

KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas

Erika Rimkutė, Vidas Daudaravičius

Anotacija. Didėjant informacinių technologijų plėtrai, spartėjant kalbos kompiuterizavimo darbams, iškilio būtinybė kurti didelius anotuotus tekstynus tam, kad būtų galima pasinaudoti jų duomenimis pereinant į aukštesnius kalbos kompiuterizavimo lygmenis (pvz., automatinę sintaksinę ir semantinę analizę, mašininį vertimą). Straipsnyje pristatomi didelio lietuvių kalbos tekstyno automatinio morfologinio vienareikšminimo tyrimai ir anotavimo rezultatai. Remiantis statistiniais metodais, pavyko sukurti lietuvių kalbos morfologinio anotavimo priemonę, kurios vienareikšminimo tikslumas siekia 94%, ir taip išspręsti didelį lietuvių kalbos morfologinį daugiareikšmiškumą. Pateikiami statistiniai duomenys apie kalbos dalių pasiskirstymą anotuotame tekстыne, dažniausias žodžių formas ir dažniausias lemas (antraštines formas), taip pat išvardytos dažniausios kaitomos kalbos dalys, išrinktos iš morfologiškai anotuoto *Dabartinės lietuvių kalbos tekstyno*.

Raktiniai žodžiai: *morfologinis anotavimas; tekstynas; morfologinė analizė; daugiareikšmiškumas; statistinis morfologinis vienareikšminimas.*

Įvadas

Dabartinės lietuvių kalbos tekstynas – plačiai Lietuvoje naudojama duomenų bazė, reprezentatyviai atspindinti dabartinę lietuvių kalbą, įvairius jos stilius (plačiau žr. <http://donelaitis.vdu.lt/>). Šis iš 100 mln. žodžių sudarytas tekstynas 2006 m. buvo morfologiškai anotuotas. Tekstyno anotavimo darbai atlikti 2005-2006 m. vykdant Lietuvos valstybinio mokslo ir studijų fondo projektą „Lietuvių kalbos išlikimas globalizacijos sąlygomis: anotuotas lietuvių kalbos tekstynas (ALKA)“. Internetu šis morfologiškai anotuotas tekstynas nėra publikuojamas.

Prieš dvidešimt metų V. Zinkevičiaus sukurtas lietuvių kalbos morfologinės analizės įrankis analizuoja tik žodžių formas ir nesprenžia morfologinio daugiareikšmiškumo problemos, kuri lietuvių kalboje yra aktuali. Atsižvelgiant į sėkmingą pasaulinę patirtį sprendžiant morfologinio daugiareikšmiškumo problemą statistiniais metodais, šiuos metodus pabandyta pritaikyti ir lietuvių kalbai. Straipsnyje pristatomi didelio lietuvių kalbos tekstyno automatinio morfologinio vienareikšminimo tyrimai ir anotavimo rezultatai. Šio straipsnio tikslas – aprašyti lietuvių kalbos morfologinio anotavimo priemonę, kurios vienareikšminimo tikslumas 94%.

Morfologinės lietuvių kalbos analizės automatizavimas

Lietuvių kalbai yra sukurta gerai veikianti automatinė morfologinės analizės programa *Lemuoklis*, žodžio formai pateikianti antraštinį pavidalą (lemą) ir gramatinę pažymą (plačiau apie programą ir jos veikimo principus žr. Zinkevičius 2000). Tačiau anotuojant su *Lemuokliu* nepavyksta išspręsti morfologinio daugiareikšmiškumo. Morfologiškai daugiareikšmės žodžių formos yra žodžiai ar jų formos, kurias išanalizavus pateikiamos dvi ar daugiau lemu, pvz.,

formai *kovų* nurodo lemas *kovas* ir *kova*, arba dvi ar daugiau morfologinių pažymų¹, pvz., *naktis* – vienaskaitos vardininkas arba daugiskaitos galininkas (plačiau apie morfologinį daugiareikšmiškumą žr. Rimkutė 2006a).

Lietuvių kalbos morfologinis daugiareikšmiškumas yra aktuali problema automatinės analizės programoms, nes beveik pusė žodžių anotuotuose tekstuose yra morfologiškai daugiareikšmiai (plačiau žr. Rimkutė 2006a). Rengiant nedidelės apimties (1 mln. žodžių) morfologiškai anotuotą tekstyną, kuris yra būtinas norint automatiškai vienareikšminti, reikia daugybės laiko. Rankiniu būdu vienareikšmintą morfologiškai anotuotą lietuvių kalbos tekstyną rengė vienas žmogus ir tam prireikė penkerių darbo metų. Pradžioje reikia priprasti prie anotavimo formato (žr. 1 lentelę); būtina nuspręsti, kokius žodžius priskirti kokiai kalbos daliai, nes yra daugybė atvejų, kai sunku iš karto nustatyti, kokią gramatinę informaciją pateikti vienam ar kitam žodžiui. Nemažai laiko sugaištama peržiūrint ir siekiant suvienodinti visus anotuotus tekstus.

Automatinis morfologinis tekstynų anotavimas

Rengiant 1 mln. žodžių morfologiškai anotuotą lietuvių kalbos tekstyną (plačiau apie šį tekstyną žr. Zinkevičius et al. 2005), suprasta, kad reikia ieškoti būdų, kaip automatiškai galima sumažinti morfologinį daugiareikšmiškumą. 2005-2006 m. trukusio Lietuvos valstybinio mokslo ir studijų fondo finansuojamo projekto metu buvo nagrinėti statisti-

¹ 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), laikoma viena morfologine pažyma.

niai lietuvių kalbos modeliai ir jų taikymas automatinio morfologinio vienareikšminimo problemai spręsti.

Statistinio morfologinio anotavimo užduotis naudojant rankiniu būdu anotuotus nedidelius tekstynus iš pirmo žvilgsnio atrodo gana paprasta – suskaičiuojami požymių dažnumai (anotavimo sistemos mokymas) ir ieškoma labiausiai tikėtina morfologinių požymių seka naujame tekste naudojant įvairius tikimybių skaičiavimo būdus. Tokių morfologinio anotavimo priemonių kūrimo anglų, čekų ir kitoms kalboms patirtis rodo, kad tekstynų naudojimas, kuriant automatinę lietuvių kalbos morfologinio anotavimo priemonę, yra labai svarbus. Ypač svarbu tinkamai parengti pirminius duomenis (nedidelį rankiniu būdu morfologiškai anotuotą tekstyną).

Kuriant automatinio lietuvių kalbos morfologinio anotavimo priemonę buvo pasitelkta čekų patirtis (Hladká 2000) morfologinio anotavimo srityje. Čekų darbuose atskleidžiamas Paslėptųjų Markovo modelių (statistinių metodų) ir formaliųjų (taisyklių metodų) taikymas čekų ir anglų kalboms. Svarbu pabrėžti, kad šie metodai yra nepriklausomi nuo kalbos ir gali būti taikomi ir lietuvių kalbai. Vienintelis nuo kalbos priklausantis dalykas yra nedidelis morfologiškai anotuotas tekstynas – mokymo tekstynas. Buvo atlikti įvairūs eksperimentai, kuriuose derinami, redukuojami įvairūs čekų kalbos morfologiniai požymiai, ir buvo pasiektas anglų kalbai artimas morfologinio anotavimo tikslumas – 96%. Tačiau šis tikslumas gaunamas nurodant ne visus čekų kalbos morfologinius požymius. Nurodant visus čekų kalbos morfologinius požymius tikslumas siekia 94% (Hladká 2000).

1 lentelė. Ištrauka iš morfologiškai anotuoto lietuvių kalbos tekstyno

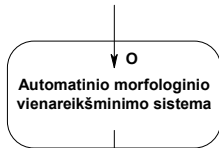
| |
|--|
| <word="Nenuostabu" lemma="nenuostabus" type="bdvr neig nelygin.l neįvardž bevr.d.gim"> ² |
| <sep=","> |
| <space> |
| <word="kad" lemma="kad" type="jngt"> |
| <space> |
| <word="muziejus" lemma="muziejus" type="dktv vyr.gim vnsk V"> |
| <space> |
| <word="susilaukia" lemma="susilaukti(-ia,-ė)" type="vksm teig sngt tiesiog.nuos esam.l vnsk IIIasm"> |
| <space> |
| <word="daugelio" lemma="daugelis" type="dktv vyr.gim vnsk K"> |
| <space> |
| <word="svečių" lemma="svečias" type="dktv vyr.gim dgsk K"> |
| <space> |
| <word="ne tik" lemma="ne tik" type="jngt"> |
| <space> |
| <word="iš" lemma="iš" type="prln"> |
| <space> |
| <word="Čikagos" lemma="Čikaga" type="tikr dktv mot.gim vnsk K"> |
| <space> |
| <word="ir" lemma="ir" type="jngt"> |
| <space> |
| <word="apylinkių" lemma="apylinkė" type="dktv mot.gim dgsk K"> |
| <sep=","> |
| <space> |
| <word="bet ir" lemma="bet ir" type="jngt"> |
| <space> |
| <word="tolimiaušių" lemma="tolimas" type="bdvr teig aukšč.l neįvardž vyr.gim dgsk K"> |
| <space> |
| <word="Amerikos" lemma="Amerika" type="tikr dktv mot.gim vnsk K"> |
| <space> |
| <word="kampelių" lemma="kampelis" type="dktv vyr.gim dgsk K"> |
| <p> |
| <word="bei" lemma="bei" type="jngt"> |
| <space> |
| <word="kitų" lemma="kitas" type="įvrd mot.gim dgsk K"> |
| <space> |
| <word="šalių" lemma="šalis" type="dktv mot.gim dgsk K"> |
| <sep=","> |

² *Word* – tai analizuojamasis žodis; *lemma* – to žodžio lema (antraštinis pavaldas, pvz., daiktavardžiams ir kitiems vardažodžiams tai vienaskaitos vardininkas, veiksmažodžiams – bendratis); *type* – gramatinės pažymos (čia pateikiama išsami gramatinė informacija apie analizuojamą žodį, pvz., kalbos dalis, giminė, nuosaka, linksnis, laikas ir pan.); pažyma <sep=",">, ">" reiškia, kad tekste toje vietoje yra kablelis; <space> reiškia tarpą tarp žodžių.

Statistinis morfologinis vienareikšminimas

Automatinio morfologinio vienareikšminimo sistemos įėjimas yra morfologiškai išanalizuoti teksto žodžiai, o išėjimas – sistemos surasta geriausia morfologinių požymių seka, atitinkanti įėjimo signalą (žr. 1 pav.).

```
<ambiguous>
<word="pamiršti" lemma="pamiršti(-ta,-o)" type="bndr">
<word="pamiršti" lemma="pamiršti(-ta,-o)" type="vksm tiesiog.nuos esam.I vnsk IIasm">
<word="pamiršti" lemma="pamiršti(-ta,-o)" type="div neveik.r būt.I neįvardž vyr.gim dgsk V">
</ambiguous>
```



```
<word="pamiršti" lemma="pamiršti(-ta,-o)" type="bndr">
```

1 pav. Morfologinio vienareikšminimo procesas.

Teksto anotavimas yra tikėtiniausios morfologinių požymių sekos naudojant Markovo modelius nustatymas. Net ir pačiuose didžiausiuose tekstuose nėra numatyti visi galimi trigramų, bigramų atvejai, nes ne visi žodžiai gali būti pavartoti. Todėl yra taikomas nerastų atvejų glotninimas, nes ieškant tikėtiniausios sekos tikimybė negali būti nulinė. Kuriant morfologinio anotavimo priemonę lietuvių kalbai, buvo pasitelkta čekų kalbos anotavimo patirtis ieškant geriausio morfologinio anotavimo metodo (Hladká 2000). Čekų naudojamas modelis yra:

$$\Gamma \approx \max_T p(w_1 | t_1) * \tilde{p}(t_1) * \tilde{p}(t_2 | t_1) * \prod_{t=3}^n \tilde{p}(w_t | t_t) * \tilde{p}(t_t | t_{t-1}, t_{t-2}),$$

$$T = t_1, t_2, \dots, t_n$$

Kuriant lietuvių kalbos anotavimo priemonę, tikimybių skaičiavimas buvo praplėstas įtraukiant ir tikėtiniausios lemos nustatymą. Tai aktualu lietuvių kalbai, nes joje pasitaiko atvejų, kai skirtingų lemų morfologiniai požymiai yra vienodi. Todėl papildomai yra įtraukiama ir lema:

$$\Gamma \approx \max_T \tilde{p}(w_1 | t_1) * \tilde{p}(w_1 | t_1) * \tilde{p}(t_1) * \tilde{p}(t_2 | t_1) * \prod_{t=3}^n \tilde{p}(w_t | t_t) * \tilde{p}(w_t | t_t) * \tilde{p}(t_t | t_{t-1}, t_{t-2}),$$

$$T = t_1, t_2, \dots, t_n$$

kur

$$\tilde{p}(w_t | t_t) = \lambda_w * p(w_t | t_t) + (1 - \lambda_w) * 1/W_t$$

$$\tilde{p}(w_t | t_t) = \lambda_{w1} * p(w_t | t_t) + (1 - \lambda_{w1}) * 1/L_{w_t}$$

$$\tilde{p}(t_t) = \lambda_{t1} * p(t_t) + (1 - \lambda_{t1}) * 1/C_T$$

$$\tilde{p}(t_t | t_{t-1}) = \lambda_{t1} * p(t_t | t_{t-1}) + \lambda_{t2} * p(t_t) + (1 - \lambda_{t1} - \lambda_{t2}) * 1/C_T$$

$$\tilde{p}(t_t | t_{t-1}, t_{t-2}) = \lambda_{t1} * p(t_t | t_{t-1}, t_{t-2}) + \lambda_{t2} * p(t_t | t_{t-1}) + \lambda_{t3} * p(t_t) + (1 - \lambda_{t1} - \lambda_{t2} - \lambda_{t3}) * 1/C_T$$

$$p(w_t | t_t) = \text{Count}(w_t | t_t) / \text{Count}(t_t)$$

$$p(t_t) = \text{Count}(t_t) / T_{\text{train}}$$

$$p(t_t | t_{t-1}) = \text{Count}(t_t, t_{t-1}) / \text{Count}(t_{t-1})$$

$$p(t_t | t_{t-1}, t_{t-2}) = \text{Count}(t_t, t_{t-1}, t_{t-2}) / \text{Count}(t_{t-1}, t_{t-2})$$

W_{t_t} yra žodžių skaičius, kurie turi t_t požymį.

C_T yra skirtingų t požymių skaičius T_{train} mokymo imtyje.

$\lambda_{w1}, \lambda_w, \lambda_{t1}, \lambda_{t2}, \lambda_{t3}, \lambda_{t1}, \lambda_{t2}, \lambda_{t3} < 1$ ir $\text{Count}(x)$ yra įvykio x dažnumas mokymo imtyje.

Naudojant šiuos modelius buvo gautas 94% visiško morfologinio anotavimo tikslumas, kuris atitinka kitoms kal-

boms taikomų modelių tikslumą mokymui naudojant 1 mln. morfologiškai anototą tekstyną. Taip pat buvo pasiektas 99% tikslumas nustatant antraštinę lietuvių kalbos žodžio formą.

Statistinio morfologinio vienareikšminimo mokymui buvo naudojamas 1 mln. žodžių apimties rankiniu būdu anototas tekstynas, kuriam būdingi tokie įvairių statistinių reikšmių kiekiai:

Skirtingų lemų – $\text{Count}(l_i) = 41\ 408$

Skirtingų žodžių formų (pažymų) – $\text{Count}(w_i, t_i) = 130\ 511$

Skirtingų žodžių (lemų) – $\text{Count}(w_i, l_i) = 121\ 634$

Pažymų unigramų – $\text{Count}(t_i) = 1449$

Pažymų bigramų – $\text{Count}(t_i, t_{i-1}) = 76\ 312$

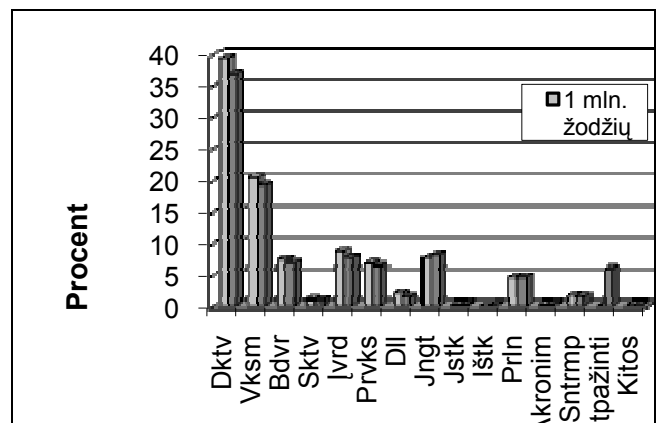
Pažymų trigramų – $\text{Count}(t_i, t_{i-1}, t_{i-2}) = 544\ 922$

Morfologiškai anototo Dabartinės lietuvių kalbos tekstyno statistiniai duomenys

Pagrindiniai duomenys apie morfologiškai anototą *Dabartinės lietuvių kalbos tekstyną* (statistiniai duomenys apie dažniausius žodžius ir jų formas pateikti straipsnio galėsiančiose lentelėse):

- tekstyno dydis – 111 745 938 žodžių;
- žodžių formų skaičius – 1 830 278;
- neatpažintų žodžių formų skaičius/ vartosenos dalis – 824387/ 5,6% vartosenos;
- atpažintų žodžių formų skaičius – 1 005 891;
- atpažintų žodžių formų antraštinių formų (lemų) skaičius – 225 319.

Suanotavus visą *Dabartinės lietuvių kalbos tekstyną*, paaiškėjo, kad kalbos dalių pasiskirstymas mažai kuo skiriasi nuo 1 mln. žodžių anototo tekstyno (žr. 2 pav.). Iš to galima daryti išvadą, kad gramatinei analizei užtenka palyginti nedidelio duomenų kiekio – 1 mln. žodžių, o didesnės apimties tekstynas turi mažai įtakos gramatinių kategorijų pasiskirstymui.



2 pav. Kalbos dalių pasiskirstymas 1 mln. ir 100 mln. žodžių morfologiškai anototuose tekstynuose.

Neišspręstos automatinės morfologinės analizės problemos

Jau minėta, kad automatinio morfologinio anotavimo programos tikslumas – 94%. Vadinasi, pavyksta išspręsti ne visus daugiareikšmiškumo atvejus. Pastebėta, kad nemažai problemų lieka, kai susiduriama su atsitiktinai sutampančiais žodžiais, kurių tam tikros formos visiškai sutampa. Dažnai pasirenkama netinkama forma, kai analizuojamos lemų *tonas* ir *tona*, *kovas* ir *kova*, *Biržai* ir *birža* ir panašaus tipo formos.

Ne visais atvejais išsprendžiamas linksnių sinkretizmas: dažniausiai painiojamas moteriškosios giminės vardažodžių vienaskaitos kilmininkas ir daugiskaitos vardininkas, pvz., *mokyklos*.

Ne visada net lingvistas gali suprasti, kuri forma – veikmažodis ar daiktavardis – pavartota tokiuose pastoviai vartojamuose junginiuose: *kovos dėl teisės likti pirmajame ešalone*; *kovos su narkotikais*; *kovos su okupantais*³. Net ir nusprendus, kad *kovos* yra daiktavardis, lieka neaišku, ar tai vienaskaitos kilmininkas ar daugiskaitos vardininkas. Žinoma, tokia problema iškyla dėl to, kad nežinomas plaatesnis kontekstas.

Nors jaustukai nėra dažnai vartojami lietuvių kalboje, vis dėlto automatinės morfologinės analizės programa vietoj santrumpos *a* nurodo, kad tai jaustukas *a*. Dažnai klaidingai morfologiškai anotuojamos santrumpos, sutampančios su romėniškai skaičiais: daugiausia problemų kelia santrumpa *V*.

Peržiūrint anotuotus ir suvienareikšmintus tekstus pastebėta, kad kai kuriais atvejais pasirenkama ne ta lema. Daugiausia problemų kyla su tokiais nekaitomais žodžiais, kaip *ir*, *tik* (šie žodžiai gali būti jungtukai, dalelytės,rieveiksmiai). Ne visada pasirenkama tinkama žodžio *vienas* lema (šis žodis gali būti įvardis, būdvardis, skaitvardis ir tikrinis daiktavardis). Tikimasi, kad ateityje patobulinus morfologinės analizės programą minėtų problemų sumažės.

Išvados

Naudojant Paslėptuosius Markovo modelius buvo gautas 94% visiško morfologinio anotavimo tikslumas, kuris atitinka kitoms kalboms taikomų modelių tikslumą mokymui naudojant 1 mln. morfologiškai anotuotą tekstyną. Taip pat buvo pasiektas 99% tikslumas nustatant antraštines lietuvių kalbos žodžių formas (lemas). Tikslumas skaičiuojamas įvertinant vienareikšminimo klaidas; nėra įskaičiuojami neatpažinti žodžiai.

Neatpažintų žodžių vartosenos kiekis sudaro 5,6% (daugiau nei 800 tūkst. žodžių formų). Norint automatiškai sėkmingai išanalizuoti šias žodžių formas reikia automatinės morfologinės analizės priemonės žodyną papildyti apie 100-150 tūkst. naujų antraštinių žodžių, t. y. beveik antra tiek, koks yra dabartinės morfologinės analizės priemonės žodyno dydis.

1 mln. morfologiškai anotuoto tekstyno užtenka nagrinėjant pagrindinius lietuvių kalbos dėsniumus, nes grama-

tinių kategorijų vartosenos pasiskirstymas yra panašus iš 1 mln. ir daugiau nei 100 mln. žodžių sudarytuose tekstynuose.

Literatūra

1. Hladká, B. (2000) Czech Language Tagging: daktaro disertacija, IFAL MFF UK, Praha.
2. Rimkutė, E. (2006a) Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekстыne: daktaro disertacija, Vytauto Didžiojo universitetas, Kaunas.
3. Rimkutė, E. (2006b) Lietuvių kalbos kolokacijų žodynas: sandara ir paskirtis, *Prace Baltystyczne*, 3, pp249-258.
4. Zinkevičius, V. (2000) Lemuoklis – morfologinei analizei, *Darbai ir Dienos* 24, pp246-273.
5. 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, pp365-370.

³ Pavyzdžiai pateikti iš morfologiškai anotuoto daiktavardinių frazių sąrašo (plačiau žr. Rimkutė 2006b).

Morphologically Annotated Corpus of Contemporary Lithuanian Language

Summary

Research of morphological disambiguation and morphological annotation of the 100 million word Lithuanian corpus are presented in the article. Statistical methods enabled to develop the automatic tool of morphological annotation for Lithuanian. The method of Hidden Markov models for morphological annotation has allowed achieving the precision of 94%, which is comparable to the precision achieved for other languages, when the 1 mln. word training corpus is used. The precision of 99% is reached for establishing headwords of Lithuanian words. The precision measure estimates only the process of disambiguation, while unrecognised words are not included in the precision test. The amount of unrecognised words makes up 5,6% of all used word-forms (more than 800,000 different word-forms). 1 million word morphological corpus is enough for the analysis of morphological phenomena in the Lithuanian language, as distribution of parts of speech in the whole 100 million word corpus does not differ significantly from the distribution in the training corpus.

Straipsnis įteiktas 2007 04
Parengtas spaudai 2007 11

Apie autorius

Erika Rimkutė, dr., Vytauto Didžiojo universiteto Humanitarinių mokslų fakulteto Kompiuterinės lingvistikos centro mokslo darbuotoja, Lietuvių kalbos katedros lektorė.

Mokslo interesų sritys: tekstynų lingvistika, kompiuterinė lingvistika, automatinė morfologinė analizė bei sintezė, morfologinis daugiareikšmiškumas ir jo ribojimas, automatinė sintaksinė analizė.

Adresas: K. Donelaičio g. 52-206, Kaunas, Lietuva.

El. paštas: e.rimkute@hmf.vdu.lt

Vidas Daudaravičius, Vytauto Didžiojo universiteto Humanitarinių mokslų fakulteto Kompiuterinės lingvistikos centro vyr. inžinierius programuotojas.

Mokslo interesų sritys: kompiuterinė lingvistika, kompiuterinės lietuvių kalbos informacinės technologijos, tekstinės informacijos saugojimas ir apdorojimas, informacijos iš teksto išgavimas, mašininis vertimas.

Adresas: K. Donelaičio g. 52-206, Kaunas, Lietuva.

El. paštas: vidas@donelaitis.vdu.lt

PRIEDAI

2 lentelė. 30 dažniausių žodžių formų.

| Eilės nr. | Žodis | Dažnumas |
|-----------|----------|----------|
| 1 | ir | 3691998 |
| 2 | kad | 1022047 |
| 3 | i | 988673 |
| 4 | su | 691396 |
| 5 | buvo | 646313 |
| 6 | iš | 641117 |
| 7 | o | 602096 |
| 8 | tai | 578021 |
| 9 | yra | 557057 |
| 10 | kaip | 542470 |
| 11 | Lietuvos | 477448 |
| 12 | ar | 462783 |
| 13 | savo | 411593 |
| 14 | jis | 376791 |
| 15 | m | 363707 |
| 16 | apie | 358297 |
| 17 | tik | 336067 |
| 18 | ne | 334210 |
| 19 | nuo | 324068 |
| 20 | jo | 304571 |
| 21 | bet | 300414 |
| 22 | jų | 291609 |
| 23 | a | 291389 |
| 24 | po | 277351 |
| 25 | dėl | 277314 |
| 26 | jau | 276094 |
| 27 | už | 273243 |
| 28 | dar | 271407 |
| 29 | bei | 261091 |
| 30 | tačiau | 258160 |

3 lentelė. 30 dažniausių lemų

| Eilės nr. | Lema | Kalbos dalis | Dažnumas |
|-----------|------------------|--------------|----------|
| 1 | Ir | jngt | 2993282 |
| 2 | jis | įvrd | 2211352 |
| 3 | būti(yra, buvo) | vksm | 1463394 |
| 4 | kad | jngt | 1000674 |
| 5 | i | prln | 988673 |
| 6 | aš | įvrd | 789989 |
| 7 | kuris | įvrd | 703458 |
| 8 | su | prln | 691396 |
| 9 | šis | įvrd | 677566 |
| 10 | Lietuva | tikr dktv | 643716 |
| 11 | iš | prln | 641117 |
| 12 | ir | prvks | 613762 |
| 13 | o | jngt | 595384 |
| 14 | tai | jngt | 526994 |
| 15 | tas | įvrd | 520113 |
| 16 | ar | jngt | 447764 |
| 17 | visas | įvrd | 446422 |
| 18 | kitas | įvrd | 437728 |
| 19 | kaip | jngt | 431629 |
| 20 | savo | įvrd | 411593 |
| 21 | metai | dktv | 395843 |
| 22 | žmogus | dktv | 372642 |
| 23 | m | sntrmp | 363707 |
| 24 | apie | prln | 358297 |
| 25 | turėti(-i, -ėjo) | vksm | 355612 |
| 26 | galėti(-i, -ėjo) | vksm | 340981 |
| 27 | nuo | prln | 324068 |
| 28 | toks | įvrd | 323254 |
| 29 | bet | jngt | 297334 |
| 30 | a | jstk | 291389 |

4 lentelė. 10 dažniausių bendrinių daiktavardžių antraštinių formų.

| Lema | Dažnumas |
|------------|----------|
| metai | 395843 |
| žmogus | 372642 |
| darbas | 260467 |
| diena | 192537 |
| respublika | 185547 |
| valstybė | 171492 |
| vaikas | 164847 |
| šalis | 161528 |
| laikas | 158459 |
| įstatymas | 154343 |

5 lentelė. 10 dažniausių veiksmažodžių antraštinių formų.

| Lema | Dažnumas |
|----------------------|----------|
| būti(yra, buvo) | 1463394 |
| turėti(-i, -ėjo) | 355612 |
| galėti(-i, -ėjo) | 340981 |
| nebūti(-ėra, -ebuvo) | 220688 |
| reikėti(-ia, -ėjo) | 162582 |
| busti(-nda, -do) | 140039 |
| sakyti(-o, -ė) | 124384 |
| norėti(-i, -ėjo) | 98248 |
| manyti(-o, -ė) | 94809 |
| negalėti(-i, -ėjo) | 90900 |

6 lentelė. 10 dažniausių būdvardžių antraštinių formų

| Lema | Dažnumas |
|-------------|----------|
| vienas | 181190 |
| didelis | 166757 |
| naujas | 165901 |
| svarbus | 104265 |
| geras | 100785 |
| aukštas | 96298 |
| valstybinis | 82554 |
| įvairus | 81471 |
| bendras | 70945 |
| mažas | 67022 |

7 lentelė. 10 dažniausių skaitvardžių antraštinių formų.

| Lema | Dažnumas |
|------------|----------|
| pirmas | 181593 |
| du | 141617 |
| antras | 89885 |
| trys | 73433 |
| abu | 47788 |
| trečias | 44150 |
| tūkstantis | 35898 |
| keturi | 30971 |
| šimtas | 23248 |
| penki | 23021 |

8 lentelė. 10 dažniausių įvardžių antraštinių formų.

| Lema | Dažnumas |
|-------|----------|
| jis | 2211352 |
| aš | 789989 |
| kuris | 703458 |
| šis | 677566 |
| tas | 520113 |
| visas | 446422 |
| kitas | 437728 |
| savo | 411593 |
| toks | 323254 |
| tu | 244812 |