ACTA LINGUISTICA LITHUANICA

XLVI (2002), 19 – 37

Dabartinės rašomosios lietuvių kalbos dažninis žodynas ir jo bazė

LAIMA GRUMADIENĖ

Lietuvių kalbos institutas

There are few electronic databases for Lithuanian. The largest are the* Frequency Dictionary of Modern Written Lithuanian*, compiled by V. Žilinskienė and L. Grumadienė, and the Lithuanian Language Corpus of Kaunas University. The former is presented in detail here: the principles of its compilation and its possible uses are discussed, and the hitherto published work in statistical linguistics based on it is briefly characterised. The main problems in connection with the databases for modern Lithuanian are: (1) their insufficient differentiation (they reflect only the written, not the spoken language), (2) their not being optimally used, as the Lithuanian scholarly community lacks the preparation needed for working with such tools, and (3) the fact that the efforts of scholars compiling them receive too little public recognition and are not valued as original scholarly achievements.

1. ELEKTRONINIŲ DUOMENŲ BAZIŲ SENĖJIMAS IR „NEREIKALINGUMAS“

Kaip tik dabar lietuvių kalbotyra yra tarsi sustingusi tigro šuolyje: po daugiausia iš struktūralizmo gauto impulso atsiradę išsiskiriantys darbai jau parašyti, tačiau ore tvyrote tvyro naujų, ne menkesnių darbų laukimas. Tam reikia ne tik kitokio požiūrio į kalbos reiškinius, bet ir naujų, dar neišbandytų metodų ir darbo priemonių.

Šiais informacinių technologijų laikais naujos kalbininkų darbo priemonės neretai yra susijusios su elektroninėmis duomenų bazėmis, o šios – su lingvostatistiniais leidiniais. Atrodytų, kad savaime aišku, jog vieni nuo kitų skiriasi, nes elektroninės bazės yra dinamiškos, su jomis galima dirbti savarankiškai, o leidiniai yra statiški, jie tėra autorių pageidavimu pateiktų duomenų sankaupa. Tačiau anksčiau, o kartais ir dabar Lietuvoje tokios bazės ir leidiniai buvo ar yra tapatinami. Tai rodo atgalinių (dar vadinamų atvirkštiniais) žodynų istorija. Pirmasis lietuviškas atgalinis žodynas, amerikiečio Davido F. Robinsono (1976) parengtas 1954 m. DLKŽ leidimo pagrindu, aprėpė apie 45 000 žodžių, jis gal nė neturėjo savo elektroninės bazės, bent autorius apie ją neužsimena. Be to, tais laikais tokia elektroninė bazė Lietuvoje būtų buvusi nė neprieinama, nes ir patį žodyną maža kas tebuvo vartęs, kadangi jis buvo nedideliu tiražu išleistas Amerikoje. Šiauliškis Juozas Korsakas (1991) atrankos principu parengė *Lietuvių kalbos inversinį*

žodyną, aprėpusį beveik 22 000 žodžių, jo pagrindas taip pat buvo DLKŽ, bet jau 1972 m. leidimas. Dirbama buvo naudojantis savo susidarytąja elektronine duomenų baze, bet pastaroji nėra viešo naudojimo (tikriausiai dėl techninių priežasčių), taigi prieinama tik spausdintos knygos pavidalu. Vidos Žilinskienės *Atgalinis dabartinės lietuvių kalbos žodynas* (1995) nėra vien DLKŽ perrašas, nes žodžiai atgaline tvarka pateikiami suskirstyti kalbos dalimis, pridedama statistinė analizė. Dabar jis iš esmės prieinamas taip pat tik spausdinta forma, nors egzistuoja ir elektroninė. Ja buvo gana nedaug pasinaudota (E. Jakaitienės vadovautai darbo grupei rašant lietuvių–norvegų žodyną, Daivos Murmulaitytės straipsniams (1998, 1999 ir kt.)). Beje, pastarasis žodynas ir analizė parengti naudojantis kompiuterine DLKŽ (1993 m.) versija ir jos statistine analize, o ne, kaip savo apžvalgoje nurodo Rūta Marcinkevičienė (2000a: 16), *Dažninio dabartinės rašomosios lietuvių kalbos žodyno* (Grumadienė, Žilinskienė 1997; 1998) baze. Bendromis Matematikos ir informatikos bei Lietuvių kalbos instituto mokslininkų pastangomis parengta DLKŽ (1993 m. leidimo) elektroninė versija buvo negausiai tiražuota (1995), nes tiesiog neturėta teisės platinti leidinius elektronine forma. Dabar Lietuvių kalbos institutas vėl ėmėsi DLKŽ, bet jau 2000 m. leidimo, kompiuterizavimo (vadovas Stasys Keinys), yra gavęs finansinę paramą iš Lietuvos mokslo fondo. Kaip žinia, 1993 m. ir 2000 m. DLKŽ leidimai maža kuo tesiskiria, todėl verčia suklusti tas faktas, kad nuspręsta ankstesne elektronine versija nesinaudoti vien dėl to, kad ji parengta jau moraliai senstelėjusioje DOS aplinkoje ir FOXPRO bazėje: vėl imtasi kompiuterio skaitliu nuskaityti jau turimą elektroninį tekstą, reikalingą tik nedidelių pataisymų, tuo didinant darbo sąnaudas ir išlaidas. Gal net ir statistinė analizė bus atliekama iš naujo?

Bet juk taip, kaip kompiuterinė 1993 m. DLKŽ leidimo versija, moraliai sensta ne viena elektroninė duomenų bazė, todėl akivaizdu, kad reikia susirūpinti jų priežiūra. Ar panašiai, kaip DLKŽ (1993) elektroninei versijai, neatsitiks ir bendromis minėtų institutų pastangomis rengiamam Terminų bankui? Taigi lingvostatistiniai leidiniai vis dar daug labiau vertinami nei brangiai kainuojančios ir daug naujoviško darbo galimybių teikiančios elektroninės duomenų bazės.

Lituanistai turėtų susitarti dėl tokių elektroninių duomenų bazių, kuriose naudojami specifiniai ženklai, pvz., senųjų raštų, transkribuotų tarmių tekstų, nes jau dabar tenka iš naujo perrinkti straipsnius su tokio tipo pavyzdžiais, jeigu dėl vienokių ar kitokių priežasčių apsisprendžiama juos pateikti ne, tarkime, Lietuvių kalbos instituto leidiniams, o kuriai nors iš Vilniaus universiteto leidinių redakcijų. Dabar leidžiami tarmių žodynai, tarmių tekstai, nors ir būdami elektroninės formos, yra tik spaustuvės modernėjimo ženklas, nes jų tekstais naudotis kaip elektronine duomenų baze neįmanoma. Taigi nesant lituanistų susitarimo dėl bendros kodavimo sistemos (pvz., UNICODE ar kitos), viešas naudojimasis elektroniniais duomenų bankais labai apsunkinamas, nekalbant jau apie tai, kad labai ribotas, o kartais ir neįmanomas tampa elektroninių technologijų, programų panaudojimas jų analizei. Negi turime jų tiek daug, kad nebepajėgiame aprėpti? Problema yra ta, kad nei kalbininkai, nei jiems talkinantys informatikos specialistai iki šiol tinkamai nesuvokia, kad duomenų bazės turi būti dinamiškos, dirbant su jomis turi būti sudarytos galimybės taikyti gausybę vis atsirandančių naujų programų ir technologijų, jos turi turėti konvertavimo į aukštesnę kompiuterinę versiją galimybę. Iki šiol dauguma Lietuvoje kuriamų elektroninių duomenų bazių sudaromos taip, tarsi tai būtų tie patys lapeliai su

kalbos duomenimis, tik surinkti kompiuterio klaviatūra. Netrukus tie duomenys gali prilygti akmenyje iškaltam dantiraščiui: naujos kartos kompiuterių ekranuose bus matyti tik „vėliavėlės" ir „kvadratukai", kurių šifravimo, nelyginant runų skaitymo, galės imtis ateities kalbininkai. Iki šiol rengtos diskusijos šiuo klausimu buvo bevaisės, bet, matyt, nereikia prarasti vilties, vėl verta kviesti diskutuoti ir tartis.

Beje, atkreiptinas dėmesys į terminus: elektroninės formos BAZĖ, BANKAS, TEKSTYNAS (KORPUSAS), ARCHYVAS. Pats madingiausias iš jų – TEKSTYNAS (KORPUSAS). Jis įvairiai apibrėžiamas, bet bendra visiems apibrėžimams yra tai, kad TEKSTYNU laikomas elektroninės formos tekstų rinkinys, todėl neretai beveik viskas, kas anksčiau vadinta duomenų bazėmis, imta vadinti tekstynais (juk atskirų tekstų dydis nereglamentuojamas). BANKO terminas dažniausiai siejamas su terminologija: paprastai terminų bankai nevadinami tekstynais. Dabar ARCHYVO terminas vis dažniau siejamas su laiko dimensija: elektroninė senų duomenų sankaupa gali būti vadinama archyvu arba baze. Taigi pats neutraliausias terminas išlieka ELEKTRONINĖ DUOMENŲ BAZĖ, todėl čia jis dažniausiai ir vartojamas (plg. Kennedy 1998: 3 tt.).

2. ELEKTRONINĖS DABARTINĖS LIETUVIŲ KALBOS DUOMENŲ BAZĖS IR LINGVOSTATISTINIAI LEIDINIAI

Elektroninės formos dabartinės lietuvių kalbos duomenų bazių pertekliaus nėra. Atitinkamai negausu ir lingvostatistinių leidinių, šiek tiek dažniau naudojamasi tik lingvostatistiniais metodais. Čia išsiskiria vienintelė fonetika ir jos rezultatais besiremianti fonologija. Gal taip atsitiko dėl pačios fonetikos prigimties, t. y. dėl to, kad dažnai kartojantis tiems patiems vienetams – garsams – tiesiog peršasi statistika grįstos išvados, o gal dėl subjektyvių veiksnių, t. y. dėl to, kad atsirado naujoviško mąstymo tyrėjų: visų pirma Aleksas Girdenis ir jo mokiniai, taip pat Antanas Pakerys. Šiaip ar taip, abu šie veiksniai lėmė lietuvių fonetikos bei fonologijos šuolį ir net savotišką atotrūkį nuo kitų kalbotyros šakų. Tiesa, pati fonetika yra tarpinė sritis tarp tiksliųjų mokslų ir lingvistikos.

Kita pagal pasikartojančių vienetų dažnumą einanti kalbotyros sritis yra morfologija. Taigi logiška būtų manyti, kad visų pirma šioje srityje jau turėtų rastis poreikis operuoti ir lingvostatistiniais duomenimis, ir elektroninėmis duomenų bazėmis. Kol kas taip neatsitiko, nors tokių darbų ir darbo priemonių jau yra. Pirmąjį lingvostatistinį leksikos ir morfologijos darbą Lietuvoje, naudodamasi ESM, parengė Vida Žilinskienė (1990). Jos *Lietuvių kalbos dažninis žodynas* remiasi vieno iš funkcinių stilių – publicistikos – duomenimis, jame pateikiama 300 000 žodžių pavartojimų analizė (t. y. tiek tų pačių leksemų, tiek ir skirtingų, o tokių būta 18 776, dėl visuotinai priimtų dažninių žodynų sudarymo taisyklių atsisakius toliau tirti 11 956 tikrinius žodžius). Visi ištisiniai tekstai, susidedantys iš atkarpų (imčių) po 1000 žodžių pavartojimų, buvo morfologiškai išanalizuoti, sukirčiuoti, po to perforuoti ir suleisti į dabar jau moraliai pasenusias ir nebenaudojamas ESM „Minsk-22". Elektroninių duomenų bankas yra, bet juo jau labai sunku būtų pasinaudoti. Gaila, bet nei leksikologine, nei morfologine, nei akcentologine vienos iš dabartinės lietuvių kalbos atmainų – publicistikos – tekstų lingvostatistine analize iki šiol tinkamai taip ir nebuvo pasinaudota nei moksliniais, nei praktiniais, ypač vadovėlių ir mokomųjų žodynų rašymo, tikslais. Vienas iš elektroninėmis priemonėmis rengtų (dalyvauta tarp-

tautinėje CHILD programoje) ir nagrinėtų sinchroninės morfologijos darbų pavyzdžių galėtų būti Inetos Savickienės (1999) daktaro disertacija, skirta vaikų kalbos morfologijai.

Beje, reikėtų paminėti, kad kitų lygmenų, stilistikos ir iš dalies sintaksės, kurių vienetų dažnumas daug mažesnis, lingvostatistiniais duomenimis besiremiančių darbų, visų pirma Audronės Bitinienės (1983, 1997, 1998 ir kt.), radosi jau prieš keliolika metų. Jau yra ir vienas kitas lingvostatistiniais duomenimis grįstas bandymas tirti dabartinę leksiką. Pastaruoju metu dažniausiai tokie darbai dirbami Vytauto Didžiojo universiteto Dabartinės lietuvių kalbos tekstyno (apie 60 000 000 žodžių pavartojimų) bazėje (pvz., Rūtos Marcinkevičienės (2000a; 2000b) atlikti kolokacijų tyrimai, Jurgitos Mikelionienės (2000) daktaro disertacija apie neologizmus ir kt.). Nedidelių elektroninių duomenų bazių pagrindu rengiama vis daugiau leksikos darbų, kartais gana specifinių, pvz., Astos Ryklienės (2000) daktaro disertacija apie interneto naršytojų vartojamą leksiką ir jos darybą, bet apskritai tokių darbų nedaug, į šį procesą menkokai yra įsitraukusi studentija (išskyrus nebent Vytauto Didžiojo universitetą). Elektroninę leksikos nagrinėjimui skirtą bazę baigia rengti Antano Pakerio vadovaujama grupė Vilniaus pedagoginiame universitete, tuomet gal ir čia padaugės tokio pobūdžio studentų darbų.

Šiame straipsnyje apsiribojama dabartinės rašytinės lietuvių kalbos tyrimams reikalingų darbo priemonių aprašymu, todėl neaptariama jau pasipylusi senųjų raštų tyrimų, besiremiančių savaja elektronine Lietuvių kalbos instituto duomenų baze, serija. Savų elektroninių duomenų bazių atsiradimas, be jokios abejonės, naują postūmį suteiks lietuvių dialektologijai: Lietuvių kalbos institute jau ruošiamasi steigti Tarmių centrą, kuriame elektronine forma būtų rengiami tarmių įrašai ir iš jų iššifruoti transkribuoti tekstai. Tame pačiame institute pradėti elektroninių onomastikos bei toponimikos duomenų bazės kūrimo darbai. Jau rengiama didžiojo *Lietuvių kalbos žodyno* elektroninė duomenų bazė turėtų tapti tikru posūkiu lietuvių leksikografijoje, daug tikimasi ir iš *Bendrinės lietuvių kalbos žodyno* elektroninės duomenų bazės. Vilniaus universiteto dėstytojai, vadovaujami Evaldos Jakaitienės, yra įsitraukę į vieną iš Europos Sąjungos projektų ir rengia elektroninį vadinamąjį universalųjį lietuvių kalbos žodžių sąrašą, kurio pagrindu turėtų pratrūkti naujoji dvikalbių žodynų serija. Tam tikru mastu tai lyg ir konkuruojanti su *Bendrinės lietuvių kalbos žodyno* redakcija firma. Daugiausia Valstybinės lietuvių kalbos komisijos jėgomis, Europos Sąjungai paakinus, pradėti rengti elektroniniai vadinamieji *Terminų* ar *Svetimžodžių portfeliai,* savotiški naujausių ir labai dažnai vartojamų įvairių sričių terminų ir žodžių samplaikų sąrašai, Europos Sąjungos platinami visoms jos narėms ir kandidatėms, kad kalbos galėtų prisitaikyti ir pasiruošti kylančiai vertimų bangai, taip pat mašininiam vertimui. Elektroninės dabartinės sakytinės kalbos duomenų bazės, kaip ir lenkų, rengiamos radijo ir televizijos kalbos įrašų pagrindu, pradžią yra padariusi grupė Vilniaus universiteto mokslininkų (vadovė Meilutė Ramonienė). Tame pačiame universitete ruošiamasi kaupti elektroninę slavizmų bazę (vadovas Vytautas Kardelis), kurioje turėtų būti tiek garso įrašų, tiek ir iš jų išrinktų bei transkribuotų pavyzdžių.

Kaip rodo patirtis, rezultatų galima bus tikėtis tik išmokus dirbti su tokio pobūdžio duomenimis, įgavus įgūdžių naudotis naujomis galimybėmis. Tuo tikslu reikalingos laboratorijos, kur šito būtų mokomasi, geriausia – nuo jaunumės. Pati elektroninių duomenų bazių prigimtis suponuoja viešumą, sutvarkytos jos turėtų būti prieinamos visiems to reikalingiems. Didžiausios lietuvių kalbos duomenų sankaupos yra Lietuvių kalbos insti-

tute, todėl modernėdamas būtent jis turėtų tapti didžiausiu elektroninių lietuvių kalbos duomenų bazių rengimo centru, viešo jų naudojimosi šaltiniu, tokio pobūdžio darbų ir idėjų traukos bei spinduliavimo centru – galbūt šį vaidmenį ir atliks Lietuvių kalbos instituto bazėje steigiamas Lituanistikos centras? Tos elektroninės duomenų bazės, kurios pateks į internetą, taps dar atviresnės. Beje, vieta, kur naudotis kompiuteriu, kai atsiranda galimybė prisijungti prie interneto, tampa nebesvarbi, svarbu tik tai, kas yra patekę į internetą. Bet tai ateities planai.

Iki šiol Lietuvoje nėra įteisinta lingvostatistinių darbų leidyba elektroniniu pavidalu, nesuderintas autorinių teisių klausimas, todėl tebėra aktualūs ir popieriuje spausdinti tokie darbai. Tačiau elektroninės duomenų bazės ir lingvostatistiniai leidiniai skiriasi iš esmės, pastarieji tėra ledkalnio viršūnė.

Toliau čia aptariamas konkretus klausimas: elektroninės dabartinės rašomosios lietuvių kalbos duomenų bazės ir lingvostatistinių žodynų santykis. Pastarųjų be bazės būti negali, tačiau bazės teikiamos galimybės toli gražu nesiriboja tik žodynų leidyba.

3. ELEKTRONINĖS DABARTINĖS RAŠOMOSIOS LIETUVIŲ KALBOS DUOMENŲ BAZĖS IR DAŽNINIAI ŽODYNAI

Elektronines kalbos duomenų bazes, savaime aišku, iš kitokių kalbos duomenų sankaupų išskiria visų pirma jų elektroninė forma ir su tuo susijusių kompiuterinių technologijų naudojimo specifika. Elektroninių duomenų bazių atsiradimas suteikė visiškai naujų galimybių tirti kalbą, bet pačios jos nesietinos tik su šiuolaikinių elektroninių laikmenų atsiradimu. Tai, kad skirtingomis priemonėmis gaunama iš esmės analogiškų rezultatų, visų pirma dažninių gramatinių formų ir leksemų charakteristikų, rodo, kad tarp šiuolaikinių elektroninių duomenų bazių ir daug senesnių jų analogų esama nemaža bendrumo. Prancūzijoje *Larousse* leidykla jau nuo 1956 m. žodynų rengimui pradėjo naudotis mašinine (senosios kartos elektroninių skaičiuojamųjų mašinų, ESM) žodynų kartoteka (Rozencvejg 1986: 82).

Klausimas dėl to, kaip sudarytinos šiuolaikinės elektroninės kalbos duomenų bazės, kilo, viena vertus, todėl, kad atsiradus naujoms technologijoms jau buvo įmanoma susidoroti su dideliu informacijos kiekiu, tuomet ėmė didėti teksto, skirto iliustruoti vieną žodį (t. y. jo konteksto), apimtis. Taip radosi naujų galimybių tirti žodžio, o vėliau ir didesnių kalbos vienetų, prezumpcijas, taigi ir sintaksę, semantiką. Imta žavėtis vis didesnėmis tekstų atkarpomis, galų gale pereita net prie ištisų baigtinių tekstų, kol to nesutramdė autorinių kūrėjų teisių įsigijimo kainos. Kita vertus, ko gero, Gutenbergo išradimo poveikiui prilygo kompiuterių atsiradimas, o svarbiausia – jų plitimas neregėtu greičiu, todėl geometrine progresija išaugus įvairiapusiams žmonių ryšiams per atstumą tiek su ta pačia kalba kalbančiais, tiek ir su kitakalbiais radosi skubus socialinis užsakymas gauti įvairialypės su kalba susijusios statistinės informacijos, kad būtų galima automatizuoti kuo daugiau bendravimo procesų: kalbų mokymo, vertimo, atpažinimo, informacijos suspaudimo, persiuntimo ir kt.

Lietuvių kalbotyra neliko nuošaly nuo šio bendro proceso, tik, žinoma, gerokai atsiliko laiko atžvilgiu. Pasaulyje per visą XX amžių, ypač antrąją jo pusę, jau tarsi buvo ruošiamasi naujam šuoliui. Visų pirma tai buvo dažniniai kalbų žodynai. Pirmieji pasirodė dar XIX a.

pabaigoje – XX a. pradžioje. Pats pirmasis – vokiečių kalbos, kurį stenografas F. Kaedingas (1898), naudodamasis lapelių kartoteka, parengė net 11 000 000 žodžių pavartojimų pagrindu. Gerokai vėliau G. Vander Beke (1929) išleido tokio pobūdžio prancūzų kalbos žodyną. H. Eatonas (1934) parengė keturių kalbų – anglų, prancūzų, vokiečių ir ispanų – tūkstančio pirmųjų dažniausių žodžių lyginamąjį žodyną. Pirmąjį anglų kalbos dažninį žodyną parengė E. Thorndike'as ir L. Lorge (1944), ispanų – V. Garcia Hozas (1953). Pirmąjį dažninį rusų kalbos žodyną Amerikoje parengė H. Josselsonas (1953), o N. P. Vacaras (1966), taip pat Amerikoje, išleido tarybinio laikotarpio rusų šnekamosios kalbos dažninį žodyną, bet iš esmės jis yra parengtas tik dramaturginių kūrinių kalbos pagrindu. E. Šteinfeldtas (Štejnfel'dt 1963) Taline išleido Puškino kalbos dažninį žodyną. Remdamasis 1 000 000 žodžių tekstynu, sudarytu iš per pusę grožinės literatūros ir vadinamosios informacinės prozos (publicistika, kanceliariniai raštai ir mokslo literatūra), dažninį rusų kalbos žodyną Upsaloje parengė L. Lönngrenas (1993). Pirmieji ne tik rusų, bet ir rumunų, italų kalbų dažniniai žodynai taip pat buvo parengti Amerikoje – A. Juillando, P. M. Edwardso, I. Juillando (1965), A. Juillando, V. Traversos (1973) (beje, A. Juillandas, E. Ch. Rodriguesas (1964) ten pat jau buvo parengę dar vieną ispanų kalbos dažninį žodyną, vėliau dar tokį parengė J. R. Alameda ir F. Cuetos (1995)). Ištisa serija dažninių žodynų buvo grindžiama Prahos lingvistinio būrelio funkcinių kalbos stilių koncepcija, taip pat Boriso Larino bendrinės kalbos samprata: pirmasis, kaip ir dera, buvo čekų kalbos, iš pradžių atskirų funkcinių stilių, o vėliau ir bendrasis, kuriuos parašė Marie Těšitelovos vadovaujamas autorių kolektyvas (J. Jelínekas, J. V. Bečka ir kt., žr. Těšitelová 1983). Anksti suskubo latviai, kaip ir čekai, iš pradžių parengę atskirų funkcinių stilių, o vėliau dviejų dalių bendrąjį žodyną – T. Jakubaite, D. Kristovska, D. Guļevska, V. Ozola, R. Prūse, A. Rubīne, N. Sika (1966–1976). Ta pačia metodika rėmėsi ir V. A. Agrajevas, V. V. Borodinas, V. M. Muratova, E. V. Tisenko, vadovaujami L. N. Zasorinos (1977), rengdami dažninį dabartinės rusų kalbos žodyną 1 000 000 žodžių pavartojimų iš keturių funkcinių stilių pagrindu. Tik vieno funkcinio stiliaus – grožinės literatūros – yra dažninis baltarusių kalbos žodynas, kurį parengė N. S. Mažeika ir A. J. Suprunas (Mažejka, Suprun 1976). Tik publicistikos stiliaus yra ir bulgarų kalbos dažninis žodynas (Todorova, Pančovska 2001). Pirmasis V. Žilinskienės (1990) parengtas lietuvių kalbos dažninis žodynas taip pat buvo tik vieno funkcinio stiliaus – publicistikos, dirbant buvo naudojamasi ta pačia čekų metodika, daug bendradarbiaujama su latviais.

Laikui bėgant, nors lietuvių kalba tuo pasigirti ir negali, pasaulyje atsirado įvairiausių dažninių žodynų: atskirų autorių kalbos, funkcinių stilių, regioninių kalbų atmainų, rašytinės ir sakytinės kalbos ir t. t. Nefunkcionalūs tapo „apskritai *n* kalbos" dažniniai žodynai, nes jų prireikė tenkinant konkrečius tikslus, taigi žodžių atranka turėjo būti daroma iš kurios nors srities tekstų, o ne iš jų visumos. Kita vertus, dažninių žodynų sudarymo problema „pateko į kitą matavimą", kai Amerikos biheivioristams ir pozityvistams ieškant vartosenos pavyzdžių radosi elektroninių tekstynų: elektroninių technologijų raida buvo sparti, todėl galima buvo ne tik kaupti, bet ir apdoroti didžiulius tekstų masyvus.

Tekstynų lingvistikos pradžia laikomi 1961 m., kai Nelsonas Francis ir Henry Kučera anonsavo ir ėmėsi kurti garsųjį elektroninį rašytinės kalbos Browno tekstyną (Brown university), Johnas Sinclairas – to paties formato LOB (Lancaster-Oslo/Birmingen) teks-

tyną; dar 1959 m. Randolphas Quirkas anonsavo rašytinės ir sakytinės kalbos tekstyną SEU (Survey of English Usage), jį tęsė Janas Svartvikas, Lundo universitete ėmęsis kurti sakytinės kalbos tekstyną LLC (London-Lund Corpus of Spoken English). Visų minėtų tekstynų apimtis tuomet buvo apie 1–1,15 mln. žodžių pavartojimų, po to atsirado daugiau panašių tekstynų (žr. Aijmer, Altenberg 1991: 1tt.). Antrosios generacijos tekstynų apimtis skaičiuojama dešimtimis milijonų, trečiosios – jau šimtais ir tūkstančiais. Pasaulyje išlieka aktuali kompiuterių programinės įrangos ir atskirų darbo programų suderinamumo problema: sukaupta jau nepaprastai didelės apimties tekstynų – šimtų milijonų žodžių (tai iš esmės bibliotekos elektronine forma), bet kaip patogiai jais naudotis, kaip suskaičiuoti elementariausius dažnius, kaip nepasimesti žodžių vandenyne? Galų gale ar kiekybė visuomet lemia kokybę? Sudarant įvairiausius dažninius žodynus daugybę kartų įsitikinta, kad pusė tirtų žodžių būna pavartota vos vieną kartą. Todėl kiekvieną kartą imantis darbo reikia gerai apsvarstyti, ar išsikeltieji uždaviniai tikrai reikalauja dirbti su gausybe medžiagos, ieškant rečiausių žodžių. Kartais, kai tiriami kurie nors gramatiniai ar žodžių darybos dėsningumai, tikrai būna neracionalu itin didinti tiriamosios aibės apimtį, viską dera įvertinti ir statistiniais kriterijais, juk jie parodys tikimybę, kada atranka pakankama daryti reprezentatyvias išvadas. Ypač atkreiptinas dėmesys į tai, iš kurių tekstų sudaroma tiriamoji aibė, t. y., kaip sudaromas tiriamojo objekto modelis. Lietuvių kalbotyroje bene daugiausia klausimų šiuo atžvilgiu dabar kelia problema, ką laikyti bendrine lietuvių kalba, kokiomis dalimis joje turėtų būti atstovaujami funkciniai kalbos stiliai ar pagal kitus kriterijus išskirti kalbos sluoksniai. Štai minėtasis Vytauto Didžiojo universiteto tekstynas sudarytas pagal vienus atrankos principus, o *Dabartinės rašomosios lietuvių kalbos dažninio žodyno* – pagal kitus, todėl ir rezultatai gaunami skirtingi. Reikia labai atidžiai ir objektyviai vertinti statistinius duomenis apie, pvz., atskiras kalbos dalis, gautus iš skirtingų tekstynų. Pastebėta, kad jau parašytuose ar rašomuose tokio pobūdžio darbuose nepakankamai skiriama dėmesio tam, kokie tekstai buvo tirti, ir daromi apibendrinimai apie „dabartinę lietuvių kalbą“. Darbas su skaičiais reikalauja preciziškumo, kitaip gaunamos klaidinančios išvados.

Straipsnyje aptariamo *Dabartinės rašomosios lietuvių kalbos dažninio žodyno* elektroninis tekstynas (1,2 mln. žodžių pavartojimų) skirtinas pirmosios generacijos tekstynams (Brown, LOB tipo) – per milijoną žodžių apimties. Su šiais garsiaisiais tekstynais jį sieja dar viena bendrybė – kaip retas kuris vėlesnis pasaulyje, ir mūsiškis yra anotuotas. Kompiuterinės technologijos analitinėms kalboms suteikė išskirtines sąlygas: specialios, bet gana plačiai paplitusios programos (ypač žinoma Oksfordo universiteto – *Oxford Concordance Program,* taip pat *KWIC, KAYE, CLAN, OCP, WordCruncher* ir kt.) leidžia gana nesunkiai parengti dažniausių žodžių sąrašus, pateikti jų konkordansijas (Leech 1991: 10; Utka 2000). Tokio pobūdžio programomis paprastai naudojasi ir Vytauto Didžiojo universiteto Dabartinės lietuvių kalbos tekstyno tyrėjai, bet tuomet galima analizuoti tik *žodžius,* t. y. leksiką, nes pristingama žinių apie žodžio galą, o jis, kaip žinia, lietuvių kalbai itin svarbus. Taigi sintetinėms kalboms sudėtingiau: juk naudojantis aprašytosiomis programomis parengiami žodžių formų, o ne antraštinių žodžių sąrašai. Reikia gana daug papildomo darbo, kol žodžių formos subendravardiklinamos ir tampa tik antraštiniais žodžiais. Tokie žodžių sąrašai, turintys ir gramatinį nagrinėjimą, jau vadinami anotuotais, pasaulyje jų nėra daug, jie labai vertinami, nes reikalauja nemažų darbo sąnaudų, todėl yra

labai brangūs (mudvi su V. Žilinskiene esame suskaičiavusios, kad mūsų atveju vien tekstų parinkimas, surinkimas kompiuterio klaviatūra, pusiau automatizuotas morfologinis jų nagrinėjimas (vadinamasis LEMAVIMAS), koregavimas ir duomenų suleidimas į kompiuterį iš vieno žmogaus pareikalautų šešerių darbo metų). Skaičiavimams ir tolesnėms korektūroms vėl reikia kelerių metų, nes bet koks netikslumas atliekant matematinius veiksmus gali virsti absurdu.

Apskritai dažninių žodynų, ypač atsiradus elektroniniams tekstynams, dabar galima susidaryti įvairiai ir įvairių – nelygu ko ieškai. Svarbiausia priežastis, kodėl rengiami vis nauji dažniniai žodynai, ypač pretenduojantys vadintis, pvz., dabartinės *n* kalbos žodynais, yra ta, kad pagrindinė lingvostatistikos, kaip ir apskirtai statistikos, sąlyga yra ta, kad *atranka turi būti reprezentatyvus tiriamosios visumos, arba aibės, modelis.* Žodžiu, nuolat suabejojama, ar tirtoji atranka tikrai atspindi numatytąją visumą. Pavyzdžiui, Vytauto Didžiojo universiteto Dabartinės lietuvių kalbos tekstyno, kuris sudarytas net iš 60 mln. žodžių pavartojimų, parengtas dažninis dabartinės lietuvių kalbos žodynas būtų kitoks nei jau išleistieji L. Grumadienės ir V. Žilinskienės (1997, 1998), nes jo rezultatai atspindėtų tekstų visumą, kurių pagrindą sudaro 70% periodinių, 26% neperiodinių ir 4% verstinių leidinių kalba (Marcinkevičienė 2000a: 16). Išleistų minėtųjų žodynų tekstai atrinkti visai kitu pamatu: tai originalūs, ne verstiniai, keturių funkcinių kalbos stilių – publicistinio, kanceliarinio, beletristinio ir mokslinio – tekstai. Taip pasielgta todėl, kad lietuvių bendrinė, ypač rašomoji kalba, iki šiol tebėra formuojama, normuojama ir kodifikuojama Prahos lingvistinio būrelio idėjų dvasia (beje, atkreiptinas dėmesys į tai, kad dabar pats STILIAUS terminas įvairiausių kalbotyros mokyklų yra gana skirtingai apibrėžiamas). Laikytasi kitų dažninių žodynų sudarytojų principų, visų pirma čekų, rusų, latvių, kurie irgi laikėsi anksčiau čia išdėstytos bendrinės kalbos – funkcinių kalbos stilių sumos – sampratos, nuostatų, taikyta jų patirtis, todėl, kaip ir jų, pasirinktos 300 000 žodžių pavartojimų atrankos iš keturių funkcinių stilių, nes manyta, kad rašomajai lietuvių kalbai dar per ankstyvi Vytauto Didžiojo universiteto tekstyno (jis, beje, atsirado vėliau nei mūsiškis) sudarymo principai, artimesni amerikiečių bendrinės (ar standartinės, kaip ten įprasta) kalbos sampratai. *Dabartinės rašomosios lietuvių kalbos dažninio žodyno* duomenų bazėje kartais pasigendama buitinės leksikos, nes pamirštama, kad ji suformuota tik viešosios rašytinės kalbos pamatu: joje nėra nei privačių laiškų, dienoraščių, nei sienų užrašų ar pan. Tačiau ir dabar į žodyną pateko šnekamosios kalbos žodžių, tarp jų ne tik svetimybių, tarmybių, bet ir netoleruotinų bendrinėje kalboje pejoratyvinės leksikos pavyzdžių, dėl ko žodyno autorės nuolat susilaukia priekaištų. Kyla klausimas, ką laikyti grožine literatūra: tie žodžiai buvo grožinės literatūros tekstuose (Lietuvoje spausdintuose daugiau nei 100 egzempliorių tiražu redaguotuose originaliuose tekstuose), o dažninio žodyno sudarymo postulatas yra neišmesti nė vieno prasminio žodžio, jeigu jis tik nėra tikrinis, skaičius, santrumpa ar korektūros klaida. Beje, duomenys rodo, kad kiekvieno iš keturių funkcinių stilių parinkus 300 000 žodžių pavartojimų atrankas, publicistiniame stiliuje rasta 284 687 prasminių žodžių (15 313 vadinamųjų „šiukšlių"), dalykiniame stiliuje – 279 505 (dėl minėtų priežasčių nepatekusių į žodyną – 20 495), moksliniame – 278 311 (21 689), beletristiniame – 290 561 (9 439). Širdis neleido išmesti tikrinio žodžio *Lietuva,* todėl jis kaip išimtis vienintelis tikrinis pateko į žodyną (jis, pasirodo, 16-as pagal dažnumą žodis, pirmasis iš daiktavardžių, jo dažnis 5151 net 614 imčių: čia lėmė oficialiųjų dokumentų,

įstaigų ir pan. pavadinimai, kuriuose dažnai būta šio žodžio, pavartoto kilmininko linksniu). Sąlygiškai tikriniais gali būti laikomi ir patekę religinių bei kitų švenčių pavadinimai (dažnai vartojami ir kaip bendriniai), pvz., *Joninės, Kūčios, Žolinė,* arba pasaulio šalių pavadinimai *Rytai, Vakarai, Pietūs, Šiaurė.*

Laikantis reprezentatyvumo principo, kiekvieno iš minėtųjų stilių buvo atrenkami tik tokie tekstai, kurie skirtini grynajam stiliui: atsisakyta paribiniais laikomų vadovėlinių tekstų, mokslinės publicistikos, siauram skaitytojų ratui skirtos publicistikos, dramos, poezijos ir pan. Publicistikos tekstus parinko Vida Žilinskienė, juos sudarė 1994–1995 m. 19-os redaguojamų centrinių dienraščių ar savaitraščių ir plačiai skaitomų žurnalų tekstai, reprezentuojantys penkias temines grupes: 1) politika, 2) pramonė ir žemės ūkis, 3) švietimas ir teisė, 4) menas ir 5) pramogos (sportas, horoskopai, orų prognozės). Dalykinio stiliaus tekstus parinko Pranas Kniūkšta ir Danutė Vainauskienė iš 1990–1995 m. leidinių ir dokumentų pagal septynias temines grupes: 1) įstatymai ir juos lydintys dokumentai, komentarai, 2) teisės dokumentai, 3) ekonomikos ir verslo žinynai bei dokumentai, 4) politologijos leidiniai, partijų dokumentai, 5) standartai ir statistikos leidiniai, 6) informacijos šaltiniai (bibliografijos, katalogai, žodynai), 7) atskirų socialinių sričių (krašto apsaugos, socialinės saugos, švietimo, kultūros ir kt. dokumentai). Mokslo stiliaus tekstai parinkti Irenos Andriukaitienės iš 1990–1995 m. Lietuvoje išleistų mokslo leidinių arba mokslo institucijų aprobuotų darbų pagal aštuonias temines grupes (mokslo kryptis): 1) agrariniai mokslai, 2) gamtos, 3) humanitariniai, 4) matematikos, 5) medicinos, 6) socialiniai, 7) technikos, 8) teologijos. Beletristikos stiliaus tekstus parinko Laima Grumadienė iš 1985–1990 m. (su nedidelėmis išimtimis) Lietuvoje išleistų grožinės literatūros prozos kūrinių, o jų šešios „teminės" grupės suskirstytos pagal rašytojų gimtines: 1) žemaičiai, 2) rytų aukštaičiai, 3) vakarų aukštaičiai, 4) pietų aukštaičiai, 5) miestiečiai, 6) išeiviai (į tikrąją pastarųjų gimtinę nebuvo atsižvelgiama).

Buvo nuspręsta, kad būtent šie tekstai gali reprezentuoti dabartinę rašomąją lietuvių kalbą. Iš jų atsitiktine tvarka (puslapiai ir eilutės parenkamos pagal atsitiktinių skaičių lentelę) būdavo pasirenkama po vieną 1000 žodžių pavartojimų apimties atkarpą, vadinamąją imtį, ir kompiuterio klaviatūra surenkama (maždaug dviejų valandų darbas). Po to imtys buvo lemuojamos, t. y. morfologiškai nagrinėjamos Vytauto Zinkevičiaus parengta MAN (*M*orfologinio *A*nalizavimo ir *N*ormalizavimo) programa. Ją autorius pritaikė specialiai šiam darbui iš savo paties seniau parengtos *kompiuterinės žodžių rašybos koregavimo programos,* 2000 m. Vytauto Didžiojo universiteto užsakymu ją šiek tiek patobulino, perkėlė į aukštesnę versiją ir pavadino *lemuokliu* (Zinkevičius 2000).

Tas darbas atrodė maždaug taip: gaunama kompiuterinė išklotinė, iš kurios reikia „išmėtyti" nereikalingas (mano pabraukta) eilutes. Pvz., jau žmogaus, ne tik mašinos apdorotos teksto atkarpos *(...) manęs visa ši komedija (...)* išklotinė tokia:

1 PAV. TEKSTO LEMAVIMO PAVYZDYS

```
    aš
    manyti
manęs       įvr vns K
manęs       vks dlv veik bk neiv vyr vns V
```

visas
visa įvr mot vns V
visa įvr mot vns I
visa įvr mot vns Š
visa įvr bvr

šis
ši įvr neįvr mot vns V

komedija
komedija dkt mot vns V
komedija dkt mot vns I
komedija dkt mot vns Š

Pirmojoje eilutėje lieka atstatytas antraštinis žodis, antrojoje (likusioje neišbrauktoje) – žodžio forma su gramatinėmis charakteristikomis. Toks darbas su 1000 žodžių užtrunka apie 4 val., bet spartesnio būdo morfologiškai išnagrinėti lietuviškus žodžius, ko gero, dar nėra. Po to ieškoma homonimų ir homoformų, imtys koreguojamos Vytauto Zinkevičiaus programa MAN-K ir suleidžiamos į kompiuterį – visa tai trunka dar valandą. Iš viso taip buvo apdorota 1 200 imčių po 1000 žodžių pavartojimų.

Atlikus šį didžiulį darbą, duomenys toliau buvo apdorojami Valentino Černiausko ir Snieguolės Meškauskienės (abu iš Matematikos ir informatikos instituto) parengtomis skaičiavimo programomis: buvo suskaičiuoti įvairūs dažniai. Spaudai buvo parengti trys žodynai, du iš jų jau išleisti (Grumadienė, Žilinskienė 1997; 1998), o trečiasis išvys dienos šviesą, tikėkimės, šiemet. Pirmieji du žodynai – tai antraštinių žodžių dažniniai sąrašai. Pirmasis – mažėjančio dažnio tvarka, antrasis – abėcėlės. Į juos atitinkamai pateko 38 337 skirtingi žodžiai. Trečiasis – žodžių formų žodynas – keliskart didesnis (31 a. l.), turintis maždaug 150 000 žodžių formų, atspindi skirtingų ir tų pačių žodžių vartojimą įvairiomis gramatinėmis formomis (jis turėtų būti pateikiamas tiek abėcėlės, tiek atgaline tvarka). Parengta dar kelios dešimtys įvairiai išdėstytų dažninių žodynų, skirtų vartoti elektronine forma: visų keturių stilių, visų kiekvieno stiliaus teminių grupių ir kt. (tiek antraštinių žodžių, tiek ir žodžių formų).

Palyginus kai kuriuos statistinius duomenis, jau dabar matyti, kad lietuvių rašomojoje kalboje daugiausia vartojama skirtingų daiktavardžių – iš visų skirtingų žodžių net 45,25% yra daiktavardžiai, gausu veiksmažodžių ir veiksmažodinių formų – 32,65%, reikšminga ir būdvardžių dalis – 15,31%, skirtingų prieveiksmių rasta 5,03%. Kitų kalbos dalių nėra nė po pusę procento (iš viso jų tėra 0,88%). Bet tai dar nereiškia, kad ir tekste jų pasiskirstymas būtent toks. Kitas rodiklis – tų pačių žodžių dažnumo tekste – rodo, kad pagrindinių kalbos dalių ir prieveiksmio procentas gerokai sumažėja tarnybinių kalbos dalių ir įvardžio sąskaita, nes tekstuose pastarieji sudaro jau daugiau nei ketvirtadalį – 26,71%. Atkreiptinas dėmesys ne tik į dažnumo rodiklį, bet ir į žodžio ar žodžio formos pasirodymo atskirose imtyse skaičių: taip galima nustatyti, ar kurio nors žodžio dažnį yra lėmęs atsitiktinai patekęs vienas tekstas su daug kartų jame pasikartojančiu žodžiu, ar vis dėlto jo dažnumas nėra atsitiktinis, nes jo dispersija įvairiose imtyse tolygi.

Leksikos įvairovė vertintina iteracijos koeficientu (žodžių pavartojimų ir antraštinių žodžių santykiu), kuris minimame žodyne lygus 20,06 (kuo šis dydis mažesnis, tuo, vadinasi, daugiau rasta skirtingų žodžių). Iš suskaičiuotų, bet nepublikuotų kiekvieno iš keturių funkcinių stilių antraštinių žodžių ir žodžių formų žodynų (apie kai kuriuos juose pastebėtus lingvostatistinius ypatumus jau rašė V. Žilinskienė (1998, 2001, spausd. ir kt.)) aiškėja, kad, pvz., grožinės literatūros leksika net dukart įvairesnė už dalykinio stiliaus tekstų leksiką (iteracijos koeficientai atitinkamai lygūs 9,53 ir 20,78). Įdomiausia čia tai, kad visų keturių stilių kartu paėmus iteracijos koeficientas gana nedaug tesiskiria nuo dalykinio stiliaus, bet labai daug – nuo grožinės literatūros stiliaus (publicistikos lygus 10,51, mokslo – 12,63). Šie rodikliai yra svarus argumentas apsisprendžiant dėl, pvz., bendrinės kalbos žodyno sudarymo principų: kokiomis dalimis turėtų būti atstovaujami funkciniai kalbos stiliai; ar iš viso dar tęstina praktika, kai bendrinė kalba siejama su funkciniais stiliais; jeigu taip, būtina atsižvelgti į statistinius funkcinių stilių rodiklius ir kt.

Įdomių duomenų gauta sugretinus Vytauto Didžiojo universiteto tekstyno pagrindu sudarytą dažniausių žodžių sąrašą (jame žodžiai fiksuoti ir skaičiuoti tomis formomis, kuriomis jie pasitaiko tekste, neatstatinėjant, priešingai nei *Dabartinės rašomosios lietuvių kalbos dažniniame žodyne,* antraštinių jų lyčių). Palyginus 50 dažniausių nekaitomų žodžių ir kaitomų žodžių formų 60 mln. žodžių lietuvių tekstyne (Utka 2000: 278) su *Dažninio žodyno* (Grumadienė, Žilinskienė 1997: 3) pirmaisiais 50 antraštinių žodžių, galima akivaizdžiai įsitikinti, kad jie iš esmės sutampa. Tik pirmajame dažniniame sąraše yra šiek tiek mažiau skirtingų žodžių, nes ten minimos įvairios to paties žodžio formos, pvz., *buvo, būti; jis, jo, jų, jos, jie,* o antrajame sąraše, t. y. žodyne, jas atitinkamai reprezentuoja *būti* ir *jis.* Į *Dažninį žodyną,* laikantis tokio tipo žodynų sudarymo principų, nepateko santrumpos, bet vis tiek iš minėtame dažniausių žodžių sąraše buvusių santrumpų *m, d, a* – atitinkamai *metai, diena, amžius* – palyginimo su *Dažninio žodyno* duomenimis aiškėja, kad pirmieji du žodžiai dažni ir žodyne – 30-as ir 57-as pagal dažnį, t. y. rangą (to paties rango žodžiais laikomi vienodą dažnumą turintys žodžiai), o *amžius* patenka tik į 3-iąjį šimtą. Beje, santrumpa ir visas šis žodis gana dažnai vartojami ir šiek tiek skirtingomis reikšmėmis. Kaip patvirtinta daugelio sintetinių kalbų dažninių žodynų sudarymo praktikos, patys dažniausi *Dabartinės rašomosios lietuvių kalbos dažniniame žodyne* yra kai kurie jungtukai (*ir, kad, o, kaip, ar, bet, bei, kai* ir kt.), prielinksniai (*į, iš, su, nuo, apie, po, už, dėl, per, prie* ir kt.), įvardžiai (*jis, tas, aš, kuris, šis, visas, kitas, savas, toks, kas, tu* ir kt.), dalelytės (*ir, ne, tik* ir kt.), populiariausi veiksmažodžiai yra *būti, galėti, turėti, nebūti, reikėti, žinoti, nustatyti* ir kt., o daiktavardžiai – *Lietuva, darbas, žmogus, metai* ir kt. Iš skaitvardžių dažniausi *vienas, pirmas, du,* o iš būdvardžių, patenkančių į pirmąjį dažniausių žodžių šimtą – *didelis, naujas, aukštas, įvairus, svarbus, pagrindinis.* Beje, ir analitinių kalbų duomenys rodo, kad, pvz., anglų kalbos tekstuose įvardis *I,* jungtukas *and* ir artikelis *the* sudaro 10% tekstų (Hillerich www: eduplace.com).

Svarbus ir tarptautinių žodžių klausimas – kiek jų ir kuriuos reikėtų dėti į bendrinės kalbos žodyną? Įsigilinus gali duoti naudos ir žinojimas to, kad, pvz., *Dabartinės rašomosios lietuvių kalbos dažninio žodyno* (1998, XXXIV–XXXVI) antraštinių žodžių sąraše rasta 27,52% tarptautinių žodžių ir kad tekstuose jie užėmė 13,34% vietos, taip pat gali būti naudingi patys tų žodžių sąrašai.

Rengiantis rašyti naują lietuvių kalbos gramatiką, svarbu žinoti ir žodžių formų statistiką: kurio tipo žodžiai kaip kaitomi, kodėl ir kurių žodžių gramatinių formų įvairovė didelė, o kitų – ne. Tokių duomenų gausu rengiamame spaudai žodžių formų žodyne. V. Žilinskienė (2001) vien apie publicistikos stilių jau paskelbė įdomių duomenų. Pasirodo, kad trys daiktavardžių linksniai sudaro 82,29% visų daiktavardžių pavartojimo atvejų (kilm. – 38,91%, vard. – 26,62%, gal. – 16,76%, o tai, kad, pvz., naud. vartojamas tik 3,88% atvejų, gal jau rodo analitiškėjimo tendencijas?). Būdvardžiai, beje, dažniausiai vartojami vard. linksniu – 36,93% atvejų. Išsiaiškinta, kad vyrauja daiktavardžių vienaskaita – 68,86% visų pavartojimų. Publicistikos tekstuose palyginti nedaug terasta padalyvių (2,34%) ir pusdalyvių (1,84%), o dalyvių – net 24,46%, bendratis vartota 15,85% atvejų.

Įdomių ir neįprastų rezultatų, ypač pedagogams ir specialistams, rengiantiems vadovėlius, mokymo programas, žodynus, gauta pažvelgus į V. Žilinskienės ir L. Grumadienės (1998; 1999a; 1999b) parengtus ir išleistus tris *Valstybinės kalbos mokėjimo kvalifikacinių kategorijų lietuvių–rusų kalbų minimaliuosius žodynus,* taip pat į kitus šešis parengtus, bet dar neišleistus analogiškus lietuvių–anglų kalbų ir lietuvių–lenkų kalbų žodynus. Čia dar kartą atkreiptinas dėmesys į tai, kad imtys rinktos iš lietuvių rašomosios kalbos tekstų, taigi šnekamosios kalbos žodžių čia galima ir pasigesti. Šie žodynėliai yra orientuoti į rašomosios kalbos mokymąsi, nes valstybinės kalbos mokėjimo kvalifikaciniai reikalavimai yra susiję su valstybinės kalbos vartojimu viešai, visų pirma oficialiomis situacijomis, ypač valstybinėse įstaigose, todėl čia ir reikalingesnė yra ne buitinė, o neutralioji ir dalykinė leksika: reikalaujama sugebėti perskaityti ir parašyti dalykinius raštus, suprasti žiniasklaidos tekstus. Ateityje rengtini ir šnekamosios kalbos minimalieji žodynai. Lietuvoje yra išleistas Vytauto Šerno (1994) vadovėlis *Kaip racionaliau mokytis kalbų*, kuriame yra ir minimalusis dvikalbis anglų–lietuvių kalbų žodynas: išverstas garsusis Ch. K. Ogdeno 1942 m. publikuotas *Basic English*, sudarytas iš 850 žodžių (jį, beje, galima rasti ir *Encyclopaedia Britannica* 3-iajame tome). Bėgant laikui vieni žodžiai sensta, kiti atsiranda, todėl tų 850 žodžių sąrašas gali atitinkamai kisti. Nustatant valstybinės lietuvių kalbos mokėjimo trijų kvalifikacinių kategorijų reikalavimus, buvo atsižvelgta ir į *Dabartinės rašomosios lietuvių kalbos dažninio žodyno* autorių nuomonę dėl minimaliųjų žodynų apimties: apsispręsta, kad I-osios kategorijos reikalavimus atitiktų apie 800 žodžių, II-osios – apie 1200, III-iosios – apie 2500 žodžių mokėjimas. Remtasi tiek visuotinai priimtu Ogdeno postulatu dėl 850 žodžių, tiek savosios ir kitų kalbų dažnumų tyrimų rezultatais, tiek kitų valstybių analogiška praktika. Tai, kad šie skaičiavimai nėra iš piršto laužti, matyti ir iš priede pateikiamų trijų lentelių, kur aiškiai matyti, kokią dalį tekstų sudaro atitinkamai 800, 1200 ir 2500 dažniausių rašomosios lietuvių kalbos žodžių: 56,66%, 65,70% ir 78,04%.

Jau tampa įprasta, kad dabar visi didieji žodynai pateikia nuorodas apie dažniausius kalbos žodžius. Pvz., anglų kalbos COBUILD žodyne (www.antimoon.com) 680 dažniausių žodžių atitinkamai žymimi 1–4 rombeliais. Longmano anglų kalbos žodyne 3000 dažniausių žodžių pažymėti atitinkamai S1, S2 (S – *spoken English*), W1, W2 (W – *written English*). Macmillano anglų kalbos žodyne dažniausi žodžiai spausdinami raudona spalva, kitų dažnumas žymimas atitinkamai 1–3 žvaigždutėmis. Artėja metas, kai, matyt, teks ir lietuvių leksikografijai tuo užsiimti.

4. IŠVADOS

Elektroninių dabartinės lietuvių kalbos duomenų bazių nėra daug, didžiausios iš jų yra V. Žilinskienės ir L. Grumadienės *Dabartinės rašomosios lietuvių kalbos dažninio žodyno* (morfologiškai anotuota ir statistiškai apdorota, 1 200 000 žodžių pavartojimų) ir Vytauto Didžiojo universiteto *Lietuvių kalbos tekstynas* (neanotuotas, apie 60 000 000 žodžių pavartojimų). Lituanistikoje dar nėra tinkamos praktikos naudotis tiek elektroninėmis dabartinės kalbos duomenų bazėmis, tiek lingvostatistiniais leidiniais. Visų pirma, per mažai pratinama taip dirbti Lietuvos aukštosiose mokyklose, bet yra ir kita medalio pusė – per maža viešai prieinamų elektroninių lietuvių kalbos duomenų bazių. Viena iš atsilikimo priežasčių yra ir ta, kad elektroninės duomenų bazės nėra įteisinamos kaip autorinis darbas. Viena iš aktualiausių problemų Lietuvoje, susijusių su elektroninėmis duomenų bazėmis, yra jų sudarymo koncepcijos silpnumas: dažnai elektroninėmis duomenų bazėmis numatoma naudotis tik kaip elektronine forma pateiktu tekstu, nenumatant kompiuterinių technologijų pritaikymo duomenims apdoroti.

1 LENTELĖ: VALST.LIETUVIŲ KALBOS MOKĖJIMO I KVALIFIKACINĖS KALBOS DALIŲ IR BENDRASIS ŽODŽIŲ DAŽNUMAS

Rangas	*Dažnis*	*dkt*	*vks*	*bdv*	*prv*	*įvr*
1–10	39582-7001	–	21262 14,38%	–	–	52221 35,31%
11–20	6877-4274	–	10582 19,96%	–	–	21221 40,02%
21–30	4113-2989	10763 28,95%	–	–	–	11048 29,71%
31–40	2950-2637	–	–	–	5619 20,35%	–
41–50	2631-2053	–	6843 30,50%	2053 9,15%	4782 21,32%	2113 9,42%
51–60	2015-1781	9284 45,22%	–	–	3867 18,83%	–
61–70	1754-1557	10014 60,41%	1739 10,49%	–	–	1593 9,62%
71–80	1548-1461	6036 40,04%	3046 20,20%	1507 10,00%	3009 19,96%	–
81–90	1460-1373	1422 10,03%	2828 19,94%	1373 9,68%	2888 20,36%	2792 19,68%
91–100	1351-1148	7421 54,63%	1305 9,61%	–	1232 9,07%	1159 8,53%
101–150	1442-905	23551 43,03%	12523 22,88%	7419 13,55%	3116 5,69%	943 1,72%
151–200	900-706	22744 52,48%	10742 24,79%	870 2,01%	4203 9,70%	1583 3,65%
201–250	702-564	14059 42,09%	6315 18,90%	4906 14,69%	3178 9,51%	3148 9,42%
251–300	562-462	17325 49,59%	7962 22,79%	1451 4,15%	3605 10,32%	1041 2,98%

KATEGORIJOS MINIMALIOJO ŽODYNO LINGVOSTATISTINĖ ANALIZĖ:

skt	*prl*	*jng*	*dll*	*išt; jst*	*Suma*
–	25050 16,93%	49371 33,38%	–	–	147904 12,33%
–	–	16902 31,87%	4320 8,15%	–	53025 4,42%
–	7156 19,25%	4102 11,03%	4113 11,06%	–	37182 3,10%
2641 9,56%	8282 29,99%	8281 29,99%	2791 10,11%	–	27614 2,30%
2116 9,43%	4526 20,18%	–	–	–	22433 1,87%
–	3589 17,48%	3793 18,47%	–	–	20533 1,71%
1557 9,39%	1673 10,09%	–	–	–	16576 1,38%
–	–	1478 9,80%	–	–	15076 1,26%
–	1427 10,06%	–	1454 10,25%	–	14184 1,18%
–	–	1237 9,11%	1230 9,05%	–	13584 1,13%
1007 1,84%	2164 3,95%	2986 5,46%	1028 1,88%	–	54737 4,56%
738 1,70%	–	900 2,08%	1557 3,59%	–	43337 3,61%
–	–	574 1,72%	1225 3,67%	–	33405 2,78%
–	1035 2,96%	512 1,47%	2007 5,74%	–	34938 2,91%

Rangas	*Dažnis*	*dkt*	*vks*	*bdv*	*prv*	*įvr*
301–350	460-374	19646	10424	2420	2073	393
		54,28%	28,80%	6,69%	5,73%	1,09%
351–400	373-313	18504	7550	3381	1347	654
		53,04%	21,64%	9,69%	3,86%	1,87%
401–450	312-260	19156	7865	3976	522	304
		57,20%	23,49%	11,87%	1,56%	0,91%
451–500	259-205	21718	9987	4806	2486	471
		53,17%	24,45%	11,76%	6,08%	1,15%

Valstybinės lietuvių kalbos I-osios kvalifikacinės kategorijos minimalusis žodynas (apie 800

2 LENTELĖ: VALST. LIETUVIŲ KALBOS MOKĖJIMO II KVALIFIKACINĖS KALBOS DALIŲ IR BENDRASIS ŽODŽIŲ DAŽNUMAS

Rangas	*Dažnis*	*dkt*	*vks*	*bdv*	*prv*	*įvr*
501–550	204-155	22494	13153	7023	2481	525
		46,97%	27,46%	14,66%	5,18%	1,10%
551–600	154-106	30834	17197	6221	4419	151
		50,91%	28,39%	10,27%	7,29%	0,25%

Valstybinės lietuvių kalbos II-osios kvalifikacinės kategorijos minimalusis žodynas (apie 1200

3 LENTELĖ: VALST. LIETUVIŲ KALBOS MOKĖJIMO III KVALIFIKACINĖS KALBOS DALIŲ IR BENDRASIS ŽODŽIŲ DAŽNUMAS

Rangas	*Dažnis*	*dkt*	*vks*	*bdv*	*prv*	*įvr*
601–650	105-56	45751	20328	12145	5822	433
		52,31%	23,25%	13,88%	6,66%	0,50%

Valstybinės lietuvių kalbos III-osios kvalifikacinės kategorijos minimalusis žodynas (apie 2500

skt	*prl*	*jng*	*dll*	*išt; jst*	*Suma*
426	–	–	812	–	36194
1,17%			2,24%		3,02%
–	–	1407	2043	–	34886
		4,03%	5,87%		2,91%
276	–	272	1118	–	33489
0,82%		0,81%	3,34%		2,79%
255	224	210	700	–	40857
0,62%	0,55%	0,51%	1,71%		3,40%

žodžių) dengia 56,66% teksto (atranka iš 1,2 mln. žodžių pavartojimų). 56,66%

KATEGORIJOS MINIMALIOJO ŽODYNO LINGVOSTATISTINĖ ANALIZĖ:

skt	*prl*	*jng*	*dll*	*išt; jst*	*Suma*
530	342	177	795	371	47891
1,11%	0,71%	0,37%	1,66%	0,78%	3,99%
520	275	132	713	109	60571
0,86%	0,45%	0,22%	1,18%	0,18%	5,05%

žodžių) dengia 65,70% teksto (atranka iš 1,2 mln. žodžių pavartojimų). 9,04%

Iš viso: (=65,70%)

KATEGORIJOS MINIMALIOJO ŽODYNO LINGVOSTATISTINĖ ANALIZĖ:

skt	*prl*	*jng*	*dll*	*išt; jst*	*Suma*
460	341	372	1697	102	87446
0,52%	0,40%	0,42%	1,94%	0,12%	7,29%

žodžių) dengia 78,04% teksto (atranka iš 1,2 mln. žodžių pavartojimų). 7,29%

Iš viso: (=78,04%)

LITERATŪRA

AIJMER, K., ALTENBERG, B. 1991: Introduction. Aijmer, K., Altenberg, B., eds., *English Corpus Linguistics. Studies in Honour of Jan Svartvik,* London–New York: Longman.

ALAMEDA, J. R., CUETOS, F. 1995: *Diccionario de frecuencias de las unidades lingüísticas del castellano,* Universidad de Oviedo.

BITINIENĖ, A. 1983: *Mokslinis stilius,* Vilnius: Aukštojo ir specialiojo mokslo ministerija.

BITINIENĖ, A. 1997: *Funkciniai stiliai: sakinio ilgis ir struktūra,* Vilnius: Vilniaus pedagoginio universiteto leidykla.

BITINIENĖ, A. 1998: Administracinis stilius ir jo sakinio ilgis. *Kalbotyra* 47 (1), 17–28.

EATON, H. 1934: Comparative frequency list on the first thousand words in English, French, German and Spanish. Coleman, A., ed., *Experiments and Studies in Modern Language Teaching,* Chicago.

GARCIA HOZ, V. 1953: *Vocabulario usual, comun y fundamental.* Madrid: Consejo Superior de Investigaciones Cientîficas – Instituto „San José de Calasanz".

GRUMADIENĖ, L., ŽILINSKIENĖ, V. 1997: *Dabartinės rašomosios lietuvių kalbos dažninis žodynas (mažėjančio dažnio tvarka),* Vilnius: Matematikos ir informatikos institutas, Lietuvių kalbos institutas.

GRUMADIENĖ, L., ŽILINSKIENĖ, V. 1998: *Dabartinės rašomosios lietuvių kalbos dažninis žodynas (abėcėlės tvarka),* Vilnius: Matematikos ir informatikos institutas, Lietuvių kalbos institutas.

HILLERICH, R., www: eduplace. com: High-frequency words and vocabulary.

JAKUBAITE, T. *e.a.* 1966–1976: *Latviešu valodas biežuma vārdnīca,* 1.–4. sējumi, Rīga: Zinātne.

JOSSELSON, H. 1953: *The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian*, Detroit: Periodical Service Co.

JUILLAND, A., RODRIGUES, E. Ch. 1964: *Frequency Dictionary of Spanish words,* The Hague: Mouton.

JUILLAND, A., EDWARDS, P. M., JUILLAND, I. 1965: *Frequency Dictionary of Roumanian Words,* The Hague: Mouton.

JUILLAND, A., TRAVERSA, V. 1973: *Frequency Dictionary of Italian Words,* The Hague: Mouton.

KAEDING, F. 1898: *Häufigkeitswörterbuch der deutschen Sprache,* Steigliz bei Berlin (Selbstverlag).

KEINYS, S., red. 1993: *Dabartinės lietuvių kalbos žodynas,* Vilnius: Mokslo ir enciklopedijų leidykla.

KEINYS, S., red. 2000: *Dabartinės lietuvių kalbos žodynas,* Vilnius: Mokslo ir enciklopedijų leidykla.

KENNEDY, G. 1998: *An Introduction to Corpus Linguistics,* London–New York: Longman.

KORSAKAS, J. 1991: *Lietuvių kalbos inversinis žodynas,* Kaunas: Šviesa.

KRUOPAS, J., red. 1954: *Dabartinės lietuvių kalbos žodynas,* Vilnius: Valstybinė politinės ir mokslinės literatūros leidykla.

KRUOPAS, J., red., 1972: *Dabartinės lietuvių kalbos žodynas,* Vilnius: Mintis.

LEECH, G. 1991: The state of the art in corpus linguistics, Aijmer, K., Altenberg, B., eds., *English Corpus Linguistics. Studies in Honour of Jan Svartvik,* London–New York: Longman.

LÖNNGREN, L. 1993: *Частотный словарь современного русского языка.* Uppsala. (Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia 32.)

MARCINKEVIČIENĖ, R. 2000a: Tekstynų lingvistika (teorija ir praktika). *Darbai ir dienos* 24, 7–64.

MARCINKEVIČIENĖ, R. 2000b: Patterns of word usage in corpus linguistics. *Kalbotyra* 49 (3), 71–80.

MAŽÈJKA, N. S. (МАЖЭЙКА, Н. С.), SUPRUN, A. Ja. 1976: *Частотны слоўнік беларускай мовы (мастацкая проза),* Мінск: Выдавецтва БДУ імя У. Леніна.

MIKELIONIENĖ, J. 2000: *Naujoji lietuvių kalbos leksika (1991–1996 m. kompiuterinio periodikos tekstyno pagrindu).* Daktaro disertacija, Kaunas: Vytauto Didžiojo universitetas.

MURMULAITYTĖ, D. 1998: Žodžių sąsajos žymėjimas „Dabartinės lietuvių kalbos žodyne" (santrumpos *žr.* vartojimas). *Lietuvių kalbotyros klausimai* 39, 240–247.

MURMULAITYTĖ, D. 1999: „Bendrinės lietuvių kalbos žodyno" kompiuterizavimo klausimu. *Lituanistica* 2 (38), 92–98.

RYKLIENĖ, A. 2000: Bendravimas internetu: kalbėjimas rašant. *Darbai ir dienos* 24, 99–107.

ROBINSON, D. F. 1976: *Lithuanian Reverse Dictionary,* Michigan: Slavica Publishers Inc.

ROZENCVEJG, V. Ju. 1986: Опыт создания национальных лексикографических служб за рубежом. Караулов Ю. Н., ред., *Машинный фонд русского языка: идеи и суждения,* Москва: Наука, 75–83.

SAVICKIENĖ, I. 1999: *Lietuvio vaiko daiktavardžio morfologija.* Daktaro disertacija, Kaunas: Vytauto Didžiojo universitetas.

ŠERNAS, V. 1994: *Kaip racionaliau mokytis kalbų (Basic English),* Vilnius: Žodynas.

ŠTEJNFEL'DT, E. A. 1963: *Частотный словарь современного русского литературного языка,* Таллин.

TĚŠITELOVÁ, M. *e.a.* 1983: *Frekvenční slovník češtiny věcného stylu*, Praha: Academia–Nakladatelství Československé Akademie Věd.

THORNDIKE, E., LORGE, L. 1944: *A Teacher's Word Book of the Twenty Thousand Words Found Most Frequently and Widely in General Reading for Children and Young People,* New York: Appleton–Century Crofts.

TODOROVA, E., PANČOVSKA, R. 2001: *Честотен речник на българската публицистика (1944–1989),* София: Пенсофт.

UTKA, A. 2000: Kalbinė įranga ir jos galimybės. *Darbai ir dienos* 24, 275–285.

VACAR, N. P. 1966: *A Word Count of Spoken Russian. The Soviet Usage,* Ohio: Ohio State University Press.

VANDER BEKE, G. 1929: *French Word Book*, New York: Macmillan.

ZASORINA, L. N., red., 1977: *Частотный словарь русского языка,* Москва: Издательство «Русский язык».

ZINKEVIČIUS, V. 2000: *Lemuoklis* – morfologinei analizei. *Darbai ir dienos* 24, 246–273.

ŽILINSKIENĖ, V. 1990: *Lietuvių kalbos dažninis žodynas,* Vilnius: Mokslas.

ŽILINSKIENĖ, V. 1995: *Atgalinis dabartinės lietuvių kalbos žodynas,* Vilnius: Matematikos ir informatikos institutas.

ŽILINSKIENĖ, V. 1998: Dažninis dabartinės rašomosios lietuvių kalbos žodynas (pirmieji tyrimo rezultatai). *Lietuvių kalbotyros klausimai* 39, 219–227.

ŽILINSKIENĖ, V., GRUMADIENĖ, L. 1998: *Valstybinės kalbos mokėjimo kvalifikacinių kategorijų lietuvių–rusų kalbų minimalusis žodynas (pirmoji kvalifikacinė kategorija),* Kaunas: Šviesa.

ŽILINSKIENĖ, V., GRUMADIENĖ, L. 1999a: *Valstybinės kalbos mokėjimo kvalifikacinių kategorijų lietuvių–rusų kalbų minimalusis žodynas (antroji kvalifikacinė kategorija),* Vilnius: UAB „Nacionalinių tyrimų centro" leidykla.

ŽILINSKIENĖ, V., GRUMADIENĖ, L. 1999b: *Valstybinės kalbos mokėjimo kvalifikacinių kategorijų lietuvių–rusų kalbų minimalusis žodynas (trečioji kvalifikacinė kategorija),* Vilnius: UAB „Nacionalinių tyrimų centro" leidykla.

ŽILINSKIENĖ, V. 2001: Lietuvių ir latvių kalbų publicistikos leksikos ir morfologijos raida statistiniu požiūriu. *Linguistica Lettica* 9, 155–168.

ŽILINSKIENĖ, V. spausd.: Lietuvių kalbos dalykinio stiliaus morfologijos statistinės charakteristikos ir jų lyginimas su atitinkamomis publicistikos charakteristikomis. *Lituanistica*.

Laima Grumadienė
Lietuvių kalbos institutas
P. Vileišio g. 5, LT–2055 Vilnius, Lietuva
laigruma@takas.lt

Gauta 2002 04 19