

KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Lietuvių kalbos terminų automatinio nustatymo galimybės

Erika Rimkutė

Anotacija. Straipsnyje trumpai apžvelgta terminologinė situacija pasaulyje ir Lietuvoje, aprašoma, kaip pasikeitė terminologų bei terminografo darbas suklestėjus kompiuterinėms technologijoms, atsiradus tekstynams; kada kilo poreikis terminus analizuoti tekstynuose; kada atsirado pirmosios automatiškai terminus nustatančios programos. Šiame darbe pabrėžiama tekstynų svarba ir nauda analizuojant terminus, juos apibrėžiant, sudarant terminų žodynus. Taip pat aprašomas terminografo darbo ciklas, kai yra naudojami vienakalbiai ar keliakalbiai tekstynai.

Straipsnyje aprašyta, kokie šiuo metu prieinami elektroniniai tekstynai ar duomenų bazės, kuriuose galima rasti lietuviškus terminus. Toliau pristatomas projektas „Švietimo ir mokslo terminų automatinis identifikavimas (ŠIMTAI)“, kuris ir paskatino pasidomėti švietimo ir mokslo terminų automatinio nustatymo galimybėmis. Taip pat apžvelgiami vieno eksperimento rezultatai, kurie pristato automatinio terminų nustatymo procesą, atskleidžia išskylančias problemas ir galimybes automatiškai nustatyti lietuvių kalbos terminus, poreikį kurti lietuvių kalbai pritaikytas automatiškai terminus nustatančias programos.

Reikšminiai žodžiai: terminologija, terminografija, tekstynas, žodynas, automatinis terminų nustatymas.

Įvadas

2009 m. dalyvavimas Nacionalinės lituanistikos plėtros 2009-2015 m. programos projekte „Švietimo ir mokslo terminų automatinis identifikavimas (ŠIMTAI)“ paskatino pasidomėti, kaip būtų galima automatiškai nustatyti vienos srities terminus. Tad šio straipsnio tikslas – pristatyti galimybes, kaip galima automatiškai nustatyti lietuvių kalbos terminus.

Tam, kad būtų galima automatiškai išgauti¹ terminus, reikalingas nemažas pasirengimas. Pirmiausia reikia sudaryti reprezentatyvų ir subalansuotą reikiamos srities tekstyną, tekstus anotuoti morfologiškai, nuspręsti, kokiais metodais (statistiniais ar lingvistiniais) remiantis sukurti automatiškai terminus išgaunantį įrankį, sukurti tą įrankį, įvertinti minėto įrankio patikimumą, iš automatiškai nustatytų galimų terminų išrinkti tikruosius analizuojamos srities terminus. Toliau naudojantis tekstynais būtų galima apibrėžti išrinktus tikruosius terminus.

Šie anksčiau minėti darbai nustatant ir apibrėžiant lietuvių kalbos švietimo ir mokslo terminus kol kas nėra įgyvendinti; sukauptas tik 2 mln. žodžių tekstynas (plačiau žr. skyrių „Švietimo ir mokslo tekstynas“) ir atliktas nedidelis eksperimentas, kurio metu bandyta automatiškai nustatyti švietimo ir mokslo terminus (žr. skyrių „Automatinis terminų nustatymas“).

Kadangi kitose šalyse jau daug nuveikta elektroninės terminografijos srityje (žr. skyrių „Terminologinės situacijos pasaulyje apžvalga“), norėta apžvelgti, kokia situacija Lietuvoje: ar turime terminų žodyną, sudarytą iš tekstynų,

kokių galima rasti elektroninių terminynų, kuriuose yra lietuviškų terminų. Šiai apžvalgai skirtas skyrius „Terminologinė situacija Lietuvoje“.

Terminologinės situacijos pasaulyje apžvalga

Kompiuterinių technologijų, kartu ir kompiuterinių tekstynų era labai aiškiai atskyrė tradicinę terminologiją bei terminografiją nuo naujosios. Pradėjo smarkiai skirtis terminologų požiūris į kelis dalykus: pačią termino sampratą, terminų šaltinius, termino skyrimo nuo netermino kriterijus, terminologo kvalifikaciją, terminų norminimą ir kt. Labai smarkiai išsiplėtus ir susipynus specialiųjų mokslų sritims, padaugėjo darbo terminologams, nes dirbant tradiciniais metodais vargiai galima suspėti paskui mokslo ir technologijų pažangą. Kaip tik todėl kitaip imta žiūrėti į terminus: labiau akcentuojama ne norma, bet vartoseną (Marcinkevičienė, 2000).

Pasikeitusi terminologinė situacija skatina ieškoti ir naujų terminijos tvarkybos būdų, nes tradiciniais būdais darosi sunku aprėpti ir sunorminti terminologijos mokslo objektą. Vakarų Europos šalyse, ypač Didžiojoje Britanijoje ir Vokietijoje, automatinis terminų atpažinimo ir išgavimo iš tekstynų būdas įsitvirtino praėjusio amžiaus paskutiniame dešimtmetyje. Informatikų K. Ahmado, K. Lauristono, M. Rogerso, lingvistų I. Meyer, J. Pearson, J. C. Sager ir jų bendraautorių tyrimai bei jų pagrindu sukurtos kompiuterinės programos (LEXTER, TERMINO ir kt.) labai palengvino terminų atpažinimą. Specializuotos terminologinės programos nuo bendrųjų teksto apdorojimo įrankių skiriasi savo dalykiniu pobūdžiu, nes jomis galima analizuoti tekстыne esančią informaciją, o tekstyną traktuoti kaip atskiros mokslo šakos žinių šaltinį. Yra kelios programų grupės:

¹ Šiame darbe angliškas terminas *term extraction* verčiama *terminų išgavimas*; *išgauti terminus* ir pan.

1) programos, skirtos analizuoti tekstą jame pateikiamos informacijos aspektu – prasminių žodžių, informacijos turtingų teksto vietų, kuriose nurodomi sąvokų tarpusavio ryšiai ir priklausomybė;

2) programos, atrenkančios terminografo reikalingas tekstyno vietas, pavyzdžiui, terminų apibrėžtis;

3) programos, padedančios analizuoti terminais įvardytas sąvokas, palengvinančios iš tekstyno gautų žinių tvarkymą (*knowledge acquisition systems*).

Pastaroji grupė, naudojama specifinių sričių dalykinėms taksonomijoms ir ontologijoms kurti, nors nauja ir perspektyvi, nėra plačiai paplitusi. Kur kas populiareesnės yra atskirų kalbų pagrindu sukurtos pirmos ir antros grupės programos (Cabré Castellví, et al., 2001; Marcinkevičienė, 2000).

Kadangi pagrindiniai pokyčiai XX a. pabaigoje terminologijoje ir terminografijoje yra susiję su tekstynų atsiradimu ir jų naudojimu minėtų sričių darbuose, kitame skyriuje aprašyta, kuo tekstynai svarbūs terminologams ir terminografams.

Tekstynų svarba terminologijos darbams

Sutelkus dėmesį į termino vartoseną, jis tiesiog pradeda rinkti iš tekstynų. Tokia termino išgavimo iš tekstų idėja pamažu suformuoja naują terminų tvarkybos metodiką, atsiranda automatinų priemonių tekste atpažinti terminą. Pakitęs požiūris į terminą ir kitokia jo tvarkybos metodika yra didelis posūkis terminologijoje (Zeller, 2005).

Kalbant apie terminų analizę tekстыne, išryškėja pagrindiniai deskriptyviosios terminologijos skirtumai nuo preskriptyviosios.² Tai termino reliatyvumas, lemiamas konteksto, terminologijos tvarkybos automatizavimas, nebe toks griežtas požiūris į termino vienareikšmiškumą, termino hierarchinės sistemos kūrimą ir praktinį funkcionavimą, termino norminimą (Zeller, 2005).

Tekstynas, būdamas terminų šaltiniu, terminografams praveria dėl daugelio priežasčių. Turbūt svarbiausia, kad iš vieno ir to paties tekstyno iškart gaunami trys terminografams reikalingi dalykai:

- 1) patys terminai,
- 2) jų definicijos, t. y. dalykinė informacija apie pačią sąvoką,
- 3) vartojimo pavyzdžiai, taigi ir vartojimo dažnumas, terminų junglumas ir pan.

Žinoma, tekstynu gali naudotis ir išankstinių terminų sąrašą turintys tyrinėtojai. Tokiu atveju jie iš tekstyno gauna dalykinę ir kalbinę informaciją. Bet jei terminai yra gaunami iš tekstyno, jie turi vieną didelį privalumą – yra sistemiški ir tarpusavyje susiję, mat atskiruose tekstuose yra pateikiami atskirų autorių vartojami terminų rinkiniai, kiekviename iš kurių terminai yra gerai apmąstyti, darniai derantys vienas prie kito. Kaip tik dėl šios priežasties – kad gaunami terminai atspindėtų atskiras sistemas – į tekstyną turi būti dedami ištisi tekstai, o ne jų ištraukos, t. y. imtys (Meyer,

et al., 1996). Ištisų tekstų tekstynas būtinas ir dėl dalykinės informacijos, kurios būtų netenkama pasitenkinus tekstų dalimis (Marcinkevičienė, 2000).

Terminografo, terminus analizuojančio tekstynuose, darbo ciklas atrodo taip: pirmajame etape iš elektroninių tekstų archyvo pagal tam tikrus principus sudaromas tekstynas, vėliau iš jo gaunami terminai, jie perkeliama į duomenų bazę, aprašomi, t. y. atskiruose duomenų bazės laukuose pateikiami iš konkordanso paimti tipiškos jų vartosenos pavyzdžiai, iš ten pat gautos definicijos. Vėliau, remiantis iš tekstyno gauta kalbine ir dalykine informacija, nurodomi su aprašumu terminu susiję žodžiai: sinonimai, antonimai, hiponimai, hiperonimai, duodamos kryžminės nuorodos į jų skiltis, termino priklausomybė dalykinei sričiai, vieta bendrojoje sąvokų sistemoje. Be to, pateikiamos gramatinės ir pragmatinės ypatybės, junglumo ar kitokie vartojimo apribojimai, informacija apie termino teiktinumą ir, jei taip sumanyta, vertimo ekvivalentai. Taigi tekstynas yra nepamainomas nuo pat pradžių, t. y. nuo pažinties su pačia mokslo šaka, iki paskutinės fazės – parengto žodyninio straipsnio. Kitose šalyse terminologai labai skeptiškai žiūri į jau turimų, ne tekstynų pagrindu rengtų popierinių žodynų kompiuterizavimą, t. y. sukėlimą į duomenų bazes (Marcinkevičienė, 2000).

Taigi akivaizdu, kad terminologijoje daug kuo gali padėti tekstynai. Kita vertus, nors jau neabejojama tekstynų svarba šiuolaikinėje kalbotyroje, nors retas kuris kalbinis tyrimas atliekamas nesinaudojant tekstynu, vis dėlto iki šiol nėra tinkamai atskleista tekstyno svarba terminografijoje. Leksikografai jau turbūt neįsivaizduoja gyvenimo be tekstyno, tačiau vis dar daug terminų žodynų, sudarytų remiantis tam tikros srities tekstais, o ne tekstynu (Meyer, et al., 1996), ypač tokia situacija išryškėja Lietuvoje (žr. skyrių „Terminologinė situacija Lietuvoje“).

Terminologinė situacija Lietuvoje

Iki šiol terminai Lietuvoje daugiausia buvo kuriami, parenkami ir aprobuojami remiantis nedideliu kiekiu tekstų, kalbos jausmu ir lietuvių kalbos normomis. Tai tradicinis, iš XIX a. paveldėtasis preskriptyvusis terminų gavimo būdas. Esant ribotai vienos srities ar šakos bendruomenei, nedideliame tekstų kiekiui, toks terminų gavimo būdas buvo patenkinamas. Šiandien, įsigalėjus informacinėms technologijoms, kuriant, skaitant, leidžiant ir verčiant didelius tekstų masyvus, šiame darbe dalyvaujant labai dideliame skaičiui terminus kuriančių žmonių (tekstų autoriai, redaktoriai, vertėjai), deskriptyvusis terminų gavimo būdas konkuruoja tradicinį.

Lietuvių terminologijos ypatumus vertinant pagal preskriptyvumo ir deskriptyvumo požymius, lietuvių, kaip ir užsienio terminologija, išgyvena preskriptyviosios ir deskriptyviosios terminologijos išsiskyrimą. Lietuvių terminologijoje jis dar nėra toks akivaizdus ir spartus kaip, pvz., Amerikoje ar Vakarų Europoje, tačiau terminologijos mokslo kitimo priežastys yra tos pačios – terminijos pradeda stokoti, gausėja nemotyvuotų verstinių terminų, terminologijos objektas darosi sunkiai valdomas, terminai skverbiasi į įvairias kitas sritis, jie kinta sparčiau nei yra tvarkomi, sunorminti sunkiai įsigali, o netaisyklingi terminai plinta tarp vartotojų. Sudėtingėjanti lietuviškosios termino-

² Preskriptyvioji terminologija – terminologija, kurios praktinė veikla nukreipta į termino kūrimą ir norminimą. Deskriptyvioji terminologija – terminologija, kurios praktinė veikla nukreipta į termino analizę, dažniausiai – kompiuterinėmis technologijomis. Šie terminai paaiškinti remiantis Zeller (2005).

logijos situacija pirmiausia verčia keisti patį požiūrį į terminą, terminą apibrėžti ne pagal jam keliamus reikalavimus, bet tirti jo kalbinę ir nekalbinę aplinką, terminams maišantis su bendrinės kalbos žodžiais, labiau gilintis į termino atskyrimo nuo netermino problemą. Dėmesio perkėlimas nuo termino apibrėžimo, jo kūrimo bei norminimo į jau pavartotą terminą bei jo analizę keičia ir pačius terminologinio darbo principus bei metodus (Zeller, 2005).

Lietuvoje automatinio terminų atpažinimu ir su juo susijusiomis problemomis (prasminių žodžių automatinio aptikimu, jo pagrindu atliekamu tekstų klasifikavimu ir pan.) domimasi Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre.³ Vienas iš pirmųjų darbų lietuvių kalba, skirtų automatiniam terminų nustatymui, yra Irmos Zeller disertacija „Automatinis terminų atpažinimas ir apdorojimas“ (apginta 2005 m. Vytauto Didžiojo universitete). Minėtoje disertacijoje aprašomos problemos, kaip nustatyti, ar žodis ir žodžių junginys yra terminas Europos Sąjungos dokumentuose, taip pat pateikiami įvairūs eksperimentai ir naudotos automatinio terminų nustatymo programos. I. Zeller parengė lygiagretų lietuvių-vokiečių kalbų Europos Sąjungos dokumentų tekstyną, kurį apdorojo trimis programiniais įrankiais: TRADOS, COMPRENDIUM, AUTOTERM.

Kitų terminologijos darbų Lietuvoje apžvalga parengta naudojantis S. Meškauskienės pranešimu „Elektroniniai terminynai“.⁴ Pranešime vartoti trys terminai:

- 1) terminų bazė (terminologinių duomenų rinkinys, sutvarkytas pagal iš anksto nustatytą sandarą);
- 2) terminų bankas (terminų bazių rinkinys su organizacine sandara terminologiniams duomenims įrašyti, apdoroti ir platinti);
- 3) elektroninis terminynas (tai apibendrintas terminų banko pavadinimas).

Iš elektroninių terminynų, kuriuose galima rasti lietuvių kalbą, pranešėjos paminėti tokie šaltiniai:

- Daugiakalbė ES terminų bazė *IATE*,⁵ kuri skirta ES dokumentų terminijos ištekliams kaupti. Joje yra 1,4 mln. daugiakalbių straipsnių – 8,4 mln. terminų 23 oficialiomis ES kalbomis.
- Terminų bankas *EuroTermBank*,⁶ skirtas ES narių ir kitų Europos šalių terminologijos plėtrai. Jame yra 600 tūkst. terminų straipsnių, o juose 1,5 mln. terminų įvairiomis kalbomis (iš viso įtraukta daugiau kaip 27 kalbos).

³ Be kitų tyrimų, kurie aprašyti KLC svetainėje (žr. <http://donelaitis.vdu.lt>), KLC dalyvavo 6 BP projekte RESIST (Resilience for Survivability in IST <http://www.resist-noe.org>), kurio metu buvo parengtas kompiuterių saugumo ir patikimumo srities konferencijų pranešimų ir straipsnių rinkinių tekstynas, iš kurio, naudojantis Saarlundo universiteto Informacinių sistemų taikymo institute (IAI <http://www.iai.uni-sb.de/iaide/index.htm>) sukurtais įrankiais buvo gauti specifiniai terminai ir jų pagrindu sukurta ontologija, kuria naudojantis galima klasifikuoti tekstus.

⁴ Šis pranešimas pristatytas 2009 m. spalio 3 d. seminare „Elektroniniai žodynai ir terminynai“, surengtame Matematikos ir informatikos institute, žr. <http://www.ministudija.mii.lt/>.

⁵ Žr. <http://iate.europa.eu>.

⁶ Žr. <http://www.eurotermbank.com>.

- Europos švietimo sistemų tezauras *TESE*,⁷ kuris skirtas Europos švietimo ir politikos terminams kaupti. Čia įtraukti 15 kalbų terminai.

Yra sudaryti keli elektroniniai vieno žodyno terminynai: „Aiškinamasis darbuotojų saugos ir sveikatos terminų žodynas“, „Astronomijos enciklopedinis žodynas“, „Enciklopedinis kompiuterijos žodynas“, „Fizikos terminų žodynas“, „Socialinės apsaugos terminų žodynas“, „Radioelektronikos terminų žodynas“.

Iš lietuviškų elektroninių terminynų S. Meškauskienė paminėjo Lietuvos standartizacijos departamento terminų duomenų bazę,⁸ kurioje surinkti Lietuvos standartų terminai, jų apibrėžtys ir kitų kalbų atitikmenys. Kitas minėtas lietuviškas terminynas – Lietuvos Respublikos Seimo terminų glossarijus,⁹ kuris sudarytas iš *Eurovoc* – daugiakalbio, daugia-dalykio ES terminų žodyno; VDIC – vertimo, dokumentacijos ir informacijos centro terminų duomenų bazės; STD – Statistikos departamento lietuvių-anglų kalbų statistikos terminų žodyno.

Taip pat paminėtas Lietuvos Respublikos TERMINŲ BANKAS.¹⁰ Jo paskirtis – užtikrinti nuoseklų sunormintų lietuvių kalbos terminų vartojimą, ypač Lietuvos Respublikos teisės aktuose, sukurti bendrą įvairių valstybės institucijų informacinę sistemą, prie kurios galėtų prisijungti ir duomenis teikti kiti juridiniai ir fiziniai asmenys, kuria galėtų laisvai naudotis ne tik Lietuvoje, bet ir užsienyje dirbantys įvairių sričių specialistai. Šiame banke sukaupta apie 150 tūkst. terminų straipsnių.

Paskutinis S. Meškauskienės pranešime paminėtas daugiausia terminų turintis lietuviškas elektroninis terminynas – Matematikos ir informatikos instituto lietuvių kalbos terminų bazė (Lietuvių kalbos terminynas),¹¹ skirta įvairioms terminijos tvarkymo ir vartojimo reikmėms. Jame yra apie 280 tūkst. terminų iš 27 įvairių terminų žodynų.

Pranešime buvo akcentuoti elektroninių terminynų privalumai: dideli terminologinių duomenų rinkiniai; įvairios paieškos galimybės; universali terminologinių duomenų pateikimo tvarka; galimybė matyti to paties termino vartojimą skirtinguose žodynuose ar dalykinėse srityse, o tai palengvina terminų norminimą; lizdinis arba sisteminis terminų rikiavimas parodo tam tikros srities terminų sistemos būklę, padeda nustatyti trūkstamus terminus; operatyviau galima redaguoti terminologinius duomenis, skelbti naujus terminus; neribota terminų sklaida visuomenėje.

Iš pranešimo galima suprasti, kad lietuviški elektroniniai terminynai yra sudaryti terminų žodynų pagrindu. Nepaminėta, kad nors vienas terminynas būtų sudarytas naudojant atitinkamos srities tekstyną. Taigi nors galima naudoti kelis lietuviškus terminynus, kurie prieinami daugybei žmonių, vis dėlto lieka atviras tokių terminynų kokybės klausimas, t. y. ar be patogesnio ir didesnei žmonių daliai prieinamesnio naudojimo yra kokių kitų elektroninių ter-

⁷ Žr. <http://eacea.ec.europa.eu/portal/page/portal/Eurydice/TESEHome>.

⁸ Žr. http://www.lsd.lt/typo_new/index.php?id=159.

⁹ Žr. http://www3.lrs.lt/pls/term/term_p.

¹⁰ Žr. <http://terminai.vlkk.lt:10001/pls/tb/tb.search>.

¹¹ Žr. <http://www.terminynas.lt>.

minynų privalumų lyginant su tradiciniais terminų žodynais. Taip pat neakcentuotas tekstynų privalumas tikslesnėms terminų apibrėžtims nustatyti, terminų daugiareikšmiškumui išvengti. Be to, neturint specializuotų tekstynų, kuriuose išryškėtų terminų vartoseną, negalima taikyti automatinio terminų nustatymo įrankių, palengvinančių terminografų darbą ir leidžiančių gauti objektyvesnius duomenis.

Taigi iš šios lietuviškų terminynų apžvalgos matyti didelis poreikis kitais būdais nustatyti ir analizuoti terminus, t. y. terminams nustatyti, apibrėžti naudoti tekstynus, kompiuterines programas.

Švietimo ir mokslo tekstynas

Šiame straipsnyje norima pristatyti švietimo ir mokslo terminų automatinio nustatymo galimybes. Pasirinkta ši sritis, nes iki šiol Lietuvos Respublikos terminų bankui nėra pateikti švietimo ir mokslo terminai. Tirti švietimo ir mokslo terminus paskatino ir dalyvavimas Nacionalinės lituanistikos plėtros 2009-2015 m. programos finansuojamame projekte „Švietimo ir mokslo terminų automatinis identifikavimas (ŠIMTAI)“.

Pagrindinis projekto tikslas buvo specialaus tekstyno pagrindu parengti aiškinamąjį švietimo ir mokslo terminų žodyną lietuvių kalba su atitikmenimis anglų, vokiečių ir prancūzų kalbomis, tačiau neskyrus reikiamo finansavimo apsiribota tik švietimo ir mokslo tekstyno sudarymu. Tikimasi, kad ateityje pratęsus projektą bus galima įvykdyti pirminėje projekto paraiškoje numatytus darbus.

Kaip minėta, projekto metu sukauptas maždaug 2 mln. Žodžių tekstynas.¹² Į klausimą, kiek tekstų reikia surinkti, kad būtų galima pasikliauti iš jų sudaryto tekstyno duomenimis, galima atsakyti taip: „Kitaip nei bendrojo pobūdžio tekstynų atveju, kokybė čia kur kas svarbesnė nei kiekybė“ (Meyer, et al., 1996). Paprastai specialieji tekstynai dėl suprantamų priežasčių būna mažesni nei bendrojo pobūdžio. Tačiau nors ir būdami mažesni, jie sudarytini iš išties, netrumpintų tekstų. Be to, jie turi apimti visą reprezentuojamąją šaką su visomis jai priklausančiomis šakelėmis, iš anksto pasirinktą terminografų. Vis dėlto nėra aiškiai pasakyta, kokio konkretaus dydžio turi būti tekstynas, kad jame būtų galima automatiškai nustatyti terminus.

Pasak Meyer et al. (1996), dėl didelių permąnų terminologijoje terminografinis tekstynas turėtų būti atviras, t. y. turėtų būti nuolat pildomas. Švietimo ir mokslo tekstyną taip pat planuojama pildyti. Tekstyną sudaro įvairūs Švietimo ir mokslo ministerijos, kitų institucijų dokumentai, apimantys švietimą, aukštąjį mokslą, technologijų ir inovacijų politiką, nes šios sritys dažnai yra susijusios. Stengtasi sudaryti subalansuotą tekstyną, nes su tekstynais dirbantys terminografai (pvz., Meyer, et al., 1996) teigia, kad labai svarbu tinkamai subalansuoti tekstyną.

Renkant dokumentus atsižvelgta į jų kilmės ir administravimo sferos institucijų principą: svarbiausiais laikytini Lietuvos mokslo politikos formavimo ir įgyvendinimo valdančiųjų institucijų – Lietuvos Respublikos Seimo, Lietuvos

Respublikos Vyriausybės, Lietuvos Respublikos Prezidento, Švietimo ir mokslo ministerijos, Ūkio ministerijos, Užsienio reikalų ministerijos – parengti dokumentai. Toliau pagal svarbą minėtinos šios dokumentus rengiančios institucijos: ES mokslo politikos institucijos, leidžiančios oficialius dokumentus lietuvių kalba – Europos Komisija, Europos Taryba, Europos Parlamentas; Lietuvos mokslo politikos įgyvendinimo tarpinės institucijos – Lietuvos mokslo taryba, Lietuvos mokslų akademija, Lietuvos valstybinis mokslo ir studijų fonas, Studijų kokybės vertinimo centras, Tarptautinių mokslo ir technologijų plėtros programų agentūra, Švietimo mainų paramos fondas, Mokslo ir studijų stebėsenos ir analizės centras; mokslo ir studijų institucijos – universitetai ir mokslo institutai; nepriklausomos ekspertinės organizacijos – Viešosios politikos analizės ir vertinimo centras, Viešosios politikos institutas ir kt.; atrinkti mokslo politikos moksliniai ir profesiniai leidiniai.

Į švietimo ir mokslo terminų tekstyną renkami dokumentai suklasifikuoti taip:

1. Įstatymai:
 - 1) LR įstatymai ir Konstitucinio Teismo sprendimai,
 - 2) ES įstatyminės galios dokumentai,
 - 3) Tarptautinės sutartys.
2. Strateginiai dokumentai:
 - 1) Mokslinių tyrimų, aukštojo mokslo ir inovacijų strategijos, EK komunikatai, ET išvados,
 - 2) Ministerijos, EK direktoratų administruojamos programos,
 - 3) ES struktūrinės paramos programos.
3. Poįstatyminiai LR ir ES aktai:
 - 1) Pirmos eilės – nurodyti įstatymuose bei strategijose ir iš jų išplaukiantys,
 - 2) Antros eilės – įvairios tvarkos, reglamentai,
 - 3) Trečios eilės – ministro ir kitų pareigūnų įsakymai.
4. Mokslo valdymo institucijų teisės aktai.
5. Mokslo ir studijų institucijų svarbiausi mokslo srities dokumentai.
6. Ekspertinės studijos, ataskaitos, mokslo politikos leidiniai.
7. Diskusijų medžiaga.

Taigi iki šiol sukauptas 2 mln. žodžių tekstynas turėtų būti pildomas ir naudojamas tolimesniems darbams, t. y. automatiškai išgauti švietimo ir mokslo terminus, šios srities terminų žodynui sudaryti. Kaip rašyta anksčiau, iš tekstynų išgauti terminai, tekstynų pagrindu sudaryti žodynai laikomi objektyvių, patikimų tyrimų rezultatais.

Automatinis terminų nustatymas

Šiandien terminografų susikurti tekstynai vis dar daugiausia naudojami informacijai apie anksčiau turėtus terminus rinkti. Žiūrima, ar jie tebevartojami, o jei taip, tai kaip, ar nepakito jų vartosenos ypatumais, o drauge su jais ir reikšmė, ar nepasikeitė jais įvardijamų sąvokų apimtis atsiradus

¹² Manoma, kad 2 mln. žodžių tekstynas nėra pakankamo dydžio, kad jame būtų galima automatiškai išgauti terminus. Plg. Anglų-italų kalbų teisės tekstynas „Bononia Legal Corpus (BOLC)“ yra maždaug 18 mln. žodžių dydžio (Rossini Favretti, et al., 2001).

naujiems terminams. Tačiau kur kas reikalingesnis tekstynas yra naujiems terminams gauti, gal todėl terminų identifikavimas ir automatinis ar pusiau automatinis jų išgavimas šiandien yra viena iš pačių populiariausių temų elektroninėje terminografijoje. Šiuo atžvilgiu terminografas skiriasi nuo leksikografo. Pastarajam nekyla klausimas, kas yra žodis, tik kokius žodžius į kokius žodynus dėti, o terminografiui toks klausimas – kas yra terminas – kyla gana dažnai. Iš tekstyno gauti ir specialiose duomenų bazėse saugomi bei tvarkomi nauji terminai pasižymi dar viena dorybe – jie turi labai tiksliai nurodytus ir datuotus šaltinius, iš kurių galima spręsti apie terminų atsiradimą, paplitimą bei išnykimą (Marcinkevičienė, 2000).

Automatinio terminų atpažinimo taikymo sritys yra labai įvairios, tačiau pagrindinės jų yra trys:

- 1) praktinė ir teorinė terminologija,
 - 2) automatinio vertimo sritis,
 - 3) kalbinės ir dalykinės informacijos elektroniniuose tekstuose paieška (Zeller, 2005, p.29).
- Nepaisant to, kad kitose šalyse nemažai pasiekta automatiškai išgaunant terminus, vis dėlto Cabré Castellví et al. (2001, p.54) straipsnyje minima keletas problemų:
- 1) sudėtinių terminų atpažinimas, t. y. kur prasideda ir baigiasi terminologinė frazė;
 - 2) sudėtinių terminų nustatymas, t. y. ar diskurso vienetas yra terminologinė frazė ar netermininis žodžių junginys;
 - 3) identifikavimas, ar leksiniai vienetai yra terminai ar paprasti žodžiai;
 - 4) ar atpažintas terminas priklauso tiriamos srities žodynui;
 - 5) automatinis rečiau vartojamų terminų nustatymas.

I. Zeller (2005) mini dar vieną iš aktualesnių problemų – termino kontekstualumą,¹³ kuris suprantamas kaip termino savybė, atspindinti termino ir vartosenos konteksto santykį. I. Zeller (2005) teigia, kad kontekstualus terminas yra tas, kurio vartosenos kontekstą galima lengvai numanyti, žvelgiant į automatinio būdu atpažintą terminų sąrašą (pvz.: *pašarai, natris, bulvė, azotas, žemės ūkis, veterinarija, vaistų gaminy, herbicidai, žvejybos įrankis*). Tokie galimi terminai yra laikomi gana patikimais, taigi jų atranka tolesniam tyrimui nėra problemiška. Nekontekstualiems terminams yra būdinga keisti termininę reikšmę, kitaip sakant, skirtinguose kontekstuose tokie terminai gali įgauti naujų termininių reikšmių ir prarasti senąsias. Vadinasi, termino kontekstualumo tyrimu iš dalies sprendžiama pagrindinė deskriptyviosios terminologijos keliamą termino atskyrimo nuo netermino problema.

Penktoje dalyje rašyta, kad iki šiol sukauptas maždaug 2 mln. žodžių tekstynas. Toliau pristatomas automatinio švietimo ir mokslo terminų nustatymo eksperimentas rėmėsi mažesniu tekstų kiekiu: 299 866 žodžių formų (18 172 skirtingų žodžių).

¹³ Automatinis terminų atpažinimas yra deskriptyviosios terminologijos metodika, o deskriptyviojoje terminologijoje laikomasi nuostatos, kad terminas, jo statusas labai priklauso nuo vartosenos konteksto (Zeller, 2005).

Minėta, kad automatiškai terminus galima nustatyti su kuriais nors kalbai specialiai sukurta ar adaptuota programa, taip pat su jau egzistuojančiomis (nebūtinai terminams išgauti skirtomis pvz., *TRADOS*) programomis. Šiam eksperimentui naudota programa *Word Smith Tools*, su kuria automatiškai nustatyti stabiliai vartojami dvižodžiai ir trižodžiai junginiai¹⁴, sudaryti dažniniai žodžių sąrašai.

Vis dėlto šis tyrimas taip pat atskleidė būtinybę sukurti lietuvių kalbai pritaikytą automatiškai terminus išgaunantį įrankį, nes su stabiliesiems junginiams nustatyti skirtomis programomis (kurios nėra pritaikytos konkrečiai kalbai) gautų rezultatų peržiūrėjimas ir tvarkymas reikalauja daug laiko. Pvz., I. Zeller (2005) naudota ir disertacijoje aprašyta programa AUTOTERM¹⁵ ypač vertinama tyrėjų dėl to, kad antrajame automatinio terminų atpažinimo tarpsnyje yra gerokai palengvinamas galimų terminų sąrašų gryninimo darbas. Jis yra šiek tiek automatizuotas: tyrėjas peržiūri galimų terminų sąrašus tiesiogiai AUTOTERM programos lange ir pašalina negalimus terminus; AUTOTERM įsidėmi kiekvieną negalimą terminą ir, pašalinus vieną negalimą terminą, automatiškai pašalina visus junginius su tuo terminu. Taip terminų apdorojimo darbas darosi lengvesnis ir greitesnis, nes galimų terminų sąrašas automatiškai sutrumpėja.

Kaip minėta anksčiau, šiame straipsnyje pristatomame eksperimente buvo naudota ne speciali lietuvių kalbos terminams automatiškai išgauti skirta programa, todėl tarp automatiškai nustatytų žodžių junginių daug buvo dažnai vartojamų neprasmingų junginių, pvz.: *į visuomenės meninį, pasinaudoti teise iš, dabar yra laikomas*. Tarp tokių neprasmingų ir stabiliaisiais junginiais nelaikytinų buvo

- 1) nepilnųjų junginių, pvz.: *su studijomis, mokytojui ir, lavinimo mokykloje*;
- 2) junginių, tiksliau, žodžių kombinacijų, sudarytų iš morfologinių samplaikų ar modalinių žodžių, pvz.: *taip pat, iš anksto, iš dalies, turi būti*.

Tokių neprasmingų junginių ištrynimasis iš galimų terminų sąrašo užėmė daug laiko.

Automatinę žodžių junginių (nebūtinai terminų) atranką galima pagerinti pritaikius morfologinį anotatorių¹⁶ (pirminiame terminų nustatymo etape ši programa nenaudota). Tokiu būdu būtų galima automatiškai atmesti tuos jungi-

¹⁴ Kadangi pristatomi tik bandomojo tyrimo rezultatai, apsiribota analizuoti tik iš dviejų ar trijų žodžių sudarytus junginius, tarp kurių ieškota galimų terminų. Neabejotina, kad terminų yra ir tarp ilgesnių žodžių junginių. Kita vertus, stabiliai vartojamus ilgesnius nei iš keturių žodžių sudarytus junginius dažnai sudaro pavadinimai, kurie nelaikytini terminais. Kad pagrindiniai terminai pakliūva tarp dvižodžių junginių, įrodo ir I. Zeller (2005) duomenys: jos tirtame Europos Sąjungos dokumentų tekстыne dvižodžiai terminai sudaro 51 %, trižodžiai – 17 %, keturžodžiai – 2 %.

¹⁵ Ši programa yra vienintelė, kuri geba atpažinti terminus lietuvių kalbos tekste ir pagal kalbinius, ir pagal statistinius požymius. Abiejų metodų pritaikymas toje pačioje programoje yra didelis privalumas. Statistinėmis terminų atpažinimo programomis atpažįstami bet kokios kalbos terminai, tačiau tokios terminų atpažinimo programos neatsižvelgia į kalbinius galimo termino požymius, neatpažįsta žodžio kaitybinių formų, todėl atskiras žodžių formas vertina kaip atskirus žodžius. Be to, jos visiškai nepaiso sintaksės taisyklių ir dėsnų, taigi neskiria žodžių junginių, tik išskiria dažniausiai greta einančių žodžių ir laiko juos junginiais (Zeller, 2005).

¹⁶ Apie lietuvių kalbos morfologinį anotatorių žr.: http://donelaitis.vdu.lt/main.php?id=4&nr=7_1.

nius, kurie baigiasi jungtukais, prielinksniais, dalelytėmis, nes visi tokios struktūros junginiai yra nepilnieji, taigi nelaikytini terminais. Dar vienas automatinis apribojimas, kuris pagerintų automatinę terminų atranką, būtų atsisakymas tų junginių, kuriuos sudaro tik nekaitomos kalbos dalys. Atsargiai reikia vertinti tikrinius daiktavardžius: jie dažnai įeina į terminus (kita vertus, su tikriniais daiktavardžiais sudaryti įmonių, įstaigų, organizacijų pavadinimai nelaikytini terminais), bet retai kada keli tikriniai daiktavardžiai sudaro terminus.

Norint tiksliau automatiškai nustatyti terminus, neužtenka pritaikyti taisyklės, kurios atmeta anksčiau minėtus ir panašius junginius. Taip pat reikalingos taisyklės, kurios padėtų tarp daugybės žodžių junginių rasti galimus terminus. Tai galėtų būti, pvz., tokios taisyklės:

- būdvardis (paprastai įvardžiutinis) + daiktavardis (*formaliojo švietimo, aukštojo mokslo*);
- daiktavardžio kilmininkas + daiktavardis (*studijų programos, atestacijos komisija*);
- būdvardis + daiktavardžio kilmininkas + daiktavardis (*profesinio mokymo programos, vyriausiojo mokslo darbuotojo, ikimokyklinio ugdymo auklėtoju*);
- būdvardis + būdvardis + daiktavardis (*valstybinė aukštoji mokykla*);
- daiktavardžio kilmininkas + daiktavardžio kilmininkas + daiktavardis (*magistrantūros studijų programos*).

Peržiūrėjus ir atsisakius neprasmingų žodžių junginių, iš 31 096 dvižodžių junginių liko 9 172, t. y. 29,5%; iš 29 369 trižodžių junginių liko 7 161, t. y. 24,4 proc. Palikti tik prasmingi junginiai, iš kurių turbūt nemaža dalis nelaikytina švietimo ir mokslo terminais, bet galima laikyti kolo-kacijomis, pvz.: *darbo diena, vartojamos sąvokos, valstybės turtas, finansavimo forma, teisės aktai, viešojo administravimo, rajono savivaldybė, naftos produktų*.¹⁷

Analizuoti junginiai, kuriuos sudarė įvairios kaitybinės žodžių formos. Junginius sugrupavus ne pagal dažnumą, o pagal abėcėlę, šalia atsidūrė skirtingos to paties junginio kaitybinės formos, taip geriau galima matyti to paties junginio vartoseną ir nuspręsti, ar jį laikyti terminu (žr. lentelę).

1 lentelė. Junginių su žodžiu *atestacija* dažnumai švietimo ir mokslo tekстыne

Žodžių junginys	Dažnumas
ATESTACIJOS DOKUMENTAI	3
ATESTACIJOS DOKUMENTAIS	2
ATESTACIJOS DOKUMENTŲ	8
ATESTACIJOS DOKUMENTUOSE	2
ATESTACIJOS FINANSAVIMAS	3
ATESTACIJOS IŠLAIDAS	3

¹⁷ Atskirti tam tikros srities terminus nuo kolokacijų – viena iš svarbiausių terminologijos problemų. Čia reikalinga tiriamos srities eksperto pagalba. Kadangi aprašomame eksperimente nesikonsultuota su švietimo ir mokslo ekspertais (svarbiausia buvo bandyti automatiškai nustatyti terminus, o ne juos suklasifikuoti ar apibrėžti), negalima pateikti duomenų, kiek tarp automatiškai nustatytų dvižodžių ir trižodžių junginių buvo švietimo ir mokslo terminų.

ATESTACIJOS KOMISIJA	37
ATESTACIJOS KOMISIJA	9
ATESTACIJOS KOMISIJAI	9
ATESTACIJOS KOMISIJAS	2
ATESTACIJOS KOMISIJOJE	2
ATESTACIJOS KOMISIJOS	96
ATESTACIJOS LAIKOTARPIS	2
ATESTACIJOS NUOSTATAI	8
ATESTACIJOS NUOSTATŲ	9
ATESTACIJOS NUOSTATŲ:	2
ATESTACIJOS NUOSTATUS	6
ATESTACIJOS PAŽYMĖJIMAS	8
ATESTACIJOS PAŽYMĖJIME	6
ATESTACIJOS PAŽYMĖJIMŲ	15
ATESTACIJOS PAŽYMĖJIMUS	5
ATESTACIJOS PRIEŽIŪRA	2
ATESTACIJOS PRIEŽIŪRĄ	2
ATESTACIJOS PROCEDŪRA	2
ATESTACIJOS PROGRAMA	4
ATESTACIJOS PROGRAMĄ	13
ATESTACIJOS TIKSLAI	3
ATESTACIJOS TVARKĄ	3
ATESTACINIO LAIKOTARPIO	2

Dvižodžių junginių (nebūtinai švietimo ir mokslo terminų) dažnumas nuo 1552 (*Lietuvos Respublikos*), 761 (*nustatyta tvarka*), 729 (*profesinio mokymo*) iki 2 (*žemdirbystės institutas, žemės sklypas, žmogiškasis kapitalas*).

Trižodžių junginių dažnumas ne toks didelis: dažniausi rasti junginiai yra *mokslo ir studijų* (607), *Lietuvos Respublikos Vyriausybės* (285), *neformaliojo suaugusiųjų švietimo* (117), o rečiau (2 kartus pavartoti junginiai): *mokyklinio brandos egzamino, kito ilgalaikio turto, įgalio to asmens prašyme* ir kt.

Pastebėta, kad ilgesni nei trižodžiai yra institucijų, dokumentų pavadinimai, bet jie nelaikyti terminais, taigi pagrindiniai terminai turėjo patekti tarp dvižodžių ir trižodžių junginių.

Tarp trižodžių junginių pakliuvo daug dvižodžių, prie kurių buvo prisišlieję nekaitomi žodžiai, pvz.: *ir meninė veiklą, ir verslo centras, į profesinio mokymo, į brandos atestato, kaip vyresniojo mokytojo*.

Kaip jau minėta, tekstynas labai svarbus termino kontekstui tirti. Toliau pateiktuose pavyzdžiuose matyti, kaip iš konteksto galima apibrėžti terminus (paimtos ištraukos iš tekstų, kuriuose galima matyti, kokiame kontekste vartojami terminai *atestacijos nuostatai, valstybės paskola*).

MOKYTOJŲ IR PAGALBOS MOKINIUI SPECIALISTŲ (IŠSKYRUS PSICHOLOGUS) **ATESTACIJOS NUOSTATAI**¹⁸

I. BENDROSIOS NUOSTATOS

1. Mokytojų ir pagalbos mokiniui specialistų (išskyrus psichologus) **atestacijos nuostatai** (toliau – Nuostatai) reglamentuoja mokytojų, dirbančių pagal ikimokyklinio, priešmokyklinio, pradinio, pagrini-

¹⁸ Pavyzdžiuose žodžių junginiai pajuodinti mano – E. R.

dinio, vidurinio, specialiojo ugdymo, profesinio mokymo, neformaliojo švietimo programos, ir pagalbos mokiniui specialistų (specialiųjų pedagogų, tiflopedagogų, surdopedagogų, logopedų, socialinių pedagogų) (išskyrus psichologus) (toliau kartu vadinama – mokytojai ir pagalbos mokiniui specialistai) kompetencijos ir praktinės veiklos įvertinimo bei kvalifikacinių kategorijų suteikimo tvarką. Suaugusiųjų neformaliojo švietimo mokytojams ir Lietuvos Respublikos piliečiams, dirbantiems užsienio valstybių lietuvių švietimo įstaigose mokytojais ar pagalbos mokiniui specialistais, šie nuostatai yra rekomendacinio pobūdžio.

4. Lietuvos mokslo ir studijų institucijų studentai, vykdamy studijuoti užsienį pagal Erasmus programą, gali gauti **valstybės paskolas**, kurios skiriamos vadovaujantis Valstybės paskolų aukštųjų mokyklų studentams suteikimo, administravimo ir grąžinimo tvarka, patvirtinta Lietuvos Respublikos Vyriausybės 2002 m. gruodžio 21 d. nutarimu Nr. 2073 (Žin., 2002, Nr. 124-5648, 2006, Nr. 104-3977).

Lietuvos valstybinis mokslo ir studijų fondas Vyriausybės nustatyta tvarka teikia ir administruoja **valstybės paskolas** studentams, stipendijas doktorantams, kitą finansinę paramą studentams ir mokslininkams, taip pat Vyriausybei pavedus dalyvauja įgyvendinant programinį konkursinį mokslinių tyrimų ir eksperimentinės (socialinės) plėtros finansavimą iš tam tikslui skirtų valstybės biudžeto ir kitų teisėtai gautų lėšų.

1. Studentai gali gauti **valstybės paskolas**:

- 1) šio įstatymo 75 straipsnio 1 dalyje nurodytai studijų įmokai mokėti, jeigu ji nebuvo sumokėta valstybės biudžeto lėšomis;
- 2) gyvenimo išlaidoms;
- 3) dalinėms studijoms pagal tarptautines (tarpžinybines) sutartis.

Nors pristatomo eksperimento rezultatai nėra geri, vis dėlto į kylančias abejones, kiek realu automatiškai nustatyti terminus lietuvių kalbos tekstuose, galima atsakyti remiantis kitu 2005 m. atliktu tyrimu (žr. Marcinkevičienė, et al., 2005). Šiame tyrime aprašyti trys pagrindiniai etapai norint sudaryti lietuviškų frazių žodyną: išgauti kolokacijas iš tekstyno, kolokacijų grandinėles transformuoti į frazes, aprašyti frazių modelius (čia labai svarbi morfologinė analizė). Šie etapai taikytini ir automatiniam terminų išgavimui, nes daugumą keliažodžių terminų galima laikyti kolokacijomis.

Minėtame straipsnyje (Marcinkevičienė, et al., 2005) teigiama, kad apie 70 proc. tekstyno (tyrime naudotas iš 100 mln. žodžių sudarytas „Dabartinės lietuvių kalbos tekstynas“) sudaro kolokacijos (žinoma, tik dalis tų kolokacijų laikytinos terminais) ir naudojant statistinį kolokacijų ištraukimo metodą (šiame straipsnyje naudotas *Gravity counts* metodas¹⁹) galima nustatyti kolokacijų grandinėles.

I. Zeller (2005) teigia, kad automatinio būdu tekstyne gali būti aptinkami ir vienažodžiai, ir sudėtiniai terminai. Didesnė dalis terminų atpažinimo programų yra pritaikyta nustatyti sudėtinius terminus, o ne vienažodžius, nes paprastai apie tris ketvirtadalius visų tekste nustatytų terminų yra sudėtiniai.

Automatiškai nustatčius sudėtinius galimus terminus, kyla sunkumų, kaip juos atskirti nuo kolokacijų. Pasak I. Zeller (2005) statistiniais metodais pagrįstos terminų atpažinimo programos sunkiai išsprendžia minėtą problemą.

Taigi anksčiau atlikti tyrimai atskleidžia, kad automatinis kolokacijų ir terminų išgavimas naudojant automatines programas yra įmanomas. O vienažodžius terminus galima

nustatyti iš tekstyno automatiškai sudarytų dažninių žodžių sąrašų.

Apibendrinant automatinio terminų nustatymo iš tekstyno galimybes, galima teigti, kad nors parengiamieji automatinio terminų nustatymo darbai (tekstyno sudarymas, terminus automatiškai nustatančios programos sukūrimas ar adaptavimas, testavimas) užima daug laiko, bet turint sukurta automatiškai terminus nustatančią programą galima greičiau išanalizuoti didelius tekstų kiekius, taip pat analizė yra objektyvesnė; pastebimi tie dažnai vartojami stabilieji junginiai, kurie nėra apibrėžiami kaip terminai, bet dažna jų vartoseną rodo esant galimais terminais. Be to, turint tekstyną ir susidarius analizuojamų žodžių ar junginių konkordansus ir juose ieškant terminų, matyti išsamus kontekstas, kuriuo remiantis galima lengviau bei objektyviau apibrėžti terminus.

Automatinio terminų atpažinimo rezultatų efektyvumas ir praktinė nauda pirmiausia priklauso nuo turimų technologijų. Automatinis terminų atpažinimas kol kas nėra vien tik automatinio būdu atliekamas procesas, taigi jis neįmanomas be tyrėjo neautomatinio būdu atliekamo darbo. Automatinio terminų atpažinimo metodika sėkminga yra tik tada, jei derinamas automatinis ir neautomatinis darbas (Zeller, 2005).

Išvados

Apžvelgus elektroninės terminologijos situaciją pasaulyje ir Lietuvoje, matyti, kad Lietuvoje turima nemažai elektroninių resursų, kuriuose galima rasti lietuviškų terminų, bet iki šiol neturime nė vieno žodyno, elektroninio terminyno, sudaryto remiantis tekstynais.

Iš straipsnyje pristatytų bandymų automatiškai išgauti lietuvių kalbos terminus ir stabiliuosius junginius matyti, kad, norint greitai ir tiksliai išgauti tam tikros srities terminus, reikalingos programos, pritaikytos lietuvių kalbai. Kol kas nesukurta automatiškai terminus atpažįstančių programų, skirtų lietuvių kalbai. Norint sukurti tokio pobūdžio programas, reikalingi nemaži žmogiškieji ir materialieji resursai, tačiau šių programų svarba neabejotina. Kaip vienas iš privalumų paminėtina tai, kad lietuvių kalbai skirta automatiškai terminus nustatančia programa galima nustatyti įvairių sričių terminus; reikalingi tik skirtingų sričių tekstynai.

Net ir turint automatiškai terminus išgaunančius įrankius, susiduriama su problemomis, kaip nustatyti, ar žodis yra terminas, ar žodis yra analizuojamos srities terminas, kur prasideda ir baigiasi ilgesni nei vienažodžiai terminai. Padėti išspręsti šiuos neaiškumus gali tam tikros srities ekspertai. Taigi terminų nustatymas negali būti visiškai automatinis procesas.

Literatūra

1. Cabré Castellví, M. T., Estopà Bagot, R., Palatresi J. V., 2001. Automatic Term Detection. A Review of Current Systems. Recent Advances in Computational Terminology, pp.53-88.
2. Daudaravičius, V., Marcinkevičienė, R., 2004. Gravity Counts for the Boundaries of Collocations. *International Journal of Corpus Linguistics*, Volume 9 (2), pp.321-348.
3. Marcinkevičienė, R., 2000. Terminografija ir tekstynas. *Terminologija*, 6, pp.5-22.

¹⁹ Apie šį metodą žr. Daudaravičius, et al. (2004).

4. Marcinkevičienė, R., Grigonytė, G., 2005. Lexicogrammatical Patterns of Lithuanian Phrases. *The Second Baltic Conference on Human Language Technologies*, pp.299-304.
5. Meyer, I., Macintosh, K., 1996. The Corpus from a Terminographer's Viewpoint. *International Journal of Corpus Linguistics*, Volume 1 (2), pp.257-285.
6. Rossini Favretti, R., Tamburini, F., Martelli, E., 2001. Words from Bononia Legal Corpus. *International Journal of Corpus Linguistics*, special issue, pp.13-33.
7. Zeller, I., 2005. *Automatinis terminų atpažinimas ir apdorojimas*. Daktaro disertacija. Vytauto Didžiojo universitetas.

Erika Rimkutė

Possibilities of Automatic Identification of Lithuanian Terms

Summary

The article briefly presents the terminological situation in Lithuania and all over the world. It analyses how the work of people involved in this field has been changed because of the technological and computational development, emergence of corpora, etc. The issues when it became instrumental to analyze the terms in corpora and when the first computational programs automatically determining the terms have been introduced are also taken into the account. The special emphasis in this article is put on the importance of corpora for analyzing terms, defining them and structurizing dictionaries of the terms. The stages of the terminograph how it analysis mono or bilingual corpora are described in the article as well. The paper also introduces the electronical corpora or data bases which are currently available for finding Lithuanian terms and a project which has encouraged research of the possibilities of automatic identification of educational and scientific terms. The project is called Automatic Identification of Educational and Scientific Terminology (ŠIMTAI) and is supported by the Lithuanian Research Council's National Development Program of Lituanistics (2009-2015). Additionally, the results of one experiment which supplies the process of automatic terms' identification, which reveals uprising problems and possibilities to identify Lithuanian terms automatically, are reviewed in this paper.

Straipsnis įteiktas 2009 12
Parengtas spaudai 2010 03

Apie autorę

Erika Rimkutė, dr., Vytauto Didžiojo universiteto Lietuvių kalbos katedros docentė, Kompiuterinės lingvistikos centro mokslo darbuotoja.

Mokslinės veiklos sritys: tekstynų lingvistika, kompiuterinė lingvistika, automatinė morfologinė analizė ir sintezė, morfologinis daugiareikšmiškumas, automatinė sintaksinė analizė.

Adresas: Vytauto Didžiojo universitetas, Humanitarinių mokslų fakultetas, Kompiuterinės lingvistikos centras, K. Donelaičio g. 52-206, 44244 Kaunas
El. paštas: e.rimkute@hmf.vdu.lt

