Automatinis švietimo ir mokslo terminų nustatymas lingvistiniais metodais

ERIKA RIMKUTĖ
Vytauto Didžiojo universitetas

ESMINIAI ŽODŽIAI: tekstynas, švietimo ir mokslo terminai, lingvistiniai metodai, antraštinė forma, gramatinė forma

ĮVADAS

Tekstynų lingvistika ir kompiuterinė lingvistika Lietuvoje jau skaičiuoja antrą dešimtmetį – šios mokslo šakos pradžią galima sieti su Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro (žr. http://tekstynas.vdu.lt) įsteigimu 1994 m. Parengta keletas tekstynų, automatinių kalbos analizės ir sintezės programų[1]. Vis dėlto tekstynų lingvistikos principai dar menkai taikomi terminologijoje ir terminografijoje.

2010 m. straipsnyje *Lietuvių kalbos terminų automatinio nustatymo galimybės* (Rimkutė 2010) apžvelgta, kiek tekstynai naudojami sudarant įvairius terminologijos darbus Lietuvoje. Lietuvių kalbos terminų galima rasti keliose tarptautinėse duomenų bazėse, terminų bankuose; nemažai parengta elektroninių terminynų ar duomenų bazių[2], bet visi lietuviški elektroniniai terminynai iš esmės yra sudaryti terminų žodynų pagrindu. Daugelis lietuviškų žodynų parengti remiantis gana nedideliais tekstų kiekiais, kalbine intuicija, taigi galima teigti, kad Lietuvoje vis dar vyrauja preskriptyvusis terminų gavimo ir aprašymo būdas.

Straipsnyje norima pristatyti švietimo ir mokslo terminams nustatyti ir aprašyti skirto projekto[3] rezultatus, susijusius su automatiniu terminų nu-

[1] Išsamų lituanistinių išteklių, parengtų iki 2011 m., sąvadą galite rasti http://sruoga.vdu.lt/lituanistiniai-skaitmeniai-istekliai/lituanistiniu-istekliu-sarasas.

[2] Nuorodos pateiktos minėtame straipsnyje.

[3] Tai Lietuvos mokslo tarybos finansuojamas projektas *Švietimo ir mokslo terminų automatinis identifikavimas* (ŠIMTAI 2) (sutarties nr. LIT-2-44). Šio projekto svarbiausi tikslai: 1) sukurti ir išbandyti automatinio terminų atpažinimo lietuvių kalbos tekstyne metodologiją kaip naują terminijos tvarkybos instrumentą, 2) sukurti specialųjį švietimo ir mokslo tekstyną, 3) specialiojo tekstyno pagrindu parengti aiškinamąjį švietimo ir mokslo terminų žodyną ir šios srities ontologiją.

statymu, jų aprašymu. Automatiniam terminų nustatymui skirto projekto idėja kilo pastebėjus, kad esami norminamieji šaltiniai[4] yra gana neišsamūs, orientuoti į pačius bendriausius švietimo politikos terminus ir beveik nė kiek neapima mokslo politikos srities, juose nespėjama pateikti naujų, iš praktinio darbo kylančių terminų; terminų normintojai nespėja jų sunorminti. Tradicinis norminimas vis dėlto yra gana subjektyvus, todėl vartoseną ir normas reikia pagrįsti tekstynų lingvistikos metodais. Tad šiame straipsnyje bus rašoma apie Lietuvoje dar gana naują terminų atrankos ir analizės būdą, kuris galėtų gerokai palengvinti terminologų darbą, suteikti daugiau objektyvumo.

AUTOMATINIS TERMINŲ NUSTATYMAS

Norint automatiškai nustatyti ir apibrėžti tam tikros srities terminus, pasirinkta viena sritis – mokslas ir švietimas. Sukauptas 4 mln. žodžių apimties švietimo ir mokslo (toliau ŠM) tekstynas, kurį sudarant orientuotasi į dvi pagrindines temas: 1) mokslinius tyrimus ir 2) ugdymą. Iš mokslinių tyrimų srities daugiausia dėmesio skirta mokslinių tyrimų politikai ir aukštojo mokslo politikai. Iš ugdymo srities imtas povidurinis ugdymas, o didžiausias dėmesys skirtas aukštajam mokslui ir tęstiniam mokymui(si). Kaip papildoma sritis įtrauktas ir profesinis rengimas. Sudarant tekstyną siekta aprėpti įvairių žanrų ir tipų tekstus, pvz.: įstatymus, kitus aukščiausių valdžios institucijų politinius dokumentus (Lietuvos Respublikos Seimo, Lietuvos Respublikos Vyriausybės, Lietuvos Respublikos Prezidento nutarimus, pareiškimus, Konstitucinio Teismo sprendimus ir kt.); strateginius dokumentus; įstatymų įgyvendinamuosius aktus; mokslo ir ugdymo politikos įgyvendinimo institucijų ir projektų dokumentus; mokslo ir studijų institucijų dokumentus; informacinius leidinius, srities būklės tyrimus, vertinimus, galimybių studijas; spaudos pranešimus, diskusijų medžiagą.

Turint reprezentatyvų, gerai subalansuotą tiriamosios srities tekstyną[5], galima automatiškai nustatyti terminus. Šiam tikslui dažniausiai naudoja-

[4] Galima paminėti Lietuvos Respublikos terminų banką, L. Jovaišos *Enciklopedinį edukologijos terminų žodyną* (2007), TESE tezaurą [žiūrėta 2012-05-20], prieiga per internetą: http://eacea.ec.europa.eu/education/eurydice/documents/tese/pdf/teselt_005_alphabetic.pdf.

[5] Tekstynų reprezentatyvumas, dydis – vieni iš sudėtingiausių tekstynų lingvistikos klausimų. Manytina, kad ŠM tekstynas gana gerai reprezentuoja pasirinktą sritį. Labiau diskutuotina dėl jo dydžio, pvz., anglų–italų kalbų teisės tekstynas *Bononia Legal Corpus* (BOLC) yra maždaug 18 mln. žodžių dydžio (Rossini Favretti et al. 2001).

mi trejopi metodai: statistiniai, lingvistiniai ir hibridiniai statistinių bei lingvistinių metodų modeliai. Atlikti keli eksperimentai, kuriais siekta nustatyti tinkamiausius terminų nustatymo metodus. Išbandyti šie statistiniai metodai: 1) *WordSmith Tools* programos funkcija *Keyword Clusters*; 2) automatinio mokymosi metodas KEA; 3) kolokacijų išgavimo metodas naudojant įrankį LICE (plačiau žr. Grigonytė et al. 2011). Atlikus šiuos tyrimus, padaryta išvada, kad tinkamiausi lietuvių kalbai yra lingvistiniai metodai. Toliau trumpai paaiškinti šių metodų ypatumai.

Pagrindinis statistinių terminų nustatymo sistemų privalumas yra tas, kad joms nereikia didelių duomenų bazių su žmonių nurodytais terminų pavyzdžiais. Be to, jos yra nepriklausomos nuo kalbos. Statistinių sistemų pagrindinis veiklos principas remiasi tuo, kad dažnai kartu vartojami žodžiai yra susiję ir traktuotini kaip galimi terminai.

Lingvistiniai metodai dažniausiai remiasi morfologinėmis pažymomis, žodžių tvarka. Kad būtų galima pritaikyti lingvistinius metodus, reikalingi morfologiškai anotuoti tekstynai. Naudojant lingvistinius terminų nustatymo metodus, dažniausiai nustatoma, kokių gramatinių formų gali būti terminai. Nustatant gramatines formas, reikia atsižvelgti į kalbų ypatybes. Pavyzdžiui, analitinėms kalboms svarbu nustatyti terminų kalbos dalį. Lietuvių kalbos terminų kalbos dalis taip pat labai svarbi, bet reikia žinoti, ir kokių kitų gramatinių kategorijų informacija būtina automatiškai nustatant terminus. Pastebėta, kad lietuvių kalbai dar svarbios skaičiaus, linksnio kategorijos, o giminės kategorija, automatiškai nustatant terminus, nėra itin reikšminga[6].

Tyrinėtojų (pvz., Bowker 2002) nustatyta, kad beveik neįmanoma pritaikyti kitoms kalboms sukurtų terminus nustatančių įrankių, ypač tų, kurie naudoja lingvistinius metodus. Pavyzdžiui, tipiška anglų kalbos terminų struktūra yra bdv. + dkt., dkt. + dkt., o prancūzų kalbai būdingi modeliai yra dkt. + bdv., dkt. + prl. + dkt. Taigi norint taikyti lingvistinius terminų atpažinimo metodus, reikia sukurti programą, kurioje būtų atsižvelgta į tam tikros kalbos gramatinę sistemą.

Toliau šiame straipsnyje bus pristatyti bandymai automatiškai nustatyti lietuvių kalbos terminus naudojant lingvistinius metodus. Atkreiptinas

[6] Giminė tampa aktuali tais atvejais, kai atsiranda linksnių sinkretizmas, pvz., *mokslo darbuotojų*: automatinės morfologinės analizės programa formą *darbuotojų* atpažįsta ir kaip vyriškosios, ir kaip moteriškosios giminės daiktavardžio daugiskaitos kilmininką, todėl kaip antraštinę formą nurodo ir *mokslo darbuotojas*, ir *mokslo darbuotoja*. Kaip terminas turėtų būti pateikta vyriškosios giminės forma, t. y. *mokslo darbuotojas*.

dėmesys, kad galimiems terminams (toliau GT)[7] nustatyti naudota programa sukurta VDU Kompiuterinės lingvistikos centre. Šioje programoje naudotos lingvistinės taisyklės, kuriose nurodyta, kokių kalbos dalių junginiai ar žodžių formos analizuotini.

Kaip minėta, naudojant lingvistinius metodus, reikalingi morfologiškai anotuoti tekstynai, todėl pirmiausia sukauptas ŠM tekstynas buvo morfologiškai anotuotas naudojant VDU Kompiuterinės lingvistikos centre sukurtą morfologinį anotatorių[8]. Čia pateiktas automatiškai morfologiškai anotuoto teksto pavyzdys:

<word="PROJEKTŲ" lemma="projektas" type="dktv vyr.gim dgsk K">[9]
<space>
<word="FINANSAVIMO" lemma="finansavimas" type="dktv vyr.gim vnsk K">
<space>
<word="SĄLYGŲ" lemma="sąlyga" type="dktv mot.gim dgsk K">
<space>
<word="APRAŠAS" lemma="aprašas" type="dktv vyr.gim vnsk V">
<p>
<word="I" lemma="I" type="rom skaič">
<sep=".">
<space>
<word="BENDROSIOS" lemma="bendras" type="bdvr teig nelygin.l įvardž
mot.gim vnsk K">
<space>
<word="NUOSTATOS" lemma="nuostata" type="dktv mot.gim vnsk K">[10]
<p>
<number="1">
<sep=".">
<space>
<number="2007">
<sep="–">

[7] Įvairius metodus naudojančios automatiškai terminus nustatančios programos dažniausiai atpažįsta galimus terminus, terminus kandidatus (žr. Zeller 2005), nors tarp automatiškai nustatytų žodžių ar junginių neretai pasitaiko visiškai neprasmingų žodžių samplaikų ar prasmingų neterminiinių žodžių junginių, pvz., kolokacijų. Iš šių žodžių ir junginių tikruosius terminus paprastai nustato tam tikros srities ekspertas.

[8] Išsamiau žr. http://tekstynas.vdu.lt/page.xhtml?id=morphological-annotator-how-to-use.

[9] *Word* nurodo konkrečią tekste pavartotą žodžio formą, *lemma* – antraštinę formą (lemą), *type* – išsamią morfologinę informaciją.

[10] Atkreiptinas dėmesys, kad morfologiškai anotuojama automatiškai, todėl pasitaiko ir klaidų, pvz., šiame pavyzdyje dėl linksnių sinkretizmo nurodyta netinkama forma: vietoj moteriškosios giminės daugiskaitos vardininko nurodytas moteriškosios giminės vienaskaitos kilmininkas. Išsamiau apie netikslumus, atsiradusius dėl morfologinio anotatoriaus specifikos, žr. skyrių *Automatinio terminų nustatymo problemos*.

<number="2013">
<space>
<word="m" lemma="m" type="sntrmp">
<sep=".">
<space>

Automatiškai GT nustatančios programos veikimo principą trumpai ir paprastai būtų galima nusakyti taip: analizuojamas morfologiškai sužymėtas tekstas, ieškoma susijusių žodžių grandinėlių. Pirmasis kriterijus – pagal sužymėtas teksto, sakinių skirtis atrinkti nepertrauktus žodžių junginius. Dalis teksto skirčių yra pažymėtos kaip skyrybos ženklai, kitos gali būti pažymėtos kaip HTML žymos, pvz., pastraipos žyma <p>, tarpo tarp žodžių žyma <space> ir pan. Toliau reikia parengti tam tikras lingvistines taisykles, kurios būtų pritaikytos programoje. Nustatant švietimo ir mokslo terminus, parengtos ir pritaikytos šios pagrindinės taisyklės[11]:

1. Analizuoti tik tie junginiai, kuriuose yra bent vienas daiktavardis. Jei GT yra dvižodis ar ilgesnis, tai kitos kalbos dalys, esančios tokiame junginyje, gali būti būdvardžiai ir dalyviai.

2. Analizuoti tik tie junginiai, kurių paskutinis žodis būtinai turi būti daiktavardis[12].

3. Dvižodžiuose GT neanalizuoti tikriniai daiktavardžiai, nes jie dažniausiai įeina į asmenų, įmonių, institucijų ir kt. pavadinimus, kurie nelaikytini srities terminais (nustatant ilgesnius terminus į GT terminų sąrašus įtraukti ir tikriniai daiktavardžiai).

4. Iš junginių išmesti skaitmenimis užrašyti skaičiai.

5. Neanalizuoti tie žodžiai ar junginiai, kuriuos sudaro bent vienas morfologinio anotatoriaus neatpažintas žodis. Tai dažniausiai užsienio kalbos intarpai, su klaidomis parašyti žodžiai, sutrumpėjusios žodžių formos[13].

Pritaikius tokias taisykles, buvo nustatyti GT, iš kurių ekspertas[14] atrinko ŠM terminus. Šie terminai analizuoti išsamiau, jų struktūra aprašyta skyriuje *Švietimo ir mokslo terminų struktūra.*

[11] Taisyklės yra universalios ir iš esmės tinka bet kurios srities terminams automatiškai nustatyti.

[12] Galima ir kitokia terminų struktūra, pvz., daiktavardis su prišlieta bendratimi, daiktavardis su prielinksnine konstrukcija. Šio tipo terminų nedaug, todėl jie neanalizuoti. Ateityje tęsiant tyrimą reikėtų paanalizuoti ir šiame straipsnyje neįtrauktas terminų struktūras.

[13] Pritaikius tam tikrus GT atrankos apribojimus, galima prarasti vertingos informacijos. Šiame tyrime neskaičiuota, kiek neatpažinta tikrųjų terminų. Naudojant automatinio terminų atpažinimo programas, reikia susitaikyti su tuo, kad dalis informacijos bus prarasta, bet automatiškai nustatyta informacija bus lengviau apdorojama, gauti junginiai bus gramatiškai taisyklingesni.

[14] Šio projekto ekspertas yra lituanistas Giedrius Viliūnas, turintis didelę pedagoginę, administracinę, vadybinę patirtį švietimo ir mokslo srityje.

TERMINŲ ATRANKA

Pritaikius ankstesniame skyriuje aprašytą metodą, programą ir joje naudojamas taisykles, buvo nustatyta apie 11 tūkst. GT, sudarytų nuo vieno iki penkiolikos žodžių[15]. Susidūrus su tokiu dideliu duomenų kiekiu, buvo pritaikyti keli atrankos kriterijai. Pirmas – terminų ilgis: atsižvelgus į bendrąsias terminų struktūros tendencijas, nuspręsta analizuoti ne ilgesnius nei septyniažodžius terminus. Antras kriterijus – terminų pavartojimo dažnumas. Pasirinkta toliau išsamiau analizuoti vienažodžius GT, kurie ŠM tekstyne pavartoti ne mažiau kaip 20 kartų; dvižodžius GT, kurie pavartoti ne mažiau kaip 10 kartų; trižodžius GT, pavartotus ne rečiau nei 8 kartus; keturžodžius GT, pavartotus bent 6 kartus; penkiažodžius GT, kurių dažnumas bent 4 kartai; šešiažodžius GT, pavartotus bent 3 kartus, ir septyniažodžius GT, kurių dažnumas ne mažesnis nei 2 kartai.

Automatiškai skaičiuojant GT pasirodymo dažnumą, remtasi žodžių antraštinėmis (žodyninėmis) formomis, t. y. lemomis. Pritaikius šiuos atrankos kriterijus nustatyti terminai tirti išsamiau (jų struktūra pristatyta kitame skyriuje).

Ekspertui peržiūrėjus pagal anksčiau minėtus kriterijus nustatytus GT, atrinkti tikrieji ŠM terminai. Jų pasiskirstymas pagal ilgį pateiktas 1 lentelėje.

1 lentelė. Švietimo ir mokslo terminų pasiskirstymas pagal ilgį

Terminus sudarančių žodžių skaičius	Terminų kiekis, skaičius	Terminų kiekis, %
1	155	19,87
2	474	60,77
3	125	16,03
4	22	2,82
5	4	0,51
Iš viso	780	100,00

[15] Čia pateikti keli ilgiausi rasti terminai: *aprašo punkte nurodytais atvejais fondo direktoriaus sudarytos komisijos sprendimu valstybės remiamos paskolos grąžinimas paskolos gavėjui* (15 žodžių; čia praleisti skaičiai, t. y. kelintas punktas); *kredito įstaigos sudarytoje valstybės remiamos paskolos sutartyje nurodytu paskolos gavėjo gyvenamosios vietos adresu* (13 žodžių); *universiteto studentų studijų įmokų mokėjimo tvarka mokslo metams nustatoma rektoriaus įsakymu* (11 žodžių). Akivaizdžiai matyti, kad šie teksto fragmentai negali būti laikomi galimais terminais.

Iš lentelės duomenų matyti, kad apie 97 % visų ŠM terminų sudaro iš 1–3 žodžių sudaryti terminai. Tarp šešiažodžių ir septyniažodžių GT švietimo ir mokslo terminų nerasta, todėl jie toliau neanalizuoti.

ŠVIETIMO IR MOKSLO TERMINŲ STRUKTŪRA

Šiame skyriuje išsamiai aprašyta automatiškai nustatytų ir eksperto atrinktų ŠM terminų struktūra, pateikti jų gramatiniai modeliai. Terminai aprašyti nuo trumpiausių, t. y. vienažodžių, iki ilgiausių, t. y. penkiažodžių.

1. Vienažodžiai terminai. Iš viso nustatyti 7635 vienažodžiai GT, iš jų 1311 tekstyne pavartota daugiau nei 20 kartų. Toliau išsamiau analizuoti tik šie žodžiai. Iš 1311 žodžių 155 galima laikyti ŠM terminais, t. y. apie 20 %. Dažniausiai tekstyne pavartoti šie ŠM terminai: *studentas* (4208[16]), *universitetas* (3959), *studijos* (1562), *mokslas* (1508), *sritis* (1403).

2. Dvižodžiai terminai. 4 mln. žodžių tekstyne rasti 145 468 dvižodžiai GT. Kaip rašyta anksčiau, toliau analizuoti tik tie dvižodžiai GT, kurių dažnumas ne mažesnis nei 10 kartų – tokių terminų rasta 4889. Peržiūrėjus GT, galutiniame ŠM terminų sąraše liko tik 474 dvižodžiai terminai ir jie sudaro didžiąją visų analizuojamų terminų dalį – daugiau nei du trečdalius (žr. 1 lentelę.). Dvižodžius terminus galima suskirstyti į tris tipus, atsižvelgiant į jų gramatines formas.

1) *dkt. K.* + *dkt.* tipo ŠM terminų yra 250, t. y. 52,7 % visų dvižodžių terminų.

Dažniausi šios struktūros ŠM terminai, pavartoti tekstyne, yra *studijų programa* (1432), *studijų pakopa* (306), *studijų kryptis* (303), *mokslo darbuotojas* (289), *studijų kokybė* (253).

2) *bdv.* + *dkt.*[17] junginiai sudaro 200 ŠM terminų (iš jų 42,2 % dvižodžių terminų). Dažniausi ŠM terminai yra *aukštoji mokykla* (2160), *moksliniai tyrimai* (918), *aukštasis mokslas* (730), *nuotolinis mokymas* (453), *profesinis mokymas* (403).

Bdv. + *dkt.* tipo junginiai laikomi GT tuo atveju, kai būdvardis su daiktavardžiu suderintas gimine, skaičiumi ir linksniu. Priešingu atveju gau-

[16] Skliaustuose nurodomas to termino dažnumas 4 mln. žodžių apimties ŠM tekstyne.

[17] Prie dvižodžių terminų modelių linksniai nenurodomi, nes būdvardžiai ir dalyviai derinami su daiktavardžiais, o daiktavardžiai paprastai būna vienaskaitos vardininko linksnio.

nami gramatiškai netaisyklingi[18] junginiai, pvz., *glaudesnis universitetų, svarbiausias teisės, svarbiausių mokymosi, geriausio metų,* arba junginys yra taisyklingas, nors dažnai nepilnas, ir tai nėra ŠM terminai, pvz., *didžiausias pasaulyje, dinamiškiausia pasaulyje, palanku studentams.*

3) *dlv.* + *dkt.* junginių kaip ŠM terminų yra tik 24, t. y. 5 %. Dažniausi ŠM terminai: *baigiamasis darbas* (162), *pasirenkamasis dalykas* (109), *suaugusiųjų*[19] *švietimas* (65), *grįžtamasis ryšys* (63), *taikomieji tyrimai* (45).

3. Trižodžiai terminai. Iš viso nustatytas 119 881 trižodis GT. Išsamiau analizuoti 2438 junginiai, kurių dažnumas tekstyne ne mažesnis kaip 8 kartai. ŠM terminų rasta 125, t. y. 16,03 % visų terminų. Dažniausi ŠM terminai: *aukštojo mokslo institucija* (251), *profesinio mokymo įstaiga* (205), *valstybės remiama paskola* (149), *bendrojo lavinimo mokykla* (147), *profesinio mokymo institucija* (77).

Trižodžiai ŠM terminai yra septynių gramatinių modelių. Tarp trižodžių ŠM terminų dažniausi *bdv. K.* + *dkt. K.* + *dkt.*[20] tipo junginiai (jie sudaro 48 % visų trižodžių terminų), pvz.: *aukštojo mokslo institucija* (251), *profesinio mokymo įstaiga* (205), *bendrojo lavinimo mokykla* (147).

Antri tarp trižodžių ŠM terminų pagal dažnumą yra *dkt. K.* + *dkt. K.* + *dkt.* tipo junginiai (24,8 %), pvz.: *studijų krypties reglamentas* (66), *studijų krypčių aprašas* (60), *valstybės mokslo institutas* (50).

Trečias pagal dažnumą struktūrinis trižodžių ŠM terminų modelis yra *bdv. V.* + *bdv. V.* + *dkt.* (12 %), pvz.: *nacionalinė kompleksinė programa* (69), *fundamentinis mokslinis tyrimas* (61), *pirminis profesinis mokymas* (52).

Toliau pagal dažnumą galima pateikti vos po kelis pavyzdžius, turinčius struktūrinius trižodžių ŠM terminų modelius: *dkt. K.* + *dlv. V.* + *dkt.* (4 %), pvz.: *valstybės remiama paskola* (149), *valstybės pripažinta kvalifikacija* (33), *valstybės finansuojamas studentas* (21); *dkt. K.* + *bdv. V.* + *dkt.* (4 %), pvz.: *užsienio aukštoji mokykla* (41), *kolegijos akademinė taryba* (32), *doktoranto mokslinis vadovas* (20); *dlv. V. / K.* + *dkt. K.* + *dkt.* (3,2 %), pvz.: *baigiamasis kvalifikacijos vertinimas* (36), *nemokama studijų vieta* (13), *taikomojo pobūdžio tyrimai* (13); *bdv. V.* +

[18] Gramatiškai taisyklingais junginiais šiame darbe laikomi tokie junginiai, kuriuose priklausomieji žodžiai suderinti su pagrindiniu, aiškūs sintaksiniai ryšiai tarp žodžių junginio dėmenų.

[19] Prie dalyvių įtraukti ir sudaiktavardėję dalyviai.

[20] Paskutinio daiktavardžio linksnis nenurodomas, nes jis dažniausiai yra vardininkas.

dlv. V. (arba K., jei dalyvis sudaiktavardėjęs) + *dkt.* (2,4 %), pvz.: *konkursinis mokomasis dalykas* (15), *neformalus suaugusiųjų mokymas* (13), *mokslinis tiriamasis darbas* (12); *dlv. V.* + *bdv. V.* + *dkt.* (1,6 %), pvz.: *taikomoji mokslinė veikla* (20), *užsakomieji moksliniai tyrimai* (8).

4. Keturžodžiai terminai. Iš viso rastas 77 961 keturžodis GT, daugiau nei 7 kartus pavartotų junginių – 760. Iš jų ekspertas nustatė 22 ŠM terminus (jie sudaro 2,82 % visų terminų). Patys dažniausi šioje terminų grupėje junginiai yra *mokslo ir studijų institucija* (568), *mokslo ir technologijų parkas* (62), *bendroji nacionalinė kompleksinė programa* (40), *darbo rinkos profesinis mokymas* (40).

ŠM tekstyne pavartoti keturžodžiai terminai yra devynių skirtingų gramatinių modelių. Kadangi keturžodžių terminų pavartota nedaug, todėl ir kiekvieną modelį iliustruoja vos po kelis ar tik po vieną pavyzdį. Čia pateikti pagal dažnumą sugrupuoti visi keturžodžių ŠM terminų struktūriniai modeliai ir nurodoma po vieną tą modelį iliustruojantį pavyzdį:

bdv. V. + *dkt. K.* + *dkt. K.* + *dkt.* (22,7 % visų keturžodžių terminų): *tarptautinė mokslo duomenų bazė* (13);
bdv. V. + *bdv. K.* + *dkt. K.* + *dkt.* (18,2 %): *bendrasis universitetinio lavinimo dalykas* (8);
bdv. K. + *dkt. K.* + *dlv. V.* + *dkt.* (13,6 %): *specialiųjų poreikių turintis mokinys* (8);
dkt. K. + *jng.* + *dkt. K.* + *dkt.* (13,6 %): *mokslo ir studijų institucija* (568);
bdv. K. + *dkt. K.* + *bdv. V.* + *dkt.* (9,1 %): *aukšto lygio moksliniai tyrimai* (18);
dlv. V. + *bdv. V.* + *dkt. K.* + *dkt.* (9,1 %): *pripažinta tarptautinė duomenų bazė* (11);
bdv. V. + *bdv. V.* + *bdv. V.* + *dkt.* (4,5 %): *bendroji nacionalinė kompleksinė programa* (40);
dkt. K. + *bdv. K.* + *dkt. K.* + *dkt.* (4,5 %): *valstybės mokslinių tyrimų įstaiga* (31);
dkt. K. + *dkt. K.* + *dkt. K.* + *dkt.* (4,5 %): *technologijos mokslų studijų sritis* (28);

5. Penkiažodžiai terminai. Nustatyti 49 768 penkiažodžiai GT, daugiau nei 5 kartus pavartotų junginių rasta 601, iš jų ŠM terminų – 4, t. y. 0,51 % visų terminų. Iš tokio nedidelio kiekio terminų sunku daryti apibendrinimus apie jų struktūrą, todėl terminai struktūriškai neskirstyti. Visi ŠM tekstyne pavartoti penkiažodžiai terminai: *moksliniai tyrimai ir eksperimentinė plėtra* (124), *moksliniai tyrimai ir technologinė plėtra* (67), *moksliniai tyrimai ir technologijų plėtra* (13), *aukšto lygio mokslinių tyrimų centras* (7).

6. Šešiažodžiai ir septyniažodžiai terminai. Jau minėta, kad tarp šešiažodžių ir septyniažodžių GT terminų nerasta nė vieno ŠM termino, nors iš viso nustatyti 30 298 šešiažodžiai GT (iš jų daugiau nei 3 kartus pavartotų junginių rasta 750) ir 17 815 septyniažodžių GT (iš jų daugiau nei 2 kartus pavartotų junginių rasta 2608).

ŠVIETIMO IR MOKSLO TERMINŲ PALYGINIMAS SU LIETUVOS RESPUBLIKOS TERMINŲ BANKO TERMINAIS

ŠM tekstyne pavartotus terminus galima palyginti su Lietuvos Respublikos terminų banke pateiktais aprobuotais švietimo ir mokslo terminais[21], jų yra 74. Šių terminų struktūra labai įvairi: nuo vienažodžių iki iš 18 žodžių sudarytų junginių. Labai ilgų terminų, sudarytų iš 15 ir 18 žodžių, rasta tik po vieną. Ilgiausias pateiktas terminas - *žemdirbių, miško savininkų, vietos veiklos grupių narių, kitų kaimo gyventojų ir jų konsultantų neformaliojo tęstinio profesinio mokymo programa.* Diskutuotina, ar programos pavadinimas laikytinas terminu.

Palyginus ŠM tekstyno ir Terminų banko švietimo ir mokslo terminų struktūrą, matyti, kad abiejuose šaltiniuose patys dažniausi vienažodžiai–trižodžiai terminai. Tiesa, Terminų banko vienažodžiai–trižodžiai terminai sudaro 86,5 %, o ŠM tekstyne pavartoti tokio paties ilgio terminai – 97 % (žr. 1 lentelę). Terminų banke keturžodžių terminų yra kiek daugiau nei pristatomame tekstyne, penkiažodžių nepavartota, iš 6, 7, 8 ir 9 žodžių sudarytų terminų Terminų banke vos po vieną ar du, iš 10 žodžių sudarytų terminų nėra, bet pateikti jau minėti iš 15 ir 18 žodžių sudaryti terminai (žr. 2 lentelę).

Abejotina dėl keleto Terminų banke pateiktų terminų priskyrimo švietimo ir mokslo sričiai, pvz., *lietuvis, pasaulio lietuvių bendruomenė, reemigracija, tautybė.* Iš gauto terminų sąrašo matyti, kad švietimo ir mokslo sritis nėra visiškai apimta, pvz., kaip profesinio ugdymo terminas pateiktas *traktorininkas*, bet nėra *kombainininko, virėjo, staliaus* ir pan., ypač trūksta mokslo politiką atspindinčių terminų. Taigi šiame straipsnyje aptariamo tyrimo metu nustatyti terminai turėtų bent šiek tiek užpildyti švietimo ir mokslo srities terminijos spragą.

[21] Aprobuotų švietimo ir mokslo srities terminų sąrašas 2010 m. gautas iš Valstybinės lietuvių kalbos komisijos. Lyginant tekstyno ir terminų banko duomenis remiamasi būtent šiuo sąrašu.

2 lentelė. Lietuvos Respublikos terminų banke pateiktų švietimo ir mokslo terminų pasiskirstymas pagal ilgį

Terminus sudarančių žodžių skaičius	Terminų kiekis, skaičius	Terminų kiekis, %
1	16	21,6
2	31	41,9
3	17	23,0
4	3	4,1
5	0	0
6	1	1,4
7	1	1,4
8	1	1,4
9	2	2,7
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	1	1,4
18	1	1,4
Iš viso	74	100

AUTOMATINIO TERMINŲ NUSTATYMO PROBLEMOS

Įvairiais metodais nustatant terminus, kyla dvi pagrindinės problemos: 1) kokiõs formos yra gramatiškai taisyklingi terminai, ypač tai aktualu nustatant antraštinę formą; 2) kaip atskirti bendrąsias frazes, bendruosius terminus nuo analizuojamos srities terminų. Su panašiomis problemomis susiduria ir kitų šalių mokslininkai (plg. Cabré Castellví et al 2001).

Dažniausiai antrą problemą galima išspręsti tik neautomatiškai, t. y. tos srities ekspertas turi peržiūrėti GT ir nustatyti, kurie terminai skiriami analizuojamai sričiai. Galima pritaikyti ir tam tikrus filtrus, t. y. sudaryti neanalizuotinų žodžių (angliškai *stop list*) sąrašą ir jį naudoti atmetant analizuojamos srities terminais negalinčius eiti žodžius. Toks filtravimas buvo pritaikytas ir šiame tyrime. Keletas pavyzdžių: ŠM terminais nelaikyti tie žodžiai ar junginiai, kurių paskutinis žodis yra laiką nurodantis

daiktavardis, pvz.: *metai, mėnuo, diena, savaitė, rytas, vakaras, sausis, vasaris, pirmadienis, antradienis* ir pan. Terminais nelaikytini matą reiškiantys žodžiai, pvz.: *milimetras, metras, centimetras, kilometras, mylia, gramas, kilogramas, tona*.

Galima atmesti junginius, kuriuose yra bendriesiems administraciniams terminams būdingų žodžių junginių, pvz.: *įstatymo straipsnio, įstatymo straipsnio dalyje, įstatymų nustatyta (tvarka), įsakymo punktu patvirtintu, teisės aktų nustatyta tvarka, respublikos vyriausybės, respublikos valstybės, nuostatų punktas, nutarimų punktas, ministro įsakymas, ministro nustatyta tvarka, ministro patvirtintas, ministro sudarytas*.

Bandant kuo tiksliau automatiškai atskirti ŠM terminus nuo bendrųjų, buvo sudaryti du lyginamieji tokios pačios apimties kaip ir ŠM tekstynas, t. y. 4 mln. žodžių, tekstynai. Pirmasis iš lyginamųjų tekstynų sudarytas iš įvairių žanrų ir stilių tekstų; jo sandara panaši į *Dabartinės lietuvių kalbos tekstyną* (įeina publicistika, grožinė literatūra, administracinė literatūra, mokslinė literatūra). Antrasis lyginamasis tekstynas sudarytas iš administracinių tekstų, bet ne ŠM srities. Kelta prielaida, kad kuo terminas bendresnis, tuo jis dažniau vartojamas bendrajame ir (ar) administraciniame tekstyne, o kuo terminas specifiškesnis, tuo jis retesnis bendrajame ir (ar) administraciniame tekstyne. Ši hipotezė iš dalies pasitvirtino, pvz., ŠM sričiai būdingi terminai *judumas, kompetentingumas, nubyrėjimas, siekinys* nerasti lyginamuosiuose tekstynuose, o tokie terminai, kaip *studentas, universitetas, vadovėlis*, dažni ir ŠM, ir lyginamuosiuose tekstynuose. Šį tyrimą reikėtų tęsti, kad būtų galima pateikti patikimus duomenis apie tai, kiek kelių tekstynų lyginimas gali padėti bent iš dalies automatiškai atskirti bendruosius terminus nuo analizuojamos srities.

Toliau išsamiau bus aptarta pirmoji problema, t. y. kaip automatiškai nustatyti taisyklingą terminų (ypač keliažodžių) lemą, kaip į vieną lemą suvesti tekstyne skirtingomis kaitybinėmis formomis pavartotus žodžius ir junginius.

Atlikus tyrimą paaiškėjo, kad ne visada lengva nustatyti antraštinę termino formą. Automatiškai nustatant lemą, pvz., junginio *aukštoji mokykla* gaunamas rezultatas *aukštas mokykla*. Tokia forma nurodyta dėl to, kad morfologinis anotatorius būdvardžiams kaip lemą nurodo nelyginamojo laipsnio vyriškosios giminės vienaskaitos vardininko formą (išsamiau apie anotatoriaus specifiką žr. Zinkevičius 2000). Keliažodžių terminų lemoms

nustatyti labai padėtų sintaksinės analizės programa. Deja, realiai veikiančios sintaksinės analizės programos kol kas nėra[22]. Vadinasi, nustatant lemas reikalingos tam tikros papildomos taisyklės – šiuo atveju reikia nurodyti, kad jei lemą sudaro daiktavardis, tai su to daiktavardžio gimine turi būti suderintas būdvardis arba dalyvis; jei būdvardis arba dalyvis yra įvardžiuotiniai, ši kategorija turi būti išlaikyta ir lemoje.

Daug klausimų kyla dėl skaičiaus kategorijos, pvz., pirmasis termino *studentų atstovybė* dėmuo turi būti daugiskaitos formos, terminų *akademiniai įgūdžiai, auditorinės darbo valandos* visi dėmenys turi būti daugiskaitos formos, o terminų *fakulteto taryba, universiteto autonomija, bendrasis priėmimas, mokslinis leidinys* abu dėmenys yra vienaskaitos formos. Žinoma, žmogui nesunku nustatyti, kur reikalinga daugiskaitos, o kur vienaskaitos forma, bet tą padaryti automatiškai labai sudėtinga. Tiesa, tekstyne išryškėja, kad kai kurie žodžiai beveik visada vartojami tik daugiskaitos forma (pvz.: *asignavimai, duomenys, studijos, rūmai, žinios, pinigai, pareigos*).

Nustatyti tinkamą skaičiaus formą sunku ir dėl to, kad kai kurie žodžiai yra homonimai, pvz., *studija* (mokslo veikalas) ir *studijos* (studijavimas). Dėl šios priežasties automatiškai netinkamai nustatytos šių terminų lemos: *bakalauro studija, rezidentūros studija, magistro studija, dieninės studija, nuotolinės studija*. Kaip minėta, kai kurie daiktavardžiai dažniausiai vartojami tik daugiskaitos formos, nors žodyne kaip lema nurodyta vienaskaitos forma. Šis žodyninės informacijos ir realios vartosenos neatitikimas lėmė netaisyklingai nustatytas terminų *mokinio pasiekimas, švietimo resursas, praktinis gebėjimas, fizinis mokslas, akademinis užsiėmimas* lemas (turėtų būti: *mokinio pasiekimai, švietimo resursai, praktiniai gebėjimai, fiziniai mokslai, akademiniai užsiėmimai*).

Dar viena problema – *substantiva mobilia* daiktavardžiai, pvz.: *abiturientas – abiturientė, abonentas – abonentė*. Kadangi tekstyne buvo pavartotų tiek vyriškosios, tiek moteriškosios giminės formų, todėl automatiškai nurodytos abiejų giminių lemos. Kad kaip ŠM terminas būtų palikta tik vyriškosios giminės lema, reikėjo pritaikyti taisyklę, kuria remiantis vyriškosios ir moteriškosios giminės *substantiva mobilia* tipo daiktavardžiai suvesti į vyriškosios giminės lemą.

[22] Automatinės sintaksinės analizės srityje paminėtini D. Šveikauskienės 2009, E. Rimkutės 2006 darbai. Automatinės sintaksinės analizės programa kuriama VDU Kompiuterinės lingvistikos centre ir iš dalies jau taikoma analizuojant ŠM terminus.

Automatiškai nustatant lemas, kartais nurodoma netinkama žodžių tvarka, pvz., nurodytos tokios antraštinės formos: *tipų studija* (turi būti *studijų tipas*), *lygmens kvalifikacija* (turi būti *kvalifikacijų lygmuo*).

Dėl to, kad nustatant lemas nepritaikyta sintaksinė analizė, ne visada pažyminiai suderinami su pažymimuoju žodžiu, pvz., *tarptautinis (= tarptautinio) lygio mokslas*, ar atsiranda kitokių sintaksės pažeidimų, pvz.: *paskolų gyvenimo išlaidos* (= *paskola gyvenimo išlaidoms*).

Keblumų kyla ir nustatant terminų ribas, pvz., ar *dėstytojų ir mokslininkų kvalifikacijos bei kompetencijos atitiktis* yra vienas terminas, ar čia reikia skirti kelis terminus:

1) dėstytojų kvalifikacijos bei kompetencijos atitiktis,
2) mokslininkų kvalifikacijos bei kompetencijos atitiktis,
3) dėstytojų ir mokslininkų kvalifikacijos atitiktis,
4) dėstytojų ir mokslininkų kompetencijos atitiktis,
5) dėstytojų kvalifikacija,
6) mokslininkų kvalifikacija,
7) dėstytojų kompetencija,
8) mokytojų kompetencija,
9) kvalifikacijos atitiktis,
10) kompetencijos atitiktis.

Nustatant lemas ir apskritai atpažįstant terminus sunkumų kyla ir dėl morfologinio anotatoriaus specifikos, kai iš kelių galimų kalbos dalių ar gramatinių formų nustatoma netinkama. Pavyzdžiui, jei būdvardis su šalia esančiu daiktavardžiu suderintas gimine, linksniu ir skaičiumi, tai didelė tikimybė, kad toks junginys bus gramatiškai taisyklingas. Iš dviejų daiktavardžių sudaryti junginiai, kurių pirmasis dėmuo nėra kilmininko linksnio, ne visada gramatiškai taisyklingi. Juos reikia atidžiau peržiūrėti, taigi sugaišti daugiau laiko. Vadinasi, nuo to, kokia kalbos dalis nurodyta, priklauso, ar junginys pateks tarp GT, ar jis bus gramatiškai taisyklingas. Keletas šį klaidos tipą iliustruojančių pavyzdžių: *metinėms* (gali būti bdv. ir dkt., morfologinis anotatorius nurodė dkt.) *ataskaitoms; keliais* (gali būti įv. ir dkt., parinktas dkt.) *procentais; bendrais* (gali būti bdv. ir dkt., parinktas dkt.) *teiginiais; uždarus* (gali būti bdv. ir dkt., parinktas dkt.) *posėdžius*.

Dėl netinkamai nurodytos kalbos dalies, dėl to, kad žodžiai sintaksiškai neanalizuojami, iš pirmo žvilgsnio kaip gramatiškai netaisyklingas atrodo morfologiškai anotuotame tekstyne rastas junginys *žemiausias balas*, nes

pirmajam žodžiui nurodytos tokios žymos: bdv., aukšč. l., neįvardž., vyr. g., vns. V., antrajam: dkt., mot. g., dgs. G. ir kaip lema nurodyta *bala*. Taigi matyti, kad šiuo atveju blogai nustatyta daiktavardžio lema – turi būti *balas*. Vadinasi, anksčiau minėtas junginys yra gramatiškai taisyklingas.

Šiame skyriuje apžvelgti sunkumai, su kuriais dažniausiai susidurta automatiškai nustatant ŠM terminus. Apie šias problemas ir jų sprendimo būdus bus rašoma kituose darbuose.

IŠVADOS

Automatinio terminų nustatymo programą ir metodiką dar reikia tobulinti. Vis dėlto pirmieji rezultatai vertintini gana gerai. Kaip matyti iš pateiktų pavyzdžių, automatinė analizė išryškina ne tik specifines, techninės, bet ir problemas, su kuriomis susiduria ir terminologai, pvz., skaičiaus kategorijos pasirinkimas.

Patobulinus morfologinį anotatorių, pritaikius automatinės sintaksės analizės programą, automatiškai terminus būtų galima nustatyti lengviau ir patikimiau.

Nepaisant kylančių sunkumų, automatiškai terminus nustatanti programa gerokai pagreitina GT nustatymo procesą. Kitas svarbus automatinio terminų nustatymo proceso privalumas – duomenų analizės objektyvumas.

Nustatyta, kad 97 % automatiškai atrinktų ir peržiūrėtų švietimo ir mokslo terminų ilgis yra 1–3 žodžių. Daugiausia dvižodžių terminų – jie sudaro beveik 61 % visų terminų.

Dažniausi švietimo ir mokslo terminų struktūriniai modeliai yra dkt. K. + dkt., bdv. + dkt., bdv. K. + dkt. K. + dkt., dkt. K. + dkt. K. + dkt. (gali būti įvairūs paskutinio daiktavardžio linksniai).

LITERATŪRA

Bowker L. 2002: *Computer-aided translation technology – a practical introduction*. University of Ottawa Press.

Cabré Castellví M. T., Estopà Bagot R., Palatresi J. V. 2001: Automatic term detection. A review of current systems. – *Recent Advances in Computational Terminology*, 53–88.

Grigonytė G., Rimkutė E., Utka A., Boizou L. 2011: Experiments on Lithuanian Term Extraction. – *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011. NEALT Proceedings Series,* Vol. 11, 82–89.

Rimkutė E. 2006: *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne.* Daktaro disertacija.

Rimkutė E. 2010: Lietuvių kalbos terminų automatinio nustatymo galimybės. – *Kalbų studijos* 16, 71–78.

Rossini Favretti R., Tamburini F., Martelli E. 2001: Words from Bononia Legal Corpus. – *International Journal of Corpus Linguistics,* special issue, 3–33.

Šveikauskienė D. 2009: *Kalbos vientisinių sakinių automatinė sintaksinė analizė.* Daktaro disertacija.

Zeller I. 2005: *Automatinis terminų atpažinimas ir apdorojimas.* Daktaro disertacija. Vytauto Didžiojo universitetas.

Zinkevičius V. 2000: Lemuoklis – morfologinei analizei. – *Darbai ir dienos* 24, 245–273.

AUTOMATIC IDENTIFICATION OF SCIENCE AND EDUCATION TERMS USING LINGUISTIC METHODS

The paper deals with possibilities and problems of automatic Lithuanian term extraction. Specifically, linguistic methods are discussed in the domain of Education and Science.

Other researchers have shown that it is almost impossible to have language-independent term extraction tools; this is especially true for tools which are based on linguistic rules. Therefore a linguistic term extraction tool should incorporate methods that would deal with a language's grammatical system. This paper presents a tool developed at the Centre of Computational Linguistics of Vytautas Magnus University that employs linguistic rules for extracting domain-specific terminology.

In order to extract domain-specific terms automatically, some preparatory work should be completed: compilation of domain-specific corpus (a corpus of four million words has been compiled for this research), morphological annotation of the corpus, formulation of appropriate linguistic rules, and creation of methodology for filtering out irrelevant word combinations.

The paper presents the linguistic rules that have been used for the extraction of Education and Science terms and the results of the extraction procedure. The identified terms are contrasted with approved terms in the Term Bank of the Republic of Lithuania.

Some specific problems of automatic term extraction are discussed in the paper, e.g. number and case agreement of terms in a multi word term.

Gauta 2012-05-17

Erika Rimkutė
Vytauto Didžiojo universitetas
K. Donelaičio g. 52, LT-44248 Kaunas
E. paštas e.rimkute@hmf.vdu.lt