

CLARIN-LT metinis seminaras: kalbos technologijos, ištekliai, tyrimai
Vytauto Didžiojo universitetas, 2025 m. gruodžio 10 d.

Jurgis Pakerys (Vilniaus universitetas)

Virginijus Dadurkevičius (Vytauto Didžiojo universitetas)

Tekstyno lemavimas ir žodžių darybos tyrimai

Projektą Lietuvių kalbos priesaginių daiktavardžių darybinis
produktyvumas Jungtinio tekstyno duomenimis finansavo
Lietuvos mokslo taryba (LMTLT), sutarties Nr. S-LIP-22-6

Planas

1. Jungtinis lietuvių kalbos tekstynas
2. Darybos produktyvumo vertinimas tekstyno duomenimis
3. Tekstyno medžiagos rengimo etapai:
 - Automatinis lemavimas
 - Papildomas pusiau automatinis lemavimas
 - Pamatinių žodžių nustatymas
4. Atvirai skelbsimi duomenys
5. Išvados
6. Literatūra

1. Jungtinis lietuvių kalbos tekstynas

Jungtinis lietuvių kalbos tekstynas, apie 1,3 mlrd. žodžių
(Dadurkevičius 2020)

Patekstynis	Apimtis
VU mašininio vertimo tekstynas (internetiniai tekstai), 2014 (nėra viešos prieigos, šiuo metu rengiama)	779 154 268
Lietuvos Respublikos Seimo dokumentai, 2011 (e-seimas.lrs.lt)	443 114 936
VDU Dabartinės lietuvių kalbos tekstynas, 2008 (tekstynas.vdu.lt)	112 575 876

2. Darybos produktyvumo vertinimas tekstyno duomenimis

Mums **reikia tekstyno lemų sąrašo**, iš kurio gautume:

- Vedinių skaičių, angl. *types*
- Vedinių (formų) dažnį, angl. *tokens*
- Vedinių hapaksų skaičių (= vedinių, kurių dažnis = 1, skaičių)

Baayen 1992, 1993; žr. apžvalgas Baayen 2009; Gaeta & Ricca 2015: 844–849, Dal & Namer 2016: 73–76

3. Tekstyno medžiagos rengimas

Tekstyno medžiagos rengimo etapai:

1. Automatinis lemavimas
2. Papildomas pusiau automatinis lemavimas
3. Darybinio skaidumo nustatymas

Žr. plačiau: Pakerys et al. 2024, 2025 (spausdinama)

3.1. Automatinis lemavimas

Automatinis tekstyno lemavimas atliktas „Hunspell“ platformos pagrindu sukurtu lemavimo įrankiu (Dadurkevičius 2017)

- Į akis krito įtartinai **per mažas darių tipų hapaksų skaičius**, pavyzdžiui, veikėjų priesagų *-toj-*, *-ėj-*, *-ik-* (žr. kitą skaidrę)
- Todėl **kilo abejonių ir dėl bendro visų (didesnio dažnio) lemu skaičiaus**
- Plg. apie minėtojo lemavimo įrankio žodyno plėtrą: Miglė Žemrietė, Tikslesnio morfologinio anotatoriaus link: neatpažintų žodžių tyrimas, *Jono Jablonskio konferencija*, LKI, 2025

Veikėjų pavadinimai (automatinis lemavimas)

V – iš viso

V₁ – hapaksai

m – vyr. g.

f – mot. g.

Priesaga	Automatinis lemavimas	
	V	V ₁
<i>-toj-as</i> (m)	632	11
<i>-toj-a</i> (f)	547	11
<i>-ėj-as</i> (m)	251	2
<i>-ėj-a</i> (f)	258	3
<i>-ik-as</i> (m)	293	2
<i>-ik-ė</i> (f)	159	6

3.2. Pusiau automatinis lemavimas

Potencialių vedinių formų paieška pagal baigmenį PRIESAGA + GALŪNĖ, pvz.: *-toj-as, -toj-o, -toj-ui..., -toj-a, -toj-os, -toj-ai..., -ėj-as, -ėj-o, -ėj-ui..., -ėj-a, -ėj-os, -ėj-ai..., -ik-as, -ik-o, -ik-ui..., -ik-ė, -ik-ės, -ik-ei...,*

Formų jungimas į lemas, potencialių **pamatinių žodžių nurodymas** (jei jie yra lemavimo įrankio žodyne)

Rezultatas – vedinių skaičius reikšmingai išaugo

Veikējū pavadinimai po papildomo lemovimo

V – iš viso

V₁ – hapaksai

m – vyr. g.

f – mot. g.

Priesaga	Automatinis lemovimas		Papildomas lemovimas	
	V	V ₁	V	V ₁
<i>-toj-as</i> (m)	632	11	3,305	822
<i>-toj-a</i> (f)	547	11	2,590	637
<i>-ėj-as</i> (m)	251	2	2,351	642
<i>-ėj-a</i> (f)	258	3	1,576	384
<i>-ik-as</i> (m)	293	2	911	189
<i>-ik-ė</i> (f)	159	6	857	274

3.2. Pusiau automatinis lemavimas

Nors vedinių skaičius išaugo, **dalį lemy buvo atstatyta klaidingai**

Lemy perteklius daugeliu atvejų atsirado **dėl homografijos**, pvz.:

Forma: ***apčiupinėtojas***

Interpretacija: vyr. g. vns. vard. arba mot. g. dgs. gal.

Rezultatas: **vietoje vienos gauname 2 lemas:**

apčiupinėtojas ir apčiupinėtoja

3.2. Pusiau automatinis lemavimas

Forma: *apgadintojo*

Interpretacija: vyr. g. vns. kilm.: veikėjo pavadinimo arba neveikiamajo dalyvio forma

Rezultatas: **neegzistuojanti lema** *apgadintojas*, nes kontekstas rodo, kad tekstyne pavartotas dalyvis

Forma: *atkentėjo*

Interpretacija: būt. k. I. 3 a. arba vyr. g. dktv. vns. kilm.

Rezultatas: **neegzistuojanti lema** *atkentėjas*, atstatyta pagal vksm. 3 asm. formą

3.3 Pamatinių žodžių nustatymas

Rankinė gautų sąrašų peržiūra, koreguojant ar įrašant pamatinius žodžius

Dauguma mūsų tirtų daiktavardžių skaidūs: *rašy-toj-as* ← *rašy-ti*, *raš-ikl-is* ← *rašy-ti*, *vir-tuv-ė* ← *vir-ti*, *prek-yb-a* ← *prekiauti* ir pan.

Bet plg. leksikalizacijos arba rečiau vartojamų pamatinių žodžių (ir jų reikšmių) atvejus: *tar-yb-a* ← *tar-ti(s)*, *pribuv-ėj-a* ← *pribū-ti*, *myg-tuk-as* ← *myg-ti*, *svarst-ykl-ės* ← *svarst-y-ti* ‘dažnai sverti’ ir kt.

3.3 Pamatinių žodžių nustatymas

Darybiškai neskaidžiomis laikomos ir lemos, turinčios atsitiktinai su tiriamomis priesagomis sutampančias fonemų eilutes, pvz.:

vėj-as, vaik-as, skeptik-as

Darybiškai neskaidžios bus ir dėl homografijos netinkamai atstatytos lemos, žr. anksčiau

3.3 Pamatinių žodžių nustatymas

Projekte siekėme įvertinti priesagų produktyvumą **atsižvelgdami į jomis žymimas kategorijas**, todėl stengėmės darybiškai skaidžiais laikyti tokius vedinius, kurie turi (arba gali turėti) nagrinėjamą kategoriją atitinkančią reikšmę, plg.:

kalėjimas (DLKŽ)

1. Namai laikyti asmenims, kuriems atimta laisvė.
2. **Buvimas tokiuose namuose.**

Veikēju pavadinimai po papildo- mo lemovimo ir rankinės peržiūros

V – iš viso

V₁ – hapaksai

m – vyr. g.

f – mot. g.

Priesaga	Papildomas lemovimas		Po rankinės peržiūros	
	V	V ₁	V	V ₁
<i>-toj-as</i> (m)	3,305	822	2,456	532
<i>-toj-a</i> (f)	2,590	637	2,299	525
<i>-ėj-as</i> (m)	2,351	642	686	130
<i>-ėj-a</i> (f)	1,576	384	640	128
<i>-ik-as</i> (m)	911	189	254	65
<i>-ik-ė</i> (f)	857	274	92	27

4. Atviri duomenys

Į CLARIN-LT saugyklą planuojame įkelti tokią struktūrą duomenis

	A	B	C	D
1	Vediny	Pamatas	Priesaga	Dažnis
2	įsitikinimas	įsitikinti	imas	48430
3	apdailintojas	apdailinti	tojas	1
4	pasakojimas	pasakoti	imas	47682
5	apdainuotoja	apdainuoti	toja	6
6	apdainuotojas	apdainuoti	tojas	10
7	parengimas	parengti	imas	47515
8	pasipriešinimas	pasipriešinti	imas	47311
9	prašymas	prašyti	ymas	545630
10	sukilėlė	sukilti	ėlė	18021
11	apdarytoja	apdaryti	toja	2
12	apdarytojas	apdaryti	tojas	2
13	valdymas	valdyti	ymas	512367
14	įregistravimas	įregistruoti	imas	47059
15	apdegėjas	apdegti	ėjas	2
16	pabėgėlė	pabėgti	ėlė	17824
17	apskaičiavimas	apskaičiuoti	imas	46634
18	lavinimas	lavinti	imas	46602
19	surinkimas	surinkti	imas	46548
20	apdengėja	apdengti	ėja	1

5. Išvados

- Automatinio lemavimo rezultatų žodžių darybos tyrimams gali nepakakti
- Mūsų taikytas papildomas lemavimas rezultatus reikšmingai pagerina, bet homografų skyrimo problema išlieka
- Svarbu plėtoti tekstynus, kurių lemų ir formų sąrašus būtų galima parsisiųsti tolesniam apdorojimui
- Reikia detalių lemavimo įrankių aprašų, kad žinotume jų veikimo principus ir galimybių ribas

Ačiū už dėmesį!



Literatūra

Baayen, R. Harald (1992). Quantitative Aspects of Morphological Productivity. In: Booij, Geert E. & van Marle, Jaap (eds.), *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers, 109–149 (https://doi.org/10.1007/978-94-011-2516-1_8).

Baayen, R. Harald (1993). On Frequency, Transparency, and Productivity. In: Booij, Geert E. & van Marle, Jaap (eds.), *Yearbook of Morphology 1992*. Dordrecht: Kluwer Academic Publishers, 181–208 (https://doi.org/10.1007/978-94-017-3710-4_7).

Baayen, R. Harald (2009). Corpus linguistics in morphology: Morphological productivity. In: Lüdeling, Anke & Kytö, Merja (eds.), *Corpus Linguistics: An International Handbook*, vol. 2. Berlin, New York: Mouton de Gruyter, 899–919 (<https://doi.org/10.1515/9783110213881.2.899>).

Dadurkevičius, Virginijus (2017). Lietuvių kalbos morfologija atvirojo kodo „Hunspell“ platformoje, *Bendrinė kalba* 90, 1–17 (<https://journals.iki.lt/bendrinekalba/article/view/156>).

Dadurkevičius, Virginijus (2020). Wordlist of Lemmas from the Joint Corpus of Lithuanian. *CLARIN-LT digital library in the Republic of Lithuania* (<http://hdl.handle.net/20.500.11821/41>).

Literatūra

Dal, Georgette & Namer, Fiammetta (2016). Productivity. In Hippisley, Andrew & Stump, Gregory (eds.), *The Cambridge Handbook of Morphology*. Cambridge: Cambridge University Press, 70–90 (<https://doi.org/10.1017/9781139814720.004>).

Gaeta, Livio & Ricca, Davide (2015). Productivity. In: Müller, Peter O., Ohnheiser, Ingeborg, Olsen, Susan & Rainer, Franz (eds.), *Word-Formation: An International Handbook of the Languages of Europe*, vol. 2. Berlin/Boston: De Gruyter Mouton, 842–858 (<https://doi.org/10.1515/9783110246278-003>).

Pakerys, Jurgis, Dadurkevičius, Virginijus, Navickaitė-Klišauskienė, Agnė. 2024. How lemmatisation and derivational annotation affect productivity measures: The case of deverbal agent nouns in the Joint Corpus of Lithuanian. *Valoda: nozīme un forma / Language: Meaning and Form* 15, 138–151. (<https://doi.org/10.22364/vnf.15.09>)

Pakerys, Jurgis, Dadurkevičius, Virginijus, Navickaitė-Klišauskienė, Agnė. 2025 [spausdinama]. Jungtinio lietuvių kalbos tekstyno lemavimas ir anotavimas darybiniam produktyvumui matuoti: priesaginių daiktavardžių atvejis, *Taikomoji kalbotyra*.