

# ***Dabartinės lietuvių kalbos tekstynas*** **– 10 metų kaupimo ir naudojimo patirtis**

JOLANTA KOVALEVSKAITĖ

*Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras*

The paper overviews the development of the *Corpus of the Contemporary Lithuanian Language* compiled at the Center of Computation Linguistics, Vytautas Magnus University during last 10 years.

The Lithuanian language corpus is described as a large (100 million running words in its Internet version), balanced, constantly being enlarged, unannotated corpus of full texts, aimed at various applications. The sub-products (available on <http://donelaitis.vdu.lt>) of the corpus include regularly updated word frequency lists, concordances and the list of collocations for every query.

The problematic issues of the characteristics such as size, dynamics vs. continuity and representativeness of the large general corpus of Lithuanian are discussed in a greater detail. The Lithuanian language corpus is balanced regarding the production of texts. The main texts-selection criteria for the corpora, e. g. readability, popularity, demographic indicators, etc., are presented here.

The present state of art, the exploitation (numbers of queries) and the next stage in the development of the *Corpus of the Contemporary Lithuanian Language* is discussed as well.

## **1. ĮVADAS**

Kaip žinia, tekstynas dažniausiai apibrėžiamas kaip (baigtinis ar tęstinis) didelis elektroninių tekstų rinkinys, sudarytas taip, kad kuo geriau atspindėtų kalbą ar jos atmainą.

*Dabartinės lietuvių kalbos tekstynas* buvo pradėtas kurti prieš 10 metų VDU Kompiuterinės lingvistikos centre. *Dabartinės lietuvių kalbos tekstyno* atsiradimą lėmė naujų technologijų padiktuoti kalbos tyrimo metodai, kurių taikymui buvo ir yra reikalinga autentiškų rašomosios lietuvių kalbos pavyzdžių bazė.

Viešai prieinamas internete, šis tekstynas dabar yra viena iš didžiausių lietuvių kalbos kalbinių duomenų kaupyklų, kuria naudojasi įvairių sričių specialistai, atlikdami tyrimus Lietuvoje ir užsienyje. *Dabartinės lietuvių kalbos tekstynas* yra bendro pobūdžio, neanotuotas, ištisu periodikos ir knygų tekstų, didelės temų ir žanrų įvairovės tekstynas.

Šiame straipsnyje, naudojantis dydžio, reprezentatyvumo ir tęstinumo kriterijais, aptariami probleminiai klausimai, susiję su *Dabartinės lietuvių kalbos tekstyno* sandara, trumpai pristatomos paieškos tekстыne galimybės, priešpaskutinėje dalyje supažindinama su tekstyno ateities gairėmis, o pabaigoje pateikiamos išvados.

## 2. TEKSTYNO YPATYBĖS

Iš *Dabartinės lietuvių kalbos tekstyno* darbo ir naudojimosi patirties kyla probleminių klausimų, susijusių su tekstyno dydžiu, sandara bei tekstų atrankos principais. Ši straipsnio dalis ir yra skirta šių klausimų aptarimui.

### 2.1. DYDIS IR TĖSTINUMAS

Tekstyno dydis yra vienas iš tų kriterijų, dėl kurio neretai ginčijamasi. *Dabartinės lietuvių kalbos tekstynas* savo internetine versija jau yra peržengęs 100 mln. žodžių skaičių ir yra kas metai papildomas apie 10 mln. žodžių, taigi jo dydis prilygsta *British National Corpus* apimčiai (pastarąjį sudaro 100 mln. žodžių). Kita vertus, jeigu palyginsime lietuviškąjį tekstyną su *Bank of English* (beveik 500 mln. žodžių), su vienu didžiausiu Europoje laikomu čekų kalbos tekstynu (arti 200 mln. žodžių) ar su dabartinės rašomosios vokiečių kalbos tekstynu (*Korpora geschriebener Gegenwartssprache des IDS*, beveik 2 mlrd. žodžių), kils klausimas, kurį užduoda dauguma su tekstynu susiduriančių vartotojų: kokio dydžio turi būti tekstynas, kad jo pakaktų? Ar yra riba, kurią pasiekus, galima sustoti? Anot J. Sinclairio (žr. Marcinkevičienė 2000a, 9), *maži tekstynai nėra tokie patys kaip ir dideli, tik kitos apimties. Tekstyno tekstų kiekis lemia ne tik jo kiekybę, bet ir kokybę*. Sutariama, kad kuo didesnis yra tekstynas, tuo geriau jame išryškėja kolokatų struktūros, kalbos vartosenos faktai ir dėsniumai, kuriuos sunku išvelgti mažame ir/ar specializuotame tekстыne.

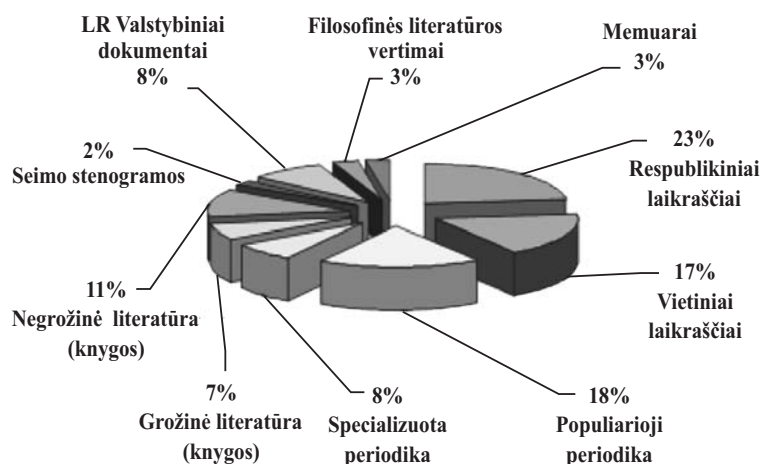
*Dabartinės lietuvių kalbos tekstyno* sudarytojai laikosi nuomonės, kad iš dviejų kruopščiai sudarytų tekstynų didesnis visada bus geresnis, nes tik jis parodys tipišką ir net rečiau vartojamą kalbos vienetų vartoseną (Marcinkevičienė 2000a, 9). Ši strategija sieja du tekstyno kriterijus – dydį ir tęstinumą, nors jokių būdu neneigia tekstyno kokybės, t. y. tekstyno reprezentatyvumo svarbos.

### 2.2. REPREZENTATYVUMAS

Vienu iš didelio tekstyno privalumų galėtume laikyti tai, jog tokiaame tekстыne labiau tikėtina rasti didesnę literatūros tipų, žanrų ir temų įvairovę. Tačiau net ir didelio tekstyno reprezentatyvumo klausimas tebėra diskusijų objektas (Čermák 2002, 269), o tai, kiek tekstynas yra reprezentatyvus, nemaža dalimi lemia tekstyno šaltinių įvairovė, kuri priklauso nuo pasirinktų atrankos kriterijų. Nepaisant to, kad tų kriterijų esama pačių įvairiausių, – elitiškumas, skaitomumas, demografiniai rodikliai, prieinamumas, subjektyvūs tekstyno rengėjų sprendimai, kriterijai, paremti aktyviaja ir pasyviaja kalbos vartoseną (plačiau žr. Marcinkevičienė 2000a, 11), praktikoje daugelis reprezentatyviais save laikančių tekstynų iš tiesų nėra sudaryti pagal aiškius kriterijus.

*Dabartinės lietuvių kalbos tekstyno* vartotojus taip pat dažnai domina klausimas apie tekstyno formavimo principus, o jeigu jie nėra aiškiai išdėstomi – tai laikoma tekstyno trūkumu. Kadangi lietuvių kalbos tekstynas kuriamas daugiausiai iš rašytinės kalbos tekstų (išskyrus LR Seimo stenogramas), jis reprezentuoja rašytinę kalbos atmainą.

Šiuo metu internete prieinamo *Dabartinės lietuvių kalbos tekstyno* sandara parodyta 1 paveiksle. Iš pradžių (nuo 1990 m.) tekstyne buvo kaupiami nepriklausomybės laikotarpio originaliosios neverstinės kalbos tekstai, vėliau imti kaupti filosofinės literatūros vertimai į lietuvių kalbą, dar vėliau tekstyno archyvas imtas pildyti kitais verstiniais grožinės ir mokslinės bei mokomosios literatūros tekstais.



1 pav. Šiuo metu internete esančio *Dabartinės lietuvių kalbos tekstyno* sudėtis

Leidinių, dedamų į *Dabartinės lietuvių kalbos tekstyną*, atrankos kriterijai įvairuoja priklausomai nuo teksto tipo – ar tai yra grožinė, ar mokslinė, populiarioji ar vietinė periodika. Pavyzdžiui, atrenkant mokslinės literatūros tipą reprezentuojančius tekstus, atsižvelgiama į elitiškumo kriterijų, o vykdant grožinės ir mokslo populiarinimo literatūros atranką, vertinamas knygos tiražas, skaitomumas, populiarumas. Atrenkant vietinės periodikos tekstus, siekiama kuo didesnės demografinės vietų (Lietuvos regionų) įvairovės, o štai kaupiant populiariają periodiką, kaip, beje, ir kitas tekstų rūšis, svarbi yra temų įvairovė. Kai kurie atrankos kriterijai jau susiformavę, pagal juos atrenkama didžioji tekstų dalis: knygos, periodinio leidinio populiarumas/skaitomumas, knygos tiražo dydis, reprezentatyvumas (kurį sudaro ir elitiškumas, ir demografiniai rodikliai). Kiti kriterijai ilgainiui keičiasi arba vietoj jų pasirenkami tinkamesni.

Vieni tyrinėtojai (pvz., Kennedy 1998, 62, žr. Marcinkevičienė 2000a, 11) tekstyno reprezentatyvumą vadina nuomonės reikalu, kiti, pavyzdžiui, J. Sinclairis (1991, 61), teigia, kad tęstinių tekstynų, koks yra ir *Dabartinės lietuvių kalbos tekstynas*, dydis atstoja kitų tekstynų subalansuotumą kaip svarbiausią tekstyno sandaros principą, kadangi didelis tekstų kiekis savaime sudaro sąlygas didesnei kalbos vienetų vartojimo įvairovei. Kitaip tariant, tam, kad būtų reprezentatyvus, tekstynas turi būti didelis, o tai dar kartą įrodo, kaip glaudžiai susiję tekstyno dydis, tęstinumas ir reprezentatyvumas.

*Dabartinės lietuvių kalbos tekstynas* laikomas subalansuotu, nes iš turimo tekstų archyvo į tekstyną pagal pasirinktas proporcijas perkeliama tik dalis tekstų (Marcinkevičienė 2000a, 16). Proporcijas lemia ne tik pačių tekstyno rengėjų pasirinkti kriterijai, bet ir leidybos tendencijos, galimybės gauti tekstus. Tačiau netgi subalansuoto ir įvairaus tekstyno reprezentatyvumas gali kelti abejonių, ypač jeigu šis tekstynas yra bendras. Tokius tekstynus sunku parengti, nes jų adresatas yra visai neaiškus. Kai sunku tiksliai nurodyti adresatą, sudėtinga nuspręsti, kokie kriterijai turi būti pasirenkami tekstyno sandarai lemti (žr. Marcinkevičienė 2000a, 17). Pabrėžtina ir tai, kad kuo mažesnis tekstynas, tuo lengviau jam pritaikyti griežtesnius reprezentatyvumo kriterijus, nes tokio tekstyno naudojimo tikslas yra aiškus. Dideliam tekstynui taikomi sandaros kriterijai gali/turi būti ne tokie griežti, nes jis yra kuriamas platesniam vartotojų ratui, įvairesniems tikslams.

Tekstyno reprezentatyvumą lengviau patikrinti negu pasiekti. Vienas iš būdų yra *type-token ratio* – tekstyno santykis tarp visų ir skirtingų žodžių<sup>1</sup>. Daugelis naudojami ir Ch. Fillmore nurodytu būdu, kurio esmė tokia: jei aš žinau kokį dalyką egzistuojant kalboje, tai jis turi atsispindėti ir tekстыne (žr. Marcinkevičienė 2000a, 12). Jei tekstyने randi ne viską, vadinasi, jį reikia didinti, įvairinti, kad ir koks reprezentatyvus jis būtų. F. Čermáko nuomone (2002, 267), netgi bilijoninės apimties tekstyną reikia didinti, kad jis būtų kiek įmanoma reprezentatyvesnis. Ir tai suprantama, nes net ir bilijono žodžių tekstynas yra per mažas, palyginti su visa kalbos įvairove.

### 3. DABARTINĖS LIETUVIŲ KALBOS TEKSTYNO GALIMYBĖS

Kaip žinia, bet kuri tekstyną sudaro ne tik tekstai, bet ir programiniai įrankiai (pvz., dažninių sąrašų generatoriai, konkordavimo programos ir kt.), nuo kurių

<sup>1</sup> *Dabartinės lietuvių kalbos tekstyno* leksinės įvairovės/žodingumo kreivė (*saturation curve*) parodo, kad tekstynui pasiekus tam tikrą apimtį, *token* (žodžių skaičiaus) didėjimas nedaro įtakos *types* (skirtingų žodžių skaičiui) (plačiau apie tai žr. Marcinkevičienė 2000a: 20, Marcinkevičienė, Bielinskienė, Daudaravičius, Rimkutė 2004).

priklauso naudojimosi tekstynu sąlygos ir pritaikymo galimybės (plačiau apie kalbinės programinės įrangos rūšis žr. Utkā 2000).

Administruojant tekstyną yra naudojamos ir specialiomis tekstyno priežiūrai skirtomis programomis. Darbui su VDU *Dabartinės lietuvių kalbos tekstynu* naudojama universali programinė įranga – Mike'o Scotto parengtas programinių įrankių paketas *WordSmith Tools* ir programa CUE (*Corpus Universal Examiner*) bei įranga, pritaikyta specialiai lietuvių kalbai (plačiau žr. Marcinkevičienė 2000a, 17). Kuriant tekstyną, daugumą iš leidyklų gautų tekstų reikia perkelti į kitus formatus, paruošti kompiuterinei analizei. Tokie darbai reikalauja didelių laiko sąnaudų, todėl kai kurie jų yra palengvinami specialiai techniniams darbams sukurtomis programomis: viena jų yra V. Daudaravičiaus parengta internetinių laikraščių tekstų kompiliavimo ir apdorojimo programa *Kolektorius*, automatiškai papildanti tekstyną naujais teksta.

Naudojant Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre sukurtą žodžio paieškos ir statistikos programą, galima atlikti automatinę tekstyno analizę ir gauti žodžių, žodžių formų dažninius sąrašus (*frequency lists*) (plačiau apie dažninių sąrašų tipus ir apie tai, kaip juos pritaikyti žodžio reikšmės analizei žr. Marcinkevičienė 2000a, Utkā 2000).

Su tekstyno konkordavimo programa galima rinkti informaciją apie pasirinktų kalbos vienetų distribuciją: ši programa randa ir stabilius žodžių junginius (kolokacijas<sup>2</sup>), paruošia konkordansus<sup>3</sup>.

Pavyzdžiui, pasirinkus žodį *laikas* galima automatiškai atlikti šio žodžio (ir jo formų) analizę. Programa gali iš pradžių pateikti dažnumų sąrašą, po to konkordansą arba tik konkordansą, arba tik stabilų junginį.

1 lentelėje matyti, kaip atrodo žodžio *laikas* dažnumų sąrašas kiekvienoje tekstyno dalyje.

1 lentelė. Žodžio *laikas* dažnumų sąrašas

Tekstyno dalis	Žodis	Pavartojimo skaičius	Bendras pavartojimo skaičius	Iš viso žodžių
Respublikinė periodika	LAIKAS	4994	4994	24803732
Vietinė periodika	LAIKAS	3355	3355	16918103
Populiarioji periodika	LAIKAS	3440	3440	16582983
Specializuota periodika	LAIKAS	1702	1702	9760811

<sup>2</sup> Kolokacija – tai mažiau už idiomias sustabarėję žodžių junginiai, sudarantys labai didelę kalbos apyvartos dalį (plačiau apie kolokacijas ir fraziškumo (*idiomaticy*) principą žr. Marcinkevičienė 1995, 1997a, 2000a, 2000b).

<sup>3</sup> Generuojant konkordansą, gaunami visi atskirų leksemos formų vartojimo atvejai minimaliame vienos eilutės kontekste (plačiau apie konkordansų tyrimo ypatumus ir konkordavimo programų galimybes žr. Marcinkevičienė 1997a, 2000a, Utkā 2000).



Be to, galima atlikti ir skirtingų žodžio *laikas* formų vartosenos tyrimą, paieškos laukelyje naudojant tam tikrus sutartinius simbolius (plačiau žr. <http://donelaitis.vdu.lt>).

Atliekant stabiliųjų žodžių junginių analizę, galima nurodyti tik vieną ieškomą žodį ar jo dalį, be to, nurodyti kontekstinį žodį, kuris gali būti aptinkamas kairėje (*left collocates*), dešinėje (*right collocates*) arba abiejose pagrindinio ieškomo žodžio pusėse (3 lentelėje pateikiama keletas žodžio *laikas* kolokatų).

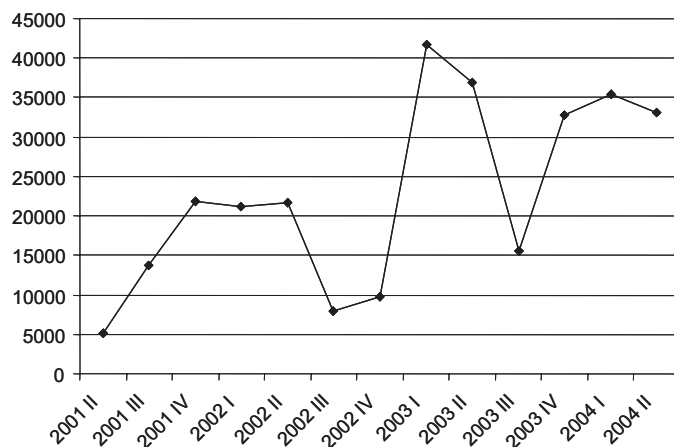
**3 lentelė. Stabiliųjų junginių su žodžiu *laikas* sąrašo dalis pagal pasirinktą santykinio dažnumo koeficientą 50 (apie šį koeficientą plačiau žr. <http://donelaitis.vdu.lt>, paieška tekстыne)**

Stabilus junginys	Dažnumas
ar laikas	27
artėja laikas	35
ateina laikas	233
ateis laikas	474
atėjęs laikas	31
atėjo laikas	794
atlikimo laikas	53
atsipirkimo laikas	40
bausmės laikas	71
bei laikas	27
bet laikas	134
bėga laikas	54
bėgo laikas	42
blogas laikas	49
bus laikas	31
buvo laikas	222

PASTABA: prie kiekvieno stabiliojo žodžių junginio pateikiamas to junginio pavartojimo dažnumas

*Dabartinės lietuvių kalbos tekstyno* duomenų bazė tinka ir tekstynais paremtai kalbos analizei (*corpus-based approach*), kurią naudoja ne tik tekstynų lingvistikos, bet ir kitų mokslo sričių tyrėjai, ir tekstyno inspiruotai analizei (*corpus-driven approach*), kurią dažniau renkasi tekstynų lingvistai. Apie lietuvių kalbos tekstyno analizės priemones, darbo su tekstynais ypatumus bei plačias tekstyno pritaikymo galimybes savo straipsniuose išsamiai kalba R. Marcinkevičienė (1997a, 1997b, 2000a), A. Utkā (2000), V. Daudaravičius (2001).

Įdomu tai, jog *Dabartinės lietuvių kalbos tekstyno* pritaikymo įvairovę rodo vis augantis jo vartotojų bei paieškų skaičius. Statistika rodo, kad spartus vartotojų skaičiaus augimas prasidėjo nuo 2001 m. pabaigos, o 2002 m. pabaigoje 2003 m. pradžioje, palyginti su 2001 metais, jis jau buvo išaugęs dvigubai (iki 200 vartotojų per mėnesį) (žr. Marcinkevičienė, Bielinskienė, Daudaravičius, Rimkutė 2004). Paieškų skaičiaus kitimus galima matyti 2 paveiksle.



**2 pav. Paieškų skaičius tekstyne ketvirčiais (nuo 2001 m. II ketvirčio iki 2004 m. II ketvirčio)**

Tyrimų institutai, universitetai, mokslo centrai, privačios kompanijos vis dažniau naudojami tekstyno privalumais – autentiška kalba, objektyvia ir greita analize, gausiais duomenimis.

#### **4. TEKSTYNO ATEITIES GAIRĖS**

Vertinant didelio tekstyno privalumus, pravartu išvelgti ir naudotis ir mažų specializuotų tekstynų galimybėmis. Todėl lietuvių kalbos tekstyno sudarytojai, neatsisakydami didelio tekstyno idėjos, kartu su didžiuoju tekstynu baigia parengti keletą mažesnių tekstynų, iš kurių vienas yra morfologiškai anotuotas (1 mln. žodžių) (plačiau žr. Zinkevičius, Daudaravičius, Rimkutė 2005), o kitas – paralelus, pritaikytas vertimų analizei (Marcinkevičienė, Bielinskienė, Daudaravičius, Rimkutė 2004). Didįjį (DLKT) tekstyną artimiausiu metu žadama papildyti naujausiais duomenimis, sukaupti daugiau sakinės kalbos atmainą reprezentuojančių tekstų bei internetinės kalbos tekstų.

Kadangi tekstyno vartotojams svarbi informacija apie tekstyną sudarančių tekstų bibliografiją, *Dabartinės lietuvių kalbos tekstyno* rengėjai netrukus planuoja taikyti naują indeksavimo sistemą, kuri leistų prie kiekvienos konkordanso eilutės matyti ne tik informaciją apie tai, koks tekstas (kuriame buvo rastas užklausoje nurodytas žodis) pavadinimas, autorius, leidykla etc., bet ir tai, kokiam literatūros tipui, žanrui, temai priklauso šis tekstas.

## 5. IŠVADOS

Kalbos kompiuterizavimas yra sudėtingas procesas, tačiau sparčiai tobulėjančios kalbinės programinės įrangos rūšys leidžia naujais metodais tyrinėti kalbos reiškinių, tobulinti tekstyną ir jo pagrindu kurti ir daugeliui kitų mokslo šakų naudingus, autentiškos kalbos vartosenos tyrimais pagrįstus, produktus – leksines duomenų bazes, specializuotus tekstynus, kompiuterinius leksikonus, kontekstinius žodynus ir pan.

Per daugiau negu dešimt metų atliekant tekstyno pildymo, administravimo bei analizės darbus, įgyta daug patirties, kaupiama vis solidesnė duomenų bazė, o tai savo ruožtu sudaro vis daugiau sąlygų Lietuvoje plėtoti tekstynų lingvistikai. Per dešimtmetį *Dabartinės lietuvių kalbos tekstynas* tapo visuotinai pripažintu įvairialypiu duomenų šaltiniu, Lietuvos interneto kultūros dalimi.

## LITERATŪRA

- ČERMÁK, FRANTIŠEK. 2002. Today's corpus linguistics. Some open questions. *International Journal of Corpus Linguistics* 7:2. 265–282.
- DAUDARAVIČIUS, VIDAS. 2001. Interneto teksto paruošimas automatinei analizei, in: *Proceedings of 6th conference of master and doctoral students*, Vytautas Magnus University. 2001 04 26.
- MARCINKEVIČIENĖ, RŪTA. 1995. Kolokacija: tyrimo kryptys, metodai, mokyklos. *Lituanistica* 2. 40–54.
- MARCINKEVIČIENĖ, RŪTA. 1997a. Klausimas dėl *klausimo*, arba ką gali kompiuterinis tekstynas. *Darbai ir Dienos* 5. 19–37.
- MARCINKEVIČIENĖ, RŪTA. 1997b. Tekstynų lingvistika ir lietuvių kalbos tekstynas. *Lituanistica* 1 (29). 58–78.
- MARCINKEVIČIENĖ, RŪTA. 2000a. Tekstynų lingvistika: teorija ir praktika. *Darbai ir Dienos* 24. 7–64.
- MARCINKEVIČIENĖ, RŪTA. 2000b. Patterns of word usage in corpus linguistics. *Kalbotyra* 49 (3). 71–80.
- MARCINKEVIČIENĖ, RŪTA & BIELINSKIENĖ, AGNĖ & DAUDARAVIČIUS, VIDAS & RIMKUTĖ, ERIKA. 2004. Corpora of Lithuanian Language Technologies. *Proceedings of the First Baltic Conference Human Language Technologies. The Baltic Perspective*. Riga. 21–24.
- UTKA, ANDRIUS. 2000. Kalbinė įranga ir jos galimybės. *Darbai ir Dienos* 24. 275–284.
- ZINKEVIČIUS, VYTAUTAS & DAUDARAVIČIUS, VIDAS & RIMKUTĖ, ERIKA. 2005 (atiduota spausdinti). The morphologically annotated Lithuanian Corpus. *Proceedings of the Second Baltic Conference Human Language Technologies*, Tallinn.