Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

Sakytinės lietuvių kalbos tekstynas – natūralios vartosenos tyrimų šaltinis

Laura Kamandulytė-Merfeldienė
Vytauto Didžiojo universitetas
laura.kamandulyte@vdu.lt

Anotacija

Straipsnyje pristatomas *Sakytinės lietuvių kalbos tekstynas*, skirtas spontaninės ir parengtos kalbos analizei. Šiuo metu (2017 m.) *Sakytinės lietuvių kalbos tekstyną* sudaro 265 pokalbiai, apimantys daugiau nei 380 000 žodžių. Tekstynas yra subalansuotas ir apima pokalbius, kauptus atsižvelgiant į sakytinės kalbos pobūdį ir struktūrą, ryšį tarp pašnekovų, demografinius rodiklius, socialinius pašnekovų santykius. Straipsnyje išamiai aptariami šie kriterijai, aprašoma tekstyno struktūra ir jo kūrimo etapai (įrašų kaupimas, transkribavimas, gramatinis transkripcijų anotavimas), duomenų kaupimo ir skaitmeninimo metodika, taip pat aptariamos tekstyno panaudojimo natūralios vartosenos tyrimuose galimybės, trumpai pristatomi jau atlikti tekstyno duomenimis paremti tyrimai.

Šiuo metu vykdant LMT finansuojamą projektą pagal Valstybinę lituanistinių tyrimų ir sklaidos 2016–2024 metų programą „Šiuolaikinė sakytinė lietuvių kalba: leksikos ir gramatikos tyrimas tekstynų lingvistikos metodu“ (LIP-085/2016) atliekami tekstyno analize paremti tyrimai, kuriama nauja internetinė prieiga. Tikimasi, kad 2018 m. vartotojams suteikus daugiau tekstyno duomenų analizės galimybių internete, sakytinės kalbos tyrimų padaugės ir jie apims įvairias leksikos ir gramatikos sritis.

Raktiniai žodžiai: tekstynas; sakytinė kalba; spontaninė kalba; morfologinis anotavimas; lietuvių kalba.

1. Įvadas

Pastaraisiais metais sparčiai besivystančios technologijos leidžia kaupti įvairaus dydžio duomenų bazes, garsynus bei tekstynus, taikyti įvairias analizės galimybes ir greitai atlikti tyrimus. Nepaisant to, sakytinės kalbos tekstynų plėtra nėra labai sparti, o sakytinės kalbos tyrimai dar gana riboti ir nesistemingi. Nors sakytinės kalbos tekstynų kūrimo pradžia siejama su 1952 m., kai buvo sukurtas

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

250 000 žodžių apimantis anglų kalbos tekstynas, skirtas spontaninės kalbos gramatikos analizei (Fries 1952), iki šiol tokio pobūdžio tekstynų nėra daug, o kuriami sakytinės kalbos tekstynai dažnai yra specializuoti ir skirti konkrečių tyrimų tikslams. Sakytinės kalbos duomenų kaupimas ir skaitmeninimas yra didelis, ilgai trunkantis, didelių finansinių ir žmogiškųjų išteklių reikalaujantis darbas, todėl sakytinės, ypatingai spontaninės kalbos, tekstynų kūrimas dar ir šiandien ne tik Lietuvoje, bet ir visame pasaulyje susiduria su technologiniais ir metodiniais iššūkiais.

Šio straipsnio **tikslas** – pristatyti *Sakytinės lietuvių kalbos tekstyną* – natūralią vartoseną atspindintį duomenų šaltinį, skirtą spontaninės ir parengtos kalbos analizei. Straipsnyje aptariami tekstyno kūrimo etapai, pristatoma tekstyno struktūra, aptariamos tekstyno panaudojimo natūralios vartosenos tyrimuose galimybės, trumpai pristatomi jau atlikti tyrimai.

2. Sakytinės lietuvių kalbos tekstyno ištakos

Sakytinės lietuvių kalbos tyrimų, vykdomų Vytauto Didžiojo universitete, pradžia yra susijusi su vaikų kalbos tyrimais. 1994 m. dalyvaujant tarptautiniame projekte „The Acquisition of Pre- and Protomorphology“ (vadovas W. U. Dressler, Austrijos mokslų akademija), skirtame ankstyvosios vaikų kalbos raidos tyrimams, buvo pradėti kaupti lietuvių vaikų kalbos duomenys. I. Dabašinskienė (Savickienė), P. Wojcik ir M. Smoczynska sukaupė ir suskaitmenino dviejų vaikų kalbos įrašus bei, naudodami kompiuterinę programą CHILDES (angl. Child Language Data Excange System http://childes.psy.cmu.edu/, MacWhinney 2000), sukūrė transkribuotą ir gramatiškai anotuotą šių vaikų kalbos duomenų bazę. Naudojantis šia duomenų baze – pirmuoju sakytinės lietuvių kalbos tekstynėliu – buvo atlikta nemažai tyrimų, davusių pradžią ne tik psicholingvistikos mokyklai, bet ir sakytinės kalbos tyrimams Lietuvoje: 1999 m. buvo apginta I. Dabašinskienės disertacija „Lietuvio vaiko daiktavardžio morfologija“ (Savickienė 1999), įvairiuose straipsniuose analizuotos gramatinės vaikų kalbos ypatybės, tirta suaugusiųjų, bendraujančių su vaikais, kalba (Savickienė 1998, 2001, 2002, 2003, 2006a, 2007 ir kt.). I. Dabašinskienės ir kolegų pradėta kurti lietuvių vaikų kalbos duomenų bazė vėliau buvo plečiama: šiuo metu Vytauto Didžiojo universitete kaupiamą vaikų kalbos tekstyną sudaro tekstine forma transkribuoti ir morfologiškai koduoti 5 vaikų kalbos įrašai, surinkti ilgalaikio stebėjimo metodu ir apimantys vaikų kalbos duomenis (kurie įrašinėti kelis kartus per savaitę po pusę valandos) nuo gramatinės sistemos formavimosi pradžios iki jos įsitvirtinimo (vidutiniškai nuo 1,7 m. iki 3,5 m.). Dabartiniu metu vaikų kalbos tekstyną sudaro apie 400 000 žodžių formų, jis pildomas dvynių kalbos duomenimis (vykdant LMT projektą pagal Valstybinės lituanistinių tyrimų ir sklaidos 2016–

2024 metų programą Nr. LIP-020/2016, vadovė I. Balčiūnienė).

Remiantis vaikų kalbos tekstyno duomenimis, buvo aptartos iš sukaupto tekstyno išryškėjusios suaugusiųjų, bendraujančių su vaikais, kalbos ypatybės, atkreiptas dėmesys į gramatines lietuvių kalbos ypatybes (Savickienė 2000, 2005, 2006b; Balčiūnienė 2005a, 2005b; Kamandulytė 2007, 2009, 2012 ir kt.), iki šiol neminimas teoriniuose darbuose, aprašančiuose idealiąją bendrinę kalbą, ir praktiniuose tyrimuose, tradiciškai paremtuose rašytinės kalbos analize. Siekiant aptarti vaikų ir vaikams skirtoje kalboje užfiksuotus reiškinius ar ypatybes, nustatyti jų priežastis ir skirtumus nuo įprastos kalbos, kilo poreikis analizuoti spontaninę suaugusiųjų kalbą. Siekiant sukaupti natūralios (neparengtos) spontaninės suaugusiųjų kalbos duomenis, 2006 m. VDU buvo pradėtas vykdyti LVMSF finansuojamas projektas, skirtas buitinės šnekamosios kalbos įrašams kaupti ir transkribuoti. 2007–2008 m., gavus LVMSF paramą, buvo sukurtas morfologiškai anotuotas 225 000 žodžių apimties *Sakytinės lietuvių kalbos tekstynas* (http://donelaitis.vdu.lt/sakytines-kalbos-tekstynas/), prie kurio prisidėjo įvairių Lietuvos mokslo institucijų darbuotojai[1], kaupę sakytinės kalbos duomenis įvairiuose Lietuvos regionuose. Vykdant šį projektą didelis dėmesys skirtas ne tik tekstyno kūrimo metodikai, įrašų kaupimui, transkribavimui ir gramatiniam anotavimui, bet ir programos CHILDES pritaikymui lietuvių kalbai. Automatinio kalbos apdorojimo ir analizės programa CHILDES anotuoja pagal šios programos reikalavimus transkribuotus tekstus remdamasi leksikonu – specialia forma pateikiamu žodynu, apimančiu gramatiškai anotuotas žodžių formas. Taigi šio projekto metu buvo sukurtas leksikonas (žr. 1 pav.), kurio pagrindas – VDU Kompiuterinės lingvistikos centre parengtas 65 000 sulemuotų ir morfologiškai anotuotų dažniausiai vartojamų rašytinės lietuvių kalbos formų sąrašas (plačiau apie tai žr. Utka 2005)[2], parengtas remiantis *Dabartinės lietuvių kalbos tekstynu* (http://donelaitis.vdu.lt). Naudojantis programa CHILDES ir parengtu leksikonu, automatiškai anotuojant sakytinės kalbos transkripcijas, buvo atpažįstama (gramatiškai užkoduojama) apie 60 % žodžių formų; neatpažintos žodžių formos buvo anotuojamos rankiniu būdu jas įtraukiant į leksikoną ir taip jį gausinant. Projekto metu leksikonas buvo pagausintas iki 100 000 žodžių formų.

[1] Projektas vykdytas VDU, vadovė – I. Dabašinskienė. Už bendradarbiavimą projekte dėkojame projekto partneriams – L. Vaicekauskienei, J. Girčienei, G. Tamaševičiui (Lietuvių kalbos institutas), M. Ramonienei, N. Linkevičienei, E. Petrašiūnienei (Vilniaus universitetas), D. Pagojienei, S. Sirvydienei, V. Drukteinytei (Klaipėdos universitetas), G. Kačiuškienei, R. Gedrimui (Šiaulių universitetas)

[2] Dėkojame VDU Kompiuterinės lingvistikos centro darbuotojams už bendradarbiavimą. Ypatingai dėkojame A. Utkai už pagalbą kuriant tekstyną.

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

```
angliškomis    {[scat adj:01:undef:fm:pl:ins]} "angliškas"    =angliškomis=
durniausių     {[scat adj:01:undef:sup:ms:pl:gen]}    "durnas"       =durniausių=   \
      {[scat adj:01:undef:sup:fm:pl:gen]}    "durnas"       =durniausių=
gudriausių     {[scat adj:05:undef:sup:ms:pl:gen]}    "gudrus"       =gudriausių=   \
      {[scat adj:05:undef:sup:fm:pl:gen]}    "gudrus"       =gudriausių=
nesinešioja    {[scat v:neg:ref:pres:3]}      "nešiotis"     =nesinešioja=
persiskaitau {[scat v:ref:pres:sg:1]}  "persiskaityti" =persiskaitau=
protingiausių {[scat adj:01:undef:sup:ms:pl:gen]}      "protingas"    =protingiausių= \
      {[scat adj:01:undef:sup:fm:pl:gen]}    "protingas"    =protingiausių=
verčiame       {[scat v:pres:pl:1]}   "versti"       =verčiame=
```

1 pav. Ištrauka iš *Sakytinės lietuvių kalbos tekstyno* kūrime naudojamo leksikono

Taigi 2007–2008 m. vykdyto projekto metu atlikti darbai tapo pagrindu tolimesnei tekstyno plėtrai, o sukauptas 225 000 žodžių morfologiškai anotuotas tekstynas davė pradžią pirmiesiems sistemingiems sakytinės kalbos tyrimams. Naudojantis šiuo tekstynu buvo tirtos leksinės ir morfologinės sakytinės kalbos ypatybės (Dabašinskienė 2008a, 2009; Dabašinskienė, Kamandulytė 2009; Kamandulytė, Tuškevičiūtė 2008; Kamandulytė-Merfeldienė, Godliauskas 2014; Balčiūnienė 2010; Vaicekauskienė, Dabašinskienė, Kamandulytė-Merfeldienė 2013), analizuoti lietuvių kalbotyroje neaptarti reiškiniai, pvz., pertarų vartosena sakytinėje kalboje (Kamandulytė-Merfeldienė 2014), sutrumpėjusių formų dažnumas (Dabašinskienė 2008b), galūnių kaitos atvejai paradigmose (Kamandulytė-Merfeldienė 2010), sudurtinių daiktavardžių rūšys ir tipai (Dabašinskienė, Kamandulytė-Merfeldienė 2017). Atliekant šiuos tyrimus pastebėta, kad sakytinės kalbos sintaksės ypatybės ypatingai skiriasi nuo aprašytųjų gramatikose, taigi nuspręsta kai kurias iš jų patyrinėti išsamiau.

2015 m. gavus LMT paramą pagal Nacionalinę lituanistikos plėtros 2009–2015 m. programą (sutarties nr. LIT-9-11, VDU, vadovė L. Kamandulytė-Merfeldienė), buvo atliktas dalinis sintaksinis tekstyno anotavimas: sužymėti struktūriniai ir funkciniai pasakymų tipai, nustatyta sintaksinė būdvardžių funkcija ir sudėtinių prijungiamųjų sakinių dėmenų tvarka. Šio projekto metu buvo analizuoti šie dar netirti sintaksiniai lietuvių kalbos reiškiniai: atributinių ir predikatinių junginių su būdvardžiais žodžių tvarka natūralioje spontaninėje kalboje (Kamandulytė-Merfeldienė, Balčiūnienė 2016b), sakytinės kalbos pasakymų struktūra ir funkcijos (Balčiūnienė, Kamandulytė-Merfeldienė 2016; Kamandulytė-Merfeldienė, Balčiūnienė 2016c), pažymimieji pasakymai sakytinėje kalboje (Kamandulytė-Merfeldienė 2016a).

2016–2018 m. gavus Lietuvos mokslo tarybos paramą pagal Valstybinę lituanistinių tyrimų ir

sklaidos 2016–2024 metų programą vykdomas projektas „Šiuolaikinė sakytinė lietuvių kalba: leksikos ir gramatikos tyrimas tekstynų lingvistikos metodu“ (LIP-085/2016, VDU, vadovė L. Kamandulytė-Merfeldienė), kurio metu toliau tęsiami sakytinės kalbos tyrimai, didžiausią dėmesį skiriant skirtingų sakytinės kalbos atmainų analizei. Iki šiol atliktuose tyrimuose į sakytinę kalbą žiūrėta kaip į vientisą raiškos formą, išskiriant tik parengtos kalbos (žiniasklaidos, akademinio diskurso) ir spontaninės (buitinės) kalbos skirtumus. Tačiau atliekant šiuos tyrimus pastebėta, kad pagal tai, kokiose srityse kalba yra vartojama, kokias funkcijas atlieka, taip pat atsižvelgiant į socialinius veiksnius, sakytinę kalbą reikėtų skaidyti į dar smulkesnes kalbos atmainas taip, kaip rašytinė kalba skaidoma į funkcinius stilius. Taigi šio projekto metu tekstynas gausinamas įvairias situacijas apimančiais pokalbiais (pvz., telefoniniai pokalbiai, pokalbiai darbo aplinkoje, instituciniai pokalbiai ir t. t.), leisiančiais atlikti įvairių kalbos atmainų / registrų tyrimus. 2016–2018 m. vykdomo projekto metu į *Sakytinės lietuvių kalbos tekstyną* įtraukta naujų įvairias sritis ir situacijas apimančių pokalbių. Taigi jais papildžius tekstyną, šiuo metu (2017 m.) jis apima 383 587 žodžius. Šio projekto metu taip pat buvo papildytas gramatiniam anotavimui CHILDES programa reikalingas leksikonas. Dabartiniu metu jis sudaro apie 200 000 skirtingų žodžių formų, o jį naudojant automatiškai atpažįstama (sukoduojama gramatiškai) apie 80 % sakytinės kalbos žodžių formų.

3. Sakytinės lietuvių kalbos tekstyno struktūra

Rengiant *Sakytinės lietuvių kalbos tekstyno* kūrimo metodologiją, didelis dėmesys buvo skirtas tekstyno kompozicijai, t. y. tekstyno proporcijų parinkimui. Anot R. Marcinkevičienės (2000: 7), pagrindiniai gero tekstyno požymiai – dydis ir bendras pobūdis (reprezentatyvumas). Pasaulinė praktika rodo, kad gramatiškai anotuoti sakytinės kalbos tekstynai nebūna dideli, kadangi jų kaupimas ir anotavimas užima be galo daug laiko, jam nepritaikomos rašytinės kalbos apdorojimo priemonės. O’Keeffe ir Farr (2003) teigimu, jei 5 mln. žodžių rašytinės kalbos tekstyną galime laikyti visiškai mažu, 1 mln. žodžių sakytinės kalbos tekstyną laikome labai dideliu. Nacionaliniuose kai kurių kalbų tekstynuose, kurių kūrimui skiriami dideli finansiniai ir žmogiškieji ištekliai, sakytinės kalbos dalis yra daug mažesnė nei rašytinės kalbos: pvz., *British National Corpus* sakytinė kalba sudaro 10 % viso tekstyno. Kai kurie autoriai teigia, kad kalbant apie sakytinės kalbos tekstynus, yra svarbus ne žodžių skaičius, o tekstyno struktūra ir jį sudarančių pokalbių įvairovė, apimanti įvairias situacijas ir pašnekovų tipus (McCarthy, O’Keeffe 2008). Taigi siekiant tekstyno universalumo ir tinkamumo įvairiapusei analizei, nuspręsta laikytis subalansuoto tekstyno principo ir kaupiant tekstyną atsižvelgti į kelis kriterijus: sakytinės kalbos pobūdį (privati

vs vieša kalba) ir struktūrą (dialogai *vs* polilogai), ryšį tarp pokalbyje dalyvaujančių pašnekovų (tiesioginis *vs* netiesioginis (pvz., kalba telefonu), demografinius rodiklius, socialinius pašnekovų santykius. Toliau išsamiau aptarsime šiuos kriterijus.

3.1. Sakytinės kalbos pobūdis ir struktūra

Sakytinė kalba apima įvairias kalbėjimo situacijas. Pagal jas sakytinę kalbą galime skirti į dvi atmainas: spontaninė šnekamoji kalba ir parengta sakytinė kalba (parengta viešoji kalba). Nuo spontaninio ar viešojo kalbėjimo pobūdžio priklauso įvairių kalbos lygmenų ypatybės – fonetikos, leksikos, morfologijos, sintaksės.

Skirti tekstus į spontaniškus pokalbius ir viešąsias kalbas svarbu ne tik kalbiniu, bet ir psicholingvistiniu bei sociolingvistiniu požiūriu. Kalbėdamas viešai, žmogus visuomet stengiasi kontroliuoti savo kalbėjimą: vartoti bendrinę kalbą, kalbėti taisyklingai, nepažeisti mandagaus pokalbio reikalavimų. Taigi viešųjų kalbų įrašai gali padėti atskleisti dabartinės bendrinės lietuvių kalbos būklę, nustatyti būdingiausius kalbos normų pažeidimus. Vis dėlto nereikia pamiršti, kad nors viešoji kalba parengta iš anksto, ji nepraranda spontaniškumo: mintis, nors ir apgalvota, formuluojama kalbos sakymo metu, o jos raiška yra unikali, todėl viešųjų kalbų tyrimai gali padėti nustatyti ir kalbos riktų – netaisyklingų spontaniškų pasakymų, dažniausiai atsirandančių dėl kalbėtojo jaudinimosi, patiriamo streso ir įtampos kalbant viešai, ypatybes lietuvių kalboje bei pan. Bendraudamas privačioje aplinkoje su gerai pažįstamais žmonėmis – draugais, giminaičiais, šeimos nariais, kolegomis – žmogus mažiau kontroliuoja savo kalbą, bendrauja laisviau ir drąsiau, nevengia vartoti tarmės, profesinio žargono, slengo. Todėl spontaniškų pokalbių įrašai naudingi tyrinėjant bendrinės lietuvių kalbos ir sociolektų sąveiką, kodų ir socialinių vaidmenų kaitą bei natūralaus pokalbio struktūrą.

Taigi remiantis minėtais kriterijais, *Sakytinės lietuvių kalbos tekstyne* nuspręsta kaupti ir spontaninės sakytinės kalbos, ir paruoštos viešosios kalbos duomenis, nes tokių duomenų analizė yra įdomi ne tik bendruoju lingvistiniu, bet ir sociolingvistiniu bei psicholingvistiniu požiūriu. Reikia paminėti, kad didžioji tekstyno dalis vis dėlto skiriama spontaninei kalbai, nes ši atmaina apima įvairius registrus, atspindinčius įvairias kalbėjimo situacijas, socialinius dalyvių vaidmenis ir kt. Viešasis kalbėjimas daugiausiai yra susijęs su žiniasklaidos kalba arba akademine komunikacija, todėl tekstyne daugiausiai viešosios komunikacijos pokalbių apima šias sritis.

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

Siekiant sukurti subalansuotą sakytinės lietuvių kalbos tekstyną, atsižvelgta ir į pokalbių struktūrą: dialogą, monologą, polilogą. Aprašomą tekstyną sudarantys pokalbiai apima daugiausiai dialogus ir polilogus, kurie yra būdingi spontaninei komunikacijai, mažesnę dalį sudaro monologai, būdingi viešajai komunikacijai (paskaitoms, pranešimams ir t. t.). Tačiau reikia paminėti, kad monologai neretai pereina į dialogą, pvz., diskusijas, o ir patys į tekstyną įtraukti monologai nėra skaitomos ar mintinai pasakojamos kalbos (išskyrus kelis žinių laidų įrašus), tai – paruoštos, tačiau ne skaitomos akademinės kalbos arba pokalbiai žiniasklaidoje, kuriems taip pat būdingas spontaniškumo elementas. Į tekstyną skaitomi monologai nėra įtraukti, nes manoma, kad jie daugiau atspindi rašytinę, o ne sakytinę kalbą.

3.2. Ryšys tarp pašnekovų

Komunikacijos procese dalyvauja adresantas ir adresatas/-ai, kurie bendrauja tiesiogiai arba tam tikru kanalu: žodžiu – telefonu, raštu – internetu, laiškais, raštais ir t. t. Analizuojant sakytinę kalbą svarbu suvokti, kad vienaip pašnekovai bendrauja tiesiogiai, kitaip – netiesiogiai, kai informacijos negali perteikti gestais ar mimika, pvz., telefonu (Rääbis 2007). Tiesioginis arba netiesioginis ryšys tarp pašnekovų lemia įvairių kalbos lygmenų modifikacijas – fonetikos (aiškesnė tartis, aukštesnis tonas), leksikos (konkretesnis žodynas, mažiau žodžių, neturinčių svarios komunikacinės reikšmės), morfologijos (nominatyvinių kalbos dalių gausa), sintaksės (trumpesni pasakymai). Taigi siekiant subalansuoto tekstyno principo nuspręsta, kad nedidelę (kadangi komunikacija telefonu nėra tokia dažna kaip tiesioginė komunikacija) tekstyno dalį turi sudaryti pokalbiai telefonu. Kitokios formos, kaip, pvz., internetinis diskursas arba trumposios žinutės, nėra įtraukiamos, nors kai kurie tyrėjai tai prilygina sakytinei kalbai, tačiau tekstyne yra žodinių pokalbių, vykstančių naudojant programėlę *Skype*.

3.3. Demografiniai rodikliai

Kaupiant tekstyną buvo nuspręsta atsižvelgti į pagrindinius demografinius kriterijus, nuo kurių priklauso kalbos ypatybės: gyvenamąją vietą, lytį, išsilavinimą ir amžių. Kaip minėta anksčiau, jau pradėjus kaupti tekstyną įrašinėta įvairių Lietuvos regionų gyventojų kalba. Tekstyno kūrėjai įrašinėjo pokalbius savo gyvenamajame mieste ir aplinkinėse vietovėse: Kaune, Kėdainiuose, Kaišiadoryse, Vilniuje, Šiauliuose, Klaipėdoje ir vakarinėje Lietuvoje. Vis dėlto reikia pastebėti, kad didžiojoje dalyje įrašytų pokalbių pašnekovai – miesto gyventojai. Ateityje planuojama tekstyną pagausinti pokalbiais, įrašytais kaimo vietovėse, tačiau tai yra sudėtinga, nes tekstyne

siekiama atspindėti bendrinės lietuvių kalbos vartojimą, bet ne tarmių ypatybes (tam būtų reikalingas specializuotas tekstynas, kuriamas pagal specialią metodiką).

Sekant subalansuoto tekstyno principu, pokalbiai, įtraukti į tekstyną, apima skirtingo amžiaus ir išsilavinimo pašnekovų kalbą, kadangi nuo lyties, amžiaus, išsilavinimo priklauso daugelis kalbos ypatybių. Vienaip suaugusieji kalba su suaugusiaisiais, savo ar panašaus amžiais žmonėmis, tačiau kitaip su vaikais, vyresniąja karta. Kalbėdami su vaikais, savo kalbą jie modifikuoja, paprastina ją (Kamandulytė 2005, 2007). Kalbant su vyresniąja karta, būdingos įvairios mandagumo modifikacijos, dažnesnis kai kurių įvardžių vartojimas (Rosinas 1996: 41), pagarbesnis tonas, aiškesnė šneka, tam tikras žodynas. Skiriasi ir pašnekovų, turinčių skirtingą išsilavinimą, kalbos ypatybės – kalbos taisyklingumas, raiškumas, žodingumas, leksinė įvairovė. Kalbėtojų lytis taip pat labai svarbi pragmatinei pokalbio analizei. Taigi didžiausią tekstyno dalį sudaro gausiausios demografinės grupės pagal amžių – vidutinio amžiaus – vyrų ir moterų kalba. Nedidelė tekstyno dalis skirta pokalbiams su vaikais, didesnė – pokalbiams tarp senyvo amžiaus žmonių / su senyvo amžiaus žmonėmis.

3.4. Socialiniai pašnekovų santykiai

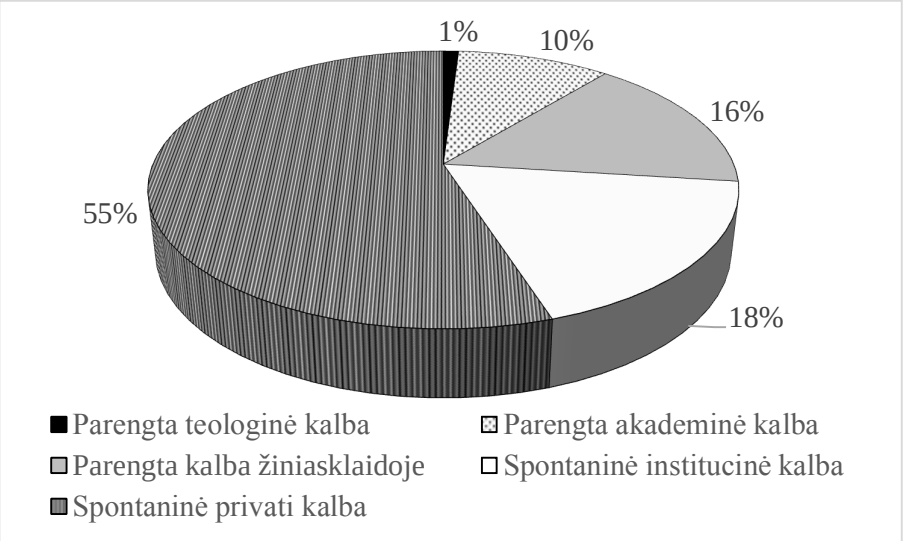
Daugelis kalbos ypatybių yra nulemtos ne tik pokalbio pobūdžio, pašnekovo lyties, amžiaus ar išsilavinimo, bet ir situacinio kalbėtojo socialinio vaidmens. Individui pereinant iš vieno vaidmens į kitą, keičiasi ir individo šneka, pavyzdžiui, įstaigos tarnautojo, direktoriaus vaidmuo daugiausiai yra susijęs su oficialiu-dalykiniu kalbėjimo stiliumi, bendradarbio, pirkėjo vaidmenims būdingesnės įprastos nerūpestingos šnekamosios kalbos priemonės, o draugo socialiniam vaidmeniui būdingas familiarusis buitinis šnekos žanras. Taigi nuspręsta, kad tekstyne kaupiami pokalbiai turėtų apimti socialiniu požiūriu skirtingus lygmenis – pokalbiai turi būti įrašinėjami įvairiose vietose įvairių situacijų metu.

Taigi siekiant subalansuoto tekstyno pricipo, nuspręsta išskirti du svarbiausius socialinius lygmenis – privačiąją erdvę ir viešąją erdvę. Privačiajai erdvei būdingos situacijos, kai kalbėtojų socialiniai santykiai yra artimi, draugiški, familiarūs. Dažniausi privačių pokalbių socialiniai vaidmenys – šeimos nario, draugo. Viešosios erdvės komunikacijai būdingi oficialūs, dalykiniai santykiai, o kalbėtojų socialinė interakcija dažniausiai vyksta įvairiose institucijose – parduotuvėje, kirpykloje, banke, mokykloje, universitete. Dažniausi viešosios erdvės pokalbių dalyvių socialiniai vaidmenys – pardavėjos, pirkėjo, tarnautojo, kliento, dėstytojo, mokytojo, studento ir t. t. Privačiosios erdvės

pokalbiai dažnai vyksta tarp gerai pažįstamų, familiariai bendraujančių asmenų, kalbančių įvairiomis temomis. Viešosios erdvės komunikacija dažna įvairiose institucijose, kalbančiųjų santykiai nėra familiarūs.

3.5. Sakytinės lietuvių kalbos tekstyno proporcijos

Apibendrinant aprašytus tekstyno kaupimo kriterijus, galima išskirti dvi pagrindines *Sakytinės lietuvių kalbos tekstyno* dalis: spontaninė kalba ir parengta kalba. Kaip jau minėta, didžiąją tekstyno dalį sudaro registrų įvairove pasižyminti spontaninė kalba, kurios pokalbius galima suskirstyti į spontaniškus privačius pokalbius (vykstančius artimoje aplinkoje tarp gerai pažįstamų kalbėtojų) ir spontaniškus institucinius pokalbius (laisvus pokalbius aptarnavimo sferoje, darbo aplinkoje ir pan.) (žr. 2 pav.). Mažesnę tekstyno dalį sudaro parengta kalba, daugiausiai apimanti akademinę komunikaciją ir komunikaciją žiniasklaidoje. Kaip minėta, nors šie pokalbiai / kalbos priskiriami viešajam kalbėjimui, į tekstyną stengtasi įtraukti tik tokius pokalbius, kurie nėra skaitomi ar sakomi mintinai (išskyrus kelias žinių laidas).

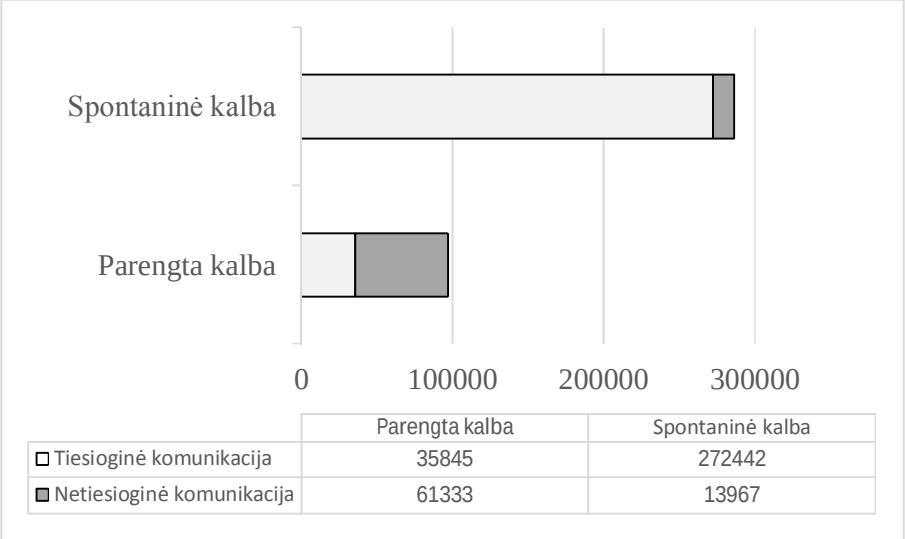


2 pav. Sakytinės lietuvių kalbos tekstyno struktūra pagal kalbos pobūdį (iš viso – 383 587 žodžiai, 2017 m.)

Pagal ryšį tarp kalbėtojų tekstyną sudarančius 265 pokalbius galima skirstyti į tiesioginius pokalbius ir netiesioginius pokalbius. Spontaniniuose pokalbiuose, sukauptuose tekstyne, didžiąją dalį sudaro tiesioginė komunikacija (204 pokalbiai), daug mažesnę – retesnė netiesioginė komunikacija, dažniausiai vykstanti telefonu (29 pokalbiai). Viešojoje kalboje tiesioginės (31 pokalbis) ir netiesioginės komunikacijos pokalbiai (28 pokalbiai) sudaro panašią dalį, nes ir tiesioginei

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

komunikacijai priskiriami akademiniai pokalbiai, ir netiesioginei komunikacijai priskiriami pokalbiai žiniasklaidos srityje (TV, radijas) yra vienodai svarbūs viešosios kalbos ypatybėms atskleisti. Tiesa, nors pokalbių skaičius čia panašus, akademinės (tiesioginės) komunikacijos dalį sudaro mažiau žodžių nei kalbos žiniasklaidoje (netiesioginės) (žr. 3 pav.).



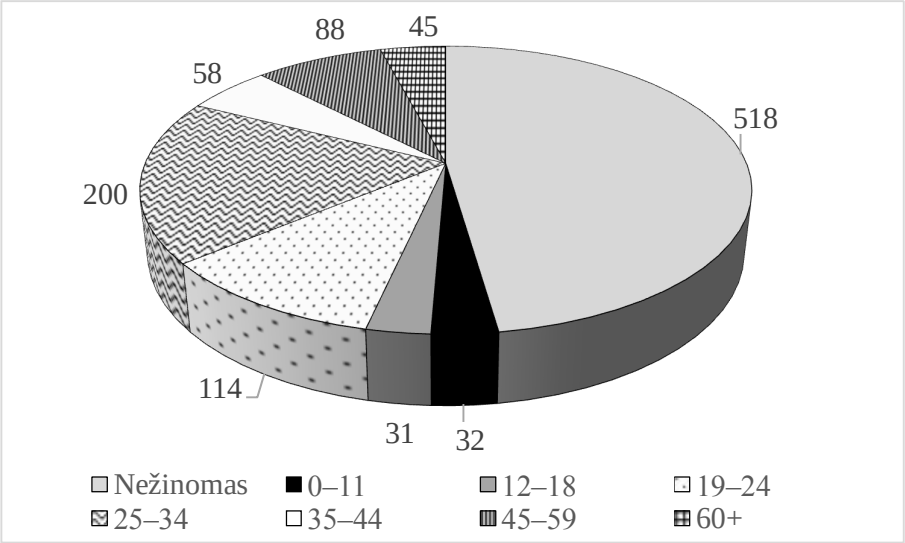
	Parengta kalba	Spontaninė kalba
Tiesioginė komunikacija	35845	272442
Netiesioginė komunikacija	61333	13967

3 pav. Sakytinės lietuvių kalbos tekstyno struktūra pagal ryšį tarp kalbėtojų (2017 m.)

Vienas iš subalansuoto tekstyno kūrimo tikslų – apimti įvairių demografinių grupių informantų kalbą. 2017 m. duomenimis, tekstyne sukaupti pokalbiai apima 1086 informantų kalbą. Tarp visų kalbėtojų vyrišosios lyties informantai sudaro 39,3 % (427 kalbėtojai), moteriškosios lyties – 60,7 % (659 kalbėtojai). Atsižvelgiant į nemažą tekstyno apimtį galima teigti, kad abiejų lyčių duomenų sukaupta pakankamai.

Kaupiant tekstyną buvo atsižvelgta į visuomenėje skaitlingiausias demografines grupes pagal amžių: siekta į tekstyną įtraukti daugiausiai pokalbių, apimančių viduriniąsias amžiaus grupes, mažiau – vaikus ir senyvo amžiaus kalbėtojus (žr. 4 pav.). Deja, ne visuose pokalbiuose buvo įmanoma nustatyti kalbėtojų amžių: dažniausiai ši problema kildavo į tekstyną įtraukiant institucinius pokalbius, pvz., parduotuvėje, banke, teatre ir t. t., kur buvo įrašinėjami įrašų atlikėjui nepažįstami žmonės. Taigi iš 1086 pašnekovų amžiaus nustatyti nebuvo galima 518 pašnekovų.

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt



4 pav. Sakytinės lietuvių kalbos tekstyno struktūra pagal kalbėtojų amžių (2017 m.)

Sakytinės lietuvių kalbos tekstyną sudaro įvairiose vietose atlikti pokalbiai, apimantys skirtingas kalbėjimo situacijas ir įvairius socialinius pašnekovų vaidmenis. Kaip matyti 1 lentelėje, tekstyną sudarantys pokalbiai daugiausiai apima namų aplinką (162 679 žodžiai), nes čia įrašyti pokalbiai pasižymi ilga trukme, temų ir situacijų įvairove. Didelė dalis pokalbių priklauso švietimo sričiai (57 837 žodžiai). Siekiant tekstyne sukurti akademinės komunikacijos dalį daug šios srities pokalbių įrašyta universitetuose, kolegijose, tačiau taip pat yra pokalbių, įrašytų ir mokyklose, vaikų darželiuose. Institucinės komunikacijos daliai priskiriami aptarnavimo srities pokalbiai, vykstantys maitimo įstaigose (restorane, kavinėje ir pan.), prekybos vietose (parduotuvėje, turguje, kioske, knygyne, vaistinėje, teatro / kino / autobusų bilietų kasoje ir pan.), paslaugas teikiančiose įstagose (kirpykloje, grožio salone, siuvykloje, banke, automobilių servise, draudimo įmonėje ir pan.), darbo vietoje ar biure. Dalis tekstyno pokalbių yra įrašyti gatvėje, kieme, bažnyčioje. Nemažą dalį tekstyno sudaro TV, radijo laidų įrašai ir pokalbiai telefonu.

1 lentelė. Sakytinės lietuvių kalbos tekstyno struktūra pagal įrašų atlikimo vietą (2017 m.)

Sritis	**Žodžių skaičius**
Švietimo įstaigos	57 837
Namai / sodyba	162 679
Aptarnavimo sritis (prekyba)	6 815
Paslaugų sritis (maitinimas)	14 455
Paslaugų sritis (gydymas, grožio paslaugos,konsultavimas, gamyba kt.)	36 019
Kiemas, gatvė	7 057
Automobilis / autobusas	2 965
Darbas / biuras	5 729

Bažnyčia	1 255
Kita	13 476
Televizija	42 516
Radijas	18 817
Telefonu, Skype	13 967
Iš viso:	383 587

Apibendrinant *Sakytinės lietuvių kalbos tekstyno* struktūrą, galima teigti, kad pokalbių įvairovė yra pankama ir suteikia galimybę analizuoti įvairius kalbinius reiškinius, nulemtus pokalbio pobūdžio ir struktūros, situacijos, vietos ir kt. ypatybių.

4. Sakytinės lietuvių kalbos tekstyno duomenų kaupimo (įrašymo, transkribavimo ir gramatinio anotavimo) procedūra

4.1. Įrašų rinkimas

Sakytinės lietuvių kalbos tekstyną sudarančių pokalbių įrašai atlikti 2006–2008 m. ir 2016–2017 m. jau minėtų projektų vykdymo laikotarpiu. Už įrašų rinkimą atsakingi mokslininkai buvo supažindinti su įrašų atlikimo taisyklėmis ir reikalavimais. Šie mokslininkai kuravo kelias tyrėjų grupes, rinkusias įrašus įvairiose Lietuvos vietose, jas mokė ir konsultavo. Iš viso dviem minėtais laikotarpiais įrašų rinkime dalyvavo daugiau nei 50 tyrėjų ir jų prižiūrimų asmenų. Visi įrašai buvo atliekami aukštos kokybės diktofonais arba kitais mobiliaisiais įrenginiais. Įrašų trukmė labai įvairi – nuo 1 minutės (dažniausiai tai pokalbiai, vykstantys aptarnavimo srityje) iki 0,5 ar net 1 valandos. Ilgesnių pokalbių įrašai, praėjus valandai, buvo nutraukiami.

Siekiant išlaikyti sakytinės kalbos spontaniškumą ir natūralumą, stengiantis, kad pokalbio dalyviai nejaustų diskomforto, nejaukumo, baimės ir bendrautų natūraliai, nuspręsta apie įrašinėjamą pokalbį pranešti tik įrašo pabaigoje. Taigi atliekant įrašą apie jį žinojo tik pokalbį įrašinėjantis asmuo. Įrašo pabaigoje apie tai informavus kitus pašnekovus, buvo jų klausiama, ar atliktas įrašas galėtų būti naudojamas tekstyno reikmėms. Informantui nesutikus, įrašas buvo ištrinamas. Reikia pastebėti, kad tokių atvejų pasitaikė nedaug.

Kad transkribuotojui būtų aiški įrašo situacija, kontekstas, socialiniai dalyvių vaidmenys, pokalbius įrašinėjantis asmuo, prieš pradėdamas (jei neįmanoma, tada baigdamas) įrašą, turėjo pasakyti įrašo datą, vietą, apibūdinti informantų socialinius vaidmenis, nurodyti amžių, kitą žinomą informaciją. Jei įrašų metu atsirasdavo daugiau pašnekovų, juos įrašinėtojas apibūdindavo įrašo pabaigoje.

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

4.2. Įrašų transkribavimas

Atlikus įrašus jie buvo transkribuojami pagal programos CHILDES (MacWhinney 2000) formato taisykles. Transkripcijos pradžioje pagal numatytą formą pateikiama metalingvistinė informacija apie įrašą ir kalbėtojus, nurodoma įrašo atlikimo ir transkribavimo data, pažymima situacija ir t. t. Remiantis CHILDES programos reikalavimais, kiekvienas pasakymas rašomas atskiroje eilutėje po santrumpos, žyminčios kalbėtoją. Toks pasakymų segmentavimas vėliau leidžia patogiai dirbti su morfologine informacija, pateikiama kiekvieno pasakymo apačioje.

Pagal programos CHILDES reikalavimus transkribuojant įrašus buvo žymimi tam tikri kalbos reiškiniai, padedantys lengviau atlikti automatinį morfologinį kodavimą arba automatinę tekstyno analizę. Pavyzdžiui, siekiant, kad sutrumpėjusios žodžių formos būtų atpažįstamos leksikone (naudojamame vėlesniame etape – gramatiniame anotavime), jos žymėtos taip: *eit(i), mokykloj(e)*. Pasitaikančias išskirtines žodžių formas – pavienius dialektizmus, nesuprantamus žodžius, specifinius vaikų kalbos žodžius ir t. t. – žymėjome ženklu @, o laužtiniuose skliaustuose pateikėme bendrinės kalbos formą, pvz., atruoda@d [: atrodo]; gelai@vz [: gerai].

Transkribuojant įrašus susidurta su keliomis specifinėmis problemomis. Esminė ir pati svarbiausia buvo susijusi su transribavimo vieneto – pasakymo – ribų nustatymu. Įprastai pagrindiniu sintaksiniu rašytinės kalbos vienetu laikomas sakinys arba žodžių junginys, tačiau sutinkama, jog sakytinės kalbos sintaksinės analizės objektu reikėtų laikyti *pasakymą* (Brown, Yule 2001: 19). Nors nedaugelis autorių kalba apie sakinio ir pasakymo skirtumus, akivaizdu, kad šių terminų vartojimas yra susijęs ne vien su skirtingomis raiškos (raštu arba žodžiu) galimybėmis. Lingvistai pastebi, kad sakytinei kalbai būdingi nebaigti sakiniai, pauzės, pakartojimai, pataisymai, minčių šuoliai, pertrūkiai ir pertraukinėjimai (Halliday 1985; Brown, Yule 2001; Liddicoat 2007; Nauckūnaitė 2003), o tai lemia neaiškias pasakymų (sakinių) ribas, kurias sunku nustatyti net ir atsižvelgiant į kontekstą, pvz., kai pašnekovas kalba labai greitai, be stabtelėjimo pasako kelias mintis arba kai pašnekovai pertraukinėja vienas kitą neleisdami pabaigti minties. Tokiais atvejais gana patogus Crystal (2008) pateikiamas apibrėžimas, kuriame pasakymas apibūdinamas kaip kalbėjimo atkarpa, prieš ir po kurios eina pauzė (tyla) arba pasikeičia kalbėtojas. Šiuo apibrėžimu buvo remtasi 2006–2008 m. vykdomų projektų metu, tačiau atliekant tekstynu paremtus tyrimus tapo akivaizdu, kad minėtas *pasakymo* apibrėžimas, nėra visiškai tinkamas atliekant ne pavienių leksemų ar gramatinių kategorijų dažnumo analizę, bet tiriant sakytinės kalbos sintaksines ar pragmatines ypatybes. Pasakymu laikydami atkarpą iki pauzės arba iki kalbėtojų kaitos, dažnai

negalime suprasti pasakymo minties (jei pasakymas nutrūksta ar nutraukiamas jos nepabaigus), nustatyti pasakymo funkcijos, apibūdinti jo struktūros ir t. t. Taigi minėtas apibrėžimas vykdant 2015 m. projektą buvo patikslintas ir praplėstas *pasakymu* laikant baigtinę intonaciją ir sąlygiškai baigtą mintį turinčią kalbėjimo atkarpą. Pagrindiniais pasakymo bruožais nuspręsta laikyti predikatumą, tipinę formaliąją struktūrinę schemą, aiškią komunikacinę funkciją, baigtinę intonaciją. Tokiu atveju, kai pašnekovo kalbėjimo tempas labai greitas, pauzių nedaug, iš kalbėjimo srauto pasakymus galima atskirti remiantis jau minėtais bruožais. Jei vieno pašnekovo mintis nepabaigiama ir kalbėjimas nutraukiamas pauzės ar į jį įsiterpusių kito pašnekvo žodžių, − tokiu atveju nutrauktą kalbėjimo atkarpą laikėme pasakymo dalimi, susijusia su vėliau einančia kalbėjimo atkarpa, ir žymėjome specialiais CHILDES programos simboliais, pvz.:

(1) *1 pašnekovas*: Ir jei jauna panelė eis į klubą <//
2 pašnekovas: Ko jai ten eit?
1 pašnekovas: <//jinai visa blizgės su šitais papuošalais.

Tiesa, reikia pripažinti, kad net ir remiantis minėtais kriterijais, ne visada tiksliai pavyko nustatyti pasakymo ribas, dažnai teko pasikliauti tyrėjų intuicija. Tačiau ši problema aktuali ne tik lietuvių kalbos, bet ir kitų kalbų tekstynų kūrėjams.

Kita su įrašų transkribavimu susijusi problema – dvižodžių kalbos dalių žymėjimas. Nors, formaliai žiūrint, rašytiniame tekste žodžius skiriame tarpeliais ar ženklais, lietuvių kalboje galima rasti daugybę žodžių, turinčių bendrą reikšmę ir vartojamų kartu, t. y. laikytinų viena leksema, pvz., *bet kas, kada nors, iš tikrųjų, iš anksto*. Rimkutė, Jarašiūnaitė ir Homola (2005: 59) savo straipsnyje tokius žodžių junginius vadina morfologinėmis samplaikomis ir teigia, kad anotuojant rašytinės *Dabartinės lietuvių kalbos tekstyną* rastos 622 tokios frazės. Taigi šių samplaikų gramatinio anotavimo problema koduojant tekstynus labai aktuali. Automatiškai anotuojant *Sakytinės lietuvių kalbos tekstyną* naudojamas leksikonas, o pagal jį automatinė gramatinio anotavimo programa tokias dvižodes kalbos dalis (dėl formalaus tarpelio žymėjimo tarp žodžių) pažymi kaip dvi atskiras kalbos dalis, pvz., nors samplaika *bet kas* yra įvardis, *bet* žymima kaip jungtukas, *kas* – kaip įvardis. Pasak kalbininkų (Rimkutė, Jarašiūnaitė, Homola 2005), tokios morfologinės samplaikos turi bendrą reikšmę ir yra nuolat vartojamos kartu, todėl transkribuojant įrašus, nuspręsta jas laikyti vienu leksiniu vienetu ir žymėti kaip vieną kalbos dalį. Remiantis minėtais autoriais, pasirinkti šie morfologinių samplaikų nustatymo kriterijai: morfologinis neskaidomumas, leksinės reikšmės vientisumas, samplaikų sustabarėjimo laipsnis, neapibrėžta leksinė aplinka, vartojimo dažnumas, gramatinės sudėties pastovumas. Žodžių junginiai, atitinkantys minėtus kriterijus, laikyti

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

morfologinėmis samplaikomis ir transkribuojant žymėti kaip vienas leksinis vienetas, pvz., *iš_tikrųjų, be_to, taip_toliau, vis_vien, vis_tiek, be_abejo, iš_esmės*. Reikia atkreipti dėmesį, kad skaičiuojant tekstyną ar jo dalį sudarančių žodžių skaičių, šios samplaikos skaičiuojamos kaip vienas žodis.

4.3. Gramatinis anotavimas

Gramatinis tekstyno žymėjimas – vienas iš svarbiausių anotuoto tekstyno rengimo etapų. *Sakytinės lietuvių kalbos tekstyno* kūrimo metu šiam etapui naudota internete laisvai prieinama programa CHILDES (MacWhinney 2000). Šios programos komanda MOR automatiškai anotuoja žodžius pagal leksikoną – jau minėtą žodžių formų sąrašą su morfologinėmis žymomis. Anotuojant naujas transkripcijas, dalis žodžių, kurių nėra leksikone, lieka neatpažinta. Jos rankiniu būdu įterpiamos į leksikoną, taip leksikonas nuolat pildomas naujais duomenimis. Šiuo metu leksikoną sudaro apie 200 000 morfologiškai aprašytų žodžių formų.

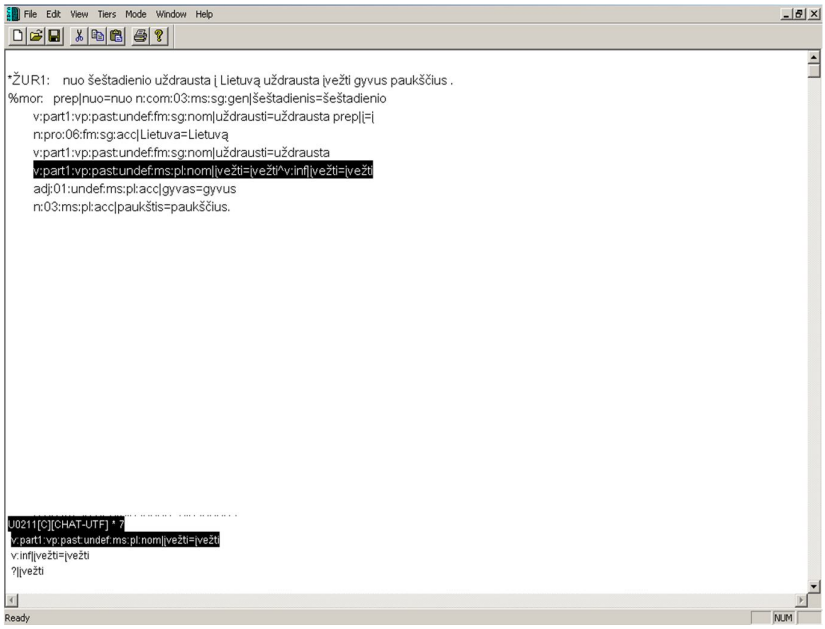
Pildant anotuotų žodžių leksikoną, gramatiškai koduojant transkripcijas, susidurta su įvairių kalbos dalių gramatinių kategorijų arba kitokių ypatybių žymėjimo problemomis. Pavyzdžiui, nemažai problemų kilo žymint sangrąžines ir įvardžiuotines daiktavardžių formas, nes jų apibrėžimai gramatikose nėra aiškūs, o pateikti pavyzdžiai yra prototipiški ir neatspindi šių žodžių įvairovės. Problemų kilo ir anotuojant neigiamuosius žodžius: visi žodžiai su priešdėliu *ne-* pažymėti kaip neigiamieji, tačiau kilo neaiškumų anotuojant tokius žodžius kaip *diskomfortas, antistresas,* nes šie priešdėliai gramatikose neaptariami. Gramatinio anotavimo metu taip pat buvo sudėtinga atskirti kai kuriuos būdvardžius nuo prieveiksmių ar dalyvių (*aišku, gera*), kilo ir kitokių gramatinių kategorijų nustatymo problemų. Reikia paminėti, kad sprendimai, susiję su lietuvių kalbotyroje dar neišspręstais klausimais, neretai buvo nulemti mokslininkų intuicijos ir jų sukauptos sakytinės kalbos tyrimų patirties.

Atlikus automatinį transkribuoto teksto kodavimą, buvo atliekami vienareikšminimo darbai. Vienareikšminimas – tai tinkamos formos parinkimas iš pateiktųjų pagal leksikoną programos lango apačioje, pavyzdžiui, programa, remdamasi leksikonu, pateikia tokią žodžio *įvežti* anotaciją:

v:part1:vp:past:undef:ms:pl:nom|įvežti=įvežti^v:inf|įvežti=įvežti.

Ženklas ^ CHILDES programoje reiškia „arba“, vadinasi, žodis *įvežti* yra arba veiksmažodžio (v) forma dalyvis (part1), arba veiksmažodžio (v) bendratis (inf). Taigi, nustačius žodžio formą, rankiniu

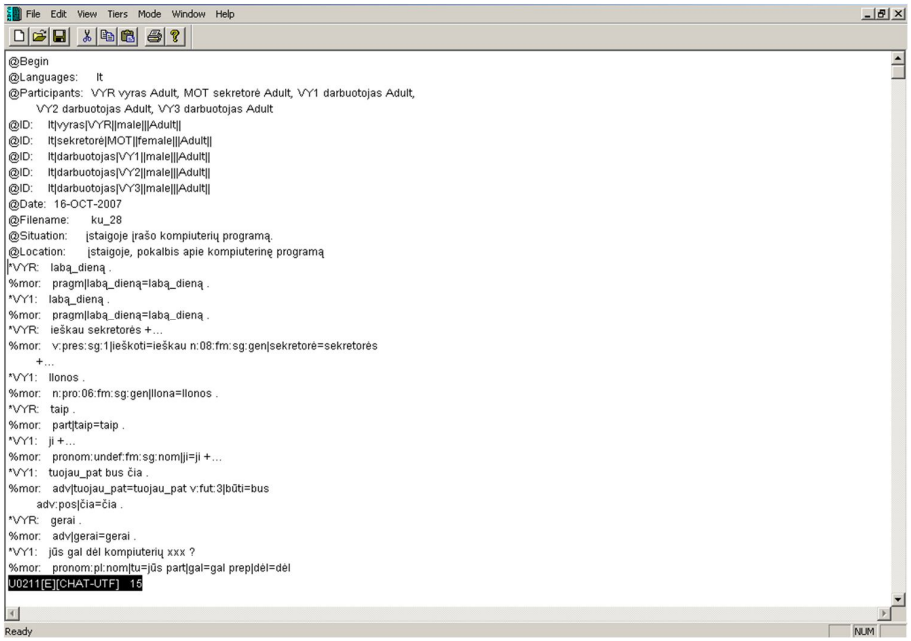
būdu parenkamas vienas iš pateikiamų variantų (automatiškai koduoto teksto ištrauka matyti 5 paveiksle).



5 pav. Automatiškai anotuoto teksto ir galimybės pasirinkti teisingą variantą langas

Vienareikšminimo etape susidurta su kalbos dalių nustatymo problema. Kai kurių kalbos dalių skyrimas nėra sudėtingas, pavyzdžiui, pagal kontekstą lengvai suprasime, ar žodis *kasa* vartojamas kaip daiktavardis, ar kaip veiksmažodis. Tačiau daug sudėtingiau gramatiškai anotuoti daugiareikšmius tarnybinius žodelius, neturinčius apibrėžtos leksinės reikšmės, nekaitomas kabos dalis, pvz., *kad, kiek, vis* ir t. t. Dažnai kalbiniuose darbuose pateikti pavyzdžiai paaiškina tik tą konkretų atvejį, bet nedaug kuo padeda susidūrus su realia vartosena (Rimkutė 2006: 63). Be to, ne visada aiškūs ir DLKŽ ar DLKG pateikiamų kalbos dalių skyrimo kriterijai. Todėl kitas svarbus ir sudėtingas gramatinio anotavimo metodologijos kūrimo etapas – nustatyti daugiareikšmių nekaitomų kalbos dalių (prielinksnių, dalelyčių, jungtukų, jaustukų) skyrimo principus. Remiantis įvairiais tyrinėtojais, nuspręsta, kad: a) norint nustatyti daugiareikšmių tarnybinių kalbos dalių gramatines kategorijas, būtina atsižvelgti į kontekstą ir jų reikšmę kontekste (Rimkutė 2006); b) norint nustatyti daugiareikšmių tarnybinių kalbos dalių gramatines kategorijas, reikia atsižvelgti į jų funkciją, pvz., dalelytė tikslina, modifikuoja įvairius žodžius, perkelia jiems savo leksinį krūvį, jungtukas jungia sakinio elementus, jaustukais žymimi žmogaus jausmai (Valeckienė 1998); c) norint atskirti dalelytes nuo jungtukų, prielinksnių, jaustukų, prieveiksmių, reikia atsižvelgti į jų ryšį su niuansuojamu žodžiu (Paulauskienė 1994: 411). Gramatiškai anotuoto ir vienareikšminto pokalbio ištrauka pateikiama 6 paveiksle.

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt



6 pav. Automatiškai anotuoto ir vienareikšminto pokalbio ištrauka

Atlikus sukauptų pokalbių transkribavimo ir gramatinio anotavimo darbus, galima atlikti automatinę anotuotų pokalbių analizę, tirti ne tik atskirų leksemų ar jų formų, bet ir įvairių gramatinių kategorijų pasiskirstymą, leksemų ir gramatinių kategorijų ryšį bei kitas ypatybes.

5. Sakytinės lietuvių kalbos tekstyno tyrimų perspektyvos

Šiuo metu (2017 m.) interneto vartotojams pateikiamas laisvai prieinamas tekstynas apima 226 174 žodžių formas. Šios internetinės tekstyno versijos vartotojai gali atlikti paiešką pagal žodį ar žodžio formą bei gauti duomenis apie pasirinktos formos dažnumą visame tekstyne arbo jo dalyje, taip pat matyti gramatinę informaciją. 2018 m. vartotojams bus suteikta daugiau galimybių. Naudodamiesi atnaujinta tekstyno versija (daugiau nei 383 000 žodžių) jie galės filtruoti rezultatus pagal įvairias kategorijas (pvz., lytį, amžių, pokalbio vietą, pokalbio pobūdį ir struktūrą), vykdyti išsamesnę paiešką.

Iki šiol, remiantis *Sakytinės lietuvių kalbos tekstyno* duomenimis, yra atlikta nemažai tyrimų, atskleidusių lietuvių kalbotyroje dar netyrinėtus reiškinius. Pavyzdžiui, Kamandulytės-Merfeldienės ir Balčiūnienės (2016b, 2016a, 2016c) atliktuose tyrimuose nustatyta, kad atributiniai ir predikatiniai junginiai sakytinėje kalboje dažnai pasižymi netipine gramatikose neaprašyta žodžių tvarka (atributas neretai vartojamas po daiktavardžio, o predikatas – prieš jį); tiriant pažymimuosius

pasakymus, pastebėta, kad spontaninėje kalboje pasitaiko teoriniuose darbuose neaprašytų atvejų, kai šalutinis pažymimojo pasakymo dėmuo pasakomas prieš antecedentą, o ir spontaninėje šnekamojoje, ir viešojoje kalboje vyrauja nežymėtoji šalutinio dėmens vieta iškart po pagrindinio dėmens antecedento − daiktavardžio arba įvardžio; analizuojant funkcinius sakinių tipus užfiksuota gramatikose neminimų klausiamųjų dalelyčių, pvz., *ko* (reikšme *kodėl*), *ką*, taip pat pastebėta klausiamųjų sakinių su dalelytėmis sakinio pabaigoje gausa (kaip ir mažesnio tekstyno analize paremtame Balčiūnienės ir Simonavičienės tyrime (2009)), užfiksuota daug tikrinamojo klausimo dalelyčių *ane, ką*, apie kurias iki šiol nekalbėta.

Kamandulytės-Merfeldienės (2014) atliktas tyrimas leido nustatyti sakytinėje kalboje vartojamus pertarus, sudaryti jų sąrašą ir aptarti skirtingas vartojimo ypatybes spontaninėje ir viešojoje kalboje. Šio tyrimo metu pastebėtas įdomus reiškinys, kad pertarai yra „užkrečiami“ ir pokalbio metu labai greitai perimami iš kito kalbėtojo. Kiti tyrimai taip pat yra atskleidę įdomių sakytinės kalbos reiškinių. Pavyzdžiui, tekstyno duomenų tyrimas parodė, kad, priešingai nei įprasta manyti ir teigti, sakytinėje kalboje vartojama nedaug naujųjų skolinių, jų įvairovė yra labai nedidelė (Vaicekauskienė, Dabašinskienė, Kamandulytė-Merfeldienė 2014), o fleksijų kaita paradigmose (*sūnams* vietoje *sūnums*) yra palyginti dažnas reiškinys (Kamandulytė-Merfeldienė 2010). Minėti tyrimai yra atskleidę ir spontaninės bei viešosios kalbos skirtumus: akivaizdu, kad viešoji kalba yra tarpinis variantas tarp sakytinės ir rašytinės kalbos. Ateityje būtų įdomu paanalizuoti skirtingų spontaninės kalbos ir viešos kalbos atmainų (registrų) ypatumus. Tikėtina, kad spontaninė kalba taip pat nevienalytė, o jos ypatybės priklauso nuo pokalbio pobūdžio, situacijos ir kitų veiksnių. Tikimasi, kad 2018 m. vartotojams suteikus daugiau tekstyno duomenų analizės galimybių internete, sakytinės kalbos tyrimų padaugės ir jie apims įvairias leksikos ir gramatikos sritis.

Duomenų šaltinis

Sakytinės lietuvių kalbos tekstynas http://donelaitis.vdu.lt/sakytines-kalbos-tekstynas/

Literatūros sąrašas

Balčiūnienė, I. 2005a. Parodomųjų įvardžių įsisavinimas. *Lituanistica* 64, 45–54.

Balčiūnienė, I. 2005b. Asmeninių įvardžių reikšmės ir funkcijų įsisavinimas. *Kalbotyra* 54 (1), 41–50.

Balčiūnienė, I., L. Kamandulytė. 2010. Dabartinės sakytinės lietuvių kalbos pasakymų ilgis ir struktūra. *Lietuvių kalba* 10. http://www.lietuviukalba.lt/index.php/lietuviu-kalba/article/view/181/147 (žiūrėta 2017–07–03).

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

Balčiūnienė, I., L. Simonavičienė. 2009. Kiekybinis klausiamųjų šnekamosios lietuvių kalbos pasakymų tyrimas. *Lietuvių kalba* 3. http://www.lietuviukalba.lt/index.php/lietuviu-kalba/article/view/17 (žiūrėta 2017–07–04).

Brown, G., G. Yule. 2001. *Discourse analysis.* Cambridge: Cambridge University Press.

Crystal, D. 2008. *A Dictionary of Linguistics and Phonetics.* Oxford: Blackwell Publishing.

Dabašinskienė, I. 2008a. Sakytinė lietuvių kalba: sociolingvistiniai ir psicholingvistiniai tyrimai. *Habilitacinis darbas*. Kaunas: Vytauto Didžiojo universitetas.

Dabašinskienė, I. 2008b. Trumpinimas ir dažnumo poveikis šnekamojoje kalboje. *Darbai ir dienos* 50, 109–117.

Dabašinskienė, I. 2009. Šnekamosios lietuvių kalbos morfologinės ypatybės. *Acta Linguistica Lithuanic*a 60, 1–15.

Dabašinskienė, I., L. Kamandulytė. 2009. Corpora of spoken Lithuanian. *Estonian papers in applied linguistics* 5, 67–77. http://donelaitis.vdu.lt/publications/Dabasinskiene_2009.pdf (žiūrėta 2017–07–03).

Dabašinskienė, I., L. Kamandulytė-Merfeldienė. 2017. The early production of compounds in Lithuanian. *Nominal compound acquisition.* W. U. Dressler, N. Ketrez, M. Kilani-Schoch (red.). 145–163. Amsterdam: John Benjamins.

Fries, C. C. 1952. *The Structure of English*. New York: Harcourt Brace.

Halliday, M. A. K. 1985. *Spoken and Written Language*. Oxford: Oxford University Press.

Kamandulytė, L. 2005. Vaikiškosios kalbos registras. *Gimtoji kalba* 7, 12–16.

Kamandulytė, L. 2007. Morphological modifications in Lithuanian child-directed speech. *Estonian Papers in Applied Linguistics* 3, 155–165. http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/ERYa3.10 (žiūrėta 2017–07–03).

Kamandulytė, L. 2009. Lietuvių kalbos būdvardžio įsisavinimas: leksinės ir morfosintaksinės ypatybės. *Daktaro disertacija*. Kaunas: VDU.

Kamandulytė, L., I. Savickienė. 2007. The Corpus of Spoken Lithuanian. *Human language technologies. The third Baltic conference proceedings*, 127–133.

Kamandulytė, L., M. Tuškevičiūtė. 2008. Būdvardžio vartojimo skirtumai sakytinės kalbos registruose. *Darbai ir dienos* 50, 91–109.

Kamandulytė-Merfeldienė, L. 2010. Daiktavardžio paradigmų produktyvumas: skolinių morfologinio įforminimo ir fleksijų varijavimo analizė. *Lietuvių kalba* 4. http://www.lietuviukalba.lt/index.php/lietuviu-kalba/article/view/23 (žiūrėta 2017–07–03).

Kamandulytė, L. 2012. Morphosyntactic features of Lithuanian adjective acquisition. *Journal of Baltic studies* 43 (2), 239–250.

Kamandulytė-Merfeldienė, L. 2014. Pertarų dažnumas ir įvairovė sakytinėje lietuvių kalboje. *Bendrinė kalba* 87. http://www.bendrinekalba.lt/Pdf/Kamandulyte_BK_87_Straipsnis.pdf (žiūrėta 2017–07–03).

Kamandulytė-Merfeldienė, L., I. Balčiūnienė. 2016a. Apie pažymimuosius pasakymus sakytinėje lietuvių kalboje. *Taikomoji kalbotyra* 8, 55–71.

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

https://taikomojikalbotyra.lt/ojs/index.php/taikomoji-kalbotyra/article/view/78 (žiūrėta 2017–07–03).

Kamandulytė-Merfeldienė, L., I. Balčiūnienė. 2016b. Atributinių ir predikatinių junginių su būdvardžiais dažnumas ir struktūra sakytinėje kalboje. *Lituanistica* 62 (2), 127–137.

Kamandulytė-Merfeldienė, L., I. Balčiūnienė. 2016c. Funkciniai pasakymų tipai sakytinėje lietuvių kalboje. *Tarp eilučių: lingvistikos, literatūrologijos, medijų erdvė: mokslinių straipsnių rinkinys*,11–29. https://eltalpykla.vdu.lt/handle/1/1389;jsessionid=4B5DE1816BBDC81FE226BC1DB4173201. (žiūrėta 2017–07–03).

Kamandulytė-Merfeldienė, L., P. Godliauskas. 2014. Creating and working with Corpus of Spoken Lithuanian. *Human language technologies - the Baltic perspective: proceedings of the 6th international conference*. A. Utka, G. Grigonytė, J. Kapočiūtė-Dzikienė, J. Vaičenonienė (Eds.). 179–183. http://ebooks.iospress.nl/volumearticle/38024 (žiūrėta 2017–07–03).

Liddicoat, A. J. 2007. *An Introduction to Conversation Analysis*. London: Continuum.

McCarthy, M. J., A. O'Keeffe. 2008. Corpora and the study of spoken language. *Corpus Linguistics. An International Handbook*. A. Lüdeling, M. Kytö (Eds.). Vol. 2. Berlin: Mouton de Gruyter, 1008–1024.

MacWhinney, B. 2000. *The CHILDES Project. Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Marcinkevičienė, R. 2000. Tekstynų lingvistika: teorija ir praktika. *Darbai ir dienos* 24, 7–64.

Nauckūnaitė, Z. 2003. Loginiai ir lingvistiniai sakytinės ir rašytinės raiškos skirtumai. *Žmogus ir žodis: didaktinė lingvistika* 5, 78–83.

O'Keeffe, A., F. Farr. 2003. Using language corpora in language teacher education: pedagogic, linguistic and cultural insights. *TESOL Quarterly* 37 (3), 389–418.

Paulauskienė, A. 1994. *Lietuvių kalbos morfologija*. Vilnius: Mokslo ir enciklopedijų leidykla.

Rääbis, A. 2007. Caller identification in insitutional telephone conversation. *Estonian Papers in Applied Linguistics* 3, 269–285.

Rimkutė, E., G. Jarašiūnaitė, P. Homola. 2005. Morfologinių samplaikų atpažinimas ir klasifikavimas. *Lituanistica* 62 (2), 58–75.

Rimkutė E. 2006. Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne. Daktaro disertacija. Kaunas: Vytauto Didžiojo universitetas.

Rosinas, A. 1996. *Lietuvių bendrinės kalbos įvardžiai.* Vilnius: Mokslo ir enciklopedijų leidykla.

Savickienė, I. 1998. The acquisition of diminutives in Lithuanian. *Antwerp papers in linguistics 95: Studies in the acquisition of number and diminutive marking*, 129–153.

Savickienė, I. 1999. Lietuvio vaiko daiktavardžio morfologija. *Daktaro disertacija.* Kaunas: VDU.

Savickienė, I. 2000. Linksniai šnekamojoje kalboje. *Darbai ir dienos* 24, 89–98.

Savickienė, I. 2001. Linksniavimo paradigmų formavimasis vaiko kalboje. *Lituanistica* 3 (47), 58–68.

Savickienė, I. 2002. The acquisition of gender. *Kalbotyra* 51 (3), 133–141.

Savickienė, I. 2003. *The Acquisition of Lithuanian Noun Morphology*. Wien: Österreichischen Akademie der Wissenschaften.

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

Savickienė, I. 2005. Linksnių vartojimo dažnumas ir daiktavardžio reikšmė. *Acta Lingusitica Lituanica* 50, 79–98.

Savickienė, I. 2006a. Linksnio kategorijos įsisavinimas. *Kalbotyra* 56 (3), 122–129.

Savickienė, I. 2006b. Komunikacinė pragmatika ir kalbėjimo situacijos tikslas. *Kalbos kultūra* 79, 256–263.

Savickienė, I. 2007. Form and meaning of diminutives in Lithuanian child language. *The Acquisition of Diminutives: a cross-linguistic perspective*. I. Savickienė, W. U. Dressler (red.). Amsterdam: John Benjamins, 13–41.

Utka, A. 2005. Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės. *Lituanistica* 61 (1), 48–55.

Vaicekauskienė, L., I. Dabašinskienė, L. Kamandulytė-Merfeldienė. 2014. Naujųjų skolinių kaitybinio ir darybinio adaptavimo modelių produktyvumas. *Taikomoji kalbotyra* 3, 1–22. https://taikomojikalbotyra.lt/ojs/index.php/taikomoji-kalbotyra/article/view/24 (žiūrėta 2017–07–03).

Valeckienė, A. 1998. *Funkcinė lietuvių kalbos gramatika*. Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Sakytinės lietuvių kalbos tekstynas – natūralios vartosenos tyrimų šaltinis

Laura Kamandulytė-Merfeldienė

Santrauka

Straipsnyje aprašomas *Sakytinės lietuvių kalbos tekstynas,* jo struktūra, kūrimo etapai (įrašų kaupimas, transkribavimas, gramatinis transkripcijų anotavimas), duomenų kaupimo ir skaitmeninimo metodika, taip pat aptariamos tekstyno panaudojimo natūralios vartosenos tyrimuose galimybės, trumpai pristatomi jau atlikti tekstyno duomenimis paremti tyrimai.

Šiuo metu (2017 m.) interneto vartotojams pateikiamas laisvai prieinamas tekstynas apima 226 174 žodžių formas. Šios internetinės tekstyno versijos vartotojai gali atlikti paiešką pagal žodį ar žodžio formą bei gauti duomenis apie pasirinktos formos dažnumą visame tekstyne arbo jo dalyje, taip pat matyti gramatinę informaciją.

2016–2017 m. vykdant LMT finansuojamą projektą „Šiuolaikinė sakytinė lietuvių kalba: leksikos ir gramatikos tyrimas tekstynų lingvistikos metodu“ (LIP-085/2016) pagal Valstybinę lituanistinių tyrimų ir sklaidos 2016–2024 metų programą, Sakytinės lietuvių kalbos tekstynas buvo pagausintas naujais duomenimis. Projekto metu taip pat kuriama nauja internetinė prieiga, suteiksianti daugiau galimybių vartotojams. Atnaujintą tekstyną sudaro 256 pokalbiai (383 587 žodžiai), apimantys 1086 kalbėtojus (659 moterys, 427 vyrai), kurių amžius nuo 3 metų iki 81 metų. Plečiant Sakytinės

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

lietuvių kalbos tekstyną, didelis dėmesys buvo skirtas tekstyno kompozicijai, t. y. tekstyno proporcijų parinkimui. Siekiant tekstyno universalumo ir tinkamumo įvairiapusei analizei, buvo laikomasi subalansuoto tekstyno principo, todėl kaupiant pokalbius atsižvelgta į kelis kriterijus: sakytinės kalbos pobūdį (privati *vs* vieša kalba) ir struktūrą (dialogai *vs* polilogai), ryšį tarp pokalbyje dalyvaujančių pašnekovų (tiesioginis *vs* netiesioginis (pvz., kalba telefonu), demografinius rodiklius, socialinius pašnekovų santykius. Taigi jau 2018 m. naudodamiesi atnaujinta tekstyno versija vartotojai galės filtruoti rezultatus pagal įvairias kategorijas (pvz., lytį, amžių, pokalbio vietą, pokalbio pobūdį ir struktūrą), vykdyti išsamesnę paiešką. Tikimasi, kad 2018 m. vartotojams suteikus daugiau tekstyno duomenų analizės galimybių internete, sakytinės kalbos tyrimų padaugės ir jie apims įvairias leksikos ir gramatikos sritis.

The Corpus of Spoken Lithuanian as a Research Source of Natural Usage

Laura Kamandulytė-Merfeldienė

Summary

The article describes the *Corpus of Spoken Lithuanian*, its structure, compilation stages (collection of the recordings, transcription, and grammatical annotation), and the methodology of data collection and digitalization; in addition, it discusses the possibilities of corpus application in the research of natural language usage and the research, which has already been carried out, using the corpus data.

At present (2017), the corpus, which is freely accessible for internet users, contains 226,174 word forms. The users of the online corpus version can perform search of a word or a word form and obtain data on the frequency of the form in the whole corpus or its part as well as see grammatical information about it.

In 2016-2017, the *Corpus of Spoken Lithuanian* was supplemented by new data resulting from the implementation of the project "Contemporary Spoken Lithuanian: A Corpus-based Analysis of Grammar and Lexis" (LIP-085/2016) financed by the Research Council of Lithuania under the programme of the State Lithuanian Studies and Dissemination Programme for 2016–2024. The project will also create a new internet access, which will provide more possibilities for the users. The updated corpus consists of 256 conversations (383,587 words) produced by 1,086 speakers (659 females and 427 males), whose age ranges from 3 to 81 years. When developing the *Corpus of*

Kamandulytė-Merfeldienė, L. 2017. Sakytinės lietuvių kalbos tekstynas. *Taikomoji kalbotyra* 9: 176–198, www.taikomojikalbotyra.lt

Spoken Lithuanian, much attention was paid to its composition, i.e. the proportions of the corpus. In order to improve the universality and suitability of the corpus for a more varied analysis, the principle of a balanced corpus was maintained; therefore, several criteria were taken into consideration when collecting the data: the nature of spoken language (private vs public speech) and structure (dialogues vs polilogues), different communication situations (direct vs indirect (e.g. a telephone conversation), demographic indicators, and social relations among the interlocutors. Therefore, in 2018, users of the updated version of the corpus will be able to filter results according to different categories, such as gender, age, place and structure of the conversation, and perform a more detailed search. It is expected that when the users are provided with more possibilities to analyse corpus data on the internet, the amount of spoken language research will increase comprising different areas of lexis and grammar.

Keywords: corpus; spoken language; spontanous speech; morphological annotation; Lithuanian.

Įteikta 2017 m. rugsėjo mėn.

Paskelbta 2017 m. gruodžio mėn.