



Filologijos
fakultetas

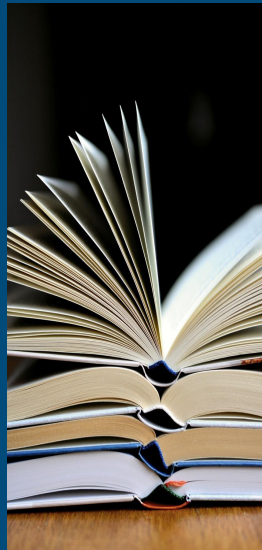
Tekstynų duomenys kaip apibendrintos kalbos vartotojo patirties indikacija: kurie tekstynai veikia geriausiai?

doc. dr. Laura Vilkaitė-Lozdienė

Teorinės prielaidos

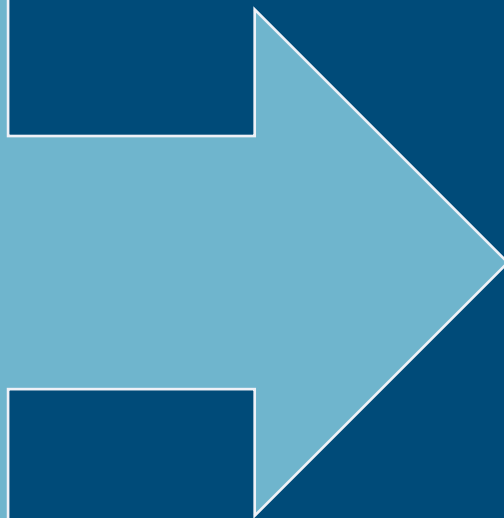
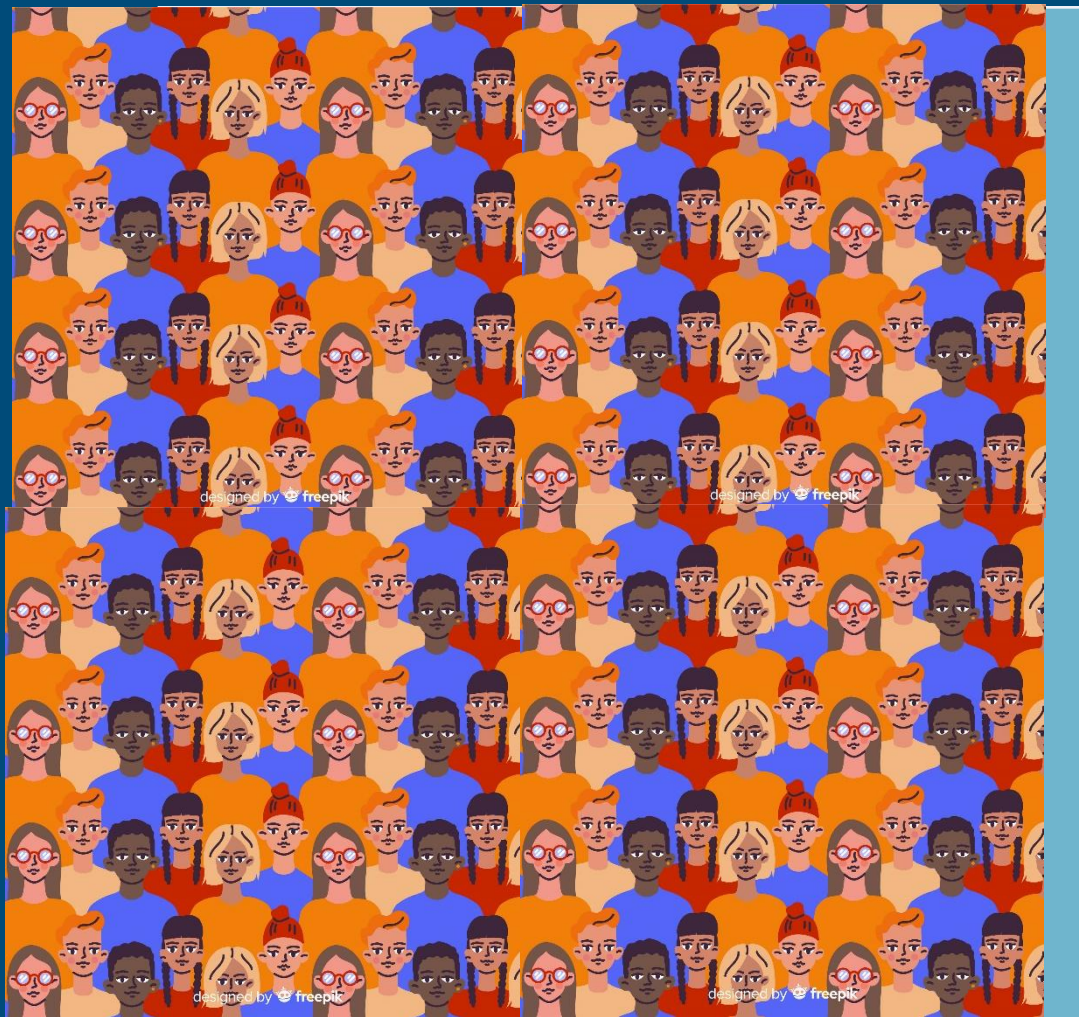


Vartojimu grįstas kalbos įsisavinimas



Tipinis kalbos vartotojas

Vilniaus universitetas
Filologijos fakultetas



Tipiška kalbinė patirtis

Vilniaus universitetas
Filologijos fakultetas



CAN BE EMAILED INTERNALLY ONLY

IMPORTANT INFORMATION ABOUT PROCEDURES FOR OPENING A NEW ACCOUNT: To help the government fight the funding of terrorism and money laundering, we require all financial institutions to obtain, verify and record information that identifies each person who opens an account. What this means for you is that we will ask you for your name, address, date of birth and other information that will allow us to identify you. We may ask to see your driver's license or other

Interest Rates and Interest Charges	Signature Bonus Rewards* or Platinum Bonus Rewards	Signature Bonus Rewards PLUS** or Platinum Bonus Rewards PLUS	Platinum
Annual Percentage Rate (APR) for Purchases	0.00% Introductory APR for 6 billing cycles. After that, your APR will be 12.24% - 23.24% based on your creditworthiness when you open your account. This APR will vary with the market based on the Prime Rate.	0.00% Introductory APR for 6 billing cycles. After that, your APR will be 10.24% - 21.24% based on your creditworthiness when you open your account. This APR will vary with the market based on the Prime Rate.	0.00% Introductory APR for 6 billing cycles. After that, your APR will be 10.24% - 21.24% based on your creditworthiness when you open your account. This APR will vary with the market based on the Prime Rate.
APR for Balance Transfers	0.00% Introductory APR for 6 billing cycles. After that, your APR will be 12.24% - 23.24% based on your creditworthiness when you open your account. This APR will vary with the market based on the Prime Rate.	0.00% Introductory APR for 6 billing cycles. After that, your APR will be 10.24% - 21.24% based on your creditworthiness when you open your account. This APR will vary with the market based on the Prime Rate.	0.00% Introductory APR for 6 billing cycles. After that, your APR will be 10.24% - 21.24% based on your creditworthiness when you open your account. This APR will vary with the market based on the Prime Rate.
APR for Cash Advances	24.24% This APR will vary with the market based on the Prime Rate.		
Penalty APR and When It Applies	Not Applicable.		
How to Avoid Paying Interest on Purchases	Your due date is 24-30 days after the close of each billing cycle. We will not charge you interest on purchases if you pay your bill by the due date each month.		
Minimum Interest Charge	If you are charged interest, the charge will be no less than \$2.00.		
For Credit Card Tips from the Consumer Financial Protection Bureau	To learn more about factors to consider when applying for or using a credit card, visit the website of the Consumer Financial Protection Bureau at http://www.consumerfinance.gov/learnmore .		



Vilniaus universitetas
Filologijos fakultetas



Dažnumas

Vienas svarbiausių kintamųjų, modeliuojant kalbos apdorojimą: dažnesni žodžiai suvokiami greičiau*

*yra papildomų sąlygų

Kokiais tekstynais remtis?

- Jei mus domina ir reti žodžiai – bent 16 milijonų žodžių tekstynas (ar 20 mln.)
- Tekstynams, didesniems nei 30 mln. žodžių, daug svarbiau yra registrai, kuriuos tekstynas apima, nebe tekstyno dydis.

(Brysbaert ir New, 2009)

Kokiais tekstynais remtis?

Jei modeliuojame studentų kalbą, geriausiai eksperimentinius duomenis atitinka:

- TV subtitrų tekstynai (Brysbaert ir New, 2009),
- socialinės medijos tekstynai (Gimenes ir New, 2016; Herdag˘ delen ir Marelli, 2017),
- blogų tekstynai (Gimenes ir New, 2016).

Vyresniems eksperimentų dalyviams tradiciniai, knygų tekstaais grįsti tekstynai, veikia geriau (Brysbaert ir Ellis, 2016).

Kokių tekstynų turime?



Dadurkevičius (2020). Wordlist of Lemmas from the Joint Corpus of Lithuanian



☰	EUR-Lex judgments 12/2016 parallel – Lithuanian	Lithuanian	16,252,111
☰	Europarl spoken parallel – Lithuanian	Lithuanian	11,424,032
☰	Lithuanian parliamentary debates (ParlaMint 2.1, CoNLL format)	Lithuanian	14,428,682
☰	Lithuanian parliamentary debates (ParlaMint 2.1)	Lithuanian	14,573,624
☰	Lithuanian Web (LithuanianWaC v2)	Lithuanian	48,650,918
☰	Lithuanian Web 2014 (ItTenTen14)	Lithuanian	778,151,979
✓	Lithuanian Web 2021 (ItTenTen21)	Lithuanian	1,772,410,416
☰	Open Parallel Corpus (OPUS) – Lithuanian	Lithuanian	29,621,940
☰	OpenSubtitles 2018 parallel – Lithuanian	Lithuanian	6,806,857

Tyrimo klausimas

Kurio iš lietuviškų, viešai prieinamų tekstynų ar dažnumo sąrašų dažnumai geriausiai atspindi kalbėtojų kalbinę patirtį?

- *OpenSubtitles 2018 parallel* – Lithuanian: filmų subtitrai
- LithuanianWaC v2 (Kilgarriff et al., 2010): interneto tekstai
- *Joint Corpus of Lithuanian*

Laura Vilkaitė-Lozdienė

Justina Bružaitė-Liseckienė



Eksperimentinis kalbos sudėtingumo tyrimas
Nr. S-MIP-24-121

Metodas



Eksperimento dalyviai

Tipiški jauni
skaitytojai

Tipiški vyresni
skaitytojai (65+)

Žmonės su
intelektu negalia

Pagyvenę
žmonės su
demencija

Negimtakalbiai

Eksperimento dalyviai

Tipiški jauni
skaitytojai

Tipiški vyresni
skaitytojai (65+)

Žmonės su
intelektu negalia

Pagyvenę
žmonės su
demencija

Negimtakalbiai

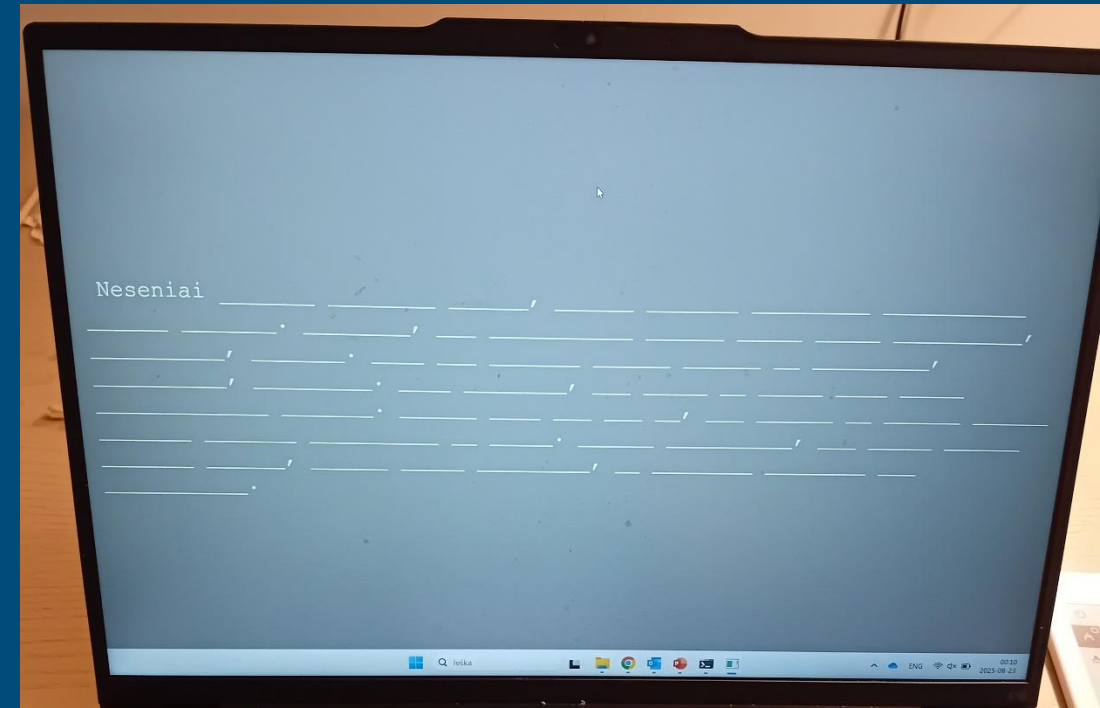
Žodžiai analizei

- Pavieniai žodžiai (pvz., *gyvenimas*, *o*, *orchidėja*...)
- $n = 118$ įprastoje versijoje
- $n = 87$ versijoje lengvai suprantama kalba
- Pavartoti tekste
- Skirtingo dažnumo, ilgio, kalbos dalių, reikšmės konkretumo

Skaitytojo kontroliuojamas skaitymas

- Vienas tekstas pasibandymui
- 9 (įprasta versija) arba 6 (lengvai suprantama kalba) trumpi pasakojimai
- Po du teksto supratimo klausimus po kiekvieno teksto

Trukmė: ~ 45 min



Neseniai

_____ atlikta _____, _____. . . .

_____ apklausa _____, ____ . . .

Vilniaus universitetas

Filologijos fakultetas

----- parodė, ---...

Vilniaus universitetas

Filologijos fakultetas

_____, kad...

Analizė

Vilniaus universitetas
Filologijos fakultetas

Tiesiniai mišrių veiksnių modeliai su R
(*lmer* paketas)

Fiksuoti efektai:

- dažnumas,
- kiti leksiniai kintamieji (žodžio ilgis, konkretumas, tikimybė, ortografiškai artimų žodžių skaičius),
- teksto eilės numeris eksperimente.

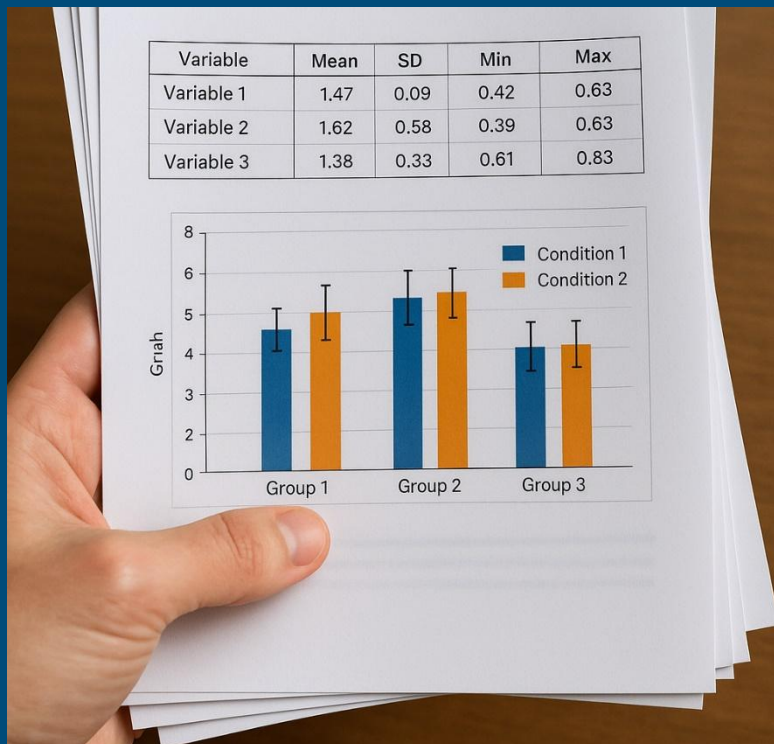
Atsitiktiniai efektai:

- atsitiktiniai laisvieji nariai dalyviui ir žodžiui.

Atskiri modeliai su dažnumais iš skirtingų tekstynų: po keturis kiekvienai dalyvių grupei

Modelių lyginimas pagal AIC, ir R^2

Rezultatai



Tipiški jauni kalbėtojai

Tekstynas	Efektas	AIC	R ²
Joint Corpus of Lithuanian (lema)	$b = -0,03, p = 0,027$	2129	0,08 (0,54)
Subtitrų tekstynas (forma)	$b = -0,03, p = 0,014$	2128	0,08 (0,54)
Interneto tekstynas (lema)	$b = -0,02, p = 0,093$	2131	0,08 (0,54)
Interneto tekstynas (forma)	$b = -0,02, p = 0,113$	2131	0,08 (0,54)

Tipiški pagyvenę kalbėtojai

Dažnumas neturi reikšmingos įtakos: rezultatas atitinka kitų kalbų rezultatus.
Dalyviai sukaupę daugiau kalbinės patirties patirties, dažnumas praranda aktualumą.

Negimtakalbiai

Tekstynas	Efektas	AIC	R ²
Joint Corpus of Lithuanian (lema)	$b = -0,09, p < 0,001$	3384	0,17 (0,56)
Subtitrų tekstynas (forma)	$b = -0,06, p < 0,001$	3393	0,17 (0,56)
Interneto tekstynas (lema)	$b = -0,07, p < 0,001$	3394	0,17 (0,56)
Interneto tekstynas (forma)	$b = -0,07, p = 0,002$	3393	0,17 (0,56)

Žmonės su intelekto negalia

Tekstynas	Efektas	AIC	R ²
Joint Corpus of Lithuanian (lema)	$b = -0,09, p < 0,001$	2452	0,07 (0,57)
Subtitrų tekstynas (forma)	$b = -0,08, p < 0,001$	2454	0,07 (0,57)
Interneto tekstynas (lema)	$b = -0,07, p < 0,001$	2459	0,07 (0,57)
Interneto tekstynas (forma)	$b = -0,07, p < 0,001$	2460	0,07 (0,57)

Vietoje išvadų

- Visų išbandytų dažnumų sąrašų dažnumai stipriai teigiamai koreliavo
- Subtitrų tekstynas, kaip ir galima tikėtis, geriausiai veikė jauniems gimtakalbiams. Geresnis (daugiau spontaniškos kalbos) subtitrų tekstynas galbūt veiktų dar geriau
- Didelio ir įvairesnės kalbą apimančio tekstyno dažnumų sąrašai veikė geriausiai, jei norime visiems tinkamo sąrašo
- Kas būtų su sakytinės kalbos tekstynu?



Filologijos
fakultetas



Laura Vilkaitė-Lozdienė

laura.vilkaite@flf.vu.lt

Ilustracijos:

<https://www.pexels.com/>

<https://pixabay.com/>

<https://www.freepik.com>