Rūta MARCINKEVIČIENĖ
Vytauto Didžiojo universitetas

TERMINOGRAFIJA IR TEKSTYNAS

Tekstynų lingvistika, su pirmuoju tekstynu atsiradusi daugiau nei prieš tris dešimtis metų, padarė perversmą bendrojoje leksikografijoje. Šiandien aiškinamieji žodynai rengiami daugiau ar mažiau remiantis tekstynais. Terminografija šiuo atžvilgiu iki dabar primiršta šalyse, toliau pažengusiose tekstynų lingvistikoje, o ką jau besakyti apie tokius naujokus kaip mes. Čia ir bus bandoma parodyti, koks svarbus, tiesiog nepamainomas dalykas yra tekstynas terminų žodynams rengti.

Kompiuterinių technologijų, podraug ir kompiuterinių tekstynų era labai aiškiai atskyrė tradicinę terminologiją ir 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 dar kai kuriuos kitus. Labai smarkiai išsiplėtus ir susipynus specialiųjų mokslų sritims, labai padaugėjo darbo terminologams, nes dirbant tradiciniais metodais vargiai begalima suspėti paskui mokslo ir technologijų pažangą. Kaip tik todėl kitaip imta žiūrėti į terminus: labiau akcentuojama ne norma, bet vartosena.

Tradicinei terminologijai, šio amžiaus pradžioje atsiradusiai dėl didžiulio technologijų proveržio, labiausiai rūpėjo įvardyti sąvokas, sutvarkyti didelę per trumpą laiką atsiradusių terminų įvairovę, susisteminti ir sunorminti atskirų mokslo šakų ir sričių terminiją. Į terminus nuo pat terminologijos kaip savarankiško mokslo atsiradimo pradžios buvo žiūrima kitaip nei į kitus žodžius. Terminologijos pradininkas Wüster'is (cituota pagal Pearson 1998: 10) teigė, jog su terminais reikia elgtis kitaip nei su bendrinės kalbos žodžiais, mat terminologija prasideda nuo sąvokų analizės. O sąvokos, žinia, egzis-

tuoja nepriklausomai ne tik nuo jas įvardijančių terminų, bet ir nuo skirtingų kalbų. Terminai yra ne kas kita kaip sąvokų vardai, idealiu atveju sąveikaujantys su pačiomis sąvokomis tokiu santykiu: viena sąvoka – vienas terminas. Terminai buvo apibrėžiami daugiausiai atsižvelgus į jais įvardytų sąvokų vietą hierarchinėje sąvokų sistemoje ir ryšius su kitomis sąvokomis. Terminijos pagrindą sudarė savos srities žinovų sukurti ir/ar sunorminti bei aprobuoti terminai, turėję aiškiai apibrėžtą, įtvirtintą, taigi ir apsaugotą reikšmę. Todėl visai nenuostabu, kad besilaikančių preskriptyviųjų nuostatų (plg. Wüster, kuriam Sein-Norm buvo kur kas svarbesnė už Ist-Norm) tradicinių terminologų visai nedomino reali atskirų mokslo šakų kalbos būklė ir autentiška terminų vartosena.

Einant nuo sąvokos prie termino ir taip sudarant kiekvienos atskiros šakos terminiją onomasiologiniu principu tuometiniams terminologams nekilo ir negalėjo kilti termino identifikavimo, tik jo kūrimo problema. Be to, anuo metu atskiros mokslo šakos dar buvo ganėtinai autonomiškos. Terminologai taip pat buvo labiau pasidaliję sritis ir išmanė palyginti izoliuotų sričių terminiją. Randantis bendriesiems dalykams, terminais ima domėtis jų atstovai, be to, į terminologiją ateina nauji žmonės, nebūtinai konkrečios mokslo srities žinovai ir/ar kalbininkai, t.y. programuotojai, terminijos duomenų bazių bei elektroninių žodynų kūrėjai ir kt. Kadangi reali kalbos vartosena ir naujų terminų skaičius jau seniai pralenkė turimus aprobuotų terminų sąrašus, dėl tradicinio kartotekų metodo gerokai nuo gyvenimo atsiliekanti sparčiai besikeičiančių mokslo šakų terminografija neleidžia naujai atėjusiems žmonėms pasikliauti seniau registruotais terminais. Jie pradedami rinkti iš tekstų, būtent tekstas tampa terminologo darbo pradžia (Rogers et al. 1994: 843). Be to, žiūrima realios jų vartosenos, kuri dažnai labai smarkiai skiriasi nuo kalbos normintojų siekiamo idealo. Čia ir tampa neaišku, ką laikyti terminu, kada žodis terminologizuojamas, kokia žodžio vartosena laikytina terminine, kaip atskirti terminą nuo netermino. Taigi terminija tampa nebe taip aiškiai apibrėžta ir izoliuota leksikos dalis. Ji sudaro gana sunkiai atskiriamą žodžių grupę, kurios centre yra prototipiniai nariai, visada suvokiami kaip

terminai, toliau eina tie, kurie tik tam tikromis aplinkybėmis, tam tikrame kontekste įgyja termino statusą. O periferijoje esama ir tokių žodžių, kurie iš pirmo žvilgsnio į terminus nėra panašūs, tačiau tokiais turėtų būti laikomi. Pavyzdžiui, tai pasakytina apie mokslo populiarinamuosiuose tekstuose aptinkamus bendrinės kalbos žodyno žodžius. Dėl šios ir kitų priežasčių terminografams labai svarbu identifikuoti terminus, išrinkti juos iš autentiškų tekstų ir aprašyti. Kaip tik todėl labai svarbūs tampa tekstynai ir darbą su jais lengvinanti programinė įranga. Dėmesys realiai vartosenai kreipia terminologus labiau aprašymo nei norminimo linkme, o preskriptyvizmo – deskriptyvizmo takoskyra dalija terminografiją į tradicinę ir šiuolaikinę.

Kodėl tekstynas

Tekstynas, būdamas terminų šaltiniu, terminografams pravartus dėl daugelio priežasčių. Už jo kūrimą ir naudojimą agituojantys autoriai nurodo pagrindines iš jų. Bene svarbiausia, kad iš vieno ir to paties tekstyno iškart gaunami trys termografams reikalingi dalykai: patys terminai, jų definicijos, t.y. dalykinė informacija apie pačią sąvoką, bei vartojimo pavyzdžiai, taigi ir vartojimo dažnumas, terminų junglumas ir pan. (Sager 1990, Meyer et al. 1996, Pearson 1998). Žinoma, tekstynu gali naudotis ir išankstinį 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: 268). Ištisų tekstų tekstynas būtinas ir dėl dalykinės informacijos, kurios būtų netenkama pasitenkinus tekstų dalimis.

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 ypatumai, o drauge su jais ir reikšmė, ar nepasikeitė jais įvardijamų sąvokų apimtis atsiradus naujiems terminams. Tačiau kur kas reikalingesnis tekstynas yra naujiems terminams gauti, gal todėl terminų identifikavimas ir automatinis ar pusiau automatinis jų gavimas 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 terminografui toks klausimas – kas yra terminas – kyla gana dažnai. Iš tekstyno gauti ir specialiose duomenų bazėse laikomi bei tvarkomi nauji terminai pasižymi dar viena ypatybe – jie turi labai tiksliai nurodytus ir datuotus šaltinius, iš kurių galima spręsti apie terminų atsiradimą, paplitimą bei išnykimą.

Jei galutinį terminografo darbo rezultatą įsivaizduosime kaip elektroninę duomenų bazę (DB) ar jos pagrindu parengtą popierinį žodyną, tai tekstynas bus svarbus visuose šios DB rengimo etapuose. Kompiuteriai apima visą terminografo darbo ciklą nuo kompiuterinių tekstynų per specialias terminų paieškos ir identifikavimo programas iki termininių duomenų bazių sudarymo, terminų aprašo ir jo platinimo elektroniniu ar tradiciniu knygos pavidalu (Ahmad et al. 1994: 267).

Terminografo darbo ciklas atrodo taip. Pirmajame etape iš elektroninių tekstų archyvo pagal tam tikrus principus sudaromas tekstynas, vėliau iš jo gaunami terminai, jie perkeliami į DB, aprašomi, t.y. atskiruose DB 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 aprašomuoju 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 (plačiau apie terminografinio aprašo ypatumus žr. Sager 1990: 142–163). Taigi tekstynas yra nepamainomas nuo pat pradžių, t.y. nuo pažinties su pačia

mokslo šaka, iki paskutinės fazės – parengto žodyninio straipsnio. Šiandieniniai terminologai labai skeptiškai žiūri į jau turimų, ne tekstynų pagrindu rengtų popierinių žodynų kompiuterizavimą, t.y. sukėlimą į duomenų bazes (Sager 1990: 141).

Dar vienas motyvas rengti specialius tekstynus terminografams yra tas, kad terminografijai dar labiau nei leksikografijai yra svarbūs autentiškos kalbos pavyzdžiai. Jie kur kas svarbesni nei introspekcija, nes terminografas, kitaip nei leksikografas, nėra specialistas, jis nekuria tos kalbos atmainos, kurią aprašo, tekstų. Susidūręs su problemišku atveju, jis negali paklausti: kaip aš vartočiau šitą žodį, o tik: ką šis žodis reiškia tos srities žinovui, kaip jį vartotų mokslininkas, technologas, ekspertas. Taigi dėl menkai pritaikomo kalbos jausmo terminografai kur kas labiau priklauso nuo tekstyno nei leksikografai.Visus tekstynų vartotojus Atkins et al. skiria į tris grupes: a) tuos, kurie domisi tekstų kalba, b) tuos, kurie domisi tekstų turiniu ir c) tuos, kurie domisi tekstais kaip tekstais, t.y. jų sandara, kūrimu ir pan. (cituota pagal Meyer et al. 1996: 269). Leksikografai priklauso pirmajai grupei, o terminografai – dviem pirmosioms, nes jiems svarbios ir kalbos, ir paties dalyko žinios. Suvokus tekstyno svarbą terminologijai ir teikiamas galimybes, būtina apibrėžti, koks turėtų būti terminografo tekstynas.

Koks tekstynas

Kaip minėta, terminografui iš tekstyno reikia dviejų vienodai svarbių dalykų: kalbinės ir dalykinės informacijos, todėl svarbu kad iš jo būtų galima gauti ir viena, ir kita. Kalbos ažvilgiu tekstynas turi būti toks, kad jame aiškiai, teisingai ir, pageidautina, su definicijomis būtų pateikti terminai. Pageidautina, kad terminų būtų kuo daugiau ir kuo įvairesnių, idealiu atveju – visi, sukurti ir vartojami konkrečios kalbinės bendruomenės. Dar svarbu, kad terminai būtų vartojami gausiuose ir įvairialypiuose kontekstuose. Dalykiniu atžvilgiu į tekstyną dėtini tokie tekstai, iš kurių išryškėtų terminais įvardijamų sąvokų turinys, tarpusavio ryšiai ir priklausomybė. Labai svarbu, kad terminai tuose tekstuose būtų apibrėžti.

Dėl nurodytų priežasčių žinovai siūlo į tekstynus įtraukti visą komunikaciniu bei pragmatiniu aspektu pateiktą galimų tekstų skalę nuo paprasčiausių iki sudėtingiausių. Pearson specialiosios kalbos tekstus pagal tikimybę juose rasti terminų skiria į keturias grupes: a) tuos, kuriuose žinovo adresatas yra žiniomis jam lygiavertis partneris – toks pats žinovas, t.y. straipsniai referuojamuose moksliniuose žurnaluose, monografijos, šiuose tekstuose terminų tankis yra pats didžiausias; b) tekstai, žinovo adresuoti menkesnio nei jis pats lygio, tačiau tos pačios srities specialistui, pavyzdžiui, įvairios instrukcijos, vadovėliai, čia terminų esama kiek mažiau; c) mokslo populiarinamieji tekstai, rašyti su specialia sritimi nesusipažinusiems žmonėms. Juose sąmoningai vengiama sudėtingų ar bet kokių terminų, mieliau aiškinama, persakoma kitais žodžiais, vartojamos analogijos ir palyginimai, taigi autorius, labai norėdamas, gali apsieiti visai be terminų; d) mokytojo – mokinio komunikaciją atitinkantys tekstai: vadovėliai, mokymo priemonės. Juose terminų esama, tik jie vartojami kiek kitaip. Svarbus šių tekstų požymis – paaiškinti terminai (Pearson 1998: 36–40).

Apibūdinti terminografinio tekstyno tekstus vien komunikaciniu pragmatiniu aspektu nepakanka. Tekstyno kokybei, jo reprezentatyvumui ne mažiau svarbūs ir kiti tekstų parametrai bei proporcijos. Į klausimą, kiek tekstų reikia surinkti, kad būtų galima pasikliauti iš jų sudaryto tekstyno duomenimis, žinovai atsako taip: „Kitaip nei bendrojo pobūdžio tekstynų atveju kokybė čia kur kas svarbesnė nei kiekybė“ (Meyer et al. 1996: 268). Paprastai specialieji tekstynai dėl suprantamų priežasčių būna mažesni nei bendrojo pobūdžio. Tačiau nors ir būdami mažesni, jie sudarytini iš ištisų, netrumpintų tekstų. Be to, jie turi apimti visą reprezentuojamąją šaką su visomis jai priklausančiomis šakelėmis, iš anksto pasirinktą terminografų. Dažnai esti sunku nubrėžti aiškią atskiras mokslo šakas skiriančią liniją, ypač jei ta šaka yra susijusi su bendraisiais dalykais ir tekstyno sudarytojams kyla pagunda eiti gilyn, t.y. įtraukti daugiau žinovų žinovams rašytų, neabejotinai tai sričiai priklausančių tekstų. Todėl terminologams siūloma kreiptis į aprašomosios srities ekspertus, kad šie patartų, kokius tekstus įtraukti, kad būtų sudarytas visomis prasmėmis reprezentatyvus tekstynas, kad drauge su neįtrauktais tekstais nebūtų praleistos

sistemiškai svarbios sąvokos bei jas įvardijantys terminai, o svarbiausia – kad nedominuotų vieno sudarytojo prioritetai (Ahmad 1993: 60). Terminografinio tekstyno atveju visai netinka kartais leksikografų bendriesiems tekstynams taikomas principas: „Ką turiu, tą dedu".

Atsižvelgiant į tekstynų tipologiją bei svarbiausius tęstinumo atžvilgiu tekstynų tipus: baigtinį – tęstinį – keistinį (monitor), terminografinį tekstyną siūloma rengti kaip antrojo, o dar geriau, pastarojo pobūdžio tekstų rinkinį. Tuomet tekstynas nuolat bus papildomas naujausiais tekstais, senuosius arba paliekant (tęstinio tekstyno atveju), arba perkeliant į archyvą, jei turimas keistinis tekstynas. Tęstinis tekstynas nuolat didėtų ir nebūtinai būtų išlaikomi iš anksto pasirinkti jo sandaros principai, nes, kintant leidybos tendencijoms, naujieji tekstai galėtų nusverti į vieną ar kitą pusę. Keistinio tekstyno proporcijos išliktų nepakitusios, o naujieji tekstai užimtų tik tiek vietos, kiek buvo iš pat pradžių skirta atitinkamo pobūdžio tekstams. Bet kuriuo atveju nuolat atnaujinti tekstyną yra būtina, nes terminai sensta kur kas greičiau už kasdienės kalbos žodžius.

Apsvarsčius bendruosius tekstyno dydžio, pobūdžio bei dalykinių jo ribų klausimus ir priėmus sprendimus, toliau reikia žiūrėti, kokius konkrečius tekstus dėti į tekstyną, kad jis būtų reprezentatyvus ir subalansuotas. Kiekvienu konkrečiu atveju reikia atsižvelgti į teksto žanrą, parašymo laiką, autorių bei kalbinį statusą ir siekti kuo didesnės įvairovės. Mat ne kas kitas, o tik žanrinė įvairovė ir plati tekstų skalė nuo rimtų iki populiarių leidžia išrinkti tiek mokslinius, formalius terminus, tarp kurių gausu autorinių neologizmų, tiek populiarius, kur kas labiau paplitusius ir šnekamojoje kalboje įsitvirtinusius terminus.

Vertinant tekstų amžių, pirmenybė teikiama naujesniems, tačiau vertingi, ypač jei žiūrėsime dalykiniu aspektu, gali pasirodyti ir seniau rašyti tekstai. Autorystė taip pat turėtų būti labai įvairi, kad nenusvertų vieno ar kito smarkiai reprezentuojamo autoriaus stiliaus ypatumai. Nors mokslinių tekstų stilistinė įvairovė ir nėra labai didelė, jei palyginsime, pavyzdžiui, su grožine literatūra, bet autoriaus braižas yra svarbus ne vien kalbos vienetų vartojimo, bet ir jų atrankos prasme. Todėl nenuostabu, kad skirtingi mokslininkai, ypač jei jie priklauso skirtingoms mokslinės minties mokykloms, vartoja ki-

tokius terminus. O terminų standartizavimo, norminimo reikalams, lygiai kaip ir neologizmų paieškai labai svarbi kuo didesnė tekstyno terminų įvairovė. Paskutinis, bet ne mažiau už kitus svarbus dalykas yra ir kalbinis tekstų statusas. Pirmenybė teikiama originaliems, o ne verstiniams, gimtakalbių autorių parašytiems, o po to dar recenzuotiems ir redaguotiems tekstams.

Tekstyno analizės priemonės terminografams

Per pastarąjį dešimtmetį buvo sukurta įvairios programinės įrangos, lengvinančios leksikografams žodžių paiešką bei kitokią tekstyno analizę. Nors šiaip jau leksikografų ir terminografų darbas ir skiriasi, tačiau jie naudojasi bemaž tomis pačiomis programinėmis galimybėmis, t.y. a) žodžių dažnumo sąrašais; b) konkordansais, t.y. tiriamojo žodžio visais pavartojimo atvejais tam tikro dydžio, paprastai vienos eilutės, kontekste; bei c) junglumo partnerių statistika (Meyer et al. 1996: 274). Šios priemonės padeda tirti tekstyną kalbiniu aspektu.

Šalia kalbinių, bendrai su leksikografais naudojamų programinės įrangos priemonių palaipsniui randasi ir specialių, tik terminografams tinkamų programų. Nuo bendrųjų programų jos skiriasi nebe kalbiniu, bet dalykiniu pobūdžiu, mat jos skirtos tekstyne esančios informacijos analizei. Šiuo atveju tekstynas traktuojamas nebe kaip terminų vartojimo kontekstų visuma, bet kaip specialiųjų atskiros srities ar mokslo šakos žinių šaltinis. Čia minėtinos kelios jų grupės: a) programos, skirtos analizuoti tekstui jame pateikiamos informacijos aspektu – prasminių žodžių, informacijos turtingų teksto vietų, kuriose nurodomi sąvokų tarpusavio ryšiai ir priklausomybė, paieška ir pan.; b) programos, atrenkančios terminografui reikalingas tekstyno vietas, pavyzdžiui, terminų definicijas; c) programos, padedančios analizuoti terminais įvardytas sąvokas, palengvinančios iš tekstyno gautų žinių tvarkymą (knowlegde acquisition systems). Pastaroji grupė, nors ir visiškai nauja bei labai perspektyvi, tačiau specialiojoje literatūroje dar per menkai jai skiriama dėmesio. Kur kas daugiau rašyta apie galimybę pusiau automatiškai rinkti tekstyne esančias terminų definicijas (Pearson 1998: 121–204).

Tiek definicijoms, tiek ir sudėtiniams terminams rinkti būtinas morfologiškai anotuotas tekstynas, t.y. toks, kurio visi žodžiai sužymėti pagal jų kalbos dalis. Pradinis darbo etapas turi būti atliktas paties leksikografo: jis skaito tekstus ir renka apibrėžtus terminus. Remdamasis ta informacija vėliau sudaro morfologinius definicijų modelius – visas definuojant gaunamas kalbos dalių sekos kombinacijas. Tuomet jau kompiuteris visame tekstyne ieško tokių žodžių derinių. Taip gaunamas dar ne terminų bei jų definicijų, bet kandidatų į apibrėžtus terminus sąrašas. Jis ir pateikiamas tos srities žinovui peržiūrėti ir išbraukti nereikalingus sakinius. Kad sumažėtų bereikalingų, atsitiktinai sutampančių morfologinių derinių kiekis, kompiuteriui nurodomi dažniausiai definicijose pasitaikantys leksiniai vienetai – dažni definicijų žodžiai. Panašiai gali būti iš tekstyno gaunami ir sudėtiniai terminai, tokia yra pirmosios kartos automatinio terminų identifikavimo programų LEXTER (Bourigault 1995) bei TERMINO (Lauriston 1994) esmė.

Be abejo, taip apdorojant tekstyną, kyla šiokių tokių problemų. Visų pirma toli gražu ne visi atrinkti žodžių junginiai yra terminai. Ši problema sprendžiama sudarant ir į kompiuterį įvedant nereikalingų žodžių sąrašą (stop-list), kurie į terminų kandidatų sąrašus netraukiami. Be to, dažnai lieka neaišku, kurie žodžiai priklauso pačiam sudėtiniam terminui, o kurie – tik jo aplinkai. Bet tie dalykai ne visada aiškūs ir renkant terminus rankiniu būdu.

Pusiau automatinis terminų atrankos būdas

Norėdami pailiustruoti paprasčiausios programinės įrangos priemonių terminografui teikiamas galimybes, analizei pasirinkome V. Labučio „*Sintaksės*“ (T.I, V., 1994) elektroninį variantą ir tyrėme tekstą su Miko Scotto lingvistui skirtų įnagių paketu *WordSmith Tools, version 2.0* (Oxford, 1997). Statistinės analizės tikslas buvo gauti keletą žodžių ar jų junginių sąrašų, iš kurių vėliau būtų galima atrinkti kandidatus į terminus ir pačius terminus.

Analizuojamo teksto statistiniai rodikliai yra tokie: visą tekstą sudaro 37 319 žodžiai ir žodžių formos, atmetus pasikartojančius, lieka

9 249 žodžiai ir jų formos. Žodingumo koeficientas (type/token ratio) yra 24,78. Visame tekste yra 2 364 sakiniai, vidutinis sakinio ilgis 15,78 žodžio.

Visų pirma buvo atliktas pats paprasčiausias veiksmas – sudarytas visų teksto žodžių dažninis sąrašas, leidžiantis tarp dažniausiai vartojamų žodžių tikėtis pamatyti terminus. Kaip matyti iš pateikiamo sąrašo fragmento, tekstą sudaro žodžiai bei žodžių formos, išrikiuoti dažnumo tvarka. Kadangi tekstas neturi gramatinių pažymų, rodančių kiekvienos žodžio formos priklausymą atskiroms kalbos dalims, tai sąrašas nėra lemuotas, taigi atskirai eina to paties žodžio skirtingos gramatinės formos. Trumpumo dėlei tiek nekaitomi žodžiai, tiek skirtingos kaitomų žodžių formos toliau tekste bus vadinamos tiesiog žodžiais.

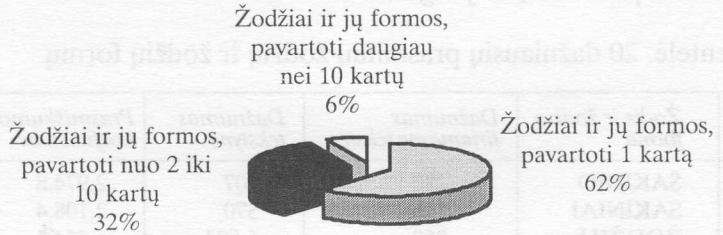
1 lentelė. 20 dažniausių teksto žodžių ir žodžių formų

Nr.	*Žodis ir žodžio forma*	*Dažnumas tiriamame tekste*	*Užimama teksto dalis (%)*
1.	IR	1,319	3.53
2.	SU	479	1.28
3.	AR	390	1.05
4.	SAKINIO	288	0.77
5.	TAI	284	0.76
6.	ŽODŽIŲ	258	0.69
7.	GALI	243	0.65
8.	O	229	0.61
9.	IŠ	206	0.55
10.	SAKINIAI	196	0.53
11.	Į	194	0.52
12.	YRA	194	0.52
13.	KAIP	187	0.50
14.	NE	182	0.49
15.	TIK	171	0.46
16.	PAGAL	170	0.46
17.	BŪTI	155	0.42
18.	JŲ	143	0.38
19.	SAKINIŲ	143	0.38
20.	JUNGINIAI	141	0.38

Kaip paprastai, tekstą sudaro palyginti nedaug žodžių, kurie vartojami daug kartų, ir daug žodžių, vartojamų retai. Vieną kartą pavartoti žodžiai sudaro 62% visų žodžių vartojimo atvejų. Paprastai jų būna pusė teksto, sulemavus sąrašą, matyt, jų tiek ir liktų, nes išvestinės žodžių formos būtų sudėtos į pagrindines. Vienkartiniai ir reti, t.y. nuo 2 iki 10, vartojimo atvejai terminologui nėra įdomūs, mat tikimybė čia rasti terminų yra nedidelė. Lieka palyginti neilga sąrašo dalis, tik 6% dažnai vartojamų teksto žodžių, iš kurių jau verta rinkti vienažodžius terminus.

Diagrama

Teksto žodžių bei žodžių formų pasiskirstymas pagal vartojimo dažnumą (pavartojimo atvejų skaičių)



Jei šie vartosenos duomenys būtų pateikti kreivės, o ne diagramos pavidalu, tai pasimatytų, kad 10 pavartojimo kartų yra labai aiški riba, slenkstis, už kurio atsidūrę žodžiai yra vartojami kur kas dažniau nei iki jo. Nuodugnesnė analizė parodė, kad iš 599 žodžių, pavartotų 10 ir daugiau kartų net 239 yra vienažodžiai terminai ar sudėtinių terminų dalys, kitaip sakant, terminografui svarbūs žodžiai. Čia priklauso kalbos dalių, linksnių, kalbos reiškinių, vienetų, kalbotyros šakų pavadinimai.

Bendrojo dažninio žodžių sąrašo privalumas yra jau minėta labai aiški vartosenos dažnumo riba, o trūkumų yra net keli. Vienas jų: sąrašą sudaro pavieniai žodžiai, o vienžodžių terminų, žinia, nėra tiek daug. Kita vertus, didžiuma naujųjų, į žodynus nepatekusių, todėl terminologus itin dominančių terminų kaip tik ir yra daugiažo-

džiai. Antrasis trūkumas: didelę sąrašo dalį užima labai dažni tarnybiniai žodžiai, terminologams taip pat neaktualūs. Pastarąjį trūkumą galima pašalinti parengus kiek kitokį, prasminių žodžių sąrašą.

Prasminiai žodžiai iš teksto gaunami šiek tiek sudėtingiau, lyginant tiriamąjį tekstą su kur kas už jį didesniu tekstynu, mūsų atveju, su 55 milijonų žodžių apimties Vytauto Didžiojo universiteto kompiuteriniu tekstynu. Jei koks žodis tiriamame tekste pasirodo sąlyginai dažniau nei vidutiniškai dideliame tekstyne, tai jis ir yra laikytinas prasminiu to teksto žodžiu. Paprasčiau sakant, jei rašoma apie kiną, tai šis žodis tekste bus minimas kur kas dažniau, todėl pateks į prasminių žodžių sąrašą. Kaip matyti iš lentelės, čia yra kur kas mažiau tarnybinių žodžių, nes jų dažnumas tekste ir tekstyne yra apylygis. Taigi į pirmą vietą iškyla unikalieji tiriamojo teksto žodžiai, mokslinio teksto atveju – terminai. Deja, šiame sąraše, kaip ir pirmajame, jie duodami pavieniui, ne junginiuose.

2 lentelė. 20 dažniausių prasminių žodžių ir žodžių formų

Nr.	*Žodis ir žodžio forma*	*Dažnumas tiriamame tekste*	*Dažnumas tekstyne*	*Prasmiškumo koeficientas*
1.	SAKINIO	288	707	2,974.8
2.	SAKINIAI	196	370	2,108.4
3.	ŽODŽIŲ	258	6,504	1,554.3
4.	SAKINIŲ	143	386	1,453.6
5.	JUNGINIAI	141	405	1,418.4
6.	DĖMENŲ	103	130	1,171.5
7.	JUNGINIŲ	103	591	909.2
8.	DĖMUO	77	124	847.4
9.	SAKINYJE	85	310	819.6
10.	DĖMENYS	73	106	815.0
11.	DĖMENS	71	84	814.2
12.	SAKINYS	91	528	801.4
13.	DĖMENIU	69	78	795.8
14.	SINTAKSĖS	68	163	704.7
15.	SAKINĮ	73	634	588.1
16.	SAKINIUS	61	315	550.1
17.	TARINIO	44	60	495.3
18.	SAKINIAIS	49	205	460.5
19.	SUDĖTINIO	42	75	455.4
20.	TARINIU	36	36	421.4

Trečioji statistiniais metodais pagrįstos programinės įrangos galimybė yra išrinkti iš tiriamojo teksto tuos žodžių junginius, kurie identiška forma yra pavartoti tekste du ar daugiau kartų. Programa leidžia aptikti net aštuonių žodžių ilgio grandines. Suprantama, tokio ilgio darinių tekste nėra daug. Kur kas daugiau esama sustabarėjusių keturžodžių, trižodžių, ypač dvižodžių junginių. Tiriamame V. Labučio tekste aštuonių žodžių junginių, pavartotų bent du ar daugiau kartų, esama 2, septynių žodžių – 6, šešių – 23, penkių – 65, keturių – 211, trijų – 756, o dviejų žodžių junginių net 2695 vienetai. Toliau pateikiamos skirtingo ilgio žodžių grandinės, juodžiau paryškinti tie junginiai, kurių dalys galėtų būti laikomos sudėtiniais terminais, o šalia duotas šios grandinės pavartojimų skaičius tekste.

AŠTUONI ŽODŽIAI

JĮ SUDARO NE MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ 2
SUDARO NE MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ FORMOS 2

SEPTYNI ŽODŽIAI

JĮ SUDARO NE MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ 2
NE MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ FORMOS 2
SUDARO NE MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ 2
ŽODŽIŲ FORMŲ TAIKYMAS PRIE KITŲ ŽODŽIŲ FORMŲ 2

ŠEŠI ŽODŽIAI

NE MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ 4
VEIKSNIU IR TARINIU EINANČIŲ ŽODŽIŲ FORMŲ 3
AKADEMINĖJE LIETUVIŲ KALBOS GRAMATIKOJE LKG III 2
DVI AR KELIOS SAVARANKIŠKŲ ŽODŽIŲ FORMOS 2
FORMŲ TAIKYMAS PRIE KITŲ ŽODŽIŲ FORMŲ 2
IR SU KAI KURIOMIS PRIELINKSNINĖMIS KONSTRUKCIJOMIS 2
SAKINIO DALĮ AR SUDĖTINIO SAKINIO DĖMENĮ 2
SU KITAIS LINKSNIAIS BEI PRIELINKSNINĖMIS KONSTRUKCIJOMIS 2
SU LINKSNIAIS IR SU PRIELINKSNINĖMIS KONSTRUKCIJOMIS 2
ŽODŽIO FORMA KAIP NORS TAIKOMA PRIE 2
ŽODŽIŲ FORMŲ TAIKYMAS PRIE KITŲ ŽODŽIŲ 2

PENKI ŽODŽIAI

MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ 4
VEIKSNIU IR TARINIU EINANČIŲ ŽODŽIŲ 4

LINKSNIAIS IR SU PRIELINKSNINĖMIS KONSTRUKCIJOMIS 3
SUDĖTINĮ SAKINĮ SUDARANČIOS PREDIKATINĖS STRUKTŪROS 3
AR KELIOS SAVARANKIŠKŲ ŽODŽIŲ FORMOS 2
ATSKIRAS PRANEŠIMAS APIE TIKROVĖS SITUACIJĄ 2
DALĮ AR SUDĖTINIO SAKINIO DĖMENĮ 2
EINA DAIKTAVARDINIO JUNGINIO PRIKLAUSOMUOJU DĖMENIU 2
FORMŲ TAIKYMAS PRIE KITŲ ŽODŽIŲ 2
GALI EITI IR SUDĖTINIO SAKINIO 2
IR TARP SUDĖTINIO SAKINIO PREDIKATINIŲ 2
IR VEIKSNIU EINANČIŲ ŽODŽIŲ FORMŲ 2
KAIP DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ FORMOS 2
KITAIS LINKSNIAIS BEI PRIELINKSNINĖMIS KONSTRUKCIJOMIS 2
LINKSNIO AR LINKSNIO SU PRIELINKSNIU 2
RYŠIO SU GRAMATINE ASMENS KATEGORIJA 2
SINTAGMINIAI SANTYKIAI IR ŽODŽIŲ JUNGINIAI 2
SKIRSTYMAS NE KARTĄ BUVO KRITIKUOJAMAS 2
SU ANTROJO LAIPSNIO ŠALUTINIU SAKINIU 2
SU KAI KURIOMIS PRIELINKSNINĖMIS KONSTRUKCIJOMIS 2
SU KITAIS LINKSNIAIS BEI PRIELINKSNINĖMIS 2
SU NAUDININKU IR SU ĮNAGININKU 2
SU VISO SAKINIO ŠALUTINIU DĖMENIU 2
TAI DAŽNIAUSIAI VEIKSMAŽODINIAI JUNGINIAI SU 2
TAIKYMAS PRIE KITŲ ŽODŽIŲ FORMŲ 2
TO PATIES LAIPSNIO ŠALUTINIAI SAKINIAI 2
TURI BENDRĄ ANTRININKĘ SAKINIO DALĮ 2
VAIDMUO VISO PRIJUNGIAMOJO SAKINIO ATŽVILGIU 2
ŽODŽIO FORMA KAIP NORS TAIKOMA 2
ŽODŽIŲ FORMŲ TAIKYMAS PRIE KITŲ 2
ŽODŽIŲ JUNGINYS AR KITA KONSTRUKCIJA 2

KETURI ŽODŽIAI

IR SU PRIELINKSNINĖMIS KONSTRUKCIJOMIS 5
IR TARINIU EINANČIŲ ŽODŽIŲ 4
KURIŲ PAGRINDINIS DĖMUO YRA 4
MAŽIAU KAIP DVIEJŲ SAVARANKIŠKŲ 4
NE MAŽIAU KAIP DVIEJŲ 4
VEIKSNIU EINANTI ŽODŽIO FORMA 4
VEIKSNIU IR TARINIU EINANČIŲ 4
DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ FORMOS 3
Į SUJUNGIAMUOSIUS IR PRIJUNGIAMUOSIUS 3
LINKSNIAIS IR SU PRIELINKSNINĖMIS 3
PAGRINDINIU DĖMENIU EINANČIO ŽODŽIO 3
SAKINĮ SUDARANČIOS PREDIKATINĖS STRUKTŪROS 3
SU GRAMATINE ASMENS KATEGORIJA 3
ŽODŽIŲ SANTYKIŲ REIŠKIMO BŪDAS 3
ANTROJO LAIPSNIO ŠALUTINIU SAKINIU 2
APIE ŠIAS SĄVOKAS ŽR 2

APIE TAI PER PRATYBAS 2
AR SUDĖTINIO SAKINIO DĖMENĮ 2
ATSKLEIDŽIA KITAS PREDIKATINIS DĖMUO 2
BENDRĄ ANTRININKĘ SAKINIO DALĮ 2
BŪDO KOKYBĖS IR KIEKYBĖS 2
BŪDVARDIS SUDARO JUNGINĮ SU 2
DAIKTAVARDINIO JUNGINIO PRIKLAUSOMUOJU DĖMENIU 2

TRYS ŽODŽIAI

GALI BŪTI IR 14
SU PRIELINKSNINĖMIS KONSTRUKCIJOMIS 11
DAUGIAU AR MAŽIAU 10
LIETUVIŲ KALBOS SINTAKSĖS 9
SAKINIAI GALI BŪTI 9
VEIKSMAŽODINIAI JUNGINIAI SU 8
DĖMENYS GALI BŪTI 7
KURIŲ PAGRINDINIS DĖMUO 7
DAIKTAVARDINIAI JUNGINIAI SU 6
NE MAŽIAU KAIP 6
PRIJUNGIMAS IR SUJUNGIMAS 6
SAKINIU GALI BŪTI 6
SAVARANKIŠKŲ ŽODŽIŲ FORMOS 6
VEIKSNIU IR TARINIU 6
DVIEJŲ SAVARANKIŠKŲ ŽODŽIŲ 5
DVIPUSĖS SĄSAJOS SANTYKIU 5

DU ŽODŽIAI

GALI BŪTI 118
IR KT. 41
LIETUVIŲ KALBOS 36
ŽODŽIŲ FORMŲ 35
TAIP PAT 34
SUDĖTINIO SAKINIO 32
SAKINIO DALIŲ 31
ŽODŽIŲ FORMOS 31
NE TIK 29
ŽODŽIŲ JUNGINIŲ 26
JUNGINIAI SU 25
PREDIKATINIŲ DĖMENŲ 24
GALI TURĖTI 23
VIS DĖLTO 23
PAGRINDINIU DĖMENIU 21
PREDIKATINĮ CENTRĄ 20
PREDIKATINIAI DĖMENYS 20
LYG IR 18
SUDĖTINIAI SAKINIAI 18

GALI SUDARYTI	17
PAGRINDINIS DĖMUO	**17**
PRIELINKSNINĖM KONSTRUKCIJOMIS	**17**
SU KILMININKU	**17**
KAIP IR	16
KALBOS SISTEMOS	**16**
PRIKLAUSOMOJO DĖMENS	**16**
SUDĖTINĮ SAKINĮ	**16**
BE TO	15
IR PAGAL	15
ŽODŽIŲ JUNGINIO	**15**

Kaip matyti iš pateiktų pavyzdžių, ilgesni žodžių junginiai labiau primena posakius ar frazes, trumpesni – sintaksinius žodžių junginius. Nevienodas yra ir jų sustabarėjimo laipsnis, išryškėjantis iš pavartojimų skaičiaus, pavyzdžiui, kai kurie dvižodžiai junginiai net tokiame neilgame tekste yra pavartoti 118 kartų. Be to, kai kurios žodžių grandinės visai sutampa su sudėtiniais terminais, pvz.: *antrojo laipsnio šalutinis sakinys, dvipusės sąsajos santykis, kalbos sistema,* o iš kitų tuos terminus reikia išrinkti, pvz.: *a t s k l e i d ž i a kitas predikatinis dėmuo, žodžio forma kaip nors t a i k o m a prie, pagrindiniu dėmeniu e i n a n č i o žodžio* ir kt. Nors ir trumpintini, jie yra savaip informatyvūs, mat predikatiniai žodžiai (pavyzdžiuose jie išretinti), priklausantys ne pačiam terminui, bet jo aplinkai, teikia svarbios informacijos apie junglumą bei vartosenos ypatumus.

Pastarasis sąrašas, kaip ir ankstesnieji, nėra lemuotas, todėl tas pats keliažodis terminas kartojasi vis kita gramatine forma. Be kita ko, jis kartojasi ir skirtingo ilgio žodžių grandžių grandinėse – nuo ilgiausios iki trumpiausios: *sudaro ne mažiau kaip dviejų savarankiškų žodžių formos, ne mažiau kaip dviejų savarankiškų žodžių formos, savarankiškų žodžių formos, žodžių formos.* Kaip ten bebūtų, dažni pasikartojimai ne trukdo, bet padeda pastebėti pačius svarbiausius, dažniausiai vartojamus sudėtinius terminus. Net jei suminėtomis pusiau automatinėmis programinės įrangos priemonėmis išrenkami ir ne visi tiriamojo teksto terminai, vis vien jos leidžia pastebėti svarbiausius, o tas faktas, kad kompiuterio sudarytus sąrašus skaito ir tvarko terminografas, garantuoja galutinio rezultato kokybę.

Išvada

Šiuolaikinė terminologija, siekianti surinkti, aprašyti ir sunorminti įvairių mokslo šakų terminus, negali apsieiti be kompiuterinio tekstyno ir su juo susijusių kitų kompiuterinio terminų banko rengimo etapų: bendrojo pobūdžio ar specialios programinės įrangos, leksinės duomenų bazės, etc. Tekstynas turi būti parengtas taip, kad apimtų kuo didesnę tekstų įvairovę ir taip reprezentuotų aprašomosios mokslo šakos kalbą. Vartojant net ir bendrojo pobūdžio, leksikografams skirtą programinę įrangą galima automatiškai gauti žodžių ir žodžių junginių – kandidatų į terminus sąrašus, iš kurių atrenkami žinomi ir nauji terminai.

Gauta 1999 06 21

Literatūra

1. Ahmad 1994 – Ahmad, K., Davies, A., Fulford, H., Rogers, M. *What is a term? The semi-automatic extraction of terms from text.* In: M. Snell-Hornby, F. Pochacker, K. Kaindl (eds.). Translation Studies: An Interdiscipline. Amsterdam – Philadelphia, 1994, P. 267–278.

2. Ahmad 1993 – Ahmad, K. *Pragmatics of Specialist Terms: The Acquisition and Representation of Terminology.* In: Petra Steffens (ed.). Machine Translation and the Lexicon. Thirds International EAMT Workshop. Heidelberg, Germany, April 26–28. 1993. Proceedings. Berlin – Heidelberg, 1995, P. 51–76.

3. Bourigault 1996 – Bourigault, D., Gonzalez-Mullier, I., Gros, C. *LEXTER, a Natural Language Processing Tool for Terminology Extraction.* In: Euralex '96 Proceedings II, Goteborg, 1996, P. 771 – 780.

4. Lauriston 1994 – Lauriston, A. *Automatic Recognition of Complex terms: Problems and the TERMINO solution.* In: Terminology. Vol 1(1), 1994, P. 147–170.

5. Meyer 1996 – Meyer, I., Mackintosh, K. *The Corpus from a Terminographer's Viewpoint.* In: International Journal of Corpus Linguistics. Vol. 1(2), 1996, P. 257–285.

6. Pearson 1998 – Pearson, J. *Terms in Context*. Amsterdam – Philadelphia, 1998.

7. Rogers 1994 – Rogers, M., Ahmad, K. *Computerised Terminilogy for Translators: the role of text.* In: M. Brekke, O. Andersen, T. Dahl, J. Myking. Applications and Implications of Current LSP Research. Proceedings of the 9th European Symposium on LSP, Bergen, Aug. 2–6, 1993. Vol. II. Bergen, 1994. P. 840–851.

8. Sager 1990 – Sager, J.C. *A Practical Course in Terminology Processing.* Amsterdam – Philadelphia, 1990.

TERMINOGRAPHY AND CORPUS

Summary

The paper deals with role of a corpus in terminology in general and terminography in particular. It argues for the necessity to use not only corpus but also general purpose and specific tools for the extraction of terminology and compilation of terminological databases. Peculiarities of the terminographer's corpus such as its contents, size and specificity are also discussed. In addition, a wide range of possible general purpose and specific tools for term extraction are presented here. Some of them, based on statistical analyses of a text, were applied for one Lithuanian text on syntax in order to show how different tools can produce different semi-automatic lists of terms.