LITUANISTICA. 2006. T. 66. Nr. 2. P. 34–55

DABARTINĖS LIETUVIŲ KALBOS GRAMATINIŲ FORMŲ VARTOSENA MORFOLOGIŠKAI ANOTUOTAME TEKSTYNE

Erika Rimkutė

Vytauto Didžiojo universitetas, K. Donelaičio g. 58, LT-44244 Kaunas, Lietuva

1. ĮVADAS

Šiame straipsnyje pristatomas Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre parengtas morfologiškai anotuotas tekstynas (toliau žymima MAT), apžvelgiami kiti anotuoti lietuvių kalbos tekstynai, pateikiami ir palyginami jų duomenys.

Galima paminėti du morfologiškai anotuotus lietuvių kalbos tekstynus. Be VDU sudaryto MAT-o (apie jį plačiau rašoma kitame skyriuje), paminėtina pirmuoju anotuotu lietuvių kalbos tekstynu laikytina L. Grumadienės ir V. Žilinskienės sukurta elektroninė duomenų bazė, naudota dažniniams žodynams rengti. Ši duomenų bazė yra prieinama tik jos sudarytojams, o MAT-ą ateityje bus galima rasti internete.

„Pirmąjį lingvostatistinį leksikos ir morfologijos darbą Lietuvoje, naudodamasi ESM, parengė Vida Žilinskienė (1990). Jos *Lietuvių kalbos dažninis žodynas* remiasi vieno iš funkcinių stilių – publicistikos – duomenimis, jame pateikiama 300 000 žodžių pavartojimo analizė (t. y. tiek tų pačių leksemų, tiek ir skirtingų, o tokių būta 18 776 [...])“ (Grumadienė, 2002, p. 21). Visi ištisiniai tekstai, susidedantys iš atkarpų (imčių) po 1000 žodžių pavartojimų, buvo morfologiškai išanalizuoti, sukirčiuoti (Grumadienė, 2002, p. 21).

L. Grumadienės ir V. Žilinskienės *Dabartinės rašomosios lietuvių kalbos dažninio žodyno* elektroninis tekstynas sudarytas iš 1,2 mln. žodžių pavartojimų. Jis sudarytas iš originalių, ne verstinių tekstų, priskirtinų keturiems funkciniams kalbos stiliams: publicistiniam, kanceliariniam, beletristiniam ir moksliniam. Buvo nuspręsta, kad minėto pobūdžio tekstai tinkamai reprezentuoja dabartinę rašomąją lietuvių kalbą. Iš jų atsitiktine tvarka būdavo pasirenkama po vieną 1000 žodžių pavartojimų apimties atkarpą, kuri būdavo surenkama kompiuterio klaviatūra, o vėliau anotuojama V. Zinkevičiaus parengta MAN (Morfologinio analizavimo ir normalizavimo) programa. Iš taip anotuotų tekstų buvo parengti trys žodynai: *Dabartinės rašomosios lietuvių kalbos dažninis žodynas* (mažėjančio dažnio tvarka), išleistas 1997, *Dabartinės rašomosios lietuvių kalbos dažninis žodynas* (abėcėlės tvarka), išleistas 1998, ir dar nepasirodęs žodžių formų žodynas, turintis maždaug 150 000 žodžių formų, atspindintis skirtingų ir tų pačių žodžių vartojimą įvairiomis gramatinėmis formomis (Grumadienė, 2002, p. 25–28). Pagal šiuos duomenis parengtas elektroni-

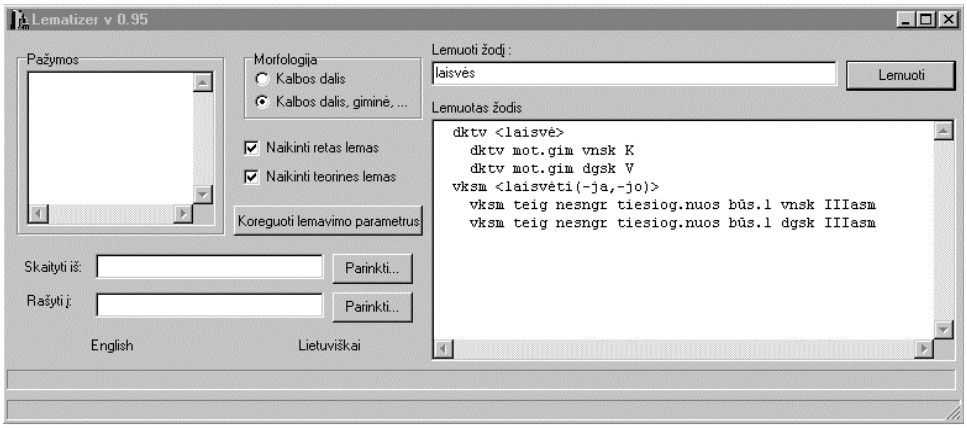
nis dažninis žodynas *Bendriniai XX a. spaudos žodžiai* (Mauricaitė et al., 2004).

Remdamasi minėta duomenų baze V. Žilinskienė yra išsamiai išnagrinėjusi kalbos dalių, gramatinių kategorijų vartoseną skirtinguose lietuvių kalbos stiliuose (žr. Žilinskienė 2001; 2002a; 2002b; 2003; 2005). Šiame straipsnyje duomenys apie lietuvių kalbos gramatines kategorijas pateikti iš 1 mln. žodžių MAT-o. Siekta parodyti bendrą dabartinės rašytinės lietuvių kalbos vaizdą, todėl nagrinėtas visas MAT, o ne atskirų stilių tekstai. Kita vertus, netgi V. Žilinskienė, jau daugelį metų tirianti gramatinių formų vartojimą skirtinguose stiliuose, pastebi, kad skirtumai yra nedideli. Labiausiai skiriasi meninio stiliaus tekstai nuo nemeninių (Žilinskienė, 2003; 2005). Žinoma, ateityje bus galima panagrinėti, kaip pasiskirsčiusios gramatinės kategorijos MAT-o skirtingų stilių tekstuose, kokie žodžiai yra dažniausi, kokią įtaką jie gali turėti tam tikrų gramatinių kategorijų vartosenai.

2. VDU KOMPIUTERINĖS LINGVISTIKOS CENTRE PARENGTAS MORFOLOGIŠKAI ANOTUOTAS TEKSTYNAS

MAT-o rengimas prasidėjo apie 2000 m. ir buvo baigtas 2005 m. pradžioje, taigi visas tvarkymas truko apie penkerius metus (2000–2003 m. MAT-o regimą rėmė Valstybinė lietuvių kalbos komisija). Tvarkymo procesas buvo gana ilgas dėl daugiareikšmių lemų bei morfologinių pažymų, taip pat dėl to, kad ne visiems žodžiams automatiškai nurodomos lemos ir morfologinės pažymos – reikėjo tvarkyti rankomis, tobulinti automatinės morfologinės analizės programą.

MAT sudarytas pusiau automatiškai: naudota Vytauto Zinkevičiaus sukurta kompiuterinė programa *Lemuoklis*, pateikianti lemas ir morfologines pažymas (plačiau apie pačią programą žr. Zinkevičius, 2000, p. 245; taip pat žr. 1 pav.).

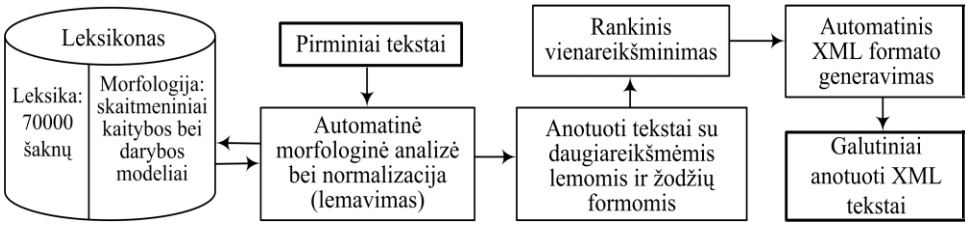


1 pav. Automatinės morfologinės analizės programos *Lemuoklis* langas

Tekstyno žymėjimą galima suskirstyti į tris etapus: pirmiausiai grynas tekstas lemuojamas – pateikiama tekstyne pavartoto žodžio antraštinė forma, t. y. lema (pvz., daiktavardis *namas*, prieveiksmis *namo*, būdvardis *tikslus*, veiksmažodis *daryti* ir pan.). Antrojo etapo metu gali būti pateiktos žodžio formos morfologinės pažymos (pvz.: *namo* – vyriškosios giminės daiktavardžio vienaskaitos kilmininkas; *tikslų* –

nelyginamojo laipsnio neįvardžiuotinio vyriškosios giminės būdvardžio vienaskaitos galininkas (lema *tikslus*) ir vyriškosios giminės daiktavardžio daugiskaitos kilmininkas (lema *tikslas*)). Toliau turėtų būti nustatomi kelis antraštinius pavidalus turintys žodžiai, t. y. vienareikšminama, nes, pavyzdžiui, forma *laisvės* gali būti sulemuota ir kaip *laisvė,* ir kaip *laisvėti*. Tam tikslui reikalinga speciali programa, kurios pagrindinė funkcija būtų dviprasmybių panaikinimas (angliškas terminas – *ambiguity resolution*) (Zinkevičius, 2000, p. 245; Marcinkevičienė, 2000, p. 22).

Visas MAT-o sudarymo ir tvarkymo procesas pavaizduotas 2 paveiksle. Kaip jau minėta, lemas ir morfologines pažymas kiekvienam žodžiui ar jo formai pateikdavo morfologinis analizatorius *Lemuoklis*. Ši programa sudaryta iš leksikono, kuris susideda maždaug iš 70 000 šaknų ir skaitmeninių kaitybos bei darybos modelių. Tai reiškia, kad analizuojant konkrečią žodžių formą nustatoma, kokia šaknis ir koks kaitybos ar darybos modelis, t. y. kiekvienas žodis išskaidomas į šaknį ir kaitybinius ar darybinius afiksus, o ne analizuojamas kaip atskiras vienetas. Tada pateikiamas rezultatas, t. y. lema ir atitinkama morfologinė informacija.



2 pav. Morfologiškai anotuoto tekstyno sudarymo ir tvarkymo etapai

Kitas etapas – rankinis daugiareikšmių lemų ir žodžių formų vienareikšminimas. Vėliau galutinai sutvarkyti tekstai generuojami į XML formatą ir dar kartą peržiūrimi, nes perkeliant anotuotą tekstą į šį formatą ne visos pažymos tiksliai pateikiamos. Taigi galutinį MAT pavidalą sudaro XML rinkmenos, kuriose yra 1 012 673 žodžiai (plačiau žr. Zinkevičius et al., 2005).

MAT-o formatas ilgainiui šiek tiek keitėsi. Pirmos lentelės kairėje pusėje pateikiamas pirminis MAT-o formatas, viduryje – anotuotas tekstas XML formatu, perdarytas iš pirminio formato, kad būtų lengviau analizuoti tekstų pažymas, o dešinėje pusėje pateikta ta pati, bet dar lingvisto neperžiūrėta ir nesutvarkyta teksto atkarpa. Kaip matyti iš 1 lentelės, pirmiausia pateikiama tekste pavartota konkreti forma, pvz., *word="dalykai"*; prie kiekvieno anotuoto tekstyno žodžio pateikiama lema, pvz., *lemma="dalykas"*, ir informacija apie kalbos dalis bei atitinkamas morfologines kategorijas, pvz., *type="dktv vyr.gim dgsk V"*[1].

2.1. Morfologiškai anotuotame tekstyne naudojamos pažymos

MAT-e naudojamos dvejopo tipo pažymos: vienos yra kalbinės (lemos, kalbos dalys, gramatinių kategorijų pažymos), kitos – nekalbinės: tai informacija apie autorius, pa-

[1] Straipsnyje vartojami ne lietuvių kalbotyroje paplitę sutrumpinimai, o tokie, kokius pateikia automatinės morfologinės analizės programa, anotuodama tekstus, dar žr. 2 ir 3 lenteles.

vadinimus, pastraipos pradžią ir pabaigą, užsienio kalbų intarpai, specialios pažymos, skirtos skyrybos ženklams, skaičiais parašytiems skaitvardžiams, pvz.: <*space*> reiškia tarpą tarp žodžių; <*p*> – naujos pastraipos pradžią; <*sep"."*> žymi tašką; <*foreign lang=en*> ... <*/foreign*> reiškia, kad tekste pasitaikė kažkoks anglų kalbos žodis ar ilgesnis teksto intarpas; <*number="3"*> žymi, kad tekste pavartotas skaitvardis *trys*, užrašytas skaičiumi.

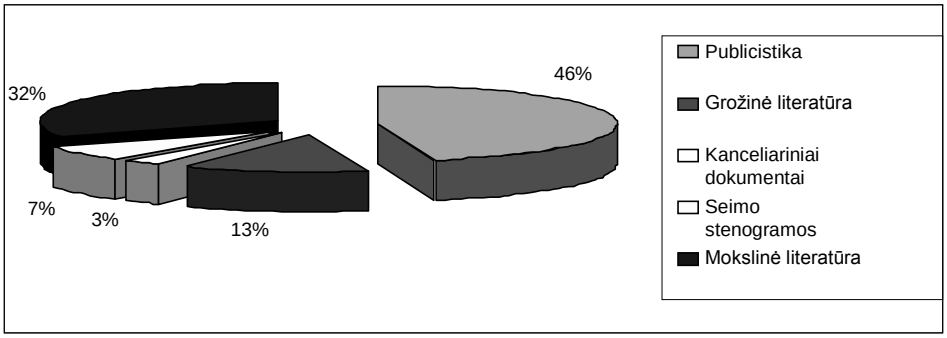
Iš 2 lentelės matyti, kad rengiant MAT-ą naudotos ne tik tradicinės kalbos dalių pažymos. Analizuojant anotuotus tekstus pastebėta, kad reikia daugiau pažymų, kad kai kurios kalbos dalys turi būti skaidomos smulkiau, pvz., skirta atskira kategorija *tikrinis daiktavardis 2*. Šiai kategorijai priskirti nekaitomi tikriniai daiktavardžiai, pvz., *Don, Van, San* ir pan. Taip pat naudota *idiomos* pažyma – tai morfologinės samplaikos, kurių negalima priskirti kuriai nors vienai kalbos daliai, pvz., *be to, ir t. t., t. y., be kita ko* (plačiau apie tai žr. Rimkutė et al., 2005).

Tvarkant MAT-ą užteko tradicinių gramatinių kategorijų (3 lentelė), išskyrus linksnio kategoriją: buvo įsivestas papildomas linksnis – iliatyvas, kuris nėra įtrauktas į *Dabartinės lietuvių kalbos gramatikoje* pateiktą linksnių sistemą kaip nykstantis ar beveik išnykęs linksnis, bet gana dažnai vartojamas *Dabartinės lietuvių kalbos tekstyne* (dėl požiūrio į iliatyvą žr. Rimkutė, 2004). Viename tekste buvo rasta aliatyvo forma (*tavęsp*), todėl į linksnių sistemą buvo įtrauktas ir šis linksnis.

2.2. Morfologiškai anotuoto tekstyno sudėtis

MAT-e siekta parinkti tekstus, kurie apimtų įvairius žanrus, suteiktų daugiau ir įvairiapusiškesnės informacijos apie morfologines formas, jų vartoseną ir dažnumą.

Didžiąją MAT-o dalį sudaro publicistikos (465 519 žodžių (46 proc.)) ir įvairiūs mokslinės literatūros tekstai (323 251 žodis (31,9 proc.)). Ji būtina, kad atsirastų kuo daugiau skirtingų žodžių ir jų formų, kad išryškėtų kuo daugiau rašytinės kalbos ypatybių. MAT-e taip pat įdėta grožinės literatūros (125 710 žodžių (12,5 proc.)), administracinio stiliaus tekstų (28 599 žodžiai (2,8 proc.)). Stengtasi apimti kuo įvairiausias kalbos atmainas, todėl į MAT-ą įdėta netgi Lietuvos Respublikos Seimo stenogramų (69 594 žodžiai (6,8 proc.)), kurios artimiausios šnekamajai kalbai (žr. 3 pav.).



3 pav. Morfologiškai anotuoto tekstyno sudėtis

2.3. Morfologiškai anotuotame tekstyne išryškėjęs morfologinis daugiareikšmiškumas

Automatinę morfologinę analizę atliekanti programa *Lemuoklis* padėjo pastebėti dažną

morfologinį daugiareikšmiškumą (toliau žymima MD, plačiau apie tai žr. Rimkutė, 2003). Minėtos programos anotuoti tekstai yra išsamus MD-o analizės objektas. Kai kurie MD-o atvejai atsirado dėl *Lemuoklio* specifikos. Automatinė morfologinė lietuvių kalbos analizės programa neatsižvelgia į kontekstą. *Lemuoklyje* taip pat neįdiegta informacija apie semantiką. Žodžių reikšmės nustatomos ne naudojantis kokiais nors žodžių formų sąrašais su nurodytais tų formų morfologiniais apibūdinimais, o turint lietuviškų žodžių šaknų sąrašą. Prie kiekvienos šaknies nurodomi atitinkami skaitmeniniai kaitybos ir darybos modeliai. V. Zinkevičiaus sukurtoje programoje taip pat nėra informacijos apie žodžių ar žodžių formų dažnines charakteristikas. Iš *Lemuoklio* apdorotų tekstų išryškėja MD dar ir dėl tos priežasties, kad analizuojamos rašytinės, be kirčio ženklų žodžių formos, pvz.: *vienas gėlės žiedas* vs. *gėlės buvo apvytusios*.

Dėl to, kad neanalizuojamas kontekstas, kad nagrinėjamos rašytinės nekirčiuotos formos, atsiranda tokių daugiareikšmiškumo atvejų, kurie atrodo visiškai nerealūs, pavyzdžiui, daiktavardžiai *padarytis, kokis*. Pirmasis žodis generuotas kaip mažybinis daiktavardis, antrasis yra nebevartojamas, bet *Dabartinės lietuvių kalbos žodyne*[2] pateiktas žodis.

Lemuoklis automatiškai iš visų daiktavardžių generuoja mažybines jų formas. Tai padaryta dėl paprastos priežasties: DLKŽ-e pateikta labai nedaug mažybinių formų ir tik tokios, kurios nuo pamatinio žodžio nutolusios semantiškai, pvz., *stiklas – stikliukas*. Kuriant morfologinės analizės programą *Lemuoklis,* pagrindinės leksikos žinios buvo imtos iš DLKŽ-o ir TŽŽ-o, bet juose dažniausiai pateikti tik pamatiniai žodžiai, pvz., nurodytas tik daiktavardis *stalas*, o kalbos vartotojai patys turi suprasti, kad iš jo galima sudaryti deminutyvus *staliukas, stalelis, staliūkštis* ir kt. Taigi norint, kad kuo daugiau formų būtų atpažįstama automatiškai, kai kuriuos darybos būdus reikėjo įtraukti kaip reguliarius, neatsižvelgiant į semantinius skirtumus. Dėl to vėliau paaiškėjo, kad buvo generuoti nerealieji homonimai, kurie sutapdavo su esamais lietuvių kalbos žodžiais, pvz., jau minėto nerealaus mažybinio daiktavardžio *padarytis* (padaryto iš daiktavardžio *padaras*) vienaskaitos šauksmininkas (*mano mažasis padaryti!*) sutampa su veiksmažodžio *padaryti* bendratimi (*dar daug reikia padaryti*).

Morfologiškai daugiareikšmių formų atsiranda ir nepateikus kitų formų DLKŽ-e, pvz., nurodyti ne visi sangrąžiniai veiksmažodžiai, veiksmažodinės kilmės daiktavardžiai su *-imas/-ymas, -umas, -ėjas, -tojas* ir kt. Dėl šių priežasčių buvo generuoti tokie nerealūs daiktavardžiai kaip *galimas, išsaugotumas, gulėjas, neišskubėjas*. Dažnai sunku suprasti, kurios formos, DLKŽ-o sudarytojų nuomone, yra darybinės, kurios kaitybinės, kurie dariniai reguliarūs, o kurie – ne (apie šiuos dalykus dar žr. GKT, 2004, p. 43–54).

Homoformų MAT-e atsiranda ir dėl kitų priežasčių: pavyzdžiui, dėl retai vartojamų ar jau išnykusių žodžių pateikimo DLKŽ-e ir TŽŽ-e. DLKŽ-e pateiktas dabartinėje lietuvių kalboje nebevartojamas daiktavardis *kokis*, t. y. „kokybė“, kuris yra įtrauktas į *Lemuoklio* leksikos duomenų bazę. Šio žodžio vienaskaitos kilmininkas (*siekiame geresnio kokio* (= kokybės)) sutampa su labai dažnu lietuvių kalbos įvardžio *koks* forma *kokio* (*o kokio darbo pageidaujate?*). Skaitantiems rišlų tekstą aišku, kada koks žodis pavartotas, bet automatinės morfologinės analizės metu analizuojami

[2] *Dabartinės lietuvių kalbos žodynas* toliau žymimas DLKŽ, o *Tarptautinių žodžių žodynas* – TŽŽ.

pavieniai žodžiai visiškai neatsižvelgiant į tekste pavartotus gretimus žodžius, nenustatant jų ryšių, todėl ir atsiranda tokių nerealiųjų homoformų. Visi šie dalykai trukdo automatinei morfologinei analizei ir reikalauja žmogaus įsikišimo tvarkant anotuotus tekstus.

3. GRAMATINIŲ FORMŲ VARTOSENA MORFOLOGIŠKAI ANOTUOTAME TEKSTYNE

Remiantis MAT-o duomenimis, galima daryti išvadas apie dažniausiai vartojamas dabartinės lietuvių kalbos gramatines formas. Palyginus su V. Žilinskienės pateiktais duomenimis, gautais išnagrinėjus publicistinio, dalykinio, grožinio ir mokslinio stiliaus tekstus (Žilinskienė, 2001; 2002a; 2002b; 2005), matyti, kad duomenys iš MAT-o yra labai panašūs. Kadangi straipsnyje aprašomas MAT yra sudarytas iš skirtingų stilių tekstų (žr. 3 pav.), todėl galima teigti, kad jis tinkamai reprezentuoja dabartinę lietuvių kalbą. Toliau šiame skyriuje pateikti statistikos duomenys, gauti ištyrus ir suskaičiavus gramatinių formų vartoseną MAT-e.

Apibendrinus iš MAT-o gautus duomenis pastebėta, kad vartojama labai nedaug skirtingų kaitybinių formų: vienai lemai tenka tik 2,34 kaitybinės formos[3] (plg. lenkų kalbai tenka 2,01 kaitybinės formos – šie duomenys pateikti remiantis IPI PAN tekstynu (Przepiórkowsky, 2004)). Kaitybinių formų skaičių mažina nekaitomos kalbos dalys, kurios vartojamos labai dažnai.

Vartojama tik 2,54 daiktavardžių, 2,63 veiksmažodžių kaitybinių formų. Didele formų įvairove pasižymi įvardžiai (8,23 formų). Nemažai vartojama skirtingų būdvardžių (4,72) ir skaitvardžių (3,55) formų. Prieveiksmių vartojama 1,52; dalelyčių – 1,03; jungtukų – 1,04; jaustukų – 1; ištiktukų – 1; prielinksnių – 1,01; akronimų – 1,02; santrumpų – 1 forma[4] (4 pav.). Taigi galima daryti išvadą, kad lietuvių kalba yra sudėtinga fleksinė kalba, pasižyminti gramatinių formų įvairove, tačiau realiai vartojama tik nedidelė tų formų dalis. Labai skiriasi gramatikose pateikiamos kalbos dalių teorinės kaitybinės paradigmos ir iš tikrųjų vartojamos formos. Pavyzdžiui, galima teigti, kad pagrindiniai vardažodžių linksniai yra vardininkas, kilmininkas ir galininkas. Kiti linksniai yra periferiniai, sudarantys tik nedidelę visų linksnių dalį.

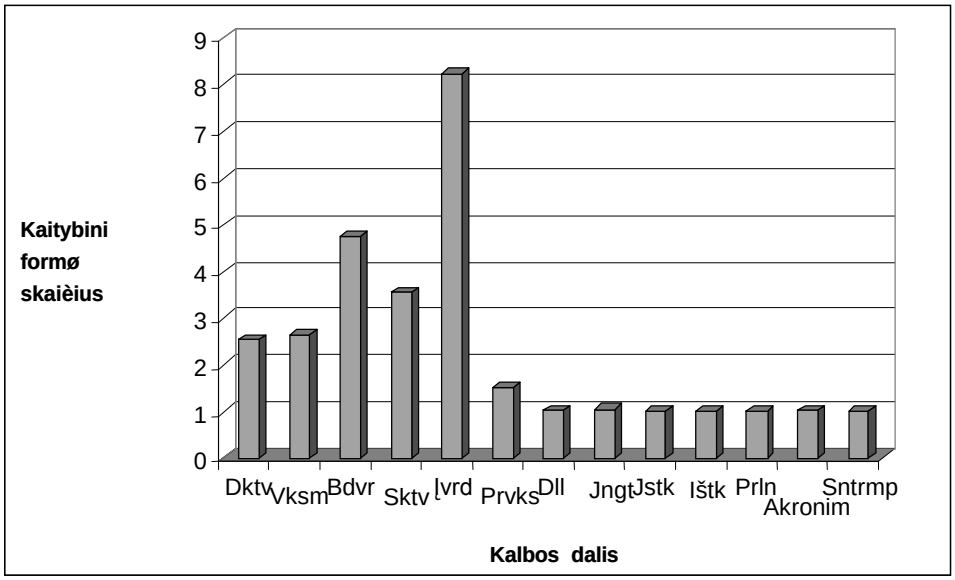
Duomenys apie tai, kiek kaitybinių formų tenka kuriai nors kalbos daliai, beveik nepriklauso nuo tekstyno dydžio. Analizuojant didesnį tekstų kiekį kaitybinių formų pasiskirstymas gali keistis labai nedaug.

Iš viso MAT-ą sudarančiose rinkmenose yra 1 012 673 žodžių[5] (ir tų pačių, ir

[3] Tai reiškia, kad, pvz., nors teoriškai, daiktavardis gali turėti 14 formų (7 vienaskaitos, 7 daugiskaitos linksnius), bet realiai vartojamos tik 2–3 formos.

[4] Nors dalelytės, jungtukai ir prielinksniai laikomi visiškai nekaitomomis kalbos dalimis (taip pat ir sąlygiškai prie kalbos dalių priskirti akronimai), bet iš pateiktų duomenų matyti, kad minėtos kalbos dalys turi po kelias formas. Taip yra dėl to, kad kai kurios dalelytės, jungtukai, prielinksniai gali turėti neigiamas formas (žr. 14, 15, 16 nuorodas) (žinoma, galima teigti, kad minėtų nekaitomų kalbos dalių neigiamos formos yra atskiri žodžiai, tačiau šiame straipsnyje nesigilinama į kalbos dalių klasifikaciją). Kai kurie akronimai gali būti linksniuojami, pvz., *ELTA, ELTOS, Sodra, Sodrai* ir pan.

[5] Čia ir toliau tekste vartojama sąvoka *žodis* reiškia ir žodį, ir žodžio formą.



4 pav. Kalbos dalims tenkančių žodžių formų pasiskirstymas morfologiškai anotuotame tekstyne

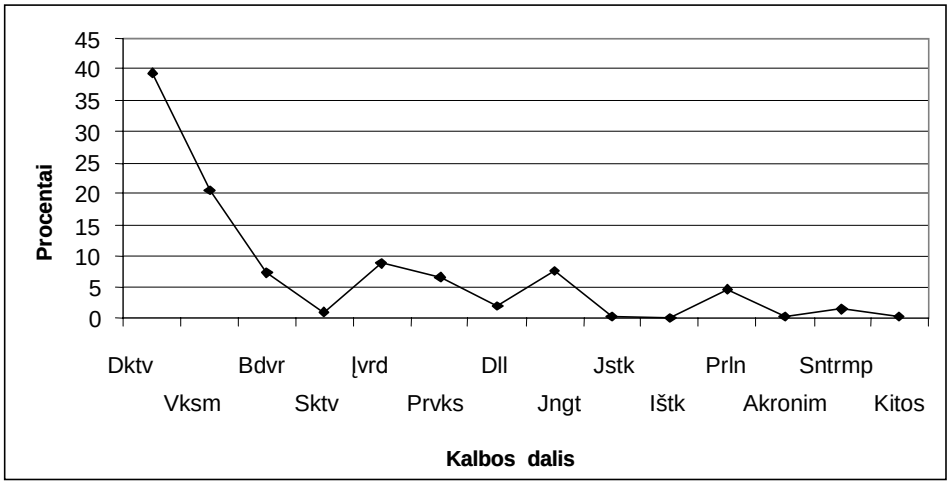
skirtingų). Lemuojant ir morfologiškai anotuojant tas rinkmenas, kai kurie žodžiai buvo pažymėti kaip vienas junginys, pvz., *bet kas, iš anksto*. Tai – morfologinės samplaikos (plačiau žr. Rimkutė et al., 2005). Jas sužymėjus kaip vieną neskaidomą junginį, sumažėjo bendras žodžių skaičius. Galutinai sutvarkytas anotuotas rinkmenas sudaro 923 397 tie patys bei skirtingi žodžiai ir 139 999 skirtingi žodžiai.

MAT-e yra 51 888 lemos. Ta pati lema, t. y. kurio nors žodžio antraštinis pavidalas, pvz., *daiktas, eiti, nuostabus, pastoviai*, MAT-e pavartota maždaug 18 kartų. Viena žodžio forma, pvz., *daiktų, eidavo, nuostabius, pastoviausiai*, MAT-e vidutiniškai pavartota 6,6 karto (apibendrinti statistikos duomenys apie MAT-ą pateikti 4 lentelėje).

Ketvirtoje lentelėje pateikti duomenys priklauso nuo tekstyno dydžio. Didesniame tekstyne, sudarytame iš žanriniu požiūriu įvairesnių tekstų, gali keistis žodžių ir lemų pavartojimo dažnumas, vienai lemai tenkančių skirtingų kaitybinių formų skaičius, duomenys, kiek vienam žodžiui tenka lemų, morfologinių pažymų, t. y. koks yra MD. Vis dėlto iš 1 mln. žodžių sudarytas MAT jau atskleidžia bendrą dabartinės rašytinės lietuvių kalbos vaizdą ir gana patikimai atspindi MD-o mastą bei proporcijas.

Lietuvių kalbos MAT-ą galima palyginti su struktūriškai panašios čekų kalbos anotuotu tekstynu. Čekų kalbos DESAM tekstyne iš viso yra 1 247 594 žodžiai. Skirtingų žodžių yra 132 447 (tik vieną kartą pavartoti yra 67 059 žodžiai). Skirtingų lemų yra 34 606 (11 759 lemos pavartotos tik vieną kartą). Skirtingų morfologinių pažymų yra 1 665. Kas dešimtas žodis čekų kalbos tekste turi mažiausiai dvi lemas. Vienam žodžiui tenka 4,21 lemos ar morfologinės pažymos (Popelînskô, 2000).

Kalbos dalių vartojimo tendencijos MAT-e sutampa su kitų tyrinėtojų duomenimis (plg. Žilinskienė, 2001; 2002a; 2002b; 2003; 2005). Daugiausia kalbos dalių sudaro daiktavardžių formos (apie 39 proc.), veiksmažodžiai sudaro apie 20 procentų. Toliau pagal dažnumą eina įvardžiai (8,7 proc.), būdvardžiai (7,33 proc.), prieveiksmiai (6,72



5 pav. Kalbos dalių pasiskirstymas morfologiškai anotuotame tekstyne

proc.), jungtukai (7,62 proc.), prielinksniai (4,65 proc.). Kitos kalbos dalys vartojamos gerokai rečiau (skaitvardžiai sudaro 0,96 proc., dalelytės – 1,97 proc., jaustukai – 0,18 proc., ištiktukai – 0,02 proc.). Taigi vardažodžiai MAT-e sudaro apie 56 proc. visų kalbos dalių (5 pav. ir 5 lentelė).

Panašius duomenis apie kalbos dalių pasiskirstymą pateikia V. Žilinskienė (2003; 2005). Pasinaudojusi *Dažniniu dabartinės rašomosios lietuvių kalbos žodynu* ir kompiuteriniais atskirų stilių (beletristinio, publicistinio, mokslinio ir dalykinio) dažniniais žodynais, V. Žilinskienė teigia, kad dažniausia kalbos dalis yra daiktavardis (jis sudaro 38,98 proc.). Veiksmažodžių yra 20,99 proc., įvardžių – 9,25 proc., jungtukų – 8,16 proc., būdvardžių – 8,14 proc. Prieveiksmiai sudaro 5,19 proc., prielinksniai – 4,93 proc., dalelytės – 3,10 proc., skaitvardžiai – 1,1 proc., jaustukai – 0,15 proc., ištiktukai – 0,03 proc. (Žilinskienė, 2003). Turbūt ir MAT-e, jeigu būtų tiriami skirtingi stiliai, paaiškėtų, kad, kaip ir V. Žilinskienės analizuotuose skirtingų stilių tekstuose, daugiau veiksmažodžių, prieveiksmių, įvardžių, dalelyčių ir prielinksnių vartojama vaizdingesniuose, emocingesniuose tekstuose. Kai svarbu perteikti faktinę informaciją, tekstuose dažniau vartojami daiktavardžiai ir būdvardžiai (Žilinskienė, 2005, p. 30).

Dažniausiai yra vartojamos vyriškosios **giminės** formos (57,2 proc.) – matyt, todėl, kad vyriškoji giminė yra nežymėtasis giminės kategorijos narys. Moteriškosios giminės formos sudaro 36,6 procento. Šiek tiek skiriasi skirtingų kalbos dalių giminių pasiskirstymas, pavyzdžiui, skirtumas tarp būdvardžių vyriškosios ir moteriškosios giminės nėra toks didelis kaip daiktavardžių (atitinkamai 54,4 proc. ir 41,5 proc. būdvardžių, 60,4 proc. ir 39,6 proc. daiktavardžių; kaip pasiskirsčiusios skirtingų kalbos dalių giminės, žr. 6 lentelę). Pasak V. Žilinskienės, visų stilių vyriškoji giminė vartojama dažniau už moteriškąją, pvz., daiktavardžių vyriškosios giminės formų skirtinguose stiliuose rasta 60 proc., būdvardžių – nuo 55 iki 57 proc. (Žilinskienė, 2003).

Kaip nežymėtasis **skaičiaus** kategorijos narys vienaskaitos formos MAT-e yra dvigubai dažnesnės negu daugiskaitos (atitinkamai 65 proc. ir 32 proc.; 7 lentelė). Panašūs

ir V. Žilinskienės pateikti duomenys: jos teigimu, visų stilių tekstuose vienaskaita vartojama dažniau už daugiskaitą, pvz., daiktavardžių vienaskaitos formų skirtinguose stiliuose rasta nuo 66 iki 72 proc., būdvardžių – nuo 58 iki 63 proc. (Žilinskienė, 2003).

Iš **linksnių** dažniausias yra vienaskaitos kilmininkas (22,8 proc.), šiek tiek rečiau vartojamos vienaskaitos vardininko formos (21,8 proc.). Panašaus dažnumo yra daugiskaitos kilmininko ir vienaskaitos galininko formos (atitinkamai 11,8 ir 10,6 proc.; 8 lentelė). Šie duomenys patvirtina V. Žilinskienės teiginį, kad visų stilių visų linksniuojamųjų kalbos dalių linksniai aiškiai skiriasi į dvi grupes: pagrindinę (vardininkas, kilmininkas, galininkas) ir periferinę (įnagininkas, vietininkas, naudininkas). Periferinei grupei dar priskirtinas ir šauksmininkas bei MAT-e atskiru linksniu laikomas gana dažnas iliatyvas.

Dažniausi daiktavardžių linksniai MAT-e pasiskirstę taip: vienaskaitos kilmininkas sudaro 28 proc. visų daiktavardžių linksnių (arba 40,2 proc. visų vienaskaitos linksnių), vienaskaitos vardininkas – 19,2 proc. visų linksnių (27,5 proc. vienaskaitos linksnių), vienaskaitos galininkas – 15,8 proc. visų daiktavardžio linksnių (15,8 proc. daiktavardžių vienaskaitos linksnių). Daugiskaitos kilmininkas sudaro 12,8 proc. visų linksnių (42,4 proc. daugiskaitos linksnių), daugiskaitos vardininkas – 7,2 proc. visų linksnių (23,7 daugiskaitos linksnių), o daugiskaitos galininkas – 5,5 proc. visų linksnių (18,2 proc. daugiskaitos linksnių) (išsamiau žr. 8 lentelę).

V. Žilinskienės straipsniuose (2001; 2002a; 2002b ir 2005) aprašyta, kaip pasiskirstę dažniausi daiktavardžių linksniai dalykiniame, publicistiniame ir moksliniame stiliuose: vienaskaitos kilmininkas (atitinkamai 45,65, 38,91 ir 46,47 proc.), vienaskaitos vardininkas (atitinkamai 25,09, 26,62 ir 22,67 proc.) ir vienaskaitos galininkas (atitinkamai 14,09, 16,76 ir 13,27 proc.). Daugiskaitos kilmininkas dalykiniame, publicistiniame ir moksliniame stiliuose sudaro atitinkamai 43,46, 41,81 ir 45,85 proc.; daugiskaitos vardininkas – atitinkamai 22,99, 23,29 ir 23,13 proc.; daugiskaitos galininkas – atitinkamai 17,26, 18,12 ir 14,41 proc. Taigi tiek iš MAT-o gautų duomenų, tiek ir iš V. Žilinskienės pateiktos informacijos matyti, kad linksnių vartosena dabartinės lietuvių kalbos įvairaus stiliaus tekstuose yra gana panaši.

Veiksmažodžių asmenuojamosios ir neasmenuojamosios formos yra pasiskirsčiusios gana tolygiai (atitinkamai 50,7 ir 49,3 proc.; 9 lentelė). Iš asmenuojamųjų formų dažniausiai vartojama tiesioginė **nuosaka** (91,5 proc.), liepiamoji nuosaka sudaro 2,7 proc., tariamoji – 5,8 proc. Iš neasmenuojamųjų formų dažniausi yra dalyviai (56,2 proc.) ir bendratys (33,6 proc.). Padalyviai sudaro 6,4 proc., pusdalyviai – 3,8 proc., būdiniai – 0,04 procento. Daugiausia MAT-e yra esamojo (52,8 proc.) ir būtojo kartinio (31,2 proc.) **laiko** (10 lentelė), taip pat trečiojo **asmens** formų (jos sudaro 81,3 proc.; 11 lentelė).

Mažai kuo nuo MAT-o duomenų skiriasi V. Žilinskienės pateikta informacija apie veiksmažodžius: jos tirtuose skirtinguose stiliuose dažniausios yra asmenuojamosios formos (nuo 36,6 iki 68,51 proc.), dalyviai (nuo 16,44 iki 37,64 proc.) ir bendratys (nuo 10,65 iki 20,25 proc.). Visų stilių tekstuose labai aiškiai vyrauja tiesioginės nuosakos formos (nuo 89 proc. dalykiniuose tekstuose iki 94 proc. moksliniuose) ir trečiojo asmens formos. Būtojo kartinio laiko formų daugiausia beletristikos tekstuose (52,48 proc.), publicistikos, mokslinio ir dalykinio stilių tekstuose vyrauja esamasis laikas (atitinkamai 46,51, 66,75, 73,72 proc.) (Žilinskienė, 2003; 2005).

Kaip ir V. Žilinskienės tirtuose dalykiniame bei publicistiniame stiliuose, taip ir MAT-e dažnesni yra neveikiamieji **dalyviai** (dalykiniame stiliuje – 81,52 proc., publicistiniame – 61,88 proc. (Žilinskienė, 2002, p. 112), MAT-e – 65,8 proc.). Veikiamieji dalyviai MAT-e sudaro 32,2 proc., o reikiamybės – 2 proc. (12 lentelė).

MAT-e daugiausia pavartota nelyginamojo **laipsnio** (jos sudaro 87,5 proc.) formų. Aukštesniojo laipsnio formos MAT-e sudaro 8 proc., aukščiausiojo – 4,5 proc. (13 lentelė). V. Žilinskienės (2002, p. 114) duomenimis, būdvardžių nelyginamasis laipsnis dalykinio stiliaus tekstuose sudaro 93 proc., publicistinio – 90,92 proc. (MAT-e – 92,3 proc.), aukštesnysis – atitinkamai 3,25 ir 4,04 proc. (MAT-e – 4 proc.), aukščiausiasis – atitinkamai 3,75 ir 5,03 proc. (MAT-e – 3,7 proc.).

Straipsnyje aptariamame MAT-e daugiausia yra neįvardžiuotinių formų (93,6 proc.), įvardžiuotinės sudaro 6,4 proc. (14 lentelė). Panašiai **apibrėžtumo** kategorija vartojama ir V. Žilinskienės tirtuose dalykinio bei publicistinio stiliaus tekstuose: įvardžiuotinės formos sudaro atitinkamai 8,96 ir 9,16 proc., neįvardžiuotinės – 91,04 ir 90,84 proc. (Žilinskienė, 2002, p. 114).

Iš 15 lentelėje pateiktų duomenų matyti, kad daugiausia yra **teigiamųjų** formų – jos sudaro 97,4 proc., o neigiamosios – 2,6 procento. Taip pat gerokai dažnesnės **kaitomos** (98,6 proc.), o ne nekaitomos (1,4 proc.) formos (16 lentelė).

Dažnesni yra **bendriniai** (89 proc.), o ne tikriniai daiktavardžiai (11 proc.; 17 lentelė), kiekiniai pagrindiniai (58 proc.) ir kelintiniai (37,7 proc.) **skaitvardžiai** (dauginiai skaitvardžiai sudaro 4 proc., o kuopiniai – 0,3 proc., dar žr. 18 lentelę).

Kaip ir reikėjo tikėtis, kaip nežymėtasis **sangrąžos** kategorijos narys didžiąją dalį sudaro nesangrąžiniai veiksmažodžiai (88,5 proc.). Sangrąžinių veiksmažodžių yra 11,5 proc. (žr. 19 lentelę).

IŠVADOS

Lietuvių kalbai iš anotuotų tekstynų kol kas yra sudaryti tik morfologiškai žymėti tekstynai. Vienas iš jų yra parengtas VDU Kompiuterinės lingvistikos centre. Šis tekstynas buvo sudarytas pusiau automatiškai, panaudojus morfologinį analizatorių *Lemuoklis*. Jame yra vartojamos kalbinės (lemos, kalbos dalys, gramatinių kategorijų pažymos) ir nekalbinės (informacija apie autorius, pavadinimus, pastraipas, užsienio kalbų intarpus ir pan.) pažymos. MAT-e nurodomos ne tik tradicinės gramatinės kategorijos (linksnis, nuosaka ir pan.), bet ir kai kurios kitos specifinei analizei būtinos pažymos, pvz., skiriami nekaitomi tikriniai daiktavardžiai, nurodomas DLKG-oje kaip atskiras linksnis neskirtas iliatyvas, skiriami trys tretieji veiksmažodžių asmenys, pažymimos iš kelių, dažniausiai nekaitomų, kalbos dalių sudarytos morfologinės samplaikos.

MAT buvo sudarytas iš publicistikos, grožinės, mokslinės literatūros tekstų, kanceliarinių dokumentų ir Lietuvos Respublikos Seimo stenogramų. MAT-e yra 1 012 673 žodžiai. Pažymėjus leksiškai gramatiškai susijusius junginius – morfologines samplaikas – MAT-ą sudaro 923 397 žodžiai, iš jų 139 999 yra skirtingi žodžiai. Iš viso MAT-e yra 51 888 lemos. Viena lema MAT-e pavartota 17,8 kartus, o viena žodžio forma – 6,6 karto.

Iš MAT-o duomenų galima daryti išvadas apie dabartinės rašytinės lietuvių kalbos gramatinių formų vartoseną. Paaiškėjo, kad vartojama palyginti labai nedaug skirtingų

kaitomų formų, pvz., vartojama tik 2,54 daiktavardžių, 2,63 veiksmažodžių kaitybinių formų. Didele formų įvairove pasižymi įvardžiai (8,23 formos). Taip pat vartojama nemažai skirtingų būdvardžių (4,72) ir skaitvardžių (3,55) formų. Vienai lietuvių kalbos lemai tenka tik 2,34 kaitybinių formų.

Kalbos dalių vartojimas MAT-e sutampa su kitų tyrinėtojų duomenimis. Daugumą kalbos dalių sudaro daiktavardžių formos (apie 39 proc.), veiksmažodžiai sudaro apie 20 procentų. Toliau pagal dažnumą eina įvardžiai (8,7 proc.), būdvardžiai (7,33 proc.), prieveiksmiai (6,72 proc.), jungtukai (7,62 proc.). Kitos kalbos dalys vartojamos gerokai rečiau.

Dažniausiai yra vartojamos vyriškosios giminės formos (57,2 proc.), moteriškosios giminės formos sudaro 36,6 procento. Vienaskaitos formos yra dvigubai dažnesnės negu daugiskaitos (atitinkamai 65 ir 32 proc.). Iš linksnių dažniausias yra vienaskaitos kilmininkas (22,8 proc.), šiek tiek rečiau vartojamas vienaskaitos vardininkas (21,8 proc.). Panašaus dažnumo yra daugiskaitos kilmininko ir vienaskaitos galininko formos (atitinkamai 11,8 ir 10,6 proc.).

Veiksmažodžių asmenuojamosios ir neasmenuojamosios formos yra pasiskirsčiusios gana tolygiai (atitinkamai 50,7 ir 49,3 proc.). Iš asmenuojamųjų formų dažniausiai vartojama tiesioginė nuosaka (91,5 proc.), iš neasmenuojamųjų – dalyviai (56,2 proc.) ir bendratys (33,6 proc.). Daugiausia MAT-e yra esamojo (52,8 proc.) ir būtojo kartinio (31,2 proc.) laiko, trečiojo asmens formų (jos sudaro 81,3 proc.).

MAT-e daugiausia pavartota nelyginamojo laipsnio (jos sudaro 87,5 proc.), neįvardžiuotinių (93,6 proc.), teigiamųjų (97,4 proc.), kaitomų (98,6 proc.) formų. Dažnesni yra bendriniai (89 proc.), o ne tikriniai daiktavardžiai, kiekiniai pagrindiniai (58 proc.) ir kelintiniai (37,7 proc.) skaitvardžiai, neveikiamieji dalyviai (65,8 proc.), nesangrąžiniai veiksmažodžiai (88,5 proc.).

Remiantis šiais duomenimis, galima daryti išvadą, kad lietuvių kalba yra fleksinė, pasižyminti sudėtingomis kaitybinėmis paradigmomis, bet realiai tekstuose vartojama tik labai nedidelė tų kaitybinių formų dalis. Taigi lietuvių kalbos formas galima skirstyti į pagrindines ir periferines. Vienų formų vyravimą kitų atžvilgiu ypač išryškina linksnio kategorija.

Remiantis tuo, kad vartojama nedaug kaitybinių formų ir kad gana dažnai vietoj vieno žodžio vartojamos dvi ar daugiau žodžių formų (pvz., *negalima – nėra galima* ir pan.), būtų galima daryti prielaidą, kad lietuvių kalba tampa analitiškesnė.

Ateityje morfologiškai anotuotas tekstynas bus naudojamas kuriant automatinę sintaksinę analizę, sudarant sintaksiškai anotuotus tekstynus. Planuojama MAT-ą padaryti visiems prieinamą internete. Turbūt tada iškils daug diskusijų, ginčytinų dalykų dėl tam tikrų žodžių ar žodžių formų priskyrimo vienai ar kitai kalbos daliai. Nelengva remiantis *Dabartinės lietuvių kalbos gramatikos* ir *Dabartinės lietuvių kalbos žodyno* duomenimis vienareikšmiškai nuspręsti, kuria kalbos dalimi laikytina viena ar kita forma. Tikimasi, kad probleminių atvejų aptarimas atkreips žodynų ir gramatikų sudarytojų dėmesį į nemažai diskutuotinų lietuvių kalbos dalykų.

1 lentelė. **Skirtingo formato morfologiškai anotuoto teksto pavyzdžiai**

Pirminis MAT-o formatas	**Antrinis MAT-o formatas**	**Tik anotuotas ir dar nevienareikšmintas tekstas**
Keli įvrd <keli> įvrd vyr.gim V dalykai dktv <dalykas> dktv vyr.gim dgsk V paakino vksm <paakinti(-a,-o)> vksm teig nesngr tiesiog. nuos būt. kart. l dgsk III asm imtis vksm <imtis(imasi ėmėsi)> bndr teig sngr šio įvrd <šis> įvrd neįvardž vyr. gim vnsk K straipsnio dktv <straipsnis> dktv vyr. gim vnsk K	<word=”Keli” lemma=”keli” type=”įvrd vyr.gim V”> <space> <word=”dalykai” lemma=”dalykas” type=”dktv vyr. gim dgsk V”> <space> <word=”paakino” lemma=”paakinti(-a,-o)” type=”vksm teig nesngr tiesiog.nuos būt. kart. l dgsk IIIasm”> <space> <word=”imtis” lemma= ”imtis(imasi,ėmėsi)” type=”bndr teig sngr”> <space> <word=”šio” lemma=”šis” ype=”įvrd neįvardž vyr. gim vnsk K”> <space> <word=”straipsnio” lemma=”straipsnis” type=”dktv vyr. gim vnsk K”>	Keli dktv <kelis> dktv vyr. gim vnsk Š įvrd <keli> įvrd vyr. gim V vksm <kelti(-elia,-ėlė)> vksm teig nesngr tiesiog. nuos esam. l vnsk II asm dalykai dktv <dalykas> dktv vyr. gim dgsk V paakino vksm <paakinti(-a,-o)> vksm teig nesngr tiesiog.nuos būt. kart. l vnsk III asm vksm teig nesngr tiesiog.nuos būt. kart. l dgsk III asm imtis vksm <imtis(imasi,ėmėsi)> bndr teig sngr dktv <imtis> dktv mot. gim vnsk V dktv mot. gim dgsk G šio įvrd <šis> įvrd neįvardž vyr. gim vnsk K straipsnio dktv <straipsnis> dktv vyr. gim vnsk K

2 lentelė. **Morfologiškai anotuotame tekstyne naudotos kalbos dalių ir kitų leksinių gramatinių vienetų pažymos**

Kalbos dalis/kiti leksiniai gramatiniai vienetai	**MAT-e naudota pažyma**
akronimas	akronim
bendratis	bndr
būdinys	būdn
būdvardis	bdvr
daiktavardis	dktv
dalelytė	dll
dalyvis	dlv
idioma	id ... (idJngt, idPrln ir kt.)
ištiktukas	ištk
įvardis	įvrd
jaustukas	jstk
jungtukas	jngt
padalyvis	padlv
prielinksnis	prln
prieveiksmis	prvks
pusdalyvis	psdlv
romėniškas skaičius	rom skaič
santrumpa	sntrmp
skaitvardis	sktv
tikrinis daiktavardis	tikr dktv
tikrinis daiktavardis 2	tikr dktv2
veiksmažodis	vksm

3 lentelė. **Morfologiškai anotuotame tekstyne naudotos gramatinių kategorijų pažymos**

Gramatinė kategorija	**Gramatinės kategorijos vertė**	**MAT-e naudota pažyma**
sangrąžiškumas	sangrąžinis	sngr
	nesangrąžinis	nesngr
teigiamumas	teigiamas	teig
	neigiamas	neig
rūšis	veikiamoji	veik. r
	neveikiamoji	neveik. r
	reikiamybės dalyviai	reikiamyb
nuosaka	tiesioginė	tiesiog. nuos
	liepiamoji	liepiam. nuos
	tariamoji	tariam. nuos
laikas	esamasis	esam. l
	būtasis kartinis	būt.kart. l
	būtasis dažninis	būt. d. l
	būtasis	būt. l
	būsimasis	būs. l
skaitvardžių tipai	kiekiniai	kiekin
	kelintiniai	kelintin
	dauginiai	daugin
	kuopiniai	kuopin
laipsnis	nelyginamasis	nelygin. l
	aukštesnysis	aukštesn. l
	aukščiausiasis	aukšč. l
	aukštėlesnis	aukštėlesn. l
apibrėžtumas	įvardžiuotinis	įvardž
	neįvardžiuotinis	neįvardž
giminė	vyriškoji	vyr.gim
	moteriškoji	mot. gim
	bevardė	bevrd. gim
	bendroji	bendr. gim
skaičius	vienaskaita	vnsk
	daugiskaita	dgsk
	dviskaita	dvisk
linksnis	vardininkas	V
	kilmininkas	K
	naudininkas	N
	galininkas	G
	įnagininkas	Įn
	vietininkas	Vt
	šauksmininkas	Š
	iliatyvas	Il
	aliatyvas	Al
asmuo	pirmas	I asm
	antras	II asm
	trečias	III asm

4 lentelė. **Pagrindiniai statistiniai duomenys apie morfologiškai anotuotą tekstyną**

MAT-ą sudarančių rinkmenų žodžių skaičius	1012673
MAT-ą sudarančių žodžių skaičius (kai buvo pažymėti leksiškai ir semantiškai susiję junginiai, t. y. morfologinės samplaikos)	923397
MAT-ą sudarančių skirtingų žodžių skaičius	139999
MAT-ą sudarančių skirtingų lemų skaičius	51888
Vidutinis vienos lemos pavartojimo MAT-e dažnumas[6]	17,8 karto
Vidutinis vieno žodžio pavartojimo MAT-e dažnumas	6,6 karto
Vienai lemai 1 mln. žodžių apimties tekstyne tenka	2,34 kaitybinių formų
Nesutvarkytus, t. y. morfologiškai daugiareikšmius, tekstus sudarančių lemų skaičius	1189209
Nesutvarkytus, t. y. morfologiškai daugiareikšmius, tekstus sudarančių morfologinių pažymų[7] skaičius	1488178
Vienam žodžiui 1 mln. žodžių apimties tekstyne tenka	1,3 lemos
Vienam žodžiui 1 mln. žodžių apimties tekstyne tenka	1,6 morfologinės pažymos
Vienai lemai 1 mln. žodžių apimties tekstyne tenka	1,25 morfologinės pažymos

5 lentelė. **Kalbos dalių pasiskirstymas morfologiškai anotuotame tekstyne**

Kalbos dalys	**Bendras visų formų skaičius**	**Proc.**	**Skirtingų lemų skaičius**	**Proc.**
1. Daiktavardžiai	363544	39,37	23413	45,12
2. Veiksmažodžiai	189279	20,5	20319	39,16
3. Būdvardžiai	67690	7,33	4577	8,82
4. Skaitvardžiai	8885	0,96	264	0,5
5. Įvardžiai	80371	8,7	124	0,24
6. Prieveiksmiai	62011	6,72	1813	3,5
7. Dalelytės	18217	1,97	112	0,22
8. Jungtukai	70342	7,62	67	0,13
9. Jaustukai	1613	0,18	74	0,14
10. Ištiktukai	138	0,02	56	0,1
11. Prielinksniai	42910	4,65	82	0,16
12. Akronimai[8]	2333	0,25	422	0,81
13. Santrumpos	14632	1,58	542	1,06
14. Kitos kalbos dalys[9]	1432	0,15	23	0,04
Iš viso	923397	100	51888	100

[6] Vidutinis vienos lemos pavartojimo MAT-e dažnumas yra 17,8 karto. Šis vienetas yra vadinamas iteracijos koeficientu ir parodo leksikos įvairovę. Kuo šis dydis mažesnis, tuo daugiau skirtingų žodžių (Grumadienė, 2002, p. 29). *Dabartinės lietuvių kalbos dažninio žodyno* iteracijos koeficientas yra 20,06, nors šio žodyno duomenų bazę sudaro 1,2 mln. žodžių (Grumadienė, 2002, p. 29). Taigi galima daryti išvadą, kad straipsnyje aptariamas morfologiškai anotuotas tekstynas, nors sudarytas tik iš 1 mln. žodžių, pasižymi įvairesne leksika nei tekstai, sudarantys *Dabartinės lietuvių kalbos dažninio žodyno* duomenų bazę.

[7] Morfologine pažyma laikomas gramatinių kategorijų pateikimas, pvz., žodžiui *naktis* gali būti pateikiamos dvi morfologinės pažymos: 1) dktv mot. gim vnsk V ir 2) dktv mot. gim dgsk G. Eilutė *dktv mot. gim vnsk V*, kurioje nurodoma kalbos dalis, giminė, skaičius ir linksnis (kitoms kalbos dalims nurodomos kitos gramatinės kategorijos), šiame straipsnyje laikoma viena morfologine pažyma.

[8] Prie kalbos dalių straipsnyje priskirti akronimai ir santrumpos.

[9] Prie „kitų kalbos dalių“ priskirtos morfologinės samplaikos, t. y. iš kelių žodžių sudaryti sintaksiškai ir semantiškai susiję junginiai, pvz.: *bet kas, kas nors, iš tolo*. Tokio pobūdžio junginiai MAT-e žymimi kaip vienas vienetas.

6 lentelė. **Giminių pasiskirstymas morfologiškai anotuotame tekstyne**

Giminė/ kalbos dalis	Dktv		Bdvr		Įvrd		Sktv		Dlv		Psdlv		Iš viso	
	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%
Vyriškoji	217670	60,4	36794	54,4	36816	45,8	5285	59,5	28784	54,8	2662	76	328011	57,2
Moteriškoji	142538	39,6	28085	41,5	19112	23,8	2766	31,1	16716	31,9	838	24	210055	36,6
Bevardė	-	-	2775	4,1	4589	5,7	448	5	6983	13,3	-	-	14795	2,6
Bendroji	138	0,04	-	-	-	-	-	-	-	-	-	-	138	0,02
Negimininiai	-	-	-	-	19854	24,7	386	4,4	-	-	-	-	20240	3,6
Iš viso	360346	100	67654	100	80371	100	8885	100	52483	100	3500	100	573239	100

7 lentelė. **Skaičiaus pasiskirstymas morfologiškai anotuotame tekstyne**

Skaičius/ kalbos dalis	Dktv		Bdvr		Įvrd		Sktv		Dlv		Psdlv		Vksm		Iš viso	
	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%
Vienaskaita	251625	69,8	40767	62,8	39404	54,8	3908	47,4	27039	59,5	2460	70,3	57208	59,7	422411	65
Daugiskaita	108718	30,2	24109	37,2	24376	33,9	987	12	18434	40,5	1040	29,7	30179	31,4	207843	32
Dviskaita	3	0,0008	3	0,005	111	0,1	-	-	1	0,002	-	-	-	-	118	0,02
Neturintys skaičiaus[10]	-	-	-	-	8075	11,2	3342	40,6	-	-	-	-	8520	8,9	19937	3
Iš viso	360346	100	64879	100	71966	100	8237	100	45474	100	3500	100	95902	100	650309	100

[10] MAT-e laikoma, kad skaičiaus kategorijos neturi, pvz., įvardžiai *kas, keletas,* skaitvardžiai *dvidešimt, trejetas, dveji*, veiksmažodžiai *reikia, norisi* ir pan.

8 lentelė. **Linksnių pasiskirstymas morfologiškai anotuotame tekstyne**

Linksnis/ kalbos dalis		Dktv		Bdvr		Įvrd		Sktv		Dlv		Iš viso	
		Sk.	%[11]	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%
	V	69234	19,2 (27,5)	16319	25,2 (40)	15250	21,2 (38,7)	1274	15,5 (32,6)	18 067	39,7 (66,8)	120144	21,8
	K	101245	28 (40,2)	10873	16,8 (26,7)	9356	13 (23,7)	917	11,1 (23,5)	3274	7,2 (12,1)	125665	22,8
	N	10022	2,8 (4)	1359	2,1 (3,3)	3366	4,7 (8,5)	194	2,4 (4,9)	581	1,3 (2,2)	15522	2,8
Vnsk	G	39633	11 (15,8)	7213	11,1 (17,7)	6922	9,6 (17,6)	1027	12,5 (26,3)	3243	7,1 (12)	58038	10,6
	Įn	14636	4 (5,8)	3259	4,8 (7,8)	2550	3,5 (6,5)	213	2,6 (5,5)	1089	2,4 (4)	21747	3,9
	Vt	15157	4,2 (6)	1675	2,6 (4,1)	1930	2,7 (4,9)	283	3,4 (7,2)	537	1,2 (2)	19582	3,6
	Š	1414	0,4 (0,6)	61	0,09 (0,2)	12	0,02 (0,03)	-	-	244	0,5 (0,9)	1731	0,3
	Il	284	0,08 (0,1)	8	0,01 (0,02)	18	0,03 (0,05)	-	-	4	0,09 (0,01)	314	0,06
	V	25840	7,2 (23,7)	8332	12,8 (34,6)	8680	12 (35,6)	261	3,2 (26,4)	10 773	23,7 (58,4)	53886	9,8
	K	46063	12,8 (42,4)	7310	11,3 (30,3)	7869	11 (32,3)	355	4,3 (36)	3404	7,5 (18,5)	65001	11,8
	N	4642	1,3 (4,3)	931	1,4 (3,9)	1875	2,6 (7,7)	27	0,3 (2,7)	561	1,2 (3)	8036	1,5
Dgsk	G	19797	5,5 (18,2)	4613	7,1 (19,1)	3912	5,4 (16,1)	201	2,4 (20,4)	2375	5,2 (12,9)	30898	5,6
	Įn	7884	2,2 (7,3)	2041	3,3 (8,5)	1227	1,7 (5)	123	1,5 (12,5)	804	1,8 (4,4)	12079	2,2
	Vt	4189	1,2 (3,8)	863	1,3 (3,6)	813	1,1 (3,3)	20	0,2 (2)	290	0,6 (1,6)	6175	1,1
	Š	300	0,08 (0,3)	19	0,03 (0,08)	-	-	-	-	227	0,5 (1,2)	546	0,1
	Il	2	0,0006 (0,002)	-	-	-	-	-	-	-	-	2	0,0004
	V	-	-	-	-	2740	3,8 (34)	1120	13,6 (33,5)	-	-	3860	0,7
	K	-	-	-	-	1631	2,3 (20,2)	805	9,8 (24)	-	-	2436	0,4
	N	-	-	-	-	567	0,8 (7)	117	1,4 (3,5)	-	-	684	0,1
Be skai-	G	-	-	-	-	2530	3,5 (31,3)	1055	12,8 (31,6)	-	-	3585	0,7
čiaus	Įn	-	-	-	-	536	0,7 (6,6)	174	2,1 (5,2)	-	-	710	0,1
	Vt	-	-	-	-	70	0,1 (0,9)	71	0,9 (2,2)	-	-	141	0,02
	Š	-	-	-	-	-	-	-	-	-	-	-	-
	Il	-	-	-	-	1	0,001 (0,01)	-	-	-	-	1	0,0002
	V	1	0,0003 (33,3)	2	0,003 (66,7)	61	0,08 (55)	-	-	-	-	64	0,01
	K	-	-	-	-	25	0,03 (22,5)	-	-	-	-	25	0,004
	N	-	-	-	-	12	0,02 (10,8)	-	-	-	-	12	0,002
Dvis-	G	2	0,0006 (66,7)	1	0,002 (33,3)	12	0,02 (10,8)	-	-	1	0,002 (100)	16	0,003
kaita[12]	Įn	-	-	-	-	1	0,001 (0,9)	-	-	-	-	1	0,0002
	Vt	-	-	-	-	-	-	-	-	-	-	-	-
	Š	-	-	-	-	-	-	-	-	-	-	-	-
	Il	-	-	-	-	-	-	-	-	-	-	-	-
Iš viso		360345	100	64879	100	71966	100	8237	100	45474	100	550901	100

9 lentelė. **Veiksmažodžių asmenuojamųjų ir neasmenuojamųjų formų pasiskirstymas morfologiškai anotuotame tekstyne**

	Asmenuojamosios formos (50,7 proc.)			**Neasmenuojamosios formos** (49,3 proc.)	
	Sk.	**Proc.**		**Sk.**	**Proc.**
Tiesioginė nuosaka	87714	91,5	**Bendratis**	31374	33,6
Liepiamoji nuosaka	2607	2,7	**Dalyviai**	52483	56,2
Tariamoji nuosaka	5586	5,8	**Pusdalyviai**	3500	3,8
			Padalyviai	5973	6,4
			Būdiniai	42	0,04
Iš viso	95907	100		93372	100

10 lentelė. **Veiksmažodžių laikų pasiskirstymas morfologiškai anotuotame tekstyne**

Vksm formos/ laikai	**Tiesioginė nuosaka**		**Dalyviai**		**Padalyviai**		**Iš viso**	
	Sk.	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**
Esamasis	46071	52,5	26563	51,6	4049	67,8	76683	52,8
Būtasis kartinis	33172	37,8	10107	19,8	1919	32,2	45198	31,2
Būtasis dažninis	1622	1,8	27	0,05	-	-	1649	1,1
Būtasis	-	-	14487	28,2	-	-	14487	10
Būsimasis	6849	7,9	258	0,5	5	0,08	7112	4,9
Iš viso	87714	100	51442	100	5973	100	145129	100

11 lentelė. **Veiksmažodžių asmenų pasiskirstymas morfologiškai anotuotame tekstyne**

Asmenys	**Vienaskaita**		**Daugiskaita**		**Be skaičiaus**		**Iš viso**	
I	6592	11,5	5183	17,2	-	-	11775	12,3
II	3998	7	2184	7,2	-	-	6182	6,4
III	46618	81,5	22812	75,6	8520	100	77950	81,3
Iš viso	57208	100	30179	100	8520	100	95907	100

[11] Pirmasis skaičius rodo, kokią procentinę dalį užima kuris nors linksnis visoje tam tikros kalbos dalies linksnio kategorijoje. Skaičius skliausteliuose – tai tam tikros kalbos dalies kurio nors linksnio procentinė dalis, kurią tas linksnis užima tarp atitinkamai vienaskaitos, daugiskaitos, dviskaitos ar be skaičiaus kategorijos linksnių.

[12] Tradiciškai skaitvardis *du* laikomas turinčiu dviskaitos formą, bet MAT-e skaičiaus kategorija buvo nurodyta tik prie kelintinių skaitvardžių, todėl ir straipsnyje neanalizuojami dviskaitos formą turintys skaitvardžiai. Daugiausia vartojama dviskaitos formą turinčių įvardžių (*mudu, judu*), pasitaikė keli vienetiniai daiktavardžių ir būdvardžių dviskaitos atvejai.

12 lentelė. **Dalyvių pasiskirstymas morfologiškai anotuotame tekstyne**

Bendras dalyvių skaičius 52 483	**Veikiamieji**		**Neveikiamieji**		**Reikiamybės**	
	Sk.	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**
	16892	32,2	34562	65,8	1029	2

13 lentelė. **Laipsnių pasiskirstymas morfologiškai anotuotame tekstyne**

Laipsniai/kalbos dalis	**Bdvr**		**Sktv**		**Dlv**		**Prvks**		**Iš viso**	
	Sk.	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**
Nelygina-masis	62414	92,3	3027	90,5	-	-	16319	73,2	81760	87,5
Aukštesnysis	2681	4	10	0,3	21	21,2	4713	21,2	7425	8
Aukštėlesnysis	3	0,004	-	-	-	-	14	0,06	17	0,02
Aukščiausiasis	2556	3,7	307	9,2	78	78,8	1250	5,6	4191	4,5
Iš viso	67654	100	3344	100	99	100	22296	100	93393	100

14 lentelė. **Įvardžiuotinių ir neįvardžiuotinių formų pasiskirstymas morfologiškai anotuotame tekstyne**

Formos/kalbos dalis	**Bdvr**		**Sktv**		**Dlv**		**Prvks**		**Iš viso**	
	Sk.	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**
Įvardžiuotinės	4901	7,2	3217	6,2	352	1	1757	53,7	10227	6,4
Neįvardžiuotinės	62753	92,8	49266	93,8	35113	99	1517	46,3	148649	93,6
Iš viso	67654	100	52483	100	35465	100	3274	100	158876	100

15 lentelė. **Teigiamųjų ir neigiamųjų formų pasiskirstymas morfologiškai anotuotame tekstyne**

Formos/ kalbos dalis	Dktv[13]		Vksm		Bdvr		Prvks		Prln[14]		Jngt[15]		Dll[16]		Iš viso	
	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%	Sk.	%
Teigiamosios	362118	99,6	174464	92,2	65132	96,2	60152	97	42794	99,7	70325	99,9	18 196	99,9	793181	97,4
Neigiamosios	1426	0,4	14815	7,8	2558	3,8	1859	3	116	0,3	17	0,1	21	0,1	20812	2,6
Iš viso	363544	100	189279	100	67690	100	62011	100	42910	100	70342	100	18 217	100	813993	100

16 lentelė. **Kaitomų ir nekaitomų vardažodžių formų pasiskirstymas morfologiškai anotuotame tekstyne**

Formos/kalbos dalis	Dktv[17]		Bdvr[18]		Įvrd[19]		Sktv[20]		Iš viso	
	Sk.	Proc.	Sk.	Proc.	Sk.	Proc.	Sk.	Proc.	Sk.	Proc.
Kaitomosios	360346	99,2	67654	99,9	76555	95,3	8682	97,7	513237	98,6
Nekaitomosios	3198	0,8	36	0,1	3816	4,7	203	2,3	7253	1,4
Iš viso	363544	100	67690	100	80371	100	8885	100	520490	100

[13] Teigiami ar neigiami dažniausiai tik veiksmažodinės kilmės daiktavardžiai, nors gali būti ir neveiksmažodinių neigiamų daiktavardžių, pvz.: *nelaimė, nedarbas* ir kt. Daugumai daiktavardžių, pvz., *namas, ranka, lėkštė* ir kt., teigiamumo ir neigiamumo priešprieša yra visai nebūdinga. Todėl, kad ji nebūtų be reikalo žymima, MAT-e prie neigiamų daiktavardžių nurodytas tik neigiamumas, o visi kiti daiktavardžiai sąlygiškai gali būti laikomi teigiamais, nors tai nepažymėta.

[14] Neigiamų prielinksnių yra tik vienas kitas. Dažniausiai vartojami prielinksniai *netoli, nepaisant*.

[15] MAT-e pasitaikė tik keli neigiami jungtukai – tai *nelyginant, nelyg*.

[16] MAT-e rasta tik keletas neigiamos dalelytės *nejaugi* vartosenos atvejų.

[17] Nekaitomais daiktavardžiais MAT-e laikomi, pvz., tokie daiktavardžiai: *taksi, ledi, ego*. Kai kurie kalbininkai teigia, kad ir nekaitomi daiktavardžiai turi giminę, kuri paaiškėja iš gretimų žodžių derinimo. Autorė sutinka su tokia nuomone, bet tam, kad būtų išvengta MD-o, MAT-e nekaitomiems daiktavardžiams nebuvo nurodomos jokios gramatinės kategorijos, išskyrus kalbos dalį.

[18] Nekaitomais būdvardžiais MAT-e laikomi, pvz., *mini, bordo*. Jiems nenurodoma nei giminė, nei skaičius, nei linksnis.

[19] Nekaitomais įvardžiais MAT-e laikomi *keliasdešimt, mano, tavo, savo* ir kt. įvardžiai. Jiems nenurodoma nei giminė, nei skaičius, nei linksnis.

[20] Nekaitomais skaitvardžiais MAT-e laikomi, pvz., *dvidešimt, septyniasdešimt* ir pan. Jiems nenurodoma nei giminė, nei skaičius, nei linksnis. Kartais šie skaitvardžiai yra linksniuojami, pvz., *dvidešimties*. Tokiu atveju jie yra laikomi kaitomais skaitvardžiais.

17 lentelė. **Tikrinių ir bendrinių daiktavardžių pasiskirstymas morfologiškai anotuotame tekstyne**

Bendras daiktavardžių skaičius 363544	**Tikriniai**		**Bendriniai**	
	Sk.	**Proc.**	**Sk.**	**Proc.**
	39890	11	323654	89

18 lentelė. **Skaitvardžių rūšių pasiskirstymas anotuotame tekstyne**

Bendras skaitvardžių skaičius 8885	**Kiekiniai pagrindiniai**		**Kiekiniai dauginiai**		**Kiekiniai kuopiniai**		**Kelintiniai**	
	Sk.	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**
	5158	58	355	4	26	0,3	3346	37,7

19 lentelė. **Sangrąžinių ir nesangrąžinių formų pasiskirstymas morfologiškai anotuotame tekstyne**

Formos/kalbos dalis	**Dktv**[21]		**Vksm**		**Iš viso**	
	Sk.	**Proc.**	**Sk.**	**Proc.**	**Sk.**	**Proc.**
Sangrąžinės	2591	28,3	21759	11,5	24350	12,3
Nesangrąžinės	6579	71,7	167520	88,5	174099	87,7
Iš viso	9170	100	189279	100	198449	100

Literatūra

GKT, 2004 – *Gramatinių kategorijų tyrimai*. Lietuvių kalbos gramatikos darbai. 2 (red. A. Holvoet, L. Semėnienė). Vilnius: Lietuvių kalbos institutas.

Grumadienė L., 2002 – Dabartinės kalbos dažninis žodynas ir jo bazė. *Acta Linguistica Lithuanica*. T. 46. P. 19–37.

Marcinkevičienė R., 2000 – Tekstynų lingvistika (teorija ir praktika). *Darbai ir Dienos*. T. 24. P. 7–64.

Mauricaitė V., Norkaitienė M., Pakerys A., Petrokienė R., 2004 – *Bendriniai XX a. spaudos žodžiai*. Elektroninis dažninis žodynas. Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Popelînskô L., Pavelek T., Ptâčnîk T., 2000 – *On Disambiguation in Czech Corpora. FIMU-RS-2000-07*, Faculty of Informatics MU Brno. – http://www.muni.cz/veda/reports/files/older/ FIMU-RS-2000-07.pdf

Przepiórkowski A., 2004 – *Korpus IPI PAN wersja wstępna*. Warszawa: Instytut podstaw informatyki PAN.

Rimkutė E., 2003 – Morfologinio daugiareikšmiškumo tipologija. *Lituanistica*. Nr. 4(56). P. 60–78.

Rimkutė E., 2004 – Dar kartą apie iliatyvą. *Kalbos kultūra*. Nr. 77. P. 124–128.

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

[21] Sangrąžinės ar nesangrąžinės formos būdingos tik veiksmažodinės kilmės daiktavardžiams.

Zinkevičius V., 2000 – *Lemuoklis* – morfologinei analizei. *Darbai ir Dienos*. T. 24. P. 246–273.

Zinkevičius V., Daudaravičius V., Rimkutė E., 2005 – The Morphologically annotated Lithuanian Corpus. *Tarptautinės konferencijos „The Second Baltic Conference on Human Language Technologies" pranešimų medžiaga*. Talinas. P. 365–370.

Žilinskienė V., 2001 – Lietuvių ir latvių kalbų publicistikos leksikos ir morfologijos pagrindinės statistinės charakteristikos. *Lituanistica*. Nr. 3(47). P. 69–79.

Žilinskienė V., 2002a – Žodžių formų vartojimas lietuvių kalbos dalykinio ir publicistinio stilių duomenimis. *Lituanistica*. Nr. 1(49). P. 106–116.

Žilinskienė V., 2002b – Gramatinių formų vartojimas lietuvių kalbos moksliniame stiliuje. *Acta Linguistica Lithuanica*. T. 46. P. 173–183.

Žilinskienė V., 2003 – Gramatinių formų vartojimas lietuvių kalbos tekstuose. *Kazimiero Jauniaus konferencijos pranešimų tezės*. Lietuvių kalbos institutas. Vilnius–Kvėdarna – http://www.lki.lt/skelbimai/Tezes.pdf

Žilinskienė V., 2005 – Vardažodžių, įvardžių ir jų gramatinių formų vartojimas lietuvių kalbos stiliuose. *Lituanistica* 4(64). P. 28–44.

Erika Rimkutė

THE USAGE OF GRAMMATICAL FORMS OF THE CONTEMPORARY LITHUANIAN LANGUAGE IN THE MORPHOLOGICALLY ANNOTATED CORPUS

Summary

This paper deals with the usage of parts of speech and their grammatical features in the morphologically annotated corpus of the Lithuanian language. This corpus was compiled and processed at the Center of Computational Linguistics of Vytautas Magnus University. The morphologically annotated corpus is a set of XML files, containing 1 million morphologically annotated running words. Each annotation for a word form contains its normalized form (lemma) and a full set of morphological properties. Non-word textual units, such as punctuation marks, spaces, paragraphs, numbers, etc. are represented in the morphologically annotated corpus by special marks.

The morphologically annotated corpus showed out that the variety of inflectional forms in real usage is not so great as in the grammatical system, since highly inflected parts of speech as verbs and nouns have less than 3 word-forms on average. Pronouns demonstrated a surprisingly big number of word forms actually used in the contemporary Lithuanian language. Overall, the tendencies for the usage of different word classes coincide with the data obtained by other researches, i.e. nouns and other nominal words have the biggest coverage (39% are nouns, 8.7% pronouns, 7.33% adverbs, 6.72% adjectives, and 20% are verbs).

The morphologically annotated corpus is of great importance for the future development of parsing tools, treebanks and other resources of the Lithuanian language.