys jose, kaip ir **tekstas** – **simboliai** tik pakeičiant vaidmenys asociacijose atsižvelgiant į pavadinimus.

Sudėjus visas 7 grupes į vieną vietą, priskyrus ir aprašius naujas tarp jų kylančias asociacijas, bei pridėjus iki šioj niekur nepriskirtus MA, belieka prikabinti prie kiekvieno MA su juo besisiejantį NLP terminą.



25 paveikslas. **7 MA grupės su pridėtais terminais (pilnas paveikslas 2 priede)**

Susidėliojus terminus pagal šiems priskirtus *meta aprašymus* (jei terminui buvo priskirti keli meta aprašymai, prie kiekvieno jų jis ir pridedamas), iš karto išryškėjo modelio trūkumai (žr. į 25 paveikslą). Tokie meta aprašymai, kaip *supaprastinimas* ir *konkretinimas* atkartoja vienas kito reikšmę, tad verta palikti tik vieną iš jų – *konkretinimas*. Kartojasi ir *žymėjimas – anotavimas*, dėl savo reikšmės panašumo, bei dėl to, kad *anotavimas* tarptautinis žodis, *žymėjimas* – bandymas jo prasmę perteikti lietuviškai.

Įdomu tai, kad jei terminui buvo priskirta keletą MA, pridėjus šiuos terminus prie sudarytos *meta aprašymų* struktūros, gaunasi du asociacijų lygmenys. Vienas, kuomet žiūrima tik į asociacijas tarp MA, kitas, kuomet žvelgiama į terminus ir kartu meta aprašymų struktūros lygmenį. Abu šie lygmenys padeda geriau suprasti terminų ryšius ir įžvelgti dar nenustatytus, beje toks grafinis vizualizavimas padeda geriau suprasti pačia NLP mokslo sritį ir jos sandarą. Papildžius šią struktūrą daugiau terminų, tai būtų matoma dar ryškiau.

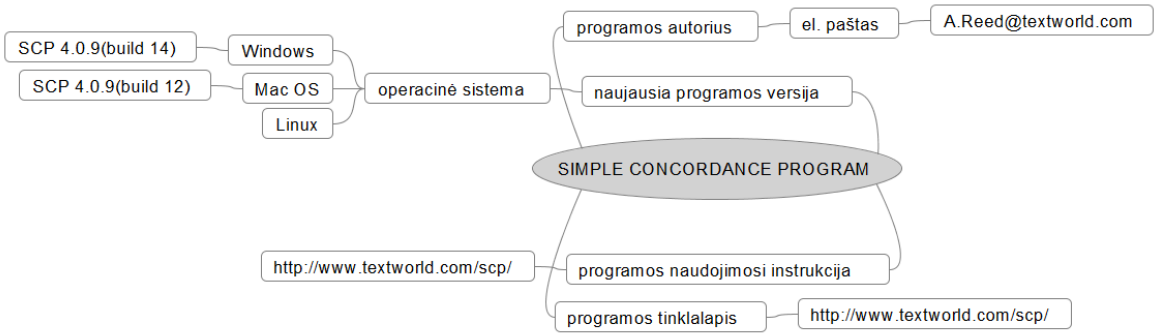
Taigi, ką davė šis bandymas ir kaip jį galima palyginti su pirmuoju. Pirmojo bandymo esmė – atsirinkus dalį NLP terminų, kiek įmanoma labiau juos struktūruoti naudojant *temų tipus* ir piešiant terminų medį. Čia asociacijos nurodomos pačia medžio struktūra, jų pačių neįvardijant. Antro tyrimo esmė – tiems patiems terminams nustatyti geriausiai juos apibūdinančius meta aprašymus ir šį kartą struktūruoti ne pačius terminus, o jų MA, remiantis *Teminių žemėlapių* sudarymo principais, nusakytos asociacijos MA grupėse, vėliau ir tarp jų, įvardinti vaidmenys asociacijose. Tiek vienu, tiek kitu atveju, siekta nustatyti geriausią NLP terminų struktūrą, visgi atrodo, jog antrasis tyrimas suteikė daugiausiai supratimo apie galimą struktūrą ir pateikė daugiausia galimų asociacijų. Verta paminėti su kokiomis problemomis susidurta: netgi dirbant su palyginti nedaug terminų, juos struktūruoti

ir nustatyti asociacijas yra daugybė variantų. Iš to iškarto kyla minčių tolesniam šio darbo plėtojimui. Sudarinėjant ontologiją, reikia struktūruoti ne tik pačius terminus, tačiau kategorizuoti ir pačias asociacijas, griežčiau nusakyti jų sudarymo principus.

9.3 Konkretūs atvejai NLP ontologijoje

Darbe jau ne kartą aptartos *asociacijos*, *vaidmenis jose*, kalbėta ir apie *temų vardus*, kuomet šie nurodomi kita kalbą (šiuo atveju – anglų arba lietuvių kalbomis), tačiau būtina pakalbėti ir apie *konkrečius atvejus* NLP ontologijoje.

Reikia apsibrėžti, *kokie konkretūs* atvejai reikalingiausi NLP ontologijoje. Vartotojui, nežinančiam ką reiškia terminas, ar norint išsamesnės jo apibrėžties, aktualu turėti nuorodą su į termino apibrėžtį vedantį tinklalapį ar duomenų bazę. Antra vertus, iš ontologijos nukreipinėti į itin išsamų aprašą taip pat nėra tikslinga, nes vartotojui prireiks laiko, kol šis šaltinyje suras jį konkrečiai dominančią informaciją. Tam, kad sutrumpėtų tokios informacijos paieška, prie termino reikia pasiūlyti *keletą konkrečių* atvejų *tipų*.

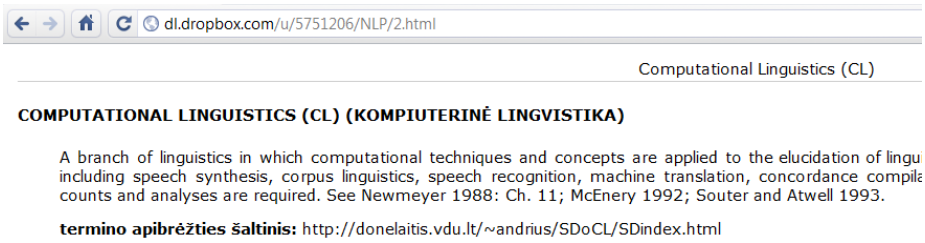


26 paveikslas. **Konkrečių atvejų pavyzdys sudarinėjant NLP ontologiją**

Kaip pavyzdys pateikiamas 26 paveikslas. Šiam pavyzdžiui panaudota tekstynų lingvistikos programa *Simple Concordance Program*. Taigi, vartotojas, ontologijoje aptikęs šią programą pirmą kartą, gali rinktis iš kelių *konkrečių atveju tipų*: nuorodos parsisiuntimui pagal konkrečia operacinę sistemą, informacijos apie programos autorių, susipažinti su programos naudojimo instrukcija, ar apsilankyti programos tinklalapyje. Kiekviena tokia kategorija netik sutaupo laiko, tačiau ir gali priminti apie kokią nors informaciją, ar sukelti minčių tolesnei paieškai.

NLP terminų ontologijoje svarbu terminams priskirti ir *konkrečius atvejus* nukreipiančias į informaciją, apie tos šakos mokslo, teorijos ar technologijos pradininkus, mokslininkus šiuo metu dirbančius ties viena ar kita technologija, aktualesnius straipsnius spaudoje ar mokslinėje literatūroje.

Svarbu nepamiršti, jog informacija, į kurią nukreipiama – sensta, keičiasi informacijos adresai internete, todėl sąryšiams reikalinga arba dažnai atnaujinama vidinė duomenų bazė, arba nuolat atnaujinamas šaltinis (pvz.: Wikipedia). 27 paveiksle pateikiamas bandymas sukurti termino apibrėžties bylą HTML formatu, pritaikius minimalų CSS (angl. Cascading Style Sheets – griežtai apibrėžta kalba aprašomas tinklalapių formatavimas, spalvos ir t.t.) byloje apsirašytą stilių ir talpinant visa tai interneto debesyje (angl. *cloud computing*).



27 paveikslas. **Apibrėžtis, į kurią nukreipiama pasinaudojus *konkretaus atvejo* nuoroda**

Tokį terminų apibrėžimų talpinimo būdą galima laikyti patikimiausiu ir nesudėtingu (kuriant nedidelės apimties ontologiją), sudarytojas bet kada gali redaguoti duomenis esančius debesyje.

Aptarus galimus NLP ontologijos modeliavimo būdus, toliau darbe kalbama apie TŽ programinę įrangą.

10. TŽ PROGRAMINĖ ĮRANGA

Apžvelgus *Teminių žemėlapių* sudarymo struktūrą, atlikus du bandymus, jais siekiant nustatyti kiek įmanoma geresnę NLP terminų struktūrą, šiame skyriuje kalbama apie programinę įrangą, naudojamą kompiuterinių ontologijų (šiuo atveju TŽ) kūrime. Plačiau aptariamos šios programos: **Ontopia Omnigator** (2009) programų paketas, **Wandora** (2010), **Onotoa** (2010), bei **Topic Map Designer** (2010). Pasirinktos būtent šios programos, nes jos yra nemokamos, bei turinčios grafinę vartotojo sąsają, dėl ko nebūtina išmanyti XTM kalbą, norint sukurti ontologiją. Aptariant programas, vertinta jų **dokumentacija**, kas svarbu besimokant kurti ontologijas, programos **paprastumas**, bet kartu ir **reikalingiausių funkcijų buvimas**. Atsižvelgta ir į metus, kuriais programa paskutinį kartą atnaujinta, pagal tai sprendžiant, ar projektas dar aktyvus. Be viso to, kiekviena iš programų bandyta paleisti Teminių žemėlapių pakete pradedantiesiems (angl. *Topic Map starter pack* (2010)) esančią Operos ontologiją.

Pirmoji bandyta: **Topic Map Designer** – kaip teigia pats programos autorius (TMD, 2010), programa nėra skirta pilnaverčiam *Teminių žemėlapių* ontologijų kūrimui ir buvo kurta tik kaip autoriaus baigiamojo darbo dalis. Dokumentacijos vos vienas puslapis, nustatymų pasirinkimo galimybė minimali, pati programa paskutinį kartą atnaujinta 2001 metais. Operos ontologijos paleisti šia programa nepavyko, tad nuspręsta, jog tolesniam šio darbo plėtojimui ir ontologijos kūrimui Topic Map Designer netinkama.

Atsisakius pirmosios programos, kita: **Onotoa**, šį kartą dokumentacija plati ir išsami, programa paskutinį kartą atnaujinta 2009 metais, aprašomi ontologijos kūrimo pavyzdžiai, tačiau pačiame programos kataloge nėra jau sukurtų ontologijų pavyzdžių. Operos ontologijos atidarymas, nors ši programa jau turi ir importavimo (angl. *import*) galimybę, nepavyko. Susidurta dar ir su tuo, jog **Onotoa** bylas išaugo naudodama savo plėtinius .ono, .tmcl, kas įneša nesuderinamumą su kita programine įranga. Yra eksportavimo (angl. *export*) funkcija, tačiau ir ja nepavyko pasinaudoti. Verdiktas – netinkama, nes savi plėtiniai, nėra pavyzdžių, pasirodė sudėtinga.

Trečioji: **Wandora**. Itin išsami dokumentacija, programos kataloge ontologijų pavyzdžiai, grafinio vizualizavimo įrankiai. Programa nuolat atnaujinama, tačiau, Operos ontologija nepasileido. Išbandžius programas akivaizdu, jog kiekviena iš jų turi savus standartus, kurie šiek tiek skiriasi nuo oficialiai pripažinto TŽ standarto. Byloms saugoti ir

atverti naudoja savo plėtinius. Operos ontologija kurta **Ontopia Omnigator** pagrindu, tad panašu, jog kitos programos būtent todėl ir neatveria jos. **Wandora** dažniausiai atnaujinama, suderinama su daugeliu formatų, geriausias vizualizavimo įrankis, turi net priedą Firefox interneto naršyklei, kurio pagalba iš tinklalapių galima išgauti norimą semantinę informaciją, ar tiesiog pagreitinti ontologijos kūrimą (nepavyko priversti veikti), tačiau kartu per sudėtinga pradedančiam. Pripažinta kaip netinkama.

Įsitikinus, jog jau trys programos netinkamos tolesniam TŽ ontologijos kūrimui, liko – **Ontopia Omnigator**. Bandant paleistį šį programų paketą, iškilo daugiausia problemų, kurių sprendimas aprašomas 3 priede. Nepaisant problemų, programa pasirodė geriausiai suprantama, turinti plačiausią dokumentaciją iš visų minėtų programų, paskutinį kartą atnaujinta 2009 metų gale. **Ontopia Omnigator** turi ir ontologijų pavyzdžių, todėl nuspręsta bandyti jau turimus terminus ir nustatytas asociacijas su vaidmenimis jose, perkelti į šia programą.

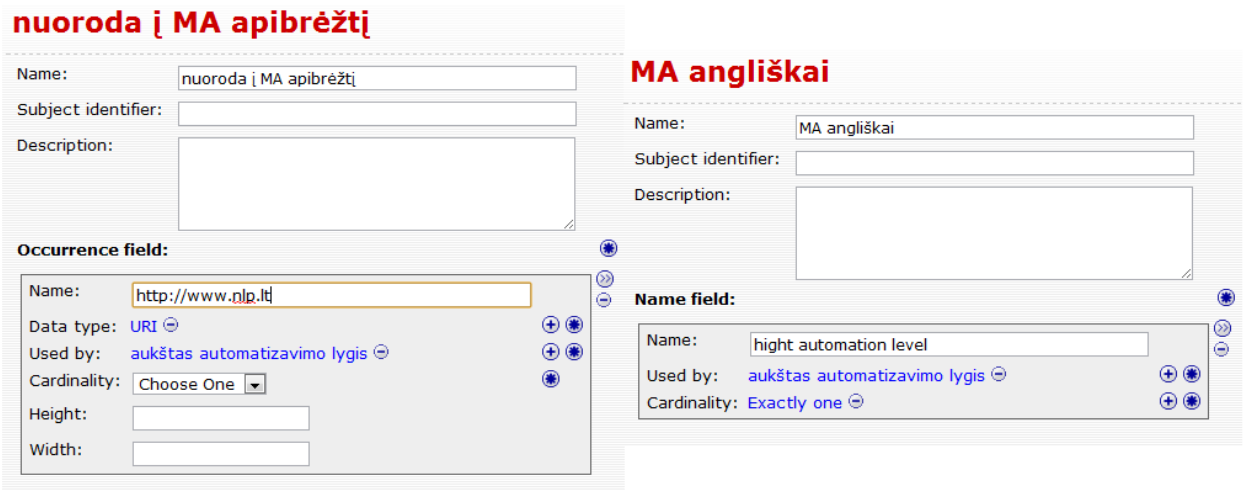
Prieš keliant visus NLP terminus į **Ontopia Omnigator** programą, pradėta nuo mažos *meta aprašymų* grupės – *AUTOMATIZAVIMO LYGIS*. Kadangi ši grupė 2 bandyme jau aprašyta, belieka tik sudėlioti viską į savo vietas (žiūrėti į 28 paveikslą). Svarbiausia nusistatyti:

28 paveikslas. **MA suvesti į Ontopia Omnigator programą**

1. Temų tipus – šiuo atveju, tai visi į *AUTOMATIZAVIMO LYGIS* grupę įeinantys *meta aprašymai*: *automatizavimo lygis, aukštas automatizavimo lygis, vidutinis automatizavimo lygis, vertimas.*
2. Asociacijų tipus – išsivardinti nagrinėjamoje MA grupėje priskirtas asociacijas: *automatizavimo_lygmuo, vertimo_automatizavimo_lygmuo*
3. Vaidmenų asociacijose tipus – *automatizavimas, aukštas, vidutinis, vertimo, aukštas lygmuo, vidutinis lygmuo.*
4. Konkrečių atvejų tipus – nuoroda į MA apibrėžtį.

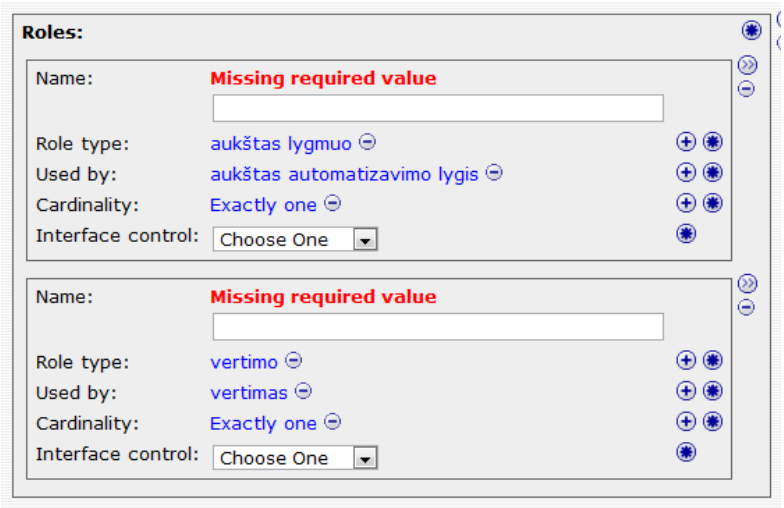
5. Vardų tipus – *MA angliškai* (nes MA pateikiami lietuviškai, o norint pademonstruoti vardų tipų arba tiesiog vardų paskirtį, geriausia tai atlikti pateikiant *temų vardus* kita kalba).

Kitas žingsnis – MA išsiverčiame į anglų kalbą ir programoje suvestiems lietuviškiems MA, priskiriame *vardo tipą* – **MA angliškai** (žr. į 29 paveikslo dešinę pusę), tad: *automatizavimo lygis* – automatization level, *aukštas automatizavimo lygis* – hight automatization level, *vidutinis automatizavimo lygis* – average automatization level, *vertimas* – translation.



29 paveikslas. ***Temų vardų* ir *konkrečių atvejų* priskyrimo langai**

Panašus veiksmas atliktas ir MA priskiriant *konkretų atvejį* – **nuoroda į MA apibrėžtį**. Programoje nurodomas konkretaus atvejo tipas (angl. *data type*) (žr. į 29 paveikslo kairę pusę), šiuo atveju **URI**, bei nurodoma, kokiam MA priskirtas atvejis (angl. *used by*), galima nustatyti ir apribojimus, jog *konkretus atvejis* gali pasikartoti viena kartą ar daugiau ir pan.



30 paveikslas. **Asociacijų priskyrimo langas**

Prisiskyrus MA *vardus*, *konkrečius atvejus*, beliko tik susieti *meta aprašymus asociacijomis*, ir nurodyti *vaidmenys* jose. Kaip matyti iš 30 paveikslo, pateikiamas dvilypis langas, nes asociacijos binarės, lange nurodomas *meta aprašymas*, tuomet jam priskiriamas *vaidmuo*, lygiai tas pats padaroma ir su kitu asociacijai priklausiančiu *meta aprašymu*.

Štai tokiu principu suvedami duomenys į **Ontopia Omnigator** programą. Visgi, visų terminų suvedimo atsisakyta, nes pats NLP modelis nėra iki galo aprašytas, o programa nepilnai perprasta, todėl tai gali būti šio darbo tąsa.

Apibendrinus gali teigti, jog darbe aptartos pagrindinės pasitaikiusios problemos, išbandyti visi ontologijos kūrimo etapai, nuo duomenų rinkimo, jų modeliavimo, asociacijų paieškos, iki duomenų vedimo į programą.

11. IŠVADOS

Ontologija filosofijoje bandoma nusakyti būtį, tuo tarpu kompiuterijoje ontologijomis aprašomos sąsajos tarp tam tikros srities sąvokų (šio darbo atveju – NPL terminai), leidžiančios vienodai suprasti šią sritį ir sukurti ryšį tarp žmonių ir mašinų.

Darbe plačiai aprašoma *Teminių žemėlapių* tipo ontologija. Būtent gilinimasis į šio tipo ontologijos struktūra, padėjo geriau suprasti pačia kompiuterinės ontologijos esmę, struktūrą ir būdą, kuriuo mus supančias sąvokas ir jas rišančias asociacijas, galima užrašyti tiek žmogui, tiek kompiuterinei sistemai suprantama kalba.

Teminių žemėlapių standartas nuo pat pradžių buvo kuriamas tam, kad perimtų knygose naudojamų indeksų (rodyklių) savybes, padėtų juos kurti, atnaujinti ir kitaip jais manipuliuoti iš didelių kiekių informacijos.

Visgi, vietoje to, kad *Teminis žemėlapis* vien kopijuotų knygose naudojamos rodyklės principus, jis šį modelį praplečia, leidžia jį įvairiai modifikuoti, naudoti įvairiapusę navigaciją. TŽ suteikia vartotojui galimybę išvengti klaidžiojimo dideliuose kiekiuose vis atsinaujinančios informacijos ir padeda rasti tai, ko reikia. Iškarto aiškėja du TŽ privalumai – sąvokų aprašymo ir jų ryšių struktūra panaši į žmogaus mastymą, o suvestą informaciją galima vaizduoti grafiškai, ką žmogus suvokia greičiau, nei skaitydamas tekstą.

Kadangi darbo tikslas: aptarti TŽ ontologijos sudarymo principus, aprašyti pagrindines problemas kuriant ontologiją ir pateikti pasiūlymų, todėl pirmiausia surinkus NLP, buvo ieškoma geriausio jų skirstymo būdo, tam, kad vėliau būtų galima nusakyti ryšius tarp jų.

Atmesta pirma hipotezė, jog geriausia ontologiją pradėti kurti jau turint nustatytą terminų struktūrą. Hipotezė nepasitvirtino, kadangi sukėlus terminus iš skirtingų šaltinių, paaiškėjo, jog sunku rasti bendrus junglumo taškus tarp šių NLP terminų.

Atsisakius didžiosios dalies terminų, nuspręsta tolesnį ontologijos kūrimą tęsti su 69 likusiais ir manomai geriausiai sritį nusakančiais terminais.

Iš jų sudarytas medis, prieš tai suskirsčius šiuos terminus į *tipus,* geriausiai juos nusakančias kategorijas. Po šio veiksmo, struktūra tapo aiškesne, tačiau patikrinta ir antra hipoteze, kuria teigiama, jog priskyrus terminams *meta aprašymus* – geriausiai kiekvieną

terminą apibūdinančius žodžius, ir šiuos sugrupavus remiantis logika ir turimomis žiniomis, struktūra taptų dar aiškesne.

Tad terminams priskirti *meta aprašymai*, šie surūšiuoti, pašalinti besikartojantys ir suskirstyti į 7 grupes. Verta nepamiršti, jog grupuoti tik MA, nebežiūrint į tai, kokiam terminui jie priklauso.

Sugrupavus MA, nusakytos asociacijos, vaidmenys jose, grupių viduje. Tuomet visos 7 grupės sudėtos į vieną vietą ir aprašytos kelios naujai pastebėtos sąsajos. Visa tai atlikus, prie kiekvieno MA gražinti terminai, iš kurių tie meta aprašymai „kilę“. Toks bandymas nusakyti, kam terminai priklauso, pasirodė geriausias. Tuo atveju, kuomet nėra aišku, kaip sąvokos jungiasi viena su kita, naudinga toms priskirti meta aprašymus.

Kalbant plačiau apie asociacijas ir vaidmenis jose, kartais geriausia kaip vaidmenis nurodyti: šaltinį, jungties tašką ar išeities tašką, ką geriausiai perteikti angliškais terminais – angl. *source*, *output*, *input*, dėl jų reikšmės apibrėžtumo, ką bandant nusakyti lietuvišku terminu sunku, nes trūksta taiklaus posakio.

Vaidmenis asociacijose galima aprašyti ir hierarchiniu principu, naudojant **superklasės** ir **subklasės** apibrėžtis. **Superklasė** – visada bus pagrindinis terminas ar *meta aprašymas*, tas savo ruožtu gali būti ir kito termino **subklase**, taip nusakant terminų išsidėstymą.

Visi šie bandymai suteikė galimybę pažvelgti į ontologijos sudarymą nuo pat pradžių. Modeliuojant galimą struktūrą, svarstant kaip pavadinti tipus, o vėliau ir asociacijas, kartu su vaidmenimis jose. Paaiškėjo, kad net iš sąlyginai nedidelio terminų kiekio, tikslią ir gerą struktūrą, ryšius tarp terminų – sunku nusakyti. Tai tik suteikė minčių, jog ontologiją tikslingiausia kurti iš pakankamai siauros srities terminijos, nes kuo struktūra sudėtingesnė, tuo kebliau ją suprasti. Tad pirmiausia būtina griežta struktūrą.

Darbe aprašomos binarės asociacijos tarp terminų, tokios kaip **pateikiama_kaip**, **išgauta_kaip, taisoma_kuo** ir pan., kartais norint nusakyti asociaciją, būtina apsirašyti ką kiekvienas terminas reiškia ir tose apibrėžtyse ieškoti bendro bruožo. Jo pasirinkimas nėra visada aiškus, arba vienintelis galimas.

Pats ontologijos kūrimas padeda jos kūrėjau geriau suvokti žmogaus mąstymą ir kaip mes jungiame vienas sąvokas prie kitų. Įrodo, jog ne viskas mintyse taip padrika, kaip iš pirmo žvilgsnio atrodo.

Naudinga įvardinti ir galimą darbo tąsą. Norint toliau plėtoti NLP terminų ontologijos kūrimą, vertėtų pradėti nuo struktūros sudarymo, kurioje tiksliai nusakyti terminų pasirinkimo kriterijai. Neturint struktūros, surinkti kiek galima daugiau terminų ir pasitelkus meta aprašymus ir šių grupavimą, apsirašyti visos srities struktūrą ir kiek galima daugiau asociacijų, tuomet pridėti prie šios struktūros terminus ir patikrinti meta aprašymų struktūros teisingumą, bei pataisyti atsiradusias klaidas. Be viso to, reikia šį sukurtą ontologijos modelį perkelti į programą, kas padėtų NLP ontologiją aprašyti XML kalba, kas suteikia galimybę ją panaudoti įvairiems taikymams. Iš bandytų programų apsistota ties Ontopia Omnigator, tačiau tai nereiškia, jog nėra geresnių alternatyvų.

Kokie gali būti NLP terminų ontologijos taikymai:

1. NLP terminų vertimo įrankis – tekste aptikus anglišką terminą, pagal ontologijoje esantį lietuvišką vertimą, galima nukreipti į lietuviškus šaltinius susijusius su terminu. Galimas ir atvirkštinis variantas.
2. Nauda studijuojantiems ar besidomintiems čia sritimi. Gerai aprašius ontologiją ir naudojant gerai apgalvotą *konkrečių atvejų* sistemą, visa informacija apie NLP atsiduria vienoje vietoje, pagrindiniai terminai, jų apibrėžtys, teorijos, nuorodos į besisiejančius šaltinius, rašytus darbus.

Ontologijų kūrimas – bene perspektyviausia informacijos apdorojimo forma, kuo daugiau sąvokų bus aprašyta ir kuo daugiau jų susieta su kitomis, tuo paprasčiau dideli informacijos kiekiai bus apdoroti. Tačiau, kaip tekste pasikeitus žodžių deriniui keičiasi reikšmė, taip ontologijų principu aprašytoms realybės sąvokoms susikeitus vietomis, statiška ontologijos struktūra tampa beverte, to išvengti padėtų automatinio apsimokymo galimybė, tačiau kol kas to nėra.

12. LITERATŪRA, ŠALTINIAI

B. Passin T. B. 2004. *Explorer's guide to the semantic web*. Greenwith: Manning Publications Co.

Butkus A. 2009: *Sparnuoti žodžiai*. Kaunas: Aesti.

Calais projektas. Prieiga internetu: http://sws.clearforest.com/calaisViewer/ (žiūrėta: 2009-12-20).

Cyc projektas. Prieiga internetu: http://www.cycfoundation.org/concepts (žiūrėta: 2009-12-20).

Davies J., Studer R., Warren P. 2006: *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. Chichester: John Wiley & Sons.

Elektroninis anglų-lietuvių kalbų žodynas Anglonas.

Free Mind programa. Prieiga internetu: http://freemind.sourceforge.net/ (žiūrėta 2010-04-10).

Google vertėjas. Prieiga internetu: http://translate.google.lt/# (žiūrėta 2010-03-02).

Horgan T., Potrc M. 2008: *Austere Realism Contextual Semantics Meets Minimal Ontology*. London: The MIT Press.

Keinys S. 2005: *Dabartinė lietuvių terminologija*. Vilnius: Lietuvių kalbos instituto leidykla.

Liddy, E.D. 2001. *Natural Language Processing*. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. Prieiga internetu: www.cnlp.org/publications/03nlp.lis.encyclopedia.pdf (žiūrėta 2010-04-29).

Linguistics Thesaurus. Prieiga internetu: http://www.dsoergel.com/775/775cAll.pdf (žiūrėta 2010-04-01).

Natural language proccesing. Prieiga internetu: http://en.wikipedia.org/wiki/Natural_language_processing (žiūrėta 2010-04-10).

Onotoa programa. Prieiga internetu: http://onotoa.topicmapslab.de/ (žiūrėta 2010-04-10).

Ontopia Omnigator programa. Prieiga internetu: http://code.google.com/p/ontopia/ (žiūrėta 2009-12-18).

Ontopia starter pack. Prieiga internetu: http://www.ontopia.net/download/ (žiūrėta 2010-05-01).

Putnam H. 2004: *Ethics without Ontology*. London: Harvard University Press.

Ramonas V. 2009: *Automatinis asmenvardžių ir vietovardžių aptikimas tekste*. Skaitmeninės lingvistikos kursinis darbas. Kaunas. Vytauto Didžiojo universitetas.

Systematic Dictionary of Corpus Linguistics. Prieiga internetu: http://donelaitis.vdu.lt/~andrius/SDoCL/SDindex.htmla (žiūrėta 2010-04-10).

Teminių žemėlapių konteksto sprendimas. Prieiga internetu: http://www.ontopia.net/topicmaps/materials/scope.htm (žiūrėta: 2009-12-18).

Teminių žemėlapių sudarymo aprašymas. Prieiga internetu: http://www.ontopia.net/topicmaps/materials/tao.html (žiūrėta: 2009-12-18).

Tildės lietuvių kalbos aiškinamasis žodynas.

TMD – Topic Map Designer programa. Prieiga internetu: http://www.topicmap-design.com/en/topicmap-designer.htm (žiūrėta 2010-04-10).

Valore P (editor). 2006: *Topics on General and Formal Ontology*. Milan. Polimetrica.

VDU vertimas. Prieiga internetu: http://vertimas.vdu.lt/twsas/ (žiūrėta 2010-04-02).

Wandora programa. Prieiga internetu: http://www.wandora.org/ (žiūrėta 2010-03-01).

XML Topic Maps. Prieiga internetu: http://www.topicmaps.org/xtm/index.html (žiūrėta 2010-03-17).

13. PRIEDAI

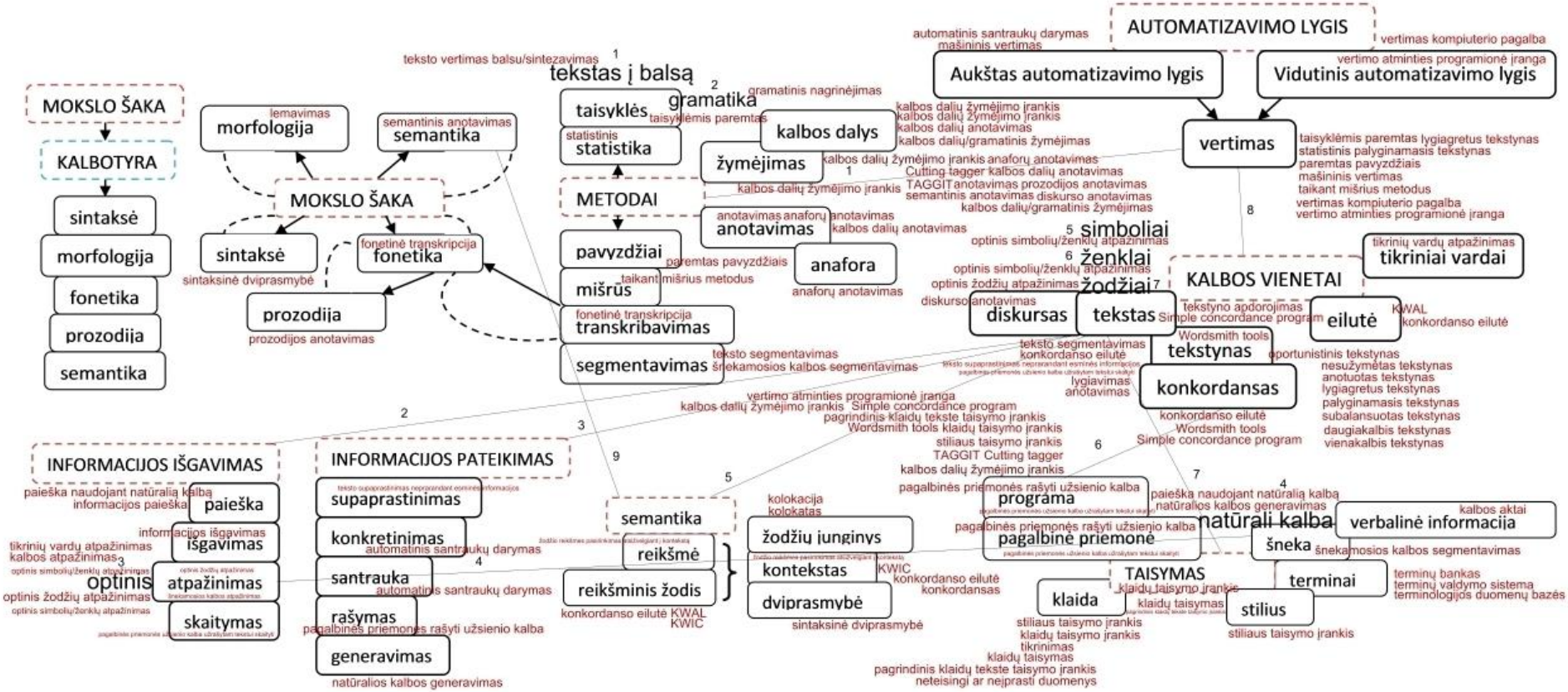
1 priedas. **Meta aprašymai, angliški terminai ir lietuviški jų vertimai**

META APRAŠYMAI	ANGLIŠKAS TERMINAS	VERTIMAS Į LIETUVIŲ KALBĄ
terminai	Terminological Data Bank (TDB)	Terminų bankas
terminai	Terminological Management System (TMS)	Terminų valdymo sistema
terminai	Terminology Databases	Terminologijos duomenų bazės
programa, žymėjimas, kalbos dalys	Part-Of-Speech Tagger	Kalbos dalių žymėjimo įrankis
vertimas, programa, vidutinis automatizavimo lygis	Translation Memory Software	Vertimo atminties programinė įranga
vertimas, aukštas automatizavimo lygis	Machine Translation (Mt)	Mašininis vertimas
vertimas, taisyklės	Rule-Based	Taisyklėmis paremtas
vertimas, statistika	Statistical	Statistinis
vertimas, pavyzdžiai	Example-Based	Paremtas pavyzdžiais
vertimas, mišrūs metodai	Hybrid MT	Taikant mišrius metodus
vertimas, vidutinis automatizavimo lygis	Computer-Assisted Translation	Vertimas kompiuterio pagalba
tekstas, tekstynas	Corpora	Tekstynai
tekstas, tekstynas	Corpus Processing	Tekstyno apdorojimas
tekstas	Alignment	Lygiavimas
tekstas, žymėjimas, anotavimas	Annotation	Anotavimas
žymėjimas, anafora, anotavimas	Anaphoric Annotation	Anaforų anotavimas
žymėjimas, kalbos dalys, anotavimas	Part-Of-Speech Tagging	Kalbos dalių anotavimas
fonetika, transkribavimas	Phonetic Transcription	Fonetinė transkripcija
žymėjimas, semantika, anotavimas	Semantic Annotation	Semantinis anotavimas
prozodija, žymėjimas, anotavimas	Prosodic Annotation	Prozodijos anotavimas
diskursas, žymėjimas, anotavimas	Discoursal Annotation	Diskurso anotavimas
tekstas, eilutė, reikšminis žodis, kontekstas	Concordance	Konkordansas

žodžių junginys	Collocate	Kolokatas
žodžių junginys	Collocation	Kolokacija
morfologija	Lemmatisation	Lemavimas
gramatika	Parsing	Gramatinis nagrinėjimas
klaida	Validation	Tikrinimas
klaida, taisymas	Text-Proofing	Klaidų taisymas
klaida, programa, taisymas	General Text Checker	Pagrindinis klaidų tekste taisymo įrankis
klaida, programa, taisymas	Spelling Checker	Klaidų taisymo įrankis
stilius, klaida, programa	Style Checker	Stiliaus taisymo įrankis
tekstynas	Monolingual Corpus	Vienakalbis tekstynas
tekstynas	Multilingual Corpus	Daugiakalbis tekstynas
tekstynas	Balanced Corpus	Subalansuotas tekstynas
vertimas, tekstynas	Comparable (Reference) Corpus	Palyginamasis tekstynas
vertimas, tekstynas	Parallel (Aligned) Corpus	Lygiagretus tekstynas
tekstynas	Opportunistic Corpus	Oportunistinis tekstynas
tekstynas	Unannotated Corpus	Nesužymėtas tekstynas
žymejimas, tekstynas, anotavimas	Annotated Corpus	Anotuotas tekstynas
tekstas, skaitymas, pagalbinė priemonė, programa	Foreign Language Reading Aid	Pagalbinės priemonės užsienio kalba užrašytam tekstui skaityti
pagalbinė priemonė, programa, rašymas	Foreign Language Writing Aid	Pagalbinės priemonės rašyti užsienio kalba
tekstas į balsą	Text-To-Speech	Teksto vertimas balsu/Sintezavimas
šneka, atpažinimas	Speech Recognition	Šnekamosios kalbos atpažinimas
paieška	Information Retrieval (IR)	Informacijos paieška
išgavimas	Information Extraction	Informacijos išgavimas
paieška, natūrali kalba	Natural Language Search	Paieška naudojant natūralią kalbą
tikriniai vardai, atpažinimas	Named Entity Recognition (NER)	Tikrinių vardų atpažinimas
generavimas, natūrali kalba	Natural Language Generation	Natūralios kalbos generavimas
aukštas automatizavimo lygis, santrauka, konkretinimas	Automatic Summarization	Automatinis santraukų darymas
tekstas, supaprastinimas	Text Simplification	Teksto supaprastinimas neprarandant esminės informacijos
atpažinimas	Language Recognition	Kalbos atpažinimas
optinis, atpažinimas, ženklai, simboliai	Optical Character Recognition	Optinis simbolių/ženklų atpažinimas
optinis, atpažinimas, žodžiai	Optical Word Recognition	Optinis žodžių atpažinimas
šneka, segmentavimas	Speech Segmentation	Šnekamosios kalbos segmentavimas
tekstas, segmentavimas	Text Segmentation	Teksto segmentavimas

kalbos dalys, gramatika, žymėjimas, anotavimas	Part-Of-Speech Tagging	Kalbos dalių/gramatinis žymėjimas
kontekstas, reikšmė	Word Sense Disambiguation	Žodžio reikšmes pasirinkimas atsižvelgiant į kontekstą
sintaksė, dviprasmybė	Syntactic Ambiguity	Sintaksinė dviprasmybė
klaida	Imperfect Or Irregular Input	Neteisingi ar neįprasti duomenys
verbalinė informacija	Speech Acts And Plans	Kalbos aktai
programa, tekstynas, konkordansas	Simple Concordance Program	Neverčiama
programa, tekstynas, konkordansas	Wordsmith Tools	Neverčiama
žymėjimas, programa	Cutting Tagger	Neverčiama
žymėjimas, programa	Taggit	Neverčiama
reikšminis žodis, eilutė	KWAL	Neverčiama
reikšminis žodis, kontekstas	KWIC	Neverčiama
tekstas, konkordansas, reikšminis žodis, eilutė, kontekstas	Concordance Line	Konkordanso eilutė
tekstas, segmentavimas	Text Segmentation	Teksto segmentavimas

2 priedas. MA grupės kartu su NLP terminais



3 priedas. **Ontopia Omnigator diegimas**

Dokumentacijoje rašoma, jog norint paleisti programą, reikalinga naujausia Java versija ir Apache Tomcat – Java pagrindu veikianti serverio programinė įranga. Pirmiausiai iš programos tinklalapio parsisiunčiamas ir išarchyvuojamas ontopia-5.0.2.zip archyvas. Išarchyvuotas programos katalogas perkeliamas į C:/ontopia-5.0.2 siekiant kiek įmanoma supaprastinti kelią iki programos katalogo atliekant tolimesnius įdiegimo darbus.

1. Kadangi programa veikia Java pagrindu, įdiegiama naujausia jos versija ir nustatomas CLASSPATH nukreipiantis į C:\ontopia-5.0.2\lib\ontopia.jar (Start/Control Panel/System/Advanced System Settings/Environment Variables/New user variable ir ten, kur rašoma Variable name: CLASSPATH, o Variable value: šis kelias C:\ontopia-5.0.2\lib\ontopia.jar)
2. Nustačius CLASSPATH ir norint įsitikinti, kad ontopia kompiuteyje aptikta, komandinėje eilutėje (nuspaudus windows+r klavišus ir iššokusioje lentelėje surinkus cmd nuspaudžiamas enter klavišas) rašoma java net.ontopia.Ontopia
3. Įsitikinus, kad viskas tvarkoje (ankstesnės komandos išvestas rezultatas: Success: All required classes found), pereinama prie sekančio etapo.
4. Nustatome, kur kompiuteryje įdiegta Java ir priskiriame JAVA_HOME. Start/Control Panel/System/Advanced System Settings/Environment Variables/New user variable ir ten, kur rašoma Variable name: JAVA_HOME, o Variable value: kelias iki Java direktorijos, šiuo atveju: C:/Program Files/Java
5. Iš: http://tomcat.apache.org/ parsisiunčiamas Apache Tomcat serverio programa, instaliuojama ir bandoma paleisti Ontopia Omnigator pagalbos failuose esančiais adresais (pvz.: http://localhost:8080/omnigator/) programinę įrangą, reikalingą ontologijų kūrimui. Būtent čia ir buvo padaryta didžiausia klaida, nes atlikus daug bandymų ir peržiūrėjus ar visi žingsniai atlikti tiksliai pagal instukciją, niekaip nepavyko paleisti šios programinės įrangos. Šios problemos sprendimas: Apache Tomcat serverį paleisti ne instaliavus kaip atskirą programą, o panaudoti ontopia kataloge jau esantį startup.bin (C:/ontopia-5.0.2/apache-tomcat/bin/startup.bin) failą paleidžiantį programą.
6. Taigi, nuo šiol Ontopia Omnigator paleidžiama naršyklėje surenkant http://localhost:8080/.