BALTISTICA XLIII(1) 2008 101–114

Aurelija USONIENĖ, Jonė GRIGALIŪNIENĖ,
Birutė RYVITYTĖ, Linas BŪTĖNAS,
Erika JASIONYTĖ
Vilniaus universitetas

LIETUVIŲ MOKSLO KALBOS TEKSTYNAS

Įvadas

Skaitmeninės bibliotekos, talpyklos (Viliūnas, Glosienė 2006), elektroniniai tekstynai – neatsiejama šiuolaikinės visuomenės gyvenimo dalis. Tai ne tik pramoga ar turiningas laisvalaikio praleidimo būdas, svarbus asmenybės vystymuisi, bet ir praktinės informacijos šaltinis, mokymo ir mokymosi bazė, mokslinių tyrimų objektas bei mokslinių tyrimų infrastruktūros dalis.

Kad paspartintų žinių ekonomikos kūrimo ir plėtros procesus Europoje, Europos Komisija 2000 metais inicijavo pirmąjį e-Europos veiksmų planą. Šiuo metu vykdomos i2010 veiksmų programos tikslas – užtikrinti informacijos ir ryšių technologijų (IRT) raidos tobulinimą ir IRT panaudojimą žmonių gyvenimo kokybei gerinti. Pagal Lietuvos Respublikos Vyriausybės 2006 m. spalio 24 d. Nr. 1048 („Valstybės žinios“, 2006 10 26, Nr. 114) nutarimą „būtina skatinti informacinių technologijų plėtrą ir diegimą visose šalies ūkio ir kultūros srityse. Kultūros srityje ypač svarbu stiprinti lietuvių šnekos atpažinimo, sintezės ir lietuvių kalbos vertimo mokslinius tyrimus, nes tik taip padėsime išlikti lietuvių kalbai modernioje elektroninėje, skaitmeninėje terpėje.“ Kaip pabrėžia Čermakas (2000, 297), čekų nacionalinio tekstyno instituto direktorius, kompiuteriu nevaldomoms kalboms, neradusioms vietos Europos informacinėje visuomenėje, gresia sumenkėjimas ir išnykimas. Kad būtų galima plėtoti informacinių technologijų krypties mokslinius tyrimus ir eksperimentinės plėtros darbus siekiant sukurti lietuvių kalbos vertimo priemones, būtina sukaupti didelės apimties autentiškos kalbinės medžiagos išteklių bazę. Iš įvairių skaitmeninės produkcijos pavyzdžių, laikomų šiuolaikinių mokslinių tyrimų infrastruktūros pagrindu, minėtini elektroniniai žodynai, bibliotekos, talpyklos, specializuoti duomenų bankai ir įvairūs tekstynai.

Pasaulyje egzistuoja didžiulė tekstynų įvairovė. Mokslinėje periodikoje lietuvių kalba yra pateikiama gana išsami ir detali tekstynų tipų apžvalga (Marcinkevičienė 2000; Rimkutė, Kovalevskaitė, Daudaravičius 2006). Pagal kalbos reiškimosi formą skiriami rašytinės ir sakytinės kalbos tekstynai, kurie gali būti bendrieji ir specialieji, sinchroniniai ir diachroniniai (dabartinės kalbos ir senųjų raštų tekstynai). Tekstynai nebūtinai reprezentuoja vienos kalbos faktus (vienakalbiai tekstynai). Labai reikalingi ir naudingi vertimo studijoms ir praktikai, kalbų mokymuisi ir mokymui yra dvikalbiai ar daugiakalbiai tekstynai. Lygiagretieji tekstynai, kurių pamatą sudaro originalo ir jo vertimo tekstų sulygiavimas papastraipiui ir pasakiniui, yra labai svarbūs šiuolaikiniams kontrastyvinės lingvistikos tyrimams. Palyginamieji tekstynai dažniausiai apima kelių kalbų tos pačios tematikos originalius tekstus. Bendrieji tekstynai reprezentuoja kalbą kaip visumą, o specialieji yra susiję su tam tikra sritimi (pvz.: medicina, teise, politika ir t. t.) arba tam tikru žanru (laikraščių straipsnių kalba, akademinė proza ir t. t.).

Yra labai daug ir įvairių specialiųjų tekstynų bei skaitmeninių vienakalbių ir daugiakalbių duomenų bazių, tezaurų ir terminų bankų, pvz.:

1. Politechninė daugiakalbė terminologinių duomenų bazė (VERBA Polytechnic and Plurilingual Terminological Database – http://catalog.elra.info);
2. Anglų, vokiečių, prancūzų, ispanų, italų, lenkų, portugalų, turkų kalbų LORETO tezauras (LORETO Thesaurus);
3. Daugiasričiai anglų, prancūzų, vokiečių, danų kalbų leksikonai (Multidomain Lexicons); ir kt. (http://catalog.elra.info/)

Iš šiuo metu kuriamų ir jau sukurtų akademinės kalbos tekstynų pavyzdžių, galima paminėti šiuos:

1. Britų akademinės šnekamosios anglų kalbos tekstynas (British Academic Spoken English (BASE) corpus – http://www.rdg.ac.uk/AcaDepts/II/base_corpus/), kaupiamas Warwicko ir Readingo universitetuose. Šį tekstyną sudaro 160 paskaitų ir 40 seminarų įrašų, padarytų keturiuose šių universitetų fakultetuose.
2. Britų akademinės rašytinės anglų kalbos tekstynas (British Academic Written English (BAWE) corpus – http://www.coventry.ac.uk/bawe), kuriamas Warwicko, Readingo ir Oxfordo Brookeso universitetuose. Surinkta daugiau kaip 500 studentų rašto darbų, kurių kiekvieno apimtis nuo 1000 iki 5000 žodžių.

3. Mičigano universiteto šnekamosios akademinės anglų kalbos tekstynas (Michigan Corpus of Academic Spoken English (MICASE) – http://www.lsa.umich.edu/eli/micase/index.htm), sudarytas iš 152 transkribuotų šnekamosios akademinės anglų kalbos įrašų, padarytų Mičigano universitete paskaitų, laboratorinių darbų, seminarų, disertacijų gynimo, interviu, susirinkimų, konsultacijų akademinės bendruomenės pokalbių su aptarnaujančiu personalu metu. Šio tekstyno dydis – 1,8 milijono žodžių. Tekstynas turi viešą prieigą internete.
4. TOEFL 2000 rašytinės ir šnekamosios akademinės anglų kalbos tekstynas (TOEFL 2000 Spoken and Written Academic Language Corpus – T2K-SWAL Corpus), kurio apimtis – 2,8 milijono žodžių. Tekstyną sudaro 490 rašto ir šnekamosios kalbos tekstų. Tekstynas reprezentuoja keturių JAV universitetų šnekamąją ir rašto kalbą, įskaitant auditorinio darbo, darbo grupių, vadovėlių ir pokalbių su aptarnaujančiu personalu kalbą.

Lietuvoje kol kas yra vos keletas tekstynų ir kalbos duomenų bei terminų bankų. Daugelis jų tik pradėti kurti. Pirmiausia minėtinas viešai prieinamas internete *Dabartinės lietuvių kalbos tekstynas*, sukauptas Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre (http://donelaitis.vdu.lt/), kuriami lygiagretieji anglų-lietuvių, čekų-lietuvių tekstynai, o VDU Regionistikos katedroje kuriamas *Bendrinės šnekamosios lietuvių kalbos tekstynas* (http://www.vdu.lt/LTcourses/?pg=41&lang=1). Nuo 2003 m. kuriamas viešas, visiems interneto vartotojams laisvai prieinamas Lietuvos Respublikos terminų bankas (http://www3.lrs.lt:10001/pls/tb/tb.search?p_sid=106425), kurio veikimą, nuolatinę priežiūrą ir duomenų atnaujinimą užtikrina Terminų banko tvarkytojai – Valstybinė lietuvių kalbos komisija kartu su Lietuvos Respublikos Seimo kanceliarija. Terminų banke teikiami terminų straipsniai su apibrėžtimis, nurodoma terminų vartojimo sritis, šaltiniai ir kartais pateikiama pavyzdžių. Neginčytina šio kalbinės informacijos šaltinio svarba ir reikalingumas, tačiau jis nėra konkurentas tekstynams, nes nepateikia terminų vartosenos, jų junglumo galimybių ir specifikos. Nepriklausomai nuo vartotojo, ar tai būtų studentai, mokytojai, mokslininkai, ar valstybės tarnautojai, šie abu minėti šaltiniai reikalingi, nes papildo vienas kitą. Mokslo kalbos tekstyno Lietuvoje nėra, o *Dabartinės lietuvių kalbos tekstyne* akademiniai tekstai sudaro nedidelę tekstyno dalį, todėl šitiek medžiagos nepakanka išsamesniems mokslo kalbos tyrimams,

nes ji negali būti laikoma reprezentatyvia, atspindinčia mokslinio diskurso žanrų įvairovę.

Kyla klausimas, kam reikalingi specialieji tekstynai, kai egzistuoja bendrieji reprezentatyvūs tekstynai? Pirma, nedideli specialieji tekstynai gali pateikti tokios informacijos, kokios nerasime dideliuose bendruosiuose tekstynuose. Specialūs tekstynai orientuoti į tam tikrą kalbos atmainą ar tam tikrą sritį, todėl pasirinktą sritį jie atspindi išsamiau negu bendrieji, į kuriuos stengiamasi sudėti visko po truputį. Antra, specialiuosiuose tekstynuose tam tikra leksika ir gramatinės konstrukcijos pasitaiko daug dažniau ir reguliariau nei bendruosiuose tekstynuose. Trečia, tokius tekstynus galima lengvai pritaikyti kalbų mokymo praktikoje.

Galbūt dėl to, kad iki šiol nėra specialaus mokslo kalbos tekstyno, lietuvių kalbos mokslinis diskursas nedaug tetyrinėtas (žr. Alaunienė 2005; Bitinienė 1983; 2000; 2005; 2007; Župerka 2001; Koženiauskienė 1999; Linkevičienė, Vilkienė 2005; Petrėnienė 2003; 2005). Tuo tarpu pasaulyje mokslinis diskursas, ypač anglų kalbos, gana plačiai tyrinėjamas (žr. detalią mokslinio diskurso tyrimų apžvalgą Grabe, Kaplan, 1996; Flowerdew 2002). Dauguma šių tyrimų remiasi gausia empirine medžiaga – tekstynais.

Lietuvių mokslo kalbos tekstynas kuriamas pagal Lietuvos valstybinio mokslo ir studijų fondo (http://www.vmsfondas.lt/) finansuojamą mokslinių tyrimų projektą „Taikomųjų uždavinių tyrimai ir realizavimas naudojant *grid* technologijas (GridTechno)“ (vadovas doc. A. Juozapavičius, VU Matematikos ir informatikos fakultetas), kuris vykdomas pagal Lietuvos Respublikos Vyriausybės 2006 m. spalio 24 d. nutarimu Nr. 1048 patvirtintą Aukštųjų technologijų plėtros 2007–2013 metų programą. Tekstyno kūrimo darbo grupė yra tarpdisciplininė, bendradarbiauja Vilniaus universiteto Filologijos fakulteto lingvistai ir Matematikos ir informatikos fakulteto informatikos specialistai. Pasaulinė patirtis rodo, kad kalbininkų ir informatikų bendradarbiavimas yra būtina sėkmingo mokslinio darbo sąlyga. Tarpdisciplininiai projektai ypač perspektyvūs ir aktualūs tapo pastaruoju metu, kai šiuolaikinės kompiuterizacijos mastai ir techninės galimybės tarsi paskatino tekstynų lingvistikos renesansą (Vargas Sierra 2005). Tekstyno sudarymo darbo grupės pagrindinis uždavinys – sukaupti kuo įvairesnius tekstus, kad šie kiek įmanoma geriau reprezentuotų šiuolaikinės lietuvių rašytinės įvairių mokslo krypčių kalbos

leksines, gramatines ir teksto lygmens, struktūrines, ypatybes. Siekiama sukurti specializuotą, sinchroninį, rašytinės lietuvių mokslo kalbos tekstyną. Tekstai bus atrenkami pagal mokslo kryptis ir tipus. Tekstyno kodavimo darbai bus atliekami remiantis 5-uoju TEI (Text Encoding Initiative (TEI) P5 http://www.tei-c.org/release/doc/tei-p5-doc/en/) rekomendacijų variantu.

Tikslas

Projekto „Lietuvių mokslo kalbos tekstynas“ tikslas – sukaupti didelės apimties, autentišką lietuvių mokslo kalbos duomenų bazę, kuri galėtų būti naudojama objektyviems, šiuolaikiniams akademinio diskurso kiekybinių ir kokybinių parametrų tyrimams, leistų atskleisti tarpdisciplininius ypatumus, žanrų įvairovės charakteristiką, suteiktų informacijos apie esamas ar galimas lietuvių mokslo kalbos savitumo, identiteto formavimo(si) ar nykimo tendencijas ir atskleistų veiksnius, lemiančius aptariamus procesus. Taikomuoju požiūriu, šis tekstynas labai reikalingas akademinio rašymo mokymui ir savarankiškam mokymuisi, nes sukaupta faktinė medžiaga nėra sukurta pavienių tyrėjų, kalbininkų – ji atspindi realią kalbinę situaciją, demonstruoja kolektyvinę intuiciją ir parodo tikrąją vartoseną.

Sandaros principai

Pačioje tekstynų sudarymo pradžioje sprendžiamas tekstyno reprezentatyvumo klausimas. Pati reprezentatyvumo sąvoka, kaip jau ne kartą buvo rašyta literatūroje (žr. Leech 1991; Kennedy 1992; Marcinkevičienė 1997; 2000; McEnery et al. 2006), yra gana polemiška ir įvairiai apibrėžiama. Kalbant apie reprezentatyvumą, visada iškyla klausimas, ką turi atspindėti tekstynas (Kennedy 1992, 62). Teiginys, kad jis turi atspindėti kalbą, jos įvairovę ar atmainą, yra neinformatyvus ir labai abstraktus. Kalba yra tokia daugialypė ir sudėtinga, kad norint ją atspindėti, reikia kuo tiksliau apsibrėžti reprezentuojamąjį objektą (žr. Biber 1993, 243). Bendrųjų ir specialiųjų tekstynų reprezentatyvumas yra matuojamas skirtingai. Bendrojo pobūdžio tekstyno reprezentatyvumą lemia kuo didesnė žanrų įvairovė, tuo tarpu specialiųjų tekstynų reprezentatyvumas matuojamas, bent jau žodyno lygmenyje, specialiosios leksikos koncentracijos, išbaigtumo laipsniu („saturation“, „degree of closure“). Norint nustatyti specialiojo tekstyno reprezentatyvumą, reikia padalyti tekstyną į vienodo dydžio segmentus. Tekstynas bus laikomas reprezentatyviu, kai naujų leksinių vienetų skaičius naujame

segmente bus apytiksliai toks pat kaip ir prieš tai buvusiame segmente (Belica 1996; McEnery 2001).

Biberis (1993, 256) teigia, kad iš anksto neįmanoma tiksliai apibrėžti, koks turi būti reprezentatyvus tekstynas. Jis siūlo pasidaryti bandomąjį tekstyną, su juo padirbėti, o ne projektuoti apimtis spekuliatyviai. Panašiu keliu buvo nutarta eiti kaupiant lietuvių mokslo kalbos tekstyną. Kaip mano Sinclairis (2005), tekstynų sudarymas nėra tikslusis mokslas ir niekas nežino, koks turėtų būti idealus tekstynas. Todėl teisingas požiūris būtų kuo tiksliau ir detaliau aprašyti tekstyno sandarą, turinį, o jau tuomet pati mokslininkų ir tyrėjų bendruomenė turėtų nuspręsti, ar gali naudotis tokiu tekstynu, ar ne. Šiandien, kol nėra griežtų mokslinių kriterijų reprezentatyvumui nustatyti, remiamasi intuicija ir įvairiais skaičiavimais (žr. McEnery & Wilson 2001, 166; Belica 1996, 61–74; Teubert 1999). Todėl labai svarbu apsibrėžti, kurių mokslo sričių ir krypčių tekstai ir kokie jų tipai bus kaupiami sudaromame tekstyne.

Siekiant Lietuvių mokslo kalbos tekstyno reprezentatyvumo, nutarta, kad jį turėtų sudaryti visų Lietuvoje plėtojamų mokslo sričių, krypčių ir šakų tekstai, parašyti lietuvių kalba. Buvo atlikta išsami duomenų šaltinių analizė (ištirti visų Lietuvos leidyklų internetiniai puslapiai, atlikta kokybinė spaudinių analizė ir jų atranka pagal mokslinę tematiką, leidžiamų mokslinių publikacijų pobūdį). Mokslinio teksto interdisciplinišKumo klausimas neturi esminės įtakos tekstyno struktūros ir turinio kokybei. Tai techninis kodavimo aspektas, kuris nelemia tekstyno reprezentatyvumo ar balanso dalykų. Lietuvoje kaip ir Europoje yra polemizuojama dėl mokslo šakų ir krypčių klasifikacijos kriterijų. Pagal Lietuvos Mokslo tarybos nutarimą Nr. VI-17, priimtą 2005 m. balandžio 11 d. (http://ml.lms.lt/) rekomenduojama mokslinių tyrimų ir eksperimentinės plėtros sritis Lietuvoje klasifikuoti vadovaujantis OECD Mokslinių tyrimų ir eksperimentinės plėtros standarto Frascati Manual 200 (ir atitinkamai 222) punktu. Siūloma skirti gamtos mokslų, inžinerijos ir technologijos, medicinos mokslų, žemės ūkio mokslų, socialinių mokslų, humanitarinių mokslų ir menų sritis.

Lietuvių mokslo kalbos tekstyno sudarymo darbo grupė nutarė remtis Lietuvos Švietimo ir mokslo ministerijos 1998 m. sausio 9 d. Nr. 30 įsakymu „Dėl mokslo sričių, krypčių ir šakų klasifikacijos“, ir tekstyno medžiaga bus tvarkoma pagal priede pateikiamą aprašą:

H – Humanitariniai mokslai (architektūra, dailėtyra, etnologija, folkloristika, filosofija, kalbotyra, literatūrologija, bibliotekininkystė, istorija, teologijos mokslai);
S – Socialiniai mokslai (teisė, politologija, ekonomika, psichologija, edukologija, vadyba);
P – Fiziniai mokslai (matematika, astronomija, fizika, chemija, geografija, geologija ir mineralogija, informatika, sistemotyra);
B – Biomedicinos mokslai (medicina, stomatologija, biologija, botanika, agronomija, zootechnika, farmacija, veterinarija, miškotyra);
T – Technologijos mokslai (energetika ir šiluminė, cheminė technologija, medžiagotyra, mechanika, metrologija ir matavimai, statyba, transporto technologija, žemės ūkio ir aplinkos, valdymas ir informatikos).

Lietuvių mokslo kalbos tekstynas yra specializuotas tekstynas, kurio tikslas – pristatyti lietuvių mokslo bendruomenės kuriamų tekstų žanrus ir kalbą. Šiuolaikinėje žanrų analizės teorijoje žanras apibrėžiamas kaip „atpažįstamas komunikacinis įvykis, kuriuo siekiama komunikacinių tikslų, suprantamų tos profesinės ar akademinės bendruomenės, kurioje jis nuolat vyksta, nariams“[1] (Bhatia 1993, 13; vert. aut.). Joks tekstynas negali ir neturi apimti visų sukurtų tekstų, nes nuolat kuriami nauji tekstai. Turėti tokį tekstyną, kuriame būtų galima rasti viską, buvo ankstyvosios tekstynų lingvistikos atstovų svajonė, tačiau dabar apie tai jau niekas nekalba. Tai reiškia, kad kuriant specializuotą tekstyną, reikia atsižvelgti į konkrečioms mokslų ar profesinėms sritims būdingų tekstų, kuriamų tam tikroje situacijoje ir atliekančių tam tikras funkcijas, įvairovę. Būtent tekstų autentiškumas ir reprezentatyvumas ir lemia tekstyno kokybę. Dėl tekstų „autentiškumo“ tekstynų lingvistikoje vis dar laužomos ietys. Kaip rašoma toliau, „kiekvienas tekstas sukuriamas konkrečioje situacijoje (autentiškoje), kuri tekstyne lyg ir prarandama“ (žr. Widdowson 2000; Stubbs 2001).

Kiekvieno teksto kūrimas ir suvokimas vyksta konkrečioje situacijoje, kuri apibrėžiama: 1) pagal komunikacinį įvykį (*field*), kuriame kalba veikia kaip ypatingas komponentas, 2) pagal komunikacijos dalyvių santykį (*tenor*) ir 3) pagal teksto kūrimo būdą (*mode*) (Halliday, Hasan 1989, 12t.). Komuni-

1 „[...] a recognisable communicative event characterised by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs“ (Bhatia 1993, 13).

kacinis įvykis lemia daugiau ar mažiau specializuoto diskurso vartojimą, todėl tai, ar tekstų kalba bus labiau techninė, ar bendrinė, priklausys nuo temos, auditorijos, komunikacinio tikslo, konkrečios situacijos, turinio abstraktumo ir pan. Mokslo bendruomenė nėra vienalytė: jai priklauso savos srities ekspertai, pakankamai profesinių žinių įgiję nariai,ir visiški naujokai, neseniai peržengę slenkstį į mokslo bendruomenę. Ir bendraudama tarpusavyje (pvz., ekspertas su ekspertu, ekspertas su naujoku), ir bendraudama su plačiąja visuomene, mokslo bendruomenė pasitelkia įvairius mokslinio diskurso žanrus. Skiriami du pagrindiniai teksto kūrimo būdai: sakytinis ir rašytinis, nors kai kurie autoriai mini net penkis: parašytas, parašytas perskaityti, parašytas kalbėti, pasakytas ir pasakytas, kad būtų užrašytas. Dar galima būtų paminėti elektroninius tekstus (pavyzdžiui, tinklalapiai, elektroniniai laiškai ir pan.).

Numatyta, kad tekstyną sudarys įvairių mokslo ir profesinių sričių tekstai, atliekantys šias funkcijas:

- *žinių kūrimo (mokslinių tyrimų)*. Šią funkciją atlieka mokslinis straipsnis (tiriamasis, eksperimentinis, apžvalginis), monografija, mokslinė ataskaita, daktaro disertacija, magistro darbas, daktaro disertacijos santrauka, habilitacijos procedūrai teikiamų mokslo darbų apžvalga, straipsnio anotacija, straipsnio santrauka, pranešimo tezės;
- *žinių perdavimo (mokymo, studijų)*: vadovėlis, mokomoji knyga, metodinė priemonė, studijų programos aprašas, dalyko kurso aprašas ;
- *žinių vertinimo (recenzavimo)*: recenzija, apžvalginis straipsnis (angl. *review article*), pratarmė;
- *žinių sklaidos (populiarinimo, reklamos)*: profesinė informacija / reklama;
- *informavimo:* kronikos.

Kadangi žanro analizės teorija dar palyginti jauna mokslas, įvairuoja panašiai apibrėžiami terminai (*registras, žanras, teksto tipas*), kuriuos reikėtų atskirti. Terminas *registras* vartojamas tada, kai kalbama apie teksto žodyną ir sintaksę, o terminas *žanras* – kai pereinama į diskurso lygmenį ir kalbama apie teksto struktūrą. Nuo registro priklauso stilistinių priemonių pasirinkimas, o žanras, kaip tikslingas komunikacinis įvykis, lemia tekstų struktūrą. Vokiškoji tradicija (Trosborg 2000, IX) skiria *Textsorten* (teksto klasė, žanras) ir *Texttyp* (teksto tipas). Lietuvių autorių (Koženiauskienė 1999) darbuose paprastai skiriami trys teksto tipai: *aprašymas, pasakojimas, samprotavimas.* Kai kurie autoriai (Nauckūnaitė 2002, 7t.) *samprotavimą* dar skaido į du

potipius: *aiškinimą* ir *argumentavimą*. Grynų teksto tipų pasitaiko retai (pvz., aprašyme gali būti panaudotas ir pasakojimas), todėl teksto tipas dažniausiai nustatomas pagal adresato intenciją (tikslą).

Tekstų žanrai tekstyne bus kaupiami pagal išorines tekstų charakteristikas: teksto komunikacinę funkciją, dalyvius ir pačią komunikacinę situaciją, kitaip sakant, pagal sociokultūrinius kriterijus. Išorinius kriterijus galima nustatyti neskaitant teksto, todėl tekstų parinkimas yra objektyvesnis. Pradinis tekstų parinkimas neišvengiamai remiasi išoriniais kriterijais, nes tik atlikus tų tekstų analizę, galima nustatyti kiekvienam tekstui būdingus lingvistinius bruožus, kurie vėliau leis nustatyti vidines tekstų charakteristikas.

Tekstynas taip pat turi sukaupti įvairaus kalbos dalykiškumo lygio diskurso pavyzdžius, t.y. nuo tekstų nespecialistams, su jiems skirtais paaiškinimais, apibrėžtimis, iki sudėtinga technine kalba pateiktos informacijos, skirtos ekspertams. Autentiški tekstai galės būti naudojami įvairiems lingvistiniams akademinio diskurso tyrimams, kurie padėtų išryškinti lietuvių mokslo kalbos identitetą ir įvairių disciplinų diskurso praktiką.

Žanro analizės teorija taip pat labai daug dėmesio skiria akademinei anglų kalbai. Daug tyrimų sulaukė įvairių disciplinų mokslinių straipsnių įžangos, tais tyrimais remiasi IMRD (Įvadas–Metodas–Rezultatai–Diskusija) modelis Feak ir Swaleso (1994) akademinio rašymo vadovėlyje magistrantams ir doktorantams. Ištirta ir kitų angliško straipsnio dalių retorinė-informacinė struktūra. Tyrinėta vadovėlių kalba (specialisto bendravimas su nespecialistu), straipsnių anotacijų kalba, retorinės strategijos recenzijose, įvadai ir literatūros apžvalgos daktaro disertacijose, magistro darbų struktūra ir t.t. Kai kuriuose tyrimuose buvo naudojamasi asmeniniais tekstynais. Lietuvoje iki šiol buvo atliekama tik pavienių mokslo kalbos tyrimų, kurie remiasi nedideliais tekstų kiekiais. Detaliai išanalizavus įvairių spaudinių tipus, juose publikuojamus mokslo darbus, buvo parengtas kaupiamų tekstų tipų sąrašas: monografija / vadovėlis / mokomoji knyga; straipsnis; daktaro disertacijos santrauka; apžvalga; recenzija; pranešimo tezės; kronikos; pratarmė (žurnalo / rinkinio); profesinė informacija; mokslinė ataskaita; dalyko kurso aprašas; studijų programos aprašas; daktaro disertacija; magistro darbas; metodinė priemonė. Šie iš anksto numatyti lietuvių mokslo kalbos tekstyno sandaros principai bei tekstų atrankos kriterijai gali būti vėliau šiek tiek pakoreguoti dėl realių aplinkybių ir problemų, su kuriomis susiduriama jau dabar.

Tekstyno kūrimo darbai ir problemos

Labai svarbus ir opus klausimas, kurį reikia spręsti tik pradėjus svarstyti tekstyno kaupimo metodikos ir struktūros dalykus, yra autorinių teisių klausimas. Tam, kad gautume leidimą naudoti tekstą tekstynui sudaryti yra nelengvas, daug laiko ir ilgų derybų reikalaujantis procesas. Pažymėtina, kad autorinės teisės yra ankstyvojoje raidos stadijoje, todėl čia daug neišspręstų klausimų, o tai labai apsunkina leidimų naudotis tekstais mokslo tiriamojo darbo tikslais (pvz., tekstynų kūrimo) procesą. Tai labai subtili teisinė problema ir sugebėjimas išlaikyti tekstyną „geros teisinės formos (in good legal health)" (Sinclair 1991, 15) reiškia didžiulį darbą. Jei autorinių teisių turėtojai suprastų, kodėl jų tekstų reikia tekstynui ir kaip tekstai galėtų būti apsaugoti nuo piratavimo, gal ir būtų galima išvengti to didžiulio neproduktyvaus darbo siekiant gauti leidimus. Kol ši problema nebus išspręsta tarptautiniu mastu, tol reikalai nepagerės. Atrodo, jog apie tai jau kalbama ir Europos Taryboje, taigi yra vilties, kad ši problema gali būti išspręsta. Leidėjai, kuriuos domina žodynų ir kitokių žinynų leidybos reikalai, jau rengia autorinių sutarčių pataisas, kurios leistų išvengti teisinių problemų ateityje.

Žinoma, sudaryti rašytinės kalbos tekstynus šiandien yra lengviau ir paprasčiau nei anksčiau, nes kol nebuvo skaitytuvų, o leidyboje – plačiai taikomų kompiuterių, tekstai tekstynams buvo renkami klaviatūra, o tam reikėjo didelių žmogaus rankų darbo sąnaudų.

Viena iš kaupiamo tekstyno paskirčių – padaryti jį prieinamą kuo didesniam vartotojų skaičiui. Šiam tikslui įgyvendinti kuriama internetinė sistema, leidžianti tekstyno kūrėjams rinkti ir išsaugoti dokumentus, o svetainės vartotojams – atlikti paiešką. Tam, kad dokumentas būtų tinkamas tekstynui, jis turi būti pritaikytas. Tekstų atgaminimas elektroninėje aplinkoje yra gana sudėtingas ir darbo atžvilgiu imlus procesas. Mokslinių tekstų pertvarka tekstynui apima du etapus:

1) tekstų kompiuterinių rinkmenų tipo keitimą ir

2) paties teksto duomenų tvarkymą, arba teksto kodavimą.

Tekstynui sudaryti būtinas *.doc, *.html, *.txt arba *.xml dokumentų formatas. Kadangi iš leidyklų ir žurnalų redakcijų gaunami tekstai beveik visada pateikiami *.pdf arba *.p65 (Page Maker) arba Adobe InDesign formatu, todėl pirmiausia turi būti atliekamas tokių tekstų konvertavimas į minėtuosius – tekstus redaguoti leidžiančius – formatus. Kiekvieną atskirą dokumentą

būtina peržiūrėti, pašalinti nereikalingas teksto dalis ir užkoduoti naudojant TEI 5 formatą (kaip jau minėta ankščiau). Paprastos paieškos sistema, kai ieškoma žodžio ar frazės, turi ignoruoti visas TEI koduotės žymas, ir atlikti paiešką tik tarp dokumento teksto žodžių. Tačiau TEI žymos reikalingos tam, kad būtų galima pažymėti teksto struktūrines dalis ir suteikti joms papildomą informaciją, pavyzdžiui pažymėti pastraipas, sakinius, antraštes, juos sunumeruoti ir pan. Tokios informacijos prireikia, jei norime pamatyti, kaip vartojamas žodis ar frazė kiekviename pirmajame pastraipos sakinyje arba tam, kad galėtume kurti automatinio teksto kodavimo sistemas, kurios pamažu pačios išmoktų teisingai koduoti dokumentus, lygindamos jau užkoduotus ir neužkoduotus tų pačių dokumentų pavyzdžius. Tam, kad žodžio ar frazės paieška vyktų sparčiai, bus diegiama teksto indeksavimo sistema. Planuojame tirti, kaip panaudoti indeksavimą ir teksto dalims su joms priklausančia struktūrine informacija, kad sudėtingos paieškos užklausos, pateikiamos tekstynui, būtų atliekamos taip pat sparčiai, kaip ir paprastos.

Padėka

Tekstyno kūrimo darbo grupė nuoširdžiai dėkoja kolegoms ir leidyklų savininkams, vadovams, darbuotojams už supratimą, geranoriškumą, patarimus ir pagalbą:

Juozui Atkočiūnui (*VGTU*); Eleonorai Dagienei (*VGTU*); Teresei ir Viliui Gužauskiams (*Homo liber*); Rimantui Jankauskui (*Medicinos teorija ir praktika*); Rimantui Kareckui (*MELI*); Rūtai Marcinkevičienei (*VDU*); Hilary Nesi, Timui Kelly'ui, Jasperui Holmesui (*Warwick universitetas*); Algiui Paulauskui ir Daivai Mickevičienei (*Gimtasis žodis*); Marijai Sniečkutei, Antanui Smetonai, Reginai Rudaitytei, Meilutei Ramonienei (*VU*); Ritai Urnėžiūtei (*Gimtoji kalba*); Lolitai Zemlienei (*KU*); Leonidui Zabulioniui ir Vidai Vaidakavičienei (*VU leidykla*); Sauliui Žukui ir Agnei Jurčiukonytei (*Baltos lankos*).

Straipsnio autoriai dėkingi recenzentams už pastabas ir siūlymus, į kuriuos buvo stengiamasi atsižvelgti redaguojant paskutinį straipsnio variantą.

BUILDING OF THE CORPUS OF ACADEMIC LITHUANIAN

Summary

The paper sets out to describe the initial stages of the design of the corpus of academic Lithuanian. Due to the increasing interest and numerous corpora-based studies in academic discourse (especially of academic English) all over the world, there is an obvious need to provide easily accessible electronic resources of academic Lithuanian to facilitate modern linguistic research, interdisciplinary studies, lexicographical practice, and terminology studies in theory and practice. The Corpus of Academic Lithuanian (CorALit) is being compiled at the University of Vilnius (Faculty of Philology and Faculty of Mathematics and Informatics). The building of the corpus is being carried out within the framework of the 2007–2013 national high-tech development programme launched by the Government of Lithuania and supervised by the Lithuanian State Science and Study Foundation (http://www.vmsfondas.lt/index.php?lang=en). The main issue in the process of corpus design is representativeness which is determined by the following factors: the number of research and study fields represented, the range of genres included (i.e. balance) and the way text chunks for each genre are selected (i.e. sampling).

The Corpus of Academic Lithuanian aims at representing the main fields of study and research developed in Lithuania and listed in Order No.30 of the Minister of Education and Science of 9 January 1998 "Concerning the Classification of Study and Research Areas, Fields and Branches" as well as the most typical genres that academic community uses for the creation, dissemination and evaluation of new knowledge and internal communication. Since at present there is no reliable scientific measure for corpus balance, the project team will have to rely on intuition and best estimates based on the experience of academic language corpora already compiled in other countries (the UK, USA, etc.). The compilation of the corpus also involves negotiations, sometimes rather time-consuming, with publishers and authors for copyright. Last but not least, technical aspects of corpus design are touched upon. The main purpose of corpus compilation is to make it easily accessible for large numbers of users, and this means changing the format of computer files and text coding in accordance with TEI P5 Guidelines. TEI P5 format will allow users to access the first synchronic corpus of written academic Lithuanian as a major resource of authentic language data via a simple internet search.

LITERATŪRA

Alaunienė Zita 2005, Akademinių tekstų struktūra ir jos raiška, *Žmogus ir žodis* 7(1), 63–67.

Belica Cyril 1996, Analysis of temporal change in corpora, *International Journal of Corpus Linguistics* 1(1), 61–74.

Bhatia Vijay 1993, *Analysing genre: language use in professional settings*, London/New York: Longman.

Biber Douglas 1993, Representativeness in corpus design, *Literary and Linguistic Computing* 8(4), 243–257.

Bitinienė Audronė 1983, *Mokslinis stilius,* Vilnius: LTSR aukštojo ir spec. vid. mokslo ministerijos Leidybinė redakcinė taryba.

Bitinienė Audronė 2000, Vientisiniai mokslinio stiliaus sakiniai, *Kalbotyra* 48–49(1), 19–27.

Bitinienė Audronė 2005, Mokslinis stilius ir jo intertekstualumas, *Žmogus ir žodis: didaktinė lingvistika* 7(1), 68-72

Čermak Fratišek 2000, „Tekstynas yra geriausias kasdienio gyvenimo atspindys...", *Darbai ir dienos* 24, 295–297.

Feak Christine, John Swales 1994, *Academic writing for graduate students*, the University of Michigan Press.

Flowerdew John (ed.) 2002, *Academic discourse*, London, New York: Longman.

Grabe William, Robert B. Kaplan 1996, *Theory and Practice of Writing: An applied linguistic perspective*. New York: Longman.

Halliday Michael, Hasan Ruqaiya 1989, *Language, context, and text: aspects of language in a social-semiotic perspective*, Oxford: Oxford University Press.

Hyland Ken 2000, *Disciplinary discourses: social interactions in academic writing*, London, New York: Longman.

Kennedy Graeme1998, *An Introduction to Corpus Linguistics,* London and New York: Longman.

Koženiauskienė Regina 1999, *Retorika: iškalbos stilistika*, Vilnius: MELI.

Leech Geoffrey 1991, The state of the art in corpus linguistics, In Aijmer Karin, Bengt Altenberg (eds.), *English Corpus Linguistics*, London, New York: Longman, 8–29.

Linkevičienė Nijolė, Loreta Vilkienė 2005, Dar kartą sintaksinės pilnaties ir pastraipos santykių klausimu, In *Žmogus ir žodis* 7(1), Vilnius: VPU, 13–16.

Marcinkevičienė Rūta 1997, Tekstynų lingvistika ir lietuvių kalbos tekstynas, *Lituanistica* 1(29), 58–78.

Marcinkevičienė Rūta 2000, Tekstynų lingvistika (Teorija ir praktika), *Darbai ir dienos* 24, 7–64.

McEnery Anthony, Andrew Wilson 2001, *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery Anthony, Richard Xiao, Yukio Tono 2006, *Corpus-based language studies. An advanced resource book*. London, New York: Routledge.

Nauckūnaitė Zita 2002, *Teksto komponavimas: rašymo procesas ir tekstų tipai*, Vilnius: Gimtasis žodis.

Petrėnienė Ona 2003, Terminų aiškinimas mokslo populiarinamuosiuose tekstuose, *Terminologija* 10, 42–54.

Petrėnienė Ona 2005, Mokslo subjektų atributų raiška tekstuose, *Žmogus ir žodis: didaktinė lingvistika*, 7 (1), 92–95.

Rimkutė Erika, Jolanta Kovalevskaitė, Vidas Daudaravičius 2006, Daugiakalbių tekstynų naudojimas ir taikymas. *Darbai ir dienos* 45, 41–62.

Sinclair John 1991, *Corpus. Concordance. Collocation*, Oxford: Oxford University Press.

Sinclair John 2005, Corpus and Text – Basic Principles, In Martin Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 1–16.

Stubbs Michael 2001, Texts, Corpora, and Problems of Interpretation: A Response to Widdowson, *Applied Linguistics* 22(2), 149–172.

Swales John 1990, *Genre analysis: English in academic and research settings*, Cambridge: Cambridge University Press.

Swales John 2004, *Research genres: exploration and applications*, Cambridge: Cambridge University Press.

Teubert Wolfgang 1999, Corpus linguistics – a partisan view, *TELRI Newsletter* 8.

Trosborg Anna (red.) 2000, *Analysing Professional Genres*, Amsterdam/Philadelphia: John Benjamins.

Vargas Sierra C. 2005, A pragmatic model of text classification for the compilation of special-purpose corpora, In J. Mateo, F. Yus (eds.), *Thistles. A homage to Brian Hughes. Essays in Memoriam* 2, 295–315.

Viliūnas Giedrius, Audronė Glosienė 2006, Institucinės talpyklos ir naujoji mokslinės komunikacijos infrastruktūrų sankloda, *Informacijos mokslai* 36, 53–67.

Widdowson Henry G. 2000, On the limitations of linguistics applied, *Applied Linguistics* 21(1), 3–25.

Župerka Kazimieras 2001, *Stilistika*, 2 papild. leid., Šiauliai: Šiaulių universitetas.

Aurelija USONIENĖ
Anglų filologijos katedra
Vilniaus universitetas
Universiteto g. 5
LT–01513 Vilnius
Lietuva
[*aurelia@usonis.lt*]

Jonė GRIGALIŪNIENĖ
Anglų filologijos katedra
Vilniaus universitetas
Universiteto g. 5
LT–01513 Vilnius
Lietuva
[*jone.grigaliuniene@gmail.com*]

Birutė RYVITYTĖ
Anglų filologijos katedra
Vilniaus universitetas
Universiteto g. 5
LT–01513 Vilnius
Lietuva
[*birute.ryvityte@flf.vu.lt*]

Linas BŪTĖNAS
Kompiuterijos katedra
Vilniaus universitetas
Naugarduko g. 24
LT–03225 Vilnius
Lietuva
[*linas.butenas@mif.vu.lt*]

Erika JASIONYTĖ
Lietuvių kalbos katedra
Vilniaus universitetas
Universiteto g. 5
*LT–01513 Vil*nius
Lietuva
[*erika.jasionyte@flf.vu.lt*]