ISSN 2335-2019 (Print), ISSN 2335-2027 (Online)
Darnioji daugiakalbystė | Sustainable Multilingualism | *17*/2020
https://doi.org/10.2478/sm-2020-0019

Jolanta Kovalevskaitė
Vytauto Didžiojo universitetas, Lietuva
Erika Rimkutė
Vytauto Didžiojo universitetas, Lietuva

MOKOMASIS LIETUVIŲ KALBOS TEKSTYNAS: NAUJAS IŠTEKLIUS BESIMOKANTIESIEMS LIETUVIŲ KALBOS

Santrauka. Straipsnyje pristatomas pirmasis mokomasis lietuvių kalbos tekstynas, t. y. vienakalbis specialusis tekstynas, skirtas mokyti(s) lietuvių kalbos kaip svetimosios. Tekstynas kuriamas vykdant projektą „Užsienio baltistikos centrų ir Lietuvos mokslo ir studijų institucijų bendradarbiavimo skatinimas". Tokio ištekliaus atsiradimą paskatino tai, kad bendrojo pobūdžio tekstynuose, pvz., *Dabartinės lietuvių kalbos tekstyne,* pateikiami duomenys ir jų dydis besimokantiesiems lietuvių kalbos kaip užsienio kalbos yra per sudėtinga mokymosi medžiaga. Mokomajame tekstyne yra autentiškos lietuvių kalbos vartosenos tekstų, kurie atrinkti pagal tokius kriterijus, kad būtų suprantami ir aktualūs skirtingų lygių besimokantiesiems. Visi tekstai pagal *Bendruosius Europos kalbų mokymosi, mokymo ir vertinimo metmenis* suklasifikuoti į A1, A2, B1 ir B2 lygius. Tekstyną sudaro skirtingų kalbos atmainų (rašytiniai ir transkribuoti sakytiniai) tekstai. Iš viso mokomąjį lietuvių kalbos tekstyną sudaro 669 000 žodžių: 111 000 žodžių A1–A2 lygio rašytinės ir natūraliosios spontaninės kalbos tekstų, 558 000 žodžių B1–B2 lygio rašytinės ir spontaninės sakytinės kalbos tekstų. Šiame straipsnyje išsamiai aprašomas rašytinės kalbos patekstynis, kurį sudaro vadovėliniai ir nevadovėliniai tekstai, šio patekstynio dydis – apie 620 000 žodžių. Vadovėlinių tekstų kalbos lygis dažniausiai būdavo aiškus, o nevadovėliniai tekstai buvo automatiškai suklasifikuoti atliekant straipsnyje aprašytą tyrimą. Vadovėliniai ir nevadovėliniai tekstai suklasifikuoti į 29 žanrus (dialogus, pasakojimus, informacinius tekstus ir kt.) ir keturias grupes pagal komunikacinius tikslus (informacinius, pažintinius, apeliacinius ir meninius). Rašytinės kalbos patekstynyje daugiausiai yra informacinių tekstų; skiriasi dažniausi vadovėlinių ir nevadovėlinių tekstų žanrai: trys dažniausi vadovėlinių tekstų žanrai yra pažintiniai tekstai, pasakojimai ir dialogai (kartu šių trijų žanrų tekstai sudaro apie 78 proc. visų vadovėlinių tekstų). Nevadovėlinių tekstų patekstynio žanrų įvairovė didesnė: didžiąją dalį (apie 73 proc.) sudaro penkių žanrų tekstai: subtitrai, informaciniai tekstai, pažintiniai tekstai, proza, patarimai. Straipsnio apibendrinamosiose pastabose paminėta, kokie laukia tolesni darbai, susiję su mokomuoju tekstynu, kaip jis gali būti panaudotas.

Pagrindinės sąvokos: mokomasis tekstynas; lietuvių kalbos kaip svetimosios mokymas(is); Bendrieji Europos kalbų mokymosi, mokymo ir vertinimo metmenys; rašytinė kalba; sakytinė kalba; automatinis tekstų klasifikavimas.

Įvadas

Vykdant projektą „Užsienio baltistikos centrų ir Lietuvos mokslo ir studijų institucijų bendradarbiavimo skatinimas" (Nr. 09.3.1-ESFA-V-709-01-0002), kuriamos besimokantiesiems lietuvių kalbos skirtos mokomosios

priemonės[1]. Vienas iš naujų išteklių yra mokomasis lietuvių kalbos tekstynas (toliau – tekstynas). Tai yra nedidelis vienakalbis specialusis tekstynas, skirtas lietuvių kalbos besimokantiems kitakalbiams. Šis tekstynas gali būti naudojamas mokant lietuvių kalbos lituanistiniuose centruose užsienyje, lietuvių kalbos vasaros kursuose užsieniečiams Lietuvos universitetuose, kitakalbius vaikus Lietuvos bendrojo lavinimo mokyklose.

Mokant užsieniečius lietuvių kalbos minimi tekstynų privalumai (Čubajevaitė, 2007; Džežulskienė, 2014); atsiranda mokomųjų knygų, parengtų naudojant prieinamus lietuvių kalbos tekstynus, pvz., bendrojo tipo *Dabartinės lietuvių kalbos tekstyną* <htpp://tekstynas.vdu.lt>. Jau ruošiamas ir besimokančiųjų lietuvių kalbos tekstynas (Ruzaitė, 2019), kuriame kaupiama skirtingų lygių besimokančiųjų medžiaga (sakytiniai ir rašytiniai besimokančiųjų tekstai), žymimos besimokančiųjų padarytos gramatikos, rašybos ir leksikos klaidos.

Šiame straipsnyje aprašomas tekstynas skiriasi nuo besimokančiųjų tekstynų (angl. *learner corpora*), kuriuose kaupiami kitakalbių produkuoti tekstai, reprezentuojantys skirtingus kalbos mokėjimo lygius (A1, A2, B1, B2 (ir aukštesnį)). Mokomajame tekstyne yra sukaupti gimtakalbių produkuoti tekstai ir tekstai iš vadovėlių, skirtų mokytis lietuvių kalbos kaip svetimosios. Mokomasis tekstynas – kalbos išteklius iš autentiškos lietuvių kalbos vartosenos tekstų, kurie būtų suprantami ir aktualūs skirtingų lygių besimokantiesiems, jų mokytojams ir dėstytojams. Tai – vienakalbis išteklius, skirtas daugiakalbei lietuvių kalbos be(si)mokančiųjų bendruomenei, sukurtas siekiant į lietuvių kalbos mokymo(si) procesą įtraukti įvairesnių išteklių, mokymo(si) metodų.

Mokomieji tekstynai dažniausiai sudaromi iš vadovėlių medžiagos, todėl pirmieji tokio tipo tekstynai vadinti *textbook corpora*, *corpora of coursebooks* (pvz., pirmasis apie 2000-uosius metus D. Biberio su bendraautoriais iš vadovėlių sukauptas tekstynas *TOEFL Spoken and Written Academic Language Corpus*, U. Römer (2004) 100 000 žodžių tekstynas su tekstais iš vokiečiakalbiams skirtų anglų kalbos vadovėlių; platesnę apžvalgą žr. Meunier et al., 2009). Kitą terminą *pedagogic corpus* (pasiūlė Willis (1993), cit. iš Meunier et al., 2009) S. Hunston (2002, p. 16) apibrėžė kaip tekstyną, kurio

[1] Žr. https://kalbu.vdu.lt/

pagrindas – visa kalba, su kuria susiduria besimokantysis. F. Meunier, C. Gouverneur (2009, p. 184–185) oponavo, kad visõs skirtingų poreikių ir skirtingiems besimokantiesiems aktualios medžiagos sukaupti tekstyne neįmanoma, ir pasiūlė mokomuoju tekstynu vadinti sakytinės ir rašytinės kalbos pakankamos apimties reprezentatyvų tekstų rinkinį, su kuriuo besimokantysis, tikėtiniausia, susidurs analizuodamas mokomąją medžiagą, bendraudamas auditorijoje, studijuodamas savarankiškai. Šio straipsnio autorės siūlo *mokomojo tekstyno* terminą ir laikosi pozicijos, kad tokiame tekstyne, be vadovėlių tekstų, turėtų būti ir iš kitų šaltinių surinktos besimokantiesiems aktualios autentiškos kalbinės medžiagos, nes vadovėlių tekstai gali būti supaprastinti, ne visada atspindėti realią vartoseną.

Vadovėlių tekstų tekstynai dažniausia kuriami kompiuterinio kalbų mokymosi (angl. *computer assisted language-learning*) tikslams: pvz., COCTAILL tekstynas, sudarytas iš švedų kalbos vadovėlių medžiagos (tekstų, užduočių ir kt.), naudojamas skirtingo lygio leksinėms ir gramatinėms kalbos ypatybėms nustatyti, kad paskui, remiantis šiais duomenimis, būtų pasirengta automatiškai nustatyti besimokančiojo kalbos lygį (Volodina et al., 2014a); prancūzų, rusų kalbų vadovėlių medžiagos tekstynai naudojami ieškant būdų kuo efektyviau automatiškai nustatyti teksto sudėtingumą (François, 2014; Batinić et al., 2016). Kaip teigia E. Volodina ir kt. (2014a, p. 129), iš gimtakalbių kalbos tekstynų (angl. *native speaker corpora*) negalima nustatyti pradedantiems arba pažengusiems besimokantiesiems būdingos kalbos požymių, nes juose yra ir paprastų, ir sudėtingų lingvistinių (leksikos, gramatikos, teksto) reiškinių, o besimokančiųjų tekstynų medžiaga yra su klaidomis, tad šie tekstynai taip pat netinka norint nustatyti konkrečiam lygiui reikalingą kalbinę kompetenciją. Taigi, galima sakyti, iš vadovėlių tekstų sudaryti tekstynai yra laikomi tinkamiausia gimtakalbių produkuota medžiaga kompiuterinio kalbų mokymosi uždaviniams įvykdyti.

Lietuvių kalbos mokomasis tekstynas pradėtas kurti pirmiausia turint omenyje vartotojų poreikius: kurti tokį išteklių paskatino užsienio baltistikos centrų dėstytojų pageidavimas, kad būtų besimokantiesiems lietuvių kalbos lengviau suprantamos medžiagos. Plg. *Dabartinės lietuvių kalbos tekstyne* pateikiami duomenys (jame yra įvairių stilių ir žanrų, skirtingų laikotarpių tekstų, tiek verstų, tiek parašytų lietuvių kalba) ir jų kiekis (140 mln. žodžių)

besimokantiesiems lietuvių kalbos kaip užsienio kalbos yra per sudėtinga mokymosi medžiaga: dar neįgudusiems lietuvių kalbos vartotojams sudėtinga savarankiškai atskirti tipines ir netipines leksikos ir gramatikos ypatybes, atpažinti tam tikro stiliaus, tam tikro formalumo lygio kalbinę raišką ir pan.

Lietuvių kalba, kaip mažiau vartojama kalba (angl. *lesser used languages*), priklauso toms kalboms, kurių mokomasi rečiau (angl. *lesser taught languages*), taigi ir mokymo priemonių bei vadovėlių jai mokytis nėra daug, be to, ne visi tekstai lengvai skaitmenizuojami. Todėl kaupiant šį tekstyną buvo numatytos dvi jo dalys: vadovėlių tekstų dalis ir kitų tekstų dalis, kurią sudarė gimtakalbių produkuoti tekstai, specialiai atrinkti atsižvelgiant į besimokančiųjų poreikius: tai – įvairių žanrų, skirtingus komunikacinius tikslus atspindintys, besimokantiesiems aktualūs nesudėtingi tekstai (sakiniai trumpesni, paprastesnės struktūros; plačiau žr. poskyrį *Tekstyno dydis*). Kuriant šį tekstyną vadovėlių tekstų dalis buvo panaudota tam, kad, nustačius konkretaus lygio tekstų rinkinių požymius, medžiagą, papildomai sukauptą ne iš vadovėlių, būtų galima automatiškai suklasifikuoti pagal lygius (šis eksperimentas aprašomas poskyryje *Automatinis tekstų klasifikavimas*).

Kaip ir įprasta mokomiesiems tekstynams, šio specializuoto tekstyno apimtis nedidelė (apie 669 000 žodžių). Vis dėlto tokio tipo tekstynuose, jeigu jie nėra sudaromi tik iš vadovėlių tekstų, svarbu atrinkti būtent besimokantiesiems aktualią rašytinę ir sakytinę medžiagą. Šio apžvalginio straipsnio **tikslas** – pristatyti lietuvių kalbos mokomojo tekstyno sudarymo principus, sandarą, tekstų pobūdį ir atrankos kriterijus; paaiškinti automatinio tekstų klasifikavimo pagal kalbos lygius rezultatus. Šiame straipsnyje išsamiai aprašoma tik rašytinė tekstyno dalis (apie 620 000 žodžių), tačiau tekstyną sudaro ir spontaninės sakytinės kalbos tekstai (jie transkribuoti, kaip ir rašytinės kalbos, automatiškai morfologiškai anotuoti).

Tekstyno sudarymo principai ir sandara

Mokomasis tekstynas parengtas tam, kad lietuvių kalbos besimokantys kitakalbiai, jų mokytojai ir dėstytojai rastų autentiškos dabartinės lietuvių kalbos vartosenos tekstų, kurie būtų suprantami ir aktualūs skirtingų lygių besimokantiesiems. Mokomąjį tekstyną sudaro gimtakalbių produkuota

rašytinė (tekstai) ir sakytinė (įrašai) kalba. Rašytinės kalbos patekstynio tekstų dalis imta iš vadovėlių, kita dalis – iš kitų šaltinių. Paminėtina, kad tekstynas (tiek jo rašytinė, tiek sakytinė dalis) sudarytas atsižvelgiant į etikos reikalavimus, keliamus mokslo išteklių kaupimui ir analizei.

Vienas iš aktualių probleminių klausimų, kuris yra susijęs su mokomaisiais (ypač iš vadovėlių sudarytais) tekstynais, – kalbos autentiškumas, t. y. kiek vadovėlių medžiagoje atspindėta kalba reprezentuoja tai kalbai būdingą realią vartoseną (žr. poskyrį *Kalbos autentiškumas*). Šioje straipsnio dalyje iš pradžių aptariamas autentiškos kalbos klausimas, toliau išsamiai aprašoma tekstyno sandara (žr. poskyrius *Tekstyno sandara, Tekstyno dydis, Tekstų žanrai*), automatinio tekstų klasifikavimo eksperimentas (žr. poskyrį *Automatinis tekstų klasifikavimas*).

Kalbos autentiškumas

Daug diskutuota, kokia kalba – autentiška ar supaprastinta – ir kiek supaprastinta kalba turėtų būti pateikiamas mokymo turinys negimtakalbiams skirtuose vadovėliuose (apžvalgą žr. Widdowson, 2003, cituojama iš Meunier et al., 2009). Kaip matyti iš U. Römer (2004) tyrimo, atotrūkis tarp vadovėliuose pateikiamos vartosenos ir realios vartosenos laikytinas vadovėlių trūkumu. Kita vertus, reikia pritarti F. Meunier ir C. Gouverneur (2009, p. 194), kad įvestis (angl. *input*) – kalbinė medžiaga, mokiniui pateikiama sakytine arba rašytine forma, – turi būti kuo autentiškesnė, bet žemesnių lygių besimokantiesiems aktualu ir net naudinga ją bent šiek tiek pritaikyti, supaprastinti.

Natūraliai, autentiškai kalbos medžiagai atrinkti naudojami tokie įrankiai, su kuriais galima automatiškai generuoti pratimus, reprezentatyvius pavyzdžius iš didelių tekstynų (Volodina et al., 2014b; tekstynų analizės įrankiai, orientuoti į paslaugas besimokantiesiems *SketchEngine for Language Learning, SKELL*[2] (anglų, vokiečių, rusų, čekų, italų, estų kalboms). Taip pat paminėtini ir supaprastintos kalbos tekstynai (Vandeghinste et al., 2019), įrankiai, skirti gimtakalbių sukurtiems tekstams paprastinti (pvz., Rennes et al., 2015).

[2] Žr. https://skell.sketchengine.co.uk/run.cgi/skell#

Mokomajame lietuvių kalbos tekstyne paimtų vadovėlių medžiagoje atsispindi įvairi kalba: vadovėlių, kuriuos rengiant remtasi tekstynais, medžiaga artimesnė autentiškai vartosenai, bet pradinių lygių vadovėliuose pasitaiko tekstų, kurie skamba ne visiškai natūraliai, tarkime, A1, A2 lygių vadovėliuose dažnas žanras yra dialogai, bet skaitant matyti, kad jie ne visai autentiški, labiau primena rašytinę, o ne spontaninę sakytinę kalbą. Vienas iš galimų būdų pasiekti, kad dialogų medžiaga būtų autentiškesnė, – naudoti transkribuotus spontaninės sakytinės kalbos įrašus. Siekiant kalbos autentiškumo, tekstyne bus pateikiami transkribuoti sakytinės lietuvių kalbos tekstai, atspindintys įvairias besimokantiesiems aktualias situacijas (žr. poskyrį *Tekstyno sandara*).

Kaip parodyta poskyryje *Tekstyno dydis*, mokomojo tekstyno rašytinės kalbos patekstynyje didelė dalis medžiagos yra ne iš vadovėlių, o iš internetinės periodikos, informacinių, pažintinių šaltinių, grožinės literatūros. Tai yra autentiškos dabartinės lietuvių kalbos vartosenos tekstai, kurie būtų suprantami ir aktualūs lietuvių kalbos besimokantiems negimtakalbiams. Atrinkti nevadovėliniai tekstai nebuvo paprastinami arba pritaikomi, siekta atrinkti besimokantiesiems aktualius tekstus pagal komunikacinius tikslus, žanrus, teksto tipus, temas. Atrinkti vadovėliniai tekstai nebuvo keičiami: jeigu jie vadovėlių autorių arba sudarytojų buvo supaprastinti arba pritaikyti, tokie ir buvo įdėti į tekstyną.

Tekstyno sandara

Mokomąjį lietuvių kalbos tekstyną sudaro A1–A2 lygio rašytinės ir natūraliosios spontaninės kalbos tekstai (111 000 žodžių: 96 000 rašytinės kalbos ir 15 000 sakytinės kalbos), B1–B2 lygio rašytinės ir spontaninės sakytinės kalbos tekstai (558 000 žodžių: 523 000 rašytinės kalbos ir 35 000 sakytinės kalbos); iš viso – 669 000 žodžių.

Kaip matyti iš šių skaičių, sakytinės kalbos medžiaga sudaro gana nedidelę dalį A1–A2 lygio tekstų, nes besimokantiesiems tinkamos medžiagos nedaug, be to, surinkti įrašų gana sunku. B1–B2 lygio vartotojams tinkamos medžiagos daugiau, todėl šią tekstyno dalį sudaro kiek didesnė sakytinės kalbos dalis. Sakytinės lietuvių kalbos tekstyno dalį sudaro įvairiose vietose įrašyti natūralūs pokalbiai, apimantys skirtingas kalbėjimo situacijas ir įvairius

socialinius pašnekovų vaidmenis. Sukaupti sakytinės kalbos duomenys apima tiesioginius aptarnavimo srities pokalbius arba pokalbius telefonu, vykstančius maitinimo įstaigose (restorane, kavinėje ir pan.), prekybos vietose (parduotuvėje, turguje, kioske, knygyne, vaistinėje, teatro, kino arba autobusų bilietų kasoje ir pan.), paslaugas teikiančiose įstaigose (kirpykloje, grožio salone, siuvykloje, banke ar pan.), taip pat pokalbius, vykstančius namų arba darbo aplinkoje[3].

Kaip minėta įvade, šiame straipsnyje rašoma tik apie rašytinės kalbos tekstyno dalį (jos dydis – 618 637 žodžiai[4]). Rašytinės kalbos patekstynį sudaro dvejopi tekstai: 1) tekstai, rinkti iš lietuvių kalbos besimokantiems kitakalbiams skirtų vadovėlių; 2) tekstai, rinkti iš populiarinamųjų ir grožinių knygų, naujienų portalų, viešųjų užrašų, instrukcijų, skelbimų, dokumentų ir kt.[5] (išsamiau apie juos rašoma poskyryje *Tekstyno dydis*).

Pirminė tekstyno sudarymo idėja buvo tai, kad renkami vadovėliniai tekstai turi atitikti keturis lygius, remiantis *Bendrųjų Europos kalbų mokymosi, mokymo ir vertinimo metmenų* (2008) lygių struktūra: A1, A2, B1, B2. Tekstai buvo suklasifikuoti taip, kaip buvo nurodyta vadovėlių autorių. Suprantama, kad aukštesnio lygio medžiagoje gali būti ir paprastesnių tekstų (atitinkančių žemesnį lygį) ir atvirkščiai – žemesnio lygio vadovėliuose gali pasitaikyti sudėtingesnės medžiagos.

Imti įvairūs skirtingiems kalbos lygiams pritaikyti vadovėliai (jų sąrašas pateiktas 2 priede). Mokomiesiems tekstynams, kurių medžiaga renkama iš vadovėlių, paprastai imamos ir užduotys, pratimai, testai (plg. Volodina et al., 2014a). Šiame tekstyne iš vadovėlių imti visi rišlūs tekstai, jie netrumpinti (pvz., dialogai, pasakojimai, pažintiniai tekstai), bet neimta užduočių (pvz., tokių, kuriose reikia įrašyti raidę, žodį), išskyrus tokias, kurios pateiktos kaip rišlus tekstas, pvz., pasakojimas apie šeimą, kurį perskaičius reikia atsakyti į teksto suvokimo klausimus.

[3] Sakytinės kalbos tekstų atranką, tvarkymą koordinavo VDU Tarpkultūrinės komunikacijos ir daugiakalbystės tyrimų centro mokslininkė doc. dr. Laura Kamandulytė-Merfeldienė.

[4] Tai – vieną ir daugiau kartų pavartotos visos žodžių formos. Už visus kiekybinius duomenis apie tekstyną dėkojame kolegai dr. Loïcui Boizou.

[5] Iš besimokantiesiems skirtų vadovėlių imti tekstai toliau vadinami **vadovėliniais**, iš kitų šaltinių – **nevadovėliniais**. Jei tekstas imtas iš gimtakalbiams lietuviams skirto vadovėlio, jis priskiriamas nevadovėliniams tekstams.

Tiek iš vadovėlių, tiek iš kitų šaltinių imti tekstai buvo suskaidyti į dokumentus, pvz., skirtingais dokumentais buvo laikomi tame pačiame vadovėlyje pateikti skirtingi dialogai, pažintiniai tekstai apie skirtingus aptariamus objektus, skirtingi knygos skyriai arba dalys. Kartais visas tekstas laikytas vienu dokumentu (jei jis trumpas, struktūriškai nesuskaidytas), kitais atvejais tas pats tekstas, pvz., knyga, kurioje yra pažintinių tekstų, įvairių patarimų, sentencijų, buvo suskaidytas į šimtus dokumentų. Iš viso tekstyne yra 3 765 tokie dokumentai.

Renkant nevadovėlinius tekstus (jų sąrašas pateiktas 2 priede), imti arba visi tekstai, pvz., apsakymai, trumpi romanai, arba didesnio kūrinio kelios dalys. Net ir tais atvejais, kai imtas ne visas grožinis kūrinys, o tik jo dalys, tos dalys netrumpintos, kad išliktų vientisas tekstas. Paminėtina, kad dauguma tekstų yra parašyti gimtakalbių lietuvių, du romanai (R. Šepetys, L. Vincės), kurių dalys yra tekstyne, versti iš anglų kalbos. Tarp kitų žanrų tekstų, pvz., pažintinių, taip pat yra verstų kūrinių. Vis dėlto šie tekstai turėtų būti suprantami negimtakalbiams, nes juose aptariamos nesudėtingos temos, patys tekstai neilgi, leksika gana paprasta, kalba taisyklinga.

Tekstyno dydis

Toliau pateikiama informacija apie tekstyno rašytinės dalies patekstynį sudarančių tekstų tipus ir kalbos lygius.

1 lentelė

Rašytinės dalies patekstynio dydis

Tekstų tipas	Žodžių skaičius	Procentai
Vadovėliniai	106 423	17,2
Nevadovėliniai	512 214	82,8
Iš viso	618 637	100

Kaip matyti iš 1 lentelės, daugiau nei 80 proc. minėto patekstynio sudaro ne iš besimokantiesiems lietuvių kalbos skirtų vadovėlių rinkti tekstai.

Kaip matyti iš 2 priedo, tekstai imti iš keliolikos vadovėlių, bet juose, ypač A1–A2 lygio, yra nedaug tekstų: vyrauja dialogai (žr. 1 priedo 1 lentelę), jie paprastai trumpi. B1 ir B2 lygių vadovėlių, skirtų mokytis lietuvių kalbos, gerokai mažiau, bet juose pateikta daugiau rišlių tekstų, tekstai ilgesni (nemažai pažintinių tekstų apie Lietuvą, jos vietoves, įžymius žmones, papročius ir pan., prozos kūrinių), taigi, sudaro didesnę vadovėlinių tekstų patekstynio dalį (žr. 2 lentelę).

2 lentelė

Vadovėliniai tekstai pagal kalbos lygius

Kalbos lygis	Žodžių skaičius	Procentai
A1	22 382	21,03
A2	15 653	14,71
B1	22 068	20,74
B2	46 320	43,52
Iš viso	106 423	100

3 lentelė

Nevadovėliniai tekstai pagal kalbos lygius

Kalbos lygis	Žodžių skaičius	Procentai
A1	20 494	4,00
A2	37 033	7,20
B1	45 927	8,97
B2	408 760	79,80
Iš viso	512 214	100

Sprendžiant iš 3 lentelėje pateiktų duomenų, labai sunku rasti nevadovėlinių tekstų, kurie atitiktų paprasčiausius kalbos lygius: A1 ir A2 (abiejų šių lygių tekstai sudaro kiek daugiau nei 11 proc. nevadovėlinių tekstų). Pradiniams lygiams aktualesnė sakytinė kalba. Didžiausia nevadovėlinių tekstų

patekstynio dalis sudaryta iš B2[6] lygio tekstų. Skaičiuojant vadovėlinius ir nevadovėlinius tekstus kartu (žr. 4 lentelę), vyrauja B2 lygio tekstai, o A1, A2 ir B1 lygio tekstai sudaro santykinai panašias dalis. Didelį B2 tekstų skaičių nulemia tai, kad automatiškai klasifikuojant daug nevadovėlinių tekstų buvo priskirta B2 lygiui (plačiau žr. poskyrį *Automatinis tekstų klasifikavimas*).

4 lentelė

Viso rašytinės dalies patekstynio dydis pagal kalbos lygius

Kalbos lygis	Žodžių skaičius	Procentai
A1	42 876	6,93
A2	52 686	8,52
B1	67 995	10,99
B2	455 080	73,56
Iš viso	618 637	100

Nemažų skirtumų išryškėja lyginant vadovėlinių ir nevadovėlinių tekstų žanrų skirtumus (plg. 1–4 lenteles 1 priede, taip pat žr. poskyrį *Tekstų žanrai*).

Tekstų žanrai

Į tekstyną įeina 29 žanrų tekstai, dauguma jų bendri vadovėliniams ir nevadovėliniams tekstams (žr. 1 priedo lenteles). Visi žanrai išvardyti ir trumpai apibūdinti 5 lentelėje.

Į tekstyną įtrauktus tekstus pagal komunikacinius tikslus (remiantis Ramonienė et al., 2016) galima suklasifikuoti taip:

- **Informaciniai**: anketos; dialogai; dokumentai; etiketės; horoskopai; informaciniai tekstai; instrukcijos; interviu; laiškai, žinutės; pasakojimai; receptai, meniu; skelbimai; tvarkaraščiai, darbotvarkės; viešieji užrašai.

[6] Atkreipiame dėmesį, kad tarp nevadovėlinių B2 lygio tekstų turbūt yra ir aukštesnio lygio tekstų.

- **Pažintiniai**: patarimai; pažintiniai tekstai; testai, užduotys.
- **Apeliaciniai**[7]: reklamos; šūkiai.
- **Meniniai**: anekdotai; dainos; eilėraščiai; linkėjimai, sveikinimai; pasakos, legendos; proza; sentencijos; subtitrai; tautosaka.

Daugiausiai yra informacinių tekstų. Kadangi mokantis kalbos svarbiausia yra informacinė-komunikacinė kalbos funkcija, todėl daugiausiai atrinkta ją atspindinčių tekstų.

5 lentelė

Tekstų žanrai

Eil. nr.	Žanro pavadinimas	Paaiškinimas
1.	Anekdotai	Trumpi juokingi pasakojimai, dažnai dialogai; stengtasi atskirti nuo dialogų žanro.
2.	Anketos	Informacija apie šeimą: sutuoktinio (-ės), vaikų vardai, amžius ir pan.; informacija, kurią reikia pateikti, pvz., norint užsiprenumeruoti periodinį leidinį: pasirinktas leidinys, pasirinktas laikotarpis, užsakovo adresas ir pan.; vizitinių kortelių imitacija iš vadovėlių, kai nurodomas asmens vardas, pavardė, einamos pareigos, darbo adresas ir pan.
3.	Dainos	Ir liaudies, ir šiuolaikinės dainos; *Tautiška giesmė*; mokyklų himnai.
4.	Dialogai	Dažniausiai iš vadovėlių paimti imitaciniai įvairių socialinių grupių žmonių pokalbiai įvairiose situacijose (parduotuvėje, kavinėje, susipažįstant, užsisakant viešbučio kambarį ir pan.), pokalbiai telefonu. Pastaba: apsakymuose, novelėse, kurie priskirti prozos žanrui, taip pat yra dialogų, bet, išlaikant kūrinio vientisumą, jie palikti tuose kūriniuose, t. y. prozos žanre. Subtitrai irgi galėtų būti laikomi dialogais, bet laikomi atskiru žanru, nes pagal komunikacinį tikslą priskiriami prie meninių.

[7] Kelionių aprašymus pagal komunikacinius tikslus priskyrėme pažintiniams tekstams, nors šie tekstai turi ir apeliaciniams tekstams būdingų ypatybių.

Eil. nr.	Žanro pavadinimas	Paaiškinimas
5.	Dokumentai	Informacija, kaip atsidaryti banko sąskaitą, gauti nedarbo arba motinystės išmoką, būsto paskolą; kas yra draudžiamas privalomuoju sveikatos draudimu; kam išduodamas leidimas gyventi Lietuvoje ir pan. Šie tekstai galėtų būti priskiriami administraciniam stiliui.
6.	Eilėraščiai	Keletas žinomų lietuvių autorių eilėraščių; jie sudaro labai mažą visų žanrų dalį, imti tik iš vadovėlių.
7.	Etiketės	Prekių etiketės, t. y. trumpa informacija apie gaminio sudėtį, skalbimo galimybes, kainą, gamintoją, dizainerį, techninius parametrus; pats prekės pavadinimas. Pateiktos darbužių, namų apyvokos daiktų, buitinės technikos, maisto produktų, kanceliarinių prekių ir kt. gaminių etiketės.
8.	Horoskopai	Dienos, mėnesio arba kito laikotarpio, tam tikro Zodiako ženklo horoskopai.
9.	Informaciniai tekstai	Įvairaus pobūdžio tekstai (apie sportą, orus, kriminalus, politikos įvykius, įžymybes ir kt.). Didžiąją dalį šio žanro tekstų sudaro tekstai, publikuoti 2014–2016 m. portale *Delfi.lt*[8]; dar priskirtos žinių laidų transkripcijos, orų prognozės (iš vadovėlių). Portalo *Delfi.lt* tekstai smulkiau neskaidyti pagal temas, nors stengtasi apimti kuo daugiau ir įvairesnių rubrikų.
10.	Instrukcijos	Informacija apie įvairių buitinių prietaisų (televizoriaus, skalbyklės, lygintuvo ir pan.) naudojimą ir priežiūrą; namų taisyklės (ką galima ir ko negalima daryti).
11.	Interviu	Ilgesni pokalbiai; nuo dialogų skiriasi tuo, kad dažniausiai pateikiami iš anksto parengti klausimai; vienas pašnekovas kalbina kitą.
12.	Kita	Prašymas priimti į darbą.
13.	Laiškai, žinutės	Vadovėliuose pateikti popierinių laiškų pavyzdžiai; iš nevadovėlinių tekstų: elektroniniai laiškai, trumposios žinutės, žinutės naujienų forumuose (šie susirašinėjimai šiek tiek ilgesni); iš dalies sutampa su dialogais, skiriasi tuo, kad tai – rašytiniai dialogai (*kalbėjimas rašant*, žr. Ryklienė, 2000), o tikrieji dialogai imituoja sakytinę kalbą. Nevadovėliniai laiškai ir žinutės labai skiriasi nuo vadovėliuose pateiktų šio žanro tekstų, nes nevadovėliniai

[8] Tai – maža dalis iš vykdant projektą „Lietuvių kalbos pastoviųjų žodžių junginių automatinis atpažinimas (PASTOVU)“ (Nr. LIP-027/2016) sukaupto *Delfi.lt* tekstyno (plačiau žr. http://mwe.lt/).

Eil. nr.	Žanro pavadinimas	Paaiškinimas
		tekstai yra natūraliosios kalbos pavyzdžiai, juose yra neformaliosios leksikos, sutrumpintų žodžių. Vadovėliniai tekstai atrodo gana dirbtini, bet lengvai suprantami besimokantiesiems.
14.	Linkėjimai, sveikinimai	Trumpi sveikinimai, linkėjimai įvairių švenčių progomis; imti tik iš vadovėlių.
15.	Pasakojimai	Neilgi rišlūs pasakojimai (dažniausiai iš vadovėlių, mokiniams skirtų diktantų), kartais nenatūralūs, pritaikytos kalbos, neatliekantys pažintinės funkcijos.
16.	Pasakos, legendos	Liaudies ir šiuolaikinių rašytojų pasakos, padavimai, legendos, sakmės.
17.	Patarimai	Praktiški patarimai, kaip padengti stalą, pagaminti maistą, suruošti vaišes ir pan.
18.	Pažintiniai tekstai	Populiariai pateikta informacija apie augalus, gyvūnus, gamtos reiškinius, Lietuvos virtuvę, gražiausias Lietuvos vietas, lankytinus objektus, ekskursijų po tam tikrus miestus aprašymai ir pan. Šiuos tekstus galima priskirti mokslo populiarinimo stiliui.
19.	Proza	Trumpi grožiniai kūriniai: apsakymai, humoreskos, trumpi romanai, romanų dalys. Prozos tekstuose gali būti įvairių kitų žanrų: dialogų, dienoraščio, sentencijų ir kt. Vis dėlto laikoma proza, nes pagal komunikacinį tikslą tai yra meniniai tekstai. Dauguma tekstų parašyti lietuvių kalba, kiti versti į lietuvių kalbą.
20.	Receptai, meniu	Receptai, kaip pasigaminti patiekalus; įvairūs meniu.
21.	Reklamos	Informacija apie poilsį tam tikrame viešbutyje, pramogas, ekskursijas (kai nurodomas konkretus paslaugos teikėjas; jei pateikiama bendro pobūdžio informacija, tada priskirta prie pažintinių tekstų arba skelbimų); socialinės reklamos.
22.	Sentencijos	Įžymių žmonių mintys. Taip pat nurodomi tų minčių autoriai.
23.	Skelbimai	Įvairaus pobūdžio skelbimai apie išnuomojamą arba perkamą butą, vyksiančią paskaitą arba kitą renginį. Taip pat įeina pažinčių skelbimai, surinkti iš 2005 m. rašytinės žiniasklaidos. Nors jie ne visai atspindi dabartinio pažinčių skelbimų žanro ypatybes, dėl kalbinės raiškos priemonių vis dėlto gali būti aktualūs besimokantiesiems lietuvių kalbos.

Eil. nr.	Žanro pavadinimas	Paaiškinimas
24.	Subtitrai	Užsienio filmų subtitrai; tai – verstiniai tekstai. Pastebėta, kad dalies jų kalba netaisyklinga, pasitaiko vulgarybių, necenzūrinės leksikos, slengo, bet, siekiant išlaikyti autentiškumą, subtitrų kalba netvarkyta.
25.	Šūkiai	Komercinių ir nekomercinių organizacijų šūkiai.
26.	Tautosaka	Smulkiosios tautosakos kūriniai: mįslės, patarlės, priežodžiai, gamtos reiškinių spėjimai, burtai, liaudies išmintis, žaidimai.
27.	Testai, užduotys	Klausimai ir paaiškinimai, ar žmogus linkęs būti verslininku, ar maitinasi taisyklingai; žurnaluose arba knygelėse vaikams pateiktos užduotys (surasti tam tikrą objektą, apibraukti, ką nors įrašyti ir pan.).
28.	Tvarkaraščiai, darbotvarkės	Įvairių institucijų darbo laikas; viešojo transporto išvykimo laikas; tam tikro asmens savaitės veiklų planas.
29.	Viešieji užrašai	Įvairiose viešosiose vietose surinkta informacija apie įstaigų, padalinių, kabinetų pavadinimus, darbo laiką; bendro pobūdžio užrašai, pvz., *atidaryta, uždaryta; darbo laikas; kasa* ir pan.

Toks žanrų pasiskirstymas išryškėjo pirmiausia renkant tekstus iš besimokantiesiems lietuvių kalbos skirtų vadovėlių, vėliau, pildant tekstyną nevadovėliniais tekstais, žanrų sąrašas buvo išplėstas, kai kurie žanrai – patikslinti. Suprantame, kad ši klasifikacija diskutuotina, bet atskleidžia tekstyną sudarančių tekstų įvairovę.

Trys dažniausi **vadovėlinių** tekstų žanrai yra pažintiniai tekstai (35,66 proc.), pasakojimai (22,83 proc.) ir dialogai (19,81 proc.) (žr. 1 priedo 1 lentelę). Kartu šių trijų žanrų tekstai sudaro 78,3 proc. visų tekstų. Toks žanrų pasiskirstymas neturėtų stebinti, nes mokant kitos kalbos įprastai supažindinama su šalies kultūra, istorija, papročiais, todėl daugiau nei trečdalį vadovėlių turinio sudaro pažintiniai tekstai. Atkreiptinas dėmesys į tai, kad skirtinguose vadovėliuose informacija, pvz., apie Lietuvą, svarbiausias šventes, įžymius žmones, pateikiama gana panašiai. Kaip matyti iš 1 priedo 1 lentelės, daugiau nei penktadalį vadovėlinių tekstų sudaro pasakojimai. Kaip minėta, tai – rišlūs, paprastai neilgi tekstai, neatliekantys pažintinės funkcijos (pvz., savo sapno pasakojimas, istorija apie apsilankymą poliklinikoje, pamąstymai,

kam lietuviams reikia Vėlinių, ir pan.). Panašią dalį sudaro ir imitaciniai pokalbiai – dialogai. Tai – tipiški tekstai pradedant mokytis kalbos, todėl jų yra nemažai (plg. su nevadovėliniais tekstais – ten dialogai sudaro tik 0,05 proc.).

Nevadovėlinių tekstų patekstynio žanrų įvairovė (žr. 1 priedo 2 lentelę) skiriasi nuo vadovėlinių tekstų; čia žanrai pasiskirstę tolygiau: daugumą (72,92 proc.) visų tekstų sudaro ne trijų žanrų (kaip vadovėlinių), o penkių žanrų tekstai. Daugiausia yra subtitrų (18,34 proc.), informacinių tekstų (16,07 proc.), pažintinių tekstų (14,68 proc.), prozos (12,97 proc.), patarimų (10,86 proc.). Kaip matyti iš pateiktų skaičių, nė vienas nevadovėlinių tekstų žanras labai nevyrauja (plg. vadovėlinių tekstų pažintinius tekstus – jie sudaro daugiau nei 35 proc.).

Kai kurių žanrų tekstų yra vadovėliuose, bet nėra nevadovėliniuose šaltiniuose, ir atvirkščiai: plg. vadovėlinių tekstų patekstynyje nėra dokumentų, sentencijų, subtitrų, šūkių, smulkiosios tautosakos. Naudotuose nevadovėliniuose šaltiniuose nepasitaikė eilėraščių, linkėjimų ir sveikinimų, tvarkaraščių ir darbotvarkių. 1 priedo 4 lentelėje matyti informacija, kiek kurio patekstynio kuriame lygyje yra tam tikro žanro tekstų (pateikiamas žodžių skaičius).

Apibendrinant **viso tekstyno** žanrų įvairovę (žr. 1 priedo 3 lentelę), galima teigti, kad nė vieno žanro tekstai nevyrauja – daugiausiai pažintinių tekstų (18,29 proc.). Pagal dažnumą toliau paminėtini šių žanrų tekstai: subtitrai (15,19 proc.), informaciniai tekstai (13,48 proc.), proza (10,88 proc.), patarimai (9,15 proc.). Sudėjus visas anksčiau minėtas tekstų žanrų grupes, matyti, kad jie sudaro beveik 67 proc. visų tekstų, o kitų žanrų tekstai – daugiau nei trečdalį viso tekstyno, taigi toks pasiskirstymas atskleidžia pakankamą žanrų įvairovę.

Automatinis tekstų klasifikavimas

Iš vadovėlių sukauptų tekstų lygis buvo žinomas, tačiau iš kitų šaltinių paimtų tekstų lygį reikėjo nustatyti. Tam tikslui buvo naudojami mašininio mokymo modeliai ir atliktas eksperimentas, kuris susidėjo iš dviejų dalių (plačiau žr. Grigonytė et al., 2018): pirma, vadovėliniai tekstai buvo panaudoti kaip medžiaga klasifikavimo algoritmui išmokyti; antra, paruoštas klasifikavimo

algoritmas buvo pritaikytas klasifikuojant tarpinio lygio tekstus, paimtus iš vadovėlių (pvz., A1–A2, B1–B2[9]), o vėliau – ir nevadovėlinius tekstus.

Automatiškai klasifikuojant tekstus buvo atsižvelgta į **paviršines teksto ypatybes**: teksto ilgį, sakinio ilgį, morfologinių formų įvairovę, vidutinį žodžių ilgį, ilgų[10] ir trumpų žodžių skaičių ir jų santykį, sakinių skaičių tekste, vidutinį sakinio ilgį tekste, vidutinį žodžių skaičių sakinyje. Šie požymiai minimi teksto sudėtingumo vertinimo tyrimuose (angl. *readability approach*). Taip pat analizuotos **gilesnės lingvistinės ypatybės**: žodingumas (kiek iš viso yra žodžių ir kiek iš jų skirtingų); daiktavardžių, prielinksnių ir dalyvių santykis viename tekste, palyginti su įvardžiais, prieveiksmiais ir veiksmažodžiais[11]; daiktavardžių ir įvardžių santykis; būdvardžių, daiktavardžių, įvardžių, skyrybos ženklų procentinė dalis tekste; vidutinis daiktavardžių, veiksmažodžių ir būdvardžių ilgis tekste, šių kalbos dalių įvairovė (t. y. kiek yra skirtingų minėtų kalbos dalių žodžių; procentinė dalis tekste šių gramatinių formų: dalyvių, padalyvių, pusdalyvių, naudininko, įnagininko, liepiamosios ir tariamosios nuosakos, bevardės giminės, aukštesniojo ir aukščiausiojo laipsnio. Kad šios gramatinės formos pradiniuose lygiuose vartojamos rečiau, išsiaiškinta analizuojant metodinėse lietuvių kalbos mokymo priemonėse pateiktą informaciją, diskutuojant su lietuvių kalbos mokančiais dėstytojais.

Kaip matyti, didžiausias dėmesys skirtas leksinėms ir morfologinėms ypatybėms. Atsižvelgta tik į vieną sintaksinį požymį – sakinio ilgį, nes iš kai kurių besimokančiųjų tekstų tyrimų matyti, kad sintaksinės savybės suteikia mažiau informacijos apie teksto sudėtingumą nei leksinės ir morfologinės savybės (kai vertinama, kas būdinga konkretaus lygio besimokančiojo produkcijai) (Hancke et al., 2013).

Po automatinio tekstų klasifikavimo eksperimento paaiškėjo, kad, išmokius mašininio mokymo algoritmą, į dviejų lygių (A1–A2 ir B1–B2) grupes buvo skirstoma tiksliau nei į keturių lygių grupes (A1, A2, B1, B2). Vadinasi,

[9] Keli tekstynui atrinkti vadovėliai, pvz., V. Stumbrienės ir A. Kaškelevičienės *Nė dienos be lietuvių kalbos* (2014), buvo skirti dviem kalbos lygiams. Šio vadovėlio pratarmėje (p. 3) rašoma: „Mokydamiesi pagal šį vadovėlį mokiniai pasieks „Slenksčio" (B1) ir „Aukštumos" (B2) kalbos mokėjimo lygius, t. y. taps savarankiškais kalbos vartotojais." Po automatinio tekstų klasifikavimo tekstai iš tokių vadovėlių buvo priskirti vienam kuriam nors kalbos lygiui.

[10] Šiame eksperimente ilgu žodžiu buvo laikomi iš 8 ir daugiau raidžių sudaryti žodžiai.

[11] Skaitomumo (angl. *readability*) tyrimai (pvz., Melin et al., 1995) atskleidė, kad šis santykis (angl. *nominal quotient*) didėja aukštesniuose kalbos lygiuose.

mokyti naudotas vadovėlinių tekstų patekstynis buvo ne visai reprezentatyvus; pirmiausia tai nulėmė nevienodas skirtingo lygio žodžių skaičius (žr. 2 lentelę). Visų ir skirtingų žodžių santykis, sakinių skaičius ir vidutinis sakinio ilgis tekste buvo naudingesnės ypatybės, palyginti su kitomis paviršinėmis ir gilesnėmis lingvistinėmis ypatybėmis (Grigonytė et al., 2018). Pastebėta, kad tekstai iš vienos knygos (jei jie suskirstyti į atskirus dokumentus) galėjo patekti į skirtingus lygius. Tai suprantama, nes jei, pvz., tekste yra sakinio ilgio sentencija ir kelių sakinių ilgio aprašymas (pvz., kaip suorganizuoti krikštynas), šie fragmentai negali būti ir nėra tokio paties sudėtingumo. Taip pat reikia pripažinti, kad dalį ne iš vadovėlių imtų tekstų, kuriuos algoritmas priskyrė B2 lygiui, būtų galima priskirti aukštesnio lygio tekstams, bet šiame tekstyne jie laikomi B2 lygio.

Vadovėlinių tekstų automatinio klasifikavimo tikslumas pagal geriausią rezultatą pateikusį mašininio mokymo modelį – 60 proc. Šis modelis panaudotas klasifikuojant nevadovėlinius tekstus, taigi tekstų lygį pavyko nustatyti maždaug 60 proc. tikslumu. Gali būti, kad aukštesnio tikslumo rodiklio nepavyko pasiekti dėl kelių priežasčių: eksperimento eiga buvo preskriptyvi, t. y. į mašininio mokymosi procesą įtrauktos ypatybės buvo pasirinktos, o ne nustatytos iš analizuotų tekstų, vadinasi, iš panaudotų ypatybių ne visos buvo vienodai svarbios, o kai kurios svarbios ypatybės galbūt buvo visai neįtrauktos. Abejonių kėlė ir vadovėlinių tekstų lygiai: jie nekeisti, bet iš eksperimento paaiškėjo, kad kai kurie vadovėliniai tekstai turėtų būti priskirti kitam (aukštesniam arba žemesniam) lygiui. Kadangi mašininio mokymo modeliai mokosi iš aiškiai suklasifikuotų duomenų (šiuo atveju – vadovėlinių tekstų, kurių lygis buvo aiškiai nurodytas), tai irgi galėjo nulemti tokius automatinio tekstų klasifikavimo rezultatus. Analizuojant tekstyną buvo pastebėta anotavimo klaidų, todėl papildomas trukdis, nulėmęs prastesnį klasifikavimo rezultatą, galėjo būti ir morfologinio anotavimo netikslumai: tekstynas buvo automatiškai morfologiškai anotuotas, pažymos neperžiūrėtos, netvarkytos; sakinių ribos taip pat buvo nustatytos automatiškai (vėliau šiek tiek patikslintos).

Tęsiant automatinio tekstų klasifikavimo tyrimus, būtų galima analizuoti ir kitus požymius, pvz., sintaksinį sudėtingumą (apskaičiuojamą įvertinant vientisinių, sudėtinių homogeniškų ir sudėtinių heterogeniškų sakinių

santykį; žr. Kalinauskaitė, 2019), įtraukti kitokių paviršinių teksto leksinių ir morfologinių ypatybių (skaičiuoti tekstą sudarančius simbolius su tarpais ir be tarpų, skiemenis, unikalius žodžius; atsižvelgti į dažniausias bigramas ir trigramas; įtraukti daugiau morfologinių ypatybių: daiktavardžių pobūdį (tikriniai ar bendriniai), giminę, skaičių, linksnį ir pan. (plg. Butkienė et al., (2019) iš viso naudoja 44 morfologines ypatybes)).

Tekstyno kalbos vienetų įvairovė ir gramatinės ypatybės

Kaip minėta poskyryje *Tekstyno sandara*, šiame straipsnyje aptariamo rašytinės kalbos patekstynio dydis – 618 637 žodžiai. Visi tekstai buvo automatiškai morfologiškai anotuoti naudojant portale <http://semantika.lt/> prieinamą morfologinį anotatorių (plačiau žr. Dadurkevičius, 2017). Morfologinės pažymos buvo pateiktos pagal Leipcigo glosavimo taisykles, prie jų buvo pridėtos kelios lietuvių kalbai būdingos pažymos: laipsnių kategorijos, neasmenuojamų veiksmažodžių formos: dalyviai, padalyviai, pusdalyviai, būdiniai. Nors naudoto morfologinio anotatoriaus kokybė gana gera (98 proc. tikslumu nustato lemą, 95,3 proc. tikslumu parenka kalbos dalį, 86,8 proc. tikslumu nustato gramatines kategorijas; žr. Kapočiūtė-Dzikienė et al., 2017), vis dėlto pasitaiko anotavimo klaidų. Planuojama ateityje jas ištaisyti.

Tekstynų lingvistikoje įprasta nurodant tekstynų dydį pateikti visų (vieną ir daugiau kartų) pavartotų žodžių formų skaičių. Iš tekstyną sudarančių 618 637 žodžių 107 967 yra skirtingos žodžių formos (jos gali būti pavartotos vieną arba daugiau kartų). Tekstyne yra 36 195 lemos (jų pasiskirstymas pagal tekstų lygius matyti 6 lentelėje). Kaitybinių formų ir lemų skaičius koreliuoja su patekstynių dydžiu: kuo didesnis patekstynis, tuo jame daugiau kaitybinių formų ir lemų.

Iš šių skaičių galima padaryti tokias išvadas apie morfologines ypatybes: vidutiniškai vienai lemai būdinga 17,1 kaitybinės formos (įskaičiuojant skirtingas ir pasikartojančias) ir 2,98 skirtingų kaitybinių formų. Tai – gana tipiški lietuvių kalbos duomenys (plg. E. Rimkutė (2006), remdamasi 1 mln. žodžių morfologiškai anotuoto tekstyno duomenimis, nustatė, kad vidutiniškai vienam žodžiui būdinga 2,34 skirtingų kaitybinių formų).

6 lentelė

Tekstyno morfologinės ir sintaksinės ypatybės

Tekstų lygis	Visi tekstyną sudarantys vienetai (žodžiai, skyrybos ir kt. ženklai)	Žodžių formos (visos formos)	Žodžių formos (skirtingos)	Lemos	Gramatinių pažymų kombinacijos	Sakiniai	Vidut. sakinio ilgis
A1	59 685	42 876	13 177	6 562	380	9 949	4,3
A2	73 945	52 686	17 439	8 782	606	9 171	5,7
B1	89 849	67 995	24 296	11 141	562	7 073	9,6
B2	589 782	455 080	89 403	30 062	1 230	58 443	7,8
Iš viso	813 261	618 637	107 967	36 195	1 287	84 636	7,3

Visame rašytinės kalbos patekstynyje rastos 1 287 gramatinių pažymų kombinacijos, t. y. tam tikroms kalbos dalims būdingas gramatines kategorijas nurodančių pažymų grupė. Pvz., ties žodžio forma *stalo* būdingos tokios morfologinės pažymos: bendrinis daiktavardis, vyriškoji giminė, vienaskaita, kilmininkas; žodžio formai *geresniuosiuose* nurodyta, kad tai – būdvardis, aukštesnysis laipsnis, įvardžiuotinė forma, vyriškoji giminė, daugiskaita, vietininkas. Tokios gramatinių kategorijų grandinėlės ir vadinamos gramatinių pažymų kombinacijomis.

Svarbu atkreipti dėmesį į tai, kad gramatinių pažymų kombinacijų daugėja atsižvelgiant į tekstų skaičių ir sudėtingumą: kaip matyti 6 lentelėje, A1 lygio tekstuose jų mažiausiai – 380, o B2 lygio tekstuose daugiausiai – 1 230. Šią tendenciją ne visai patvirtina A2 lygio gramatinių pažymų skaičius – 606 ir B1 lygio – 562. Vis dėlto tai nėra dideli skirtumai, galbūt nulemti ne visai tikslaus tekstų suskirstymo pagal kalbos lygius (žr. poskyrį *Automatinis tekstų klasifikavimas*). Galima tvirtinti, kad aukštesnio lygio tekstams būdingos sudėtingesnės kaitybinės formos ir didesnė jų įvairovė.

Tekstyne yra 84 636 sakiniai, vidutinis sakinio ilgis – 7,3 žodžio. Sakinių ilgis taip pat daugmaž koreliuoja su tekstų sudėtingumu: kaip matyti iš 6 lentelės, A1 lygio tekstų sakiniai trumpiausi – 4,3 žodžio, B1 lygio sakiniai ilgiausi – 9,6 žodžio, nors tikėtasi, kad ilgiausiais sakiniais pasižymės B2 lygio tekstai. Kaip minėta anksčiau, tai gali nulemti ne visai tikslus automatinis tekstų suklasifikavimas arba kiti dalykai: pavyzdžiui, B2 lygio nevadovėlinių tekstų nemažą dalį (18 proc.) sudaro subtitrai: jų leksika gali būti gana sudėtinga, bet

sakiniai palyginti trumpi.

Planuojama ateityje išsamiau išanalizuoti rašytinės kalbos patekstynyje pavartotų žodžių gramatinius požymius. Tikėtina, kad tai leis tiksliau įvertinti, ar aukštesnių lygių tekstuose išryškėja sudėtingesnės gramatinės formos, o tai svarbu tolesniems kompiuterinio kalbos mokymosi tyrimams siekiant kokybiškesnio automatinio tekstų klasifikavimo.

Apibendrinamosios pastabos

Parengus mokomąjį lietuvių kalbos tekstyną, kitas svarbus su tekstynu susijęs darbas – paieškos sistemos kūrimas, nes mokomiesiems tekstynams reikia numatyti ir specifinę vartotojo sąsają (Widman et al., 2011). Planuojama, kad šiame tekstyne bus galima atlikti įvairialypę paiešką: ieškoti antraštinio žodžio, žodžio formos; paieška bus galima atlikti ir pagal gramatinius požymius. Numatoma paieška ne tik pagal leksikos vienetus, bet ir pagal gramatinę informaciją (pvz., kalbos dalis). Ištisi tekstai nebus prieinami, bus pateikiamos tik jų ištraukos, vadinamieji konkordansai (juos paprastai sudaro tekstų ištraukos iki 300 simbolių), kuriuose bus rasti vartotojo užklausą atitinkantys pavyzdžiai. Planuojama numatyti galimybę matyti teksto metaduomenis (kalbos lygį, kalbos atmainą, teksto tipą (vadovėliniai ar nevadovėliniai), failo pavadinimą ir pan.) ir atlikti paiešką pasirinktuose norimo lygio ir žanro tekstuose. Galimybė atlikti paiešką bus prieinama portale <https://kalbu.vdu.lt/>.

Mokomieji tekstynai, kuriuose įtraukta tekstų iš vadovėlių, gali būti papildomai anotuoti, atsižvelgiant į kalbos mokymo proceso poreikius (angl. *pedagogical annotation*): vadovėlių užduotys, įtrauktos į tekstyną, gali būti suanotuotos kaip skaitymo, klausymo ir kt. užduotys (apie tai plačiau žr. Meunier et al., 2009, taip pat Volodina et al., 2014a). Aprašomame tekstyne iš vadovėlių buvo paimta palyginti mažai užduočių, tačiau, atsižvelgiant į besimokančiųjų poreikius, gali būti pravartu daugiareikšmių žodžių konkordansus suanotuoti reikšmėmis, kad būtų galima juos surūšiuoti semantiškai, taip pat specialiomis žymomis sužymėti frazeologizmus, kad besimokantieji galėtų aiškiai atpažinti perkeltinės reikšmės junginius ir kt.

Šiame straipsnyje aprašytas tekstynas galėtų tapti vienu iš išteklių,

naudingų rengiant įvairias leksikai ir gramatikai perprasti skirtas užduotis, mokymui(si) aktualius duomenynus. Mokomojo tekstyno pagrindu jau kuriamas mokomasis elektroninis lietuvių kalbos vartosenos leksikonas, skirtas tiek besimokantiesiems, tiek jų dėstytojams. Jame bus pateikti vartosenos modeliai, į kuriuos įeina dažnai tekstyne pavartoti antraštiniai žodžiai: bus matyti, kokias sintaksines funkcijas (subjekto, objekto, atributo ir pan.) šie žodžiai atlieka, kokių reikšmių žodžiai įeina į tuos modelius (abstraktai, asmenų pavadinimai, būseną reiškiantys, ypatybes nurodantys žodžiai ir pan.). Prie vartosenos modelių bus pateikti iš mokomojo tekstyno atrinkti reprezentatyvūs pavyzdžiai (apie vartosenos modelių metodą žr. Kovalevskaitė et al., 2019).

Tikėtina, kad lietuvių kalbos kaip svetimosios mokytojai arba dėstytojai ras įvairių būdų panaudoti šiame straipsnyje aprašytą išteklių. Išsamesnės tekstyno panaudojimo galimybės, jo pritaikymas elektroniniam lietuvių kalbos vartosenos leksikonui sudaryti bus aprašytas ateityje, kai bus sukurta tekstyno paieškos sistema ir leksikonas bus viešai prieinamas vartotojams.

Mokomasis tekstynas pirmiausia yra mokyti(s) kalbos skirtas išteklius. Vėliau, tikėtina, šį tekstyną savo tyrimams galėtų naudoti mokslininkai, atliekantys kompiuterinio lietuvių kalbos mokymosi srities tyrimus, pavyzdžiui, kaip automatiškai nustatyti besimokančiojo kalbos lygį, kaip tiksliau nustatyti tekstų sudėtingumą ir pan.

Literatūra

Batinić, D., Birzer, S., & Zinsmeister H. (2016). Creating an extensible, levelled study corpus for learners of Russian. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 38–43.

Bendrieji Europos kalbų mokymosi, mokymo ir vertinimo metmenys (2008). Vilnius. http://www.lsk.flf.vu.lt/file/BEKM.pdf

Butkienė, R., Butleris, R., Ablonskis, L., Jurgelaitis, M., Šukys, A., Vaičiukynas, E., & Žitkus, V. (2019). Lietuviškų dokumentų tekstų pagrindinės ir specializuotos statistinės analizės IT sprendimas: pranešimas. *Tarpdisciplininio lietuvių kalbos technologijų projekto „Semantika2" tarpinių rezultatų pristatymas „Dirbtinis intelektas ir lietuvių kalba: kada kompiuteris visiškai supras lietuvių rašytinę ir*

sakytinę kalbą?“, 2019 m. rugsėjo 27 d. Kaunas, Vytauto Didžiojo universitetas.

Čubajevaitė, L. (2007). Lithuanian as a foreign language. Means for effective vocabulary learning/teaching. *Kalba ir kontekstai*, *2*, 285–296.

Dadurkevičius, V. (2017). Lietuvių kalbos morfologija atvirojo kodo *Hunspell* platformoje. *Bendrinė kalba*, *90*. http://www.bendrinekalba.lt/Straipsniai/90/Dadurkevicius_BK_90_straipsnis.pdf

Džežulskienė, J. (2014). Kalbos technologijų taikymo lietuvių kaip svetimai kalbai mokyti ypatumai. *Kalbų studijos*, *24*, 106–112.

Hancke, J., & Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. *Learner Corpus Research 2013*, *Book of Abstracts*. Bergen, Norway, 54–56. http://www.sfs.uni-tuebingen.de/~dm/papers/Hancke.Meurers-13.pdf

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

François, T. (2014). An analysis of a French as a Foreign Language corpus for readability assessment. *Proceedings of the third workshop on NLP for computer-assisted language learning. NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings*, *107*, 13–32.

Grigonytė, G., Kovalevskaitė, J., & Rimkutė, E. (2018). Linguistically-motivated automatic classification of Lithuanian texts for didactic purposes. Proceedings of the Eighth International Conference Baltic HLT 2018. In K. Muischnek, & K. Müürisep (Eds.), *Frontiers in Artificial Intelligence and Applications*, *307*. IOS Press, 38–46. http://ebooks.iospress.nl/volumearticle/50302

Kalinauskaitė, D. (2019). *Lietuvių kalbos tekstų informatyvumo nustatymas*: daktaro disertacija. Kaunas: Vytauto Didžiojo universitetas.

Kapočiūtė-Dzikienė, J., Rimkutė, E., & Boizou, L. (2017). A comparison of Lithuanian morphological analyzers. *20th International Conference “Text, Speech, and Dialogue” (TSD 2017)*. Springer International Publishing AG, 47–56.

Kovalevskaitė, J., & Jancaitė, L. (2019). Vartosenos modelių analizė mokomojoje leksikografijoje: žvalgomasis tyrimas lietuvių kalbos veiksmažodžių pavyzdžiu. *Taikomoji kalbotyra*, *12*, 124–153. https://taikomojikalbotyra.lt/ojs/index.php/taikomoji-

kalbotyra/article/view/195

Meunier, F., & Gouverneur, C. (2009). New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (Ed.), *Corpora and Language Teaching*. *Studies in Corpus Linguistics, 33,* 179–201. John Benjamins Publishing Company.

Ramonienė, M., Pribušauskaitė, J., & Vilkienė, L. (2016). *Aukštuma*. Vilniaus universiteto leidykla.

Rennes, E., & Jönsson, A. (2015). A tool for automatic simplification of Swedish texts. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, 317–320.

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

Rimkutė, E. (2006). *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne*: daktaro disertacija. Vytauto Didžiojo universitetas.

Römer, U. (2004). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and Language Learners* (pp. 151–169). John Benjamins Publishing Company.

Ruzaitė, J. (2019). Learner corpora for lesser taught languages: A work-in-progress report on the Lithuanian learner corpus. *Konferencijos „Sustainable Multilingualism*" *pranešimai*, 24–25 May 2019, Kaunas, Lithuania.

Vandeghinste, V., Bulté, B., & Augustinus, L. (2019). Wablieft: An easy-to-read newspaper corpus for Dutch. *CLARIN Annual Conference 2019*, 188–191.

Volodina, E., Ildikó, P., Eide, S. R., & Heidarsson, H. (2014a). You get what you annotate: A pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *Proceedings of the third workshop on NLP for computer assisted language learning. NEALT Proceedings Series, 22 / Linköping Electronic Conference Proceedings, 107*, 128–144.

Volodina, E., Ildikó, P., Borin, L., & Lindström Tiedemann, T. (2014b). A flexible

language learning platform based on language resources and web services. *Proceedings of LREC*. 26–31 May 2014, Reykjavik, Iceland.

Widmann, J., Kohn, K., & Ziai, R. (2011). The SACODEYL search tool: Exploiting corpora for language learning purposes. In A. Frankenberg-Garcia, G. Aston, & L. Flowerdew (Eds.), *New Trends in Corpora and Language Learning* (pp. 167–178). Continuum.

1 priedas. Kiekybinė informacija apie mokomojo tekstyno rašytinės kalbos patekstynio žanrus

1 lentelė

Vadovėlinių tekstų žanrai

Žanras	Žodžių skaičius	Procentai
Anekdotai	2 356	2,21
Anketos	394	0,37
Dainos	670	0,63
Dialogai	21 085	19,81
Dokumentai	0	0,00
Eilėraščiai	143	0,13
Etiketės	21	0,02
Horoskopai	402	0,38
Informaciniai tekstai	1 062	1,00
Instrukcijos	285	0,27
Interviu	2 984	2,80
Kita	21	0,02
Laiškai, žinutės	3 559	3,34
Linkėjimai, sveikinimai	589	0,55
Pasakojimai	24 298	22,83
Pasakos, legendos	2 384	2,24
Patarimai	1 008	0,95
Pažintiniai tekstai	37 952	35,66
Proza	881	0,83
Receptai, meniu	1 196	1,12
Reklamos	827	0,78
Sentencijos	0	0,00
Skelbimai	3 368	3,16
Subtitrai	0	0,00
Šūkiai	0	0,00

Žanras	Žodžių skaičius	Procentai
Tautosaka	0	0,00
Testai	503	0,47
Tvarkaraščiai, darbotvarkės	158	0,15
Viešieji užrašai	277	0,26
Iš viso	106 423	100

2 lentelė

Nevadovėlinių tekstų žanrai

Žanras	Žodžių skaičius	Procentai
Anekdotai	1 928	0,38
Anketos	39	0,01
Dainos	3 833	0,75
Dialogai	281	0,05
Dokumentai	7 990	1,56
Eilėraščiai	0	0,00
Etiketės	4 545	0,89
Horoskopai	20 085	3,92
Informaciniai tekstai	82 336	16,07
Instrukcijos	10 176	1,99
Interviu	6 277	1,23
Kita	0	0,00
Laiškai, žinutės	18 857	3,68
Linkėjimai, sveikinimai	0	0,00
Pasakojimai	18 191	3,55
Pasakos, legendos	12 480	2,44
Patarimai	55 622	10,86
Pažintiniai tekstai	75 169	14,68
Proza	66 457	12,97
Receptai, meniu	7 089	1,38
Reklamos	2 250	0,44
Sentencijos	601	0,12
Skelbimai	33 004	6,44
Subtitrai	93 941	18,34
Šūkiai	1 263	0,25
Tautosaka	4 333	0,85
Testai	502	0,10

Žanras	Žodžių skaičius	Procentai
Tvarkaraščiai, darbotvarkės	0	0,00
Viešieji užrašai	1 467	0,29
Iš viso	512 214	100

3 lentelė

Viso tekstyno dydis pagal žanrus

Žanras	Žodžių skaičius	Procentai
Anekdotai	4 284	0,69
Anketos	433	0,07
Dainos	4 503	0,73
Dialogai	21 366	3,45
Dokumentai	7 990	1,29
Eilėraščiai	143	0,02
Etiketės	4 566	0,74
Horoskopai	20 487	3,31
Informaciniai tekstai	83 398	13,48
Instrukcijos	10 461	1,69
Interviu	9 261	1,50
Kita	21	0,003
Laiškai, žinutės	22 416	3,62
Linkėjimai, sveikinimai	589	0,10
Pasakojimai	42 489	6,87
Pasakos, legendos	14 864	2,40
Patarimai	56 630	9,15
Pažintiniai tekstai	113 121	18,29
Proza	67 338	10,88
Receptai, meniu	8 285	1,34
Reklamos	3 077	0,50
Sentencijos	601	0,10
Skelbimai	19 870	3,21
Subtitrai	93 941	15,19
Šūkiai	1 263	0,20
Tautosaka	4 333	0,70
Testai, užduotys	1 005	0,16
Tvarkaraščiai, darbotvarkės	158	0,03
Viešieji užrašai	1 744	0,28
Iš viso	618 637	100

4 lentelė

Viso tekstyno dydis pagal žanrus, tekstų tipus ir kalbos lygius

Žanras / tekstų tipas	Vadovėliniai tekstai				Nevadovėliniai tekstai				Iš viso žodžių	Proc.
	A1	A2	B1	B2	A1	A2	B1	B2		
Anekdotai	1 296[12]	649	268	143	1 004	358	166	400	4 284	0,69
Anketos	221		173		39				433	0,07
Dainos		55	615		2 651	718	140	324	4 503	0,73
Dialogai	12 729	5 159	1 284	1 913	281				21 366	3,45
Dokumentai						50	2 434	5 506	7 990	1,29
Eilėraščiai			143						143	0,02
Etiketės	21				1 491	1 152	1 605	297	4 566	0,74
Horoskopai		402						20 085	20 487	3,31

[12] Prie atitinkamų lygių žanrų nurodytas žodžių skaičius.

Žanras / tekstų tipas	Vadoveliniai tekstai				Nevadoveliniai tekstai				Iš viso žodžių	Proc.
	A1	A2	B1	B2	A1	A2	B1	B2		
Informaciniai tekstai	209	488		365		46	1 871	80 419	83 398	13,48
Instrukcijos		67	218				1 748	8 428	10 461	1,69
Interviu		83	363	2 538	88	200	522	5 467	9 261	1,50
Kita		21							21	0,00
Laiškai, žinutės	1 972	799	338	450	5 071	750	4 994	8 042	22 416	3,62
Linkėjimai, sveikinimai	88	181	320						589	0,10
Pasakojimai	4 056	1 926	6 839	11 477	6 401	6 017	720	5 053	42 489	6,87
Pasakos, legendos	92	1 395	394	503	839	1 759	830	9 052	14 864	2,40
Patarimai		204	647	157	418	908	5 047	49 249	56 630	9,15
Pažintiniai tekstai	205	2 342	7 613	27 792	250	1 644	15 216	58 059	113 121	18,29

Žanras / tekstų tipas	Vadovėliniai tekstai				Nevadovėliniai tekstai				Iš viso žodžių	Proc.
	A1	A2	B1	B2	A1	A2	B1	B2		
Proza		227		654	430	2 815	4 611	58 601	67 338	10,88
Receptai, meniu	422	532	242		66	871	2 671	3 481	8 285	1,34
Reklamos	203	396	228		347	164	1 739		3 077	0,50
Sentencijos					515	60		26	601	0,10
Skelbimai	729	431	1 880	328		16 502			19 870	3,21
Subtitrai								93 941	93 941	15,19
Šūkiai						593	670		1 263	0,20
Tautosaka					293	926	784	2 330	4 333	0,70
Testai, užduotys			503		310	33	159		1 005	0,16
Tvarkaraščiai, darbotvarkės	132	26							158	0,03

Žanras / tekstų tipas	Vadoveliniai tekstai				Nevadoveliniai tekstai				Iš viso žodžių	Proc.
	A1	A2	B1	B2	A1	A2	B1	B2		
Viešieji užrašai	7	270				1 467			1 744	0,28
Iš viso	22 382	15 653	22 068	46 320	20 494	37 033	45 927	408 760	618 637	100,00

2 priedas. Mokomojo tekstyno rašytinės kalbos patekstynio šaltiniai

1. Vadovėliai

Čubajevaitė, L., Ruzaitė, J., & Lemanaitė, G. (2014). *Takas.* Vytauto Didžiojo universitetas. Versus aureus.

Džežulskienė, J. (2005). *Lietuvių kalba kitakalbiams.* Technologija.

Džežulskienė, J. (2014). *Kalbu lietuviškai*. https://www.easylithuanian.com/

Hilbig, I., Migauskienė, R., Našlėnaitė-Eberhardt, V., Petrašiūnienė, E., Tamošaitienė, A., Valančiauskienė, A., & Vaškevičienė, L. (2010). *Sveiki atvykę!* Vilniaus universiteto leidykla.

Hilbig, I., Stumbrienė, V., & Vaškevičienė, L. (2009). *Trumpas lietuvių kalbos kursas pradedantiesiems*. Vilniaus universiteto leidykla.

Jakaitienė, E. (1994). *Lietuviškai apie Lietuvą*. Alma littera.

Kruopienė, I. (2009). *10 žingsnių. Trumpas lietuvių kalbos kursas pradedantiesiems*. Vilniaus universitetas.

Migauskienė, R. (2014). *Žingsnis.* Eugrimas.

Migauskienė, R., & Vaisėtaitė, E. (2014). *Žodis žodį veja*. Eugrimas.

Narbutas, E., Pribušauskaitė, J., Ramonienė, M., Skapienė, S., & Vilkienė, L. (2002). *Slenkstis*. Council of Europe Press.

Pribušauskaitė, J., Ramonienė, M., Skapienė, S., & Vilkienė, L. (2000). *Aukštuma.* Council of Europe Press.

Ramonienė, M., & Vilkienė, L. (1998, 1999). *Po truputį (mokytojo ir mokinio knygos)*. Baltos lankos.

Ramonienė, M., Pribušauskaitė, J., & Vilkienė, L. (2006). *Pusiaukelė.* Europos Taryba.

Stumbrienė, V., & Kaškelevičienė, A. (2002). *Nė dienos be lietuvių kalbos.* Gimtasis žodis.

Vaškevičienė, L., Kutanovienė, E., & Valančiauskienė, A. (2015). *Pažiūrėk! Paklausyk! Pasakyk!* Eugrimas.

2. Nevadovėliniai šaltiniai

2.1. Informaciniai tekstai

Delfi.lt. 2014–2016 m. įvairių rubrikų tekstai.

Dialogai apie gyvenimą, užrašyti 2008 m. Kalbėjosi Rasma Aniulytė, Asta Ziutelytė, Stasė Venčaitienė. VDU Kultūrų studijų katedros Etnologijos rankraštynas.

Dienoraštis, užrašytas 2009 m. Į rankraštyną pateikė Klara Liebutė. VDU Kultūrų studijų katedros Etnologijos rankraštynas.

Diktantai pirmokams, antrokams, trečiokams, ketvirtokams. http://mudubudu.lt

Dokumentai (informacija, kaip atsidaryti banko sąskaitą, gauti nedarbo arba motinystės išmoką, būsto paskolą; kas yra draudžiamas privalomuoju sveikatos draudimu; kam išduodamas leidimas gyventi Lietuvoje ir pan.), surinkti 2018 m. iš įvairių valstybinių įstaigų interneto svetainių.

Elektroniniai laiškai. Projekto vykdytojų asmeniniai duomenys.

Etiketės, trumpi aprašai apie prekes, surinkti 2018 m. iš įvairių interneto svetainių.

Horoskopai, surinkti 2010 m. iš įvairių interneto svetainių.

Instrukcijos (informacija apie įvairių buitinių prietaisų (televizoriaus, skalbyklės, lygintuvo ir pan.) naudojimą ir priežiūrą; namų taisyklės (ką galima ir ko negalima daryti), surinktos 2018 m. iš įvairių interneto svetainių.

Pažinčių skelbimai, surinkti 2005 m. iš laikraščių ir žurnalų *Lietuvos rytas*, *Atleisk*, *Viltys ir likimai*, *Antra pusė*, *Gyvenimiškos istorijos*.

SMS žinutės; internetinių forumų žinutės. Projekto vykdytojų asmeniniai duomenys.

Viešieji užrašai, surinkti 2018 m. iš gydymo, maitinimo, kultūros įstaigų, viešojo

transporto ir pan. viešųjų vietų.

2.2. Pažintiniai tekstai

Aleksaitė, I., & Jazbutytė, N. (2008). *Geros manieros – pusė karjeros*. Mintis.

Aleksaitė, I., & Jazbutytė, N. (2010). *Prie stalo, ant stalo, po stalu: šventė kiekvienuose namuose*. Mintis.

Flintas, Flinto bumas (žurnalas) (2008). Jūsų Flintas.

Gudzinskas, Z. (2010). *Kur uogauti Lietuvoje*. Šviesa.

Heiney, P. (2008). *Ar karvės gali nulipti laiptais? Atsakymai į keblius klausimus*. Mintis.

Imbrasienė, B. (2010). *Lietuvių kulinarijos paveldas*. Baltos lankos.

Iršėnaitė, R. (2010). *Kur grybauti Lietuvoje*. Šviesa.

Kauno turizmo informacijos centras (ekskursijų aprašai, lankytų objektų, švenčių aprašai, informacija) (2017). https://visit.kaunas.lt/lt/

Kelionių aprašymai, surinkti 2017 m. http://www.kiveda.lt/

Kelionių aprašymai, surinkti 2017 m. https://www.gruda.lt/

Kelionių aprašymai, surinkti 2017 m. https://www.makalius.lt/

Klaipėdos turizmo informacijos centras (ekskursijų aprašai, lankytų objektų, švenčių aprašai, informacija) (2017). http://www.klaipedainfo.lt/

Mergaitė (žurnalas) (2008). Egmont Lietuva.

Penki: užduotys, juokai, konkursai, prizai (2017), *2*. UAB Presa.

Pokalbiai apie Kauną, užrašyti 2009 m. Kalbėjosi Justė Vasilionytė-Stašaitienė. VDU Kultūrų studijų katedros Etnologijos rankraštynas.

Saugaus pirmoko pasas (2012). Demokratinių iniciatyvų centras.

Semaška, A. (2007). *Lietuvos keliais. Turisto žinynas*. Algimantas.

Spalvink: Ledo šalis (2017). *Dysney*, *2*, liepa–rugsėjis.

SU P.E.R. (žurnalas) (2008). Jūsų Flintas.

Tekstai, gauti iš lietuvių kalbos kaip svetimosios dėstančių dėstytojų.

Vilniaus turizmo informacijos centras (ekskursijų aprašai, lankytų objektų, švenčių aprašai, informacija) (2017). http://www.vilnius-tourism.lt/

Žilinskas, R. (2010). *Kur žvejoti Lietuvoje*. Šviesa.

2.3. Apeliaciniai tekstai

Reklaminiai ir nereklaminiai šūkiai, surinkti 2017 m. iš įvairių įmonių interneto

svetainių.

Socialinės reklamos, surinktos 2008 m. iš pakelės stendų, interneto, televizijos.

2.4. Meniniai tekstai

Ambrukaitis, J., & Pobrein, V. (2001). *Lietuvių kalba 5. Antroji knyga*. Šviesa.

Anekdotai, užrašyti 2009 m. VDU Kultūrų studijų katedros Etnologijos rankraštynas.

Beresnevičius, G. (2005). *Pabėgęs dvaras*. Lietuvos rašytojų sąjungos leidykla.

Dzvankauskaitė, I. (2009). *Lietuvių šiuolaikinių populiariųjų jaunimo dainų kalbinės ypatybės*: bakalauro darbas. Kaunas: Vytauto Didžiojo universitetas (naudotas priedas, kuriame pateiktos dainos).

Gimberis, J. (2011). *Jūs turite teisę tylėti*. Versus aureus.

Gudonytė, K. (2012). *Ida iš šešėlių sodo*. Tyto alba.

Inis, L. (2012). *Atsidaro metų durys*. Arx reklama.

Kasparavičius, K. (2009). *Baltasis dramblys. Tolimųjų kraštų istorijos*. Nieko rimto.

Kunčinas, J. (2006). *Baltųjų sūrių naktis.* Gimtasis žodis.

Mokyklų himnai, surinkti 2000–2012 m. iš įvairių interneto svetainių.

Patarlės ir priežodžiai: elektroninis sąvadas, 1998–2005. http://www.aruodai.lt/patarles/

Skeris, R. (1990). *Ką žmonės dirba visą dieną?* Vyturys.

Subtitrai (meninių filmų). http://www.subtitrai.net

Šepetys, R. (2011). *Tarp pilkų debesų*. Alma littera.

Šimaitis, V. (2008). *Komunalinis bliuzas*. Vaga.

Vincė, L. (2008). *Lenino galva ant padėklo. Amerikietės studentės dienoraštis, rašytas paskutiniais Sovietų Sąjungos gyvavimo metais*. Lietuvos rašytojų sąjungos leidykla.

Jolanta Kovalevskaitė
Vytautas Magnus University, Lithuania; jolanta.kovalevskaite@vdu.lt
Erika Rimkutė
Vytautas Magnus University, Lithuania; erika.rimkute@vdu.lt

PEDAGOGIC CORPUS OF LITHUANIAN: A NEW RESOURCE FOR LEARNING AND TEACHING LITHUANIAN AS A FOREIGN LANGUAGE

Summary. The paper aims to present the first pedagogic corpus of Lithuanian i.e. monolingual specialized corpus, prepared for learning and teaching Lithuanian in a foreign language classroom. The corpus has been collected as a result of the project "Lithuanian Academic Scheme for International Cooperation in Baltic Studies". It is motivated by the need to have a more appropriate resource which could be representative, authentic and relevant enough concerning the process of learning and teaching Lithuanian as it is known that language represented in other existing corpora of Lithuanian (e.g. Corpus of Contemporary Lithuanian, 140 m tokens) is too complex to use for learning activities. The pedagogic corpus includes authentic Lithuanian texts, selected using such criteria as a learner-relevant communicative function and genre. Spoken language as well as written language are represented in the corpus. The size of the corpus is 669.000 tokens: 111.000 tokens from texts and spoken language for A1–A2 levels, 558.000 tokens from texts and spoken language for B1–B2 levels (according to the CEFR – Common European Framework of Reference for Languages). In this paper, we aim to discuss in detail the written subpart of the corpus (containing 620.000 tokens) which includes levelled texts from coursebooks and unlevelled texts from other sources. The level-appropriate labels were assigned automatically to the texts from other sources and this text classification procedure is presented in the paper. The texts from coursebooks and other sources could be classified into 29 text types (dialogs, narratives, information, etc.) and 4 groups according to the communicative aims: informational texts, educational texts, advertising and fiction. Informational texts comprise the biggest part of the corpus; three mostly represented text types differ in coursebook texts and other sources: the most common coursebook texts are informational, narratives, and dialogs (appr. 78% of all coursebook texts). Texts from other sources are represented with richer diversity – appr. 73% of all texts from this subpart can be classified into 5 text types: subtitles, informational texts, educational texts, fiction, and advisory texts. The future work making pedagogic corpus available for learners and its possible application are presented in the closing remarks.

Keywords: pedagogic corpus; teaching Lithuanian as a foreign language; Common European Framework of Reference for Languages; written language; spoken language; automatic text classification.