

Eesti keele ühendkorpus 2023: uued(meta)andmed ja uued analüüsivõimalused

Kristina Koppel, Jelena Kallas, Madis Jürviste, Helen Kaljumäe
Eesti Keele Instituut | EKI



Eesti keele ühendkorpuste sari

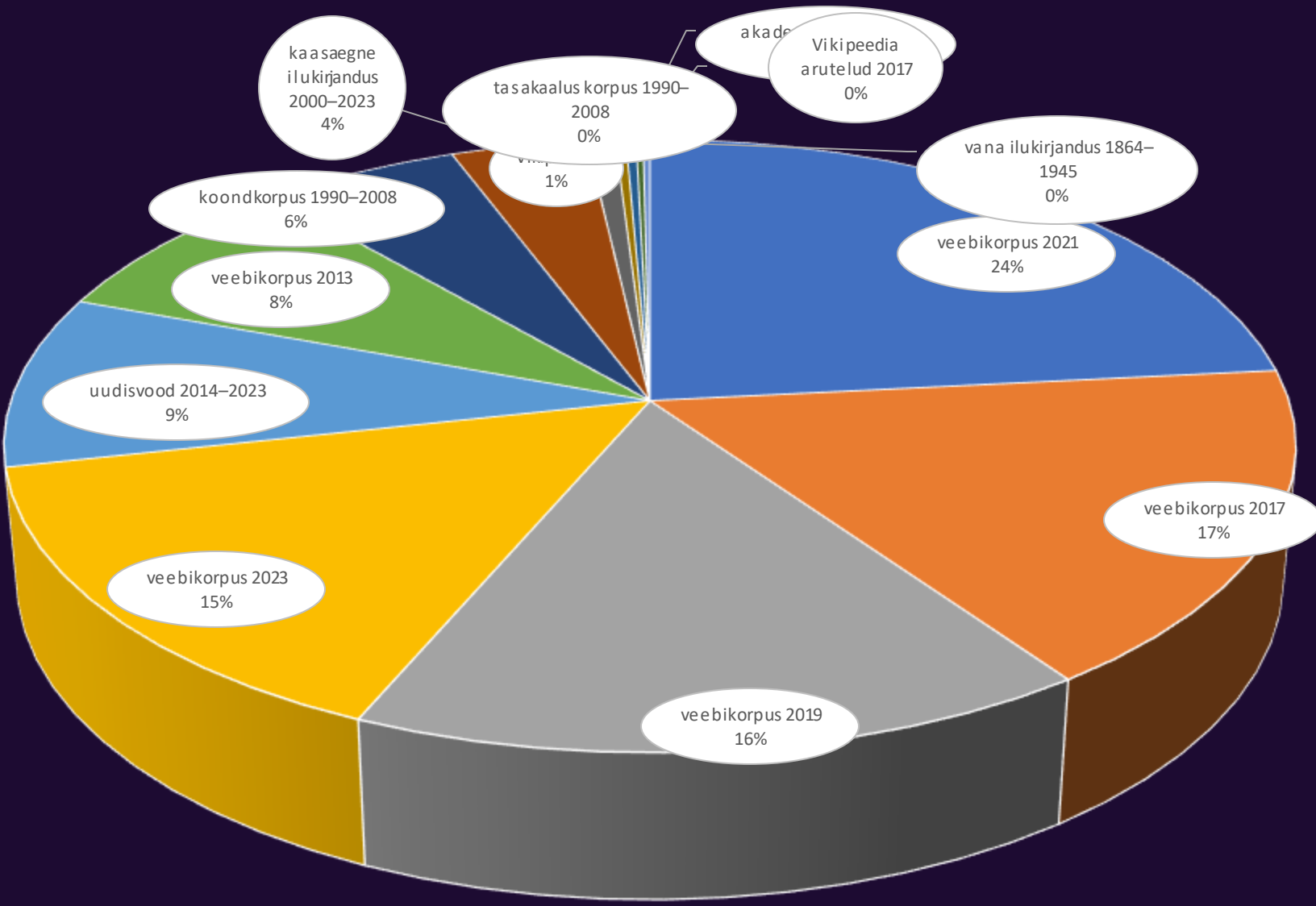
Kaasaegseim, mahukaim ja žanriliselt mitmekesiseim eestikeelsete digitekstide kogu, mis on valminud EKI ja UK/Tšehhi tarkvarafirma Lexical Computing Ltd. koostöös.

	Tekstisõnu	Lauseid	Lõike	Dokumente
eesti keele ühendkorpus 2023	3,8 mld	255 mln	83 mln	15 mln
eesti keele ühendkorpus 2021	2,9 mld	197 mln	64 mln	12 mln
eesti keele ühendkorpus 2019	1,8 mld	120 mln	35 mln	6 mln
eesti keele ühendkorpus 2017	1,3 mld	88 mln	27 mln	3 mln
eesti keele ühendkorpus 2013	563 mln	38 mln	7,5 mln	700 tuh



Eesti keele ühendkorpus 2023

	Allkorpus	Tekstisõnu
1.	veebikorpus 2021	884 524 889
2.	veebikorpus 2017	638 470 425
3.	veebikorpus 2019	613 775 513
4.	veebikorpus 2023	571 221 646
5.	ajamärgistatud uudisvood 2014–2023	344 463 562
6.	veebikorpus 2013	302 950 928
7.	TÜ koondkorpus 1990–2008	212 424 035
8.	kaasaegne ilukirjandus 2000–2023	148 559 737
9.	Vikipeedia	30 747 204
10.	TÜ tasakaalus korpus 1990–2008	11 631 801
11.	akadeemilised tekstid	11 026 315
12.	Vikipeedia arutelud 2017	75 71 584
13.	vana ilukirjandus 1864–1945	7 421 569





Mis tekstid veebikorpuses on?

Meediaportalide ja perioodikaväljaannete tekstid

- Delfi, Postimees, Õhtuleht, ERRi portaal, Saarte Hää, Eesti Ekspress, Eesti Päevaleht, Maaleht, Äripäev, Õpetajate Leht, Sirp, Horisont, Tehnikamaailm, Eesti Loodus, Pealinn, Kesknädal, Lääne Elu, Eesti Kirik, Hiiu Leht, Põhjarannik

Ametlikud ja haldustekstid

- Riigi Teataja, Euroopa Parlament, Riigikohus, Riigikogu, Vabariigi Valitsus, Vabariigi Presidendi Kantselei, Riigikontroll, Õiguskantsleri Kantselei, Välisministeerium, Rahandusministeerium, Haridus- ja Teadusministeerium, Majandus- ja Kommunikatsiooniministeerium, Keskkonnaministeerium, Kaitseministeerium, Siseministeerium, Sotsiaalministeerium, Maksu- ja Tolliamet, Politsei- ja Piirivalveamet, Statistikaamet, Töötukassa, Terviseamet, Maanteeamet, Maa-amet, Transpordiamet, Riigi Infosüsteemi Amet

Asutuste kodulehed ja veebipoed

- Tartu Ülikool, Eesti Maaülikool, Tallinna Tehnikaülikool, Eesti Keele Instituut, Kirjandusmuuseum, Eesti Kunsti Muuseum, Vanemuine, Linnateater, TV3, Apollo, LHV, Kaup24, Photopoint, Tradehouse

Blogid ja foorumid

- Mallukas, Marimell, Paljas Porgand, Päevakera, Tunne Kelam, Kaja Kallas, Henn Põlluaas, Varro Vooglaid, Yoko Alender, Urmas Paet, Simo Runnel
- Matkafoorum, Fotofoorum, Ulmefoorum, Elektroonikafoorum



Mis tekstid korpuses veel on?

Teaduskeel

- Keel ja Kirjandus, Eesti ja soome-ugri keeleteaduse ajakiri, Eesti Rakenduslingvistika Ühingu aastaraamat, Lähivõrdlusi, Methis: Studia Humaniora Estonica, LingVaria, Folklore, Eesti Arst, Eesti Haridusteaduste Ajakiri, Ajalooline Ajakiri, Mäetagused, Estonian Journal of Earth Sciences, Eesti Majanduspoliitilised Väitlused, Agraarteadus jpm

Kirjandus

- Vana ilukirjandus (1864–1945)
- Kaasaegne ilukirjandus (2000–2023)



Kirjanduse allkorpused



Klassifikatsioon:

- **ilukirjandus** (*fiction*): romaan, novell, lühijutt, luule, lasteraamat, följeton, muinasjutt, memuaar
- **mitteilukirjandus** (*non-fiction*): käsiraamat, reisikiri, spordiajalugu, kokaraamat, laulik, õpik, töövihik, aimekirjandus, olukirjandus, biograafia

Vana ilukirjandus 1864–1945 (7 421 569 tekstisõna)

- **Autorid:** Eduard Bornhöhe, Eduard Vilde, Villem Grünthal-Ridala, Elisabeth Aspe, August Kitzberg, Juhan Liiv, Jakob Liiv, August Weizenberg, Carl Robert Jakobson, Anton Hansen Tammsaare, Ado Grenzstein, Friedrich Robert Faehlmann, Johannes Barbarus jpt

Kaasaegne ilukirjandus 2000–2023 (147 273 709 tekstisõna)

- **Kirjastused:** Varrak, Petrone Print, Hea Lugu, Koolibri
- **Autorid (originaalid):** Jaan Ross, Erik Tohvri, Tõnu Õnnepalu, Dagmar Lamp, Leelo Tungal, Karl Ristikivi, Mart Laar, Art Leete, Indrek Hargla, Andres Ehin, Lembit Uustulnd, Jaan Undusk, Aimée Beekman, Andrus Kivirähk, Raimond Kaugver jpt
- **Autorid (tõlked):** Emily Brontë, Rosamunde Pilcher, Rex Stout, Jane Austen, Agatha Christie, Virginia Woolf, Umberto Eco, Justin Petrone, Roy Jenkins, Terry Pratchett, Doris Lessing, William Shakespeare, Tove Jansson, Stephen King, J. K. Rowling jpt



Kirjanduse allkorpuste metainfo

Display and count metadata

<s>Keset ulgumerd muutub silmapiir normaalse **muumitrolli** jaoks veidike liiga avaraks.</s>

✓ Source corpus	Literature Contemporary 2000–2023
✓ Token number	1644839311
✓ Document number	6632718
✓ Genre	fiction:Children's booklfiction
✓ Literature ISBN	978-9916-6-0107-5
✓ Literature original author	Tove Jansson
✓ Literature original title	Muminpappans memoarer
✓ Literature original year	1950
✓ Literature publisher	Hea Lugu
✓ Literature translated yes/no	yes
✓ Literature translation source language	SV
✓ Literature translation target language	ET
✓ Literature translator	Vladimir Beekman
✓ Title	Muumipapa memuaarid
✓ Year of publication	2020



Kirjanduse allkorpuste piirangutest

- Sisu maht: AutÕS
 - vrd Abruka raamatukogu
- Ärihuvid vs. uurijate huvid
- Ideaal: kõik RaRas deponeeritud raamatute digiversioonid?
- Tasakaaluta olek ongi „tasakaalus“ olek
- SkE ÜK kõigile vabalt ligipääsetavaks: tehniline lahendus AÕ-piirangutega tekstidele?
- Kogumistöö jätkub



Veebikorpuse kvaliteedikontroll



Veebikorpus 2023 koguti kokku 46 770 URL-ilt

5000 URL-i käsitsikontroll (96% kõigist kogutud tekstisõnadest)

Tulemus: eemaldati ligikaudu 40% veebist kogutud materjalist

URL	Tekstisõnu
blogspot.com	263 571 500
delfi.ee	179 843 943
postimees.ee	155 515 211
ohtuleht.ee	121 022 262
err.ee	103 003 956
wordpress.com	97 808 291
europa.eu	63 285 462
riigikogu.ee	61 672 165
wikipedia.org	38 323 022
soccernet.ee	38 204 936

Eesti keele veebikorpuse 2023 veebiallikate esikümme



Veebitekstide klassifitseerimine

Paralleelselt masintõkeliste tekstide tuvastamisega

Kolmetasemeline klassifikatsioon: **žanrid** (laiem klass), **teemad** (kitsam klass), **allteemad**

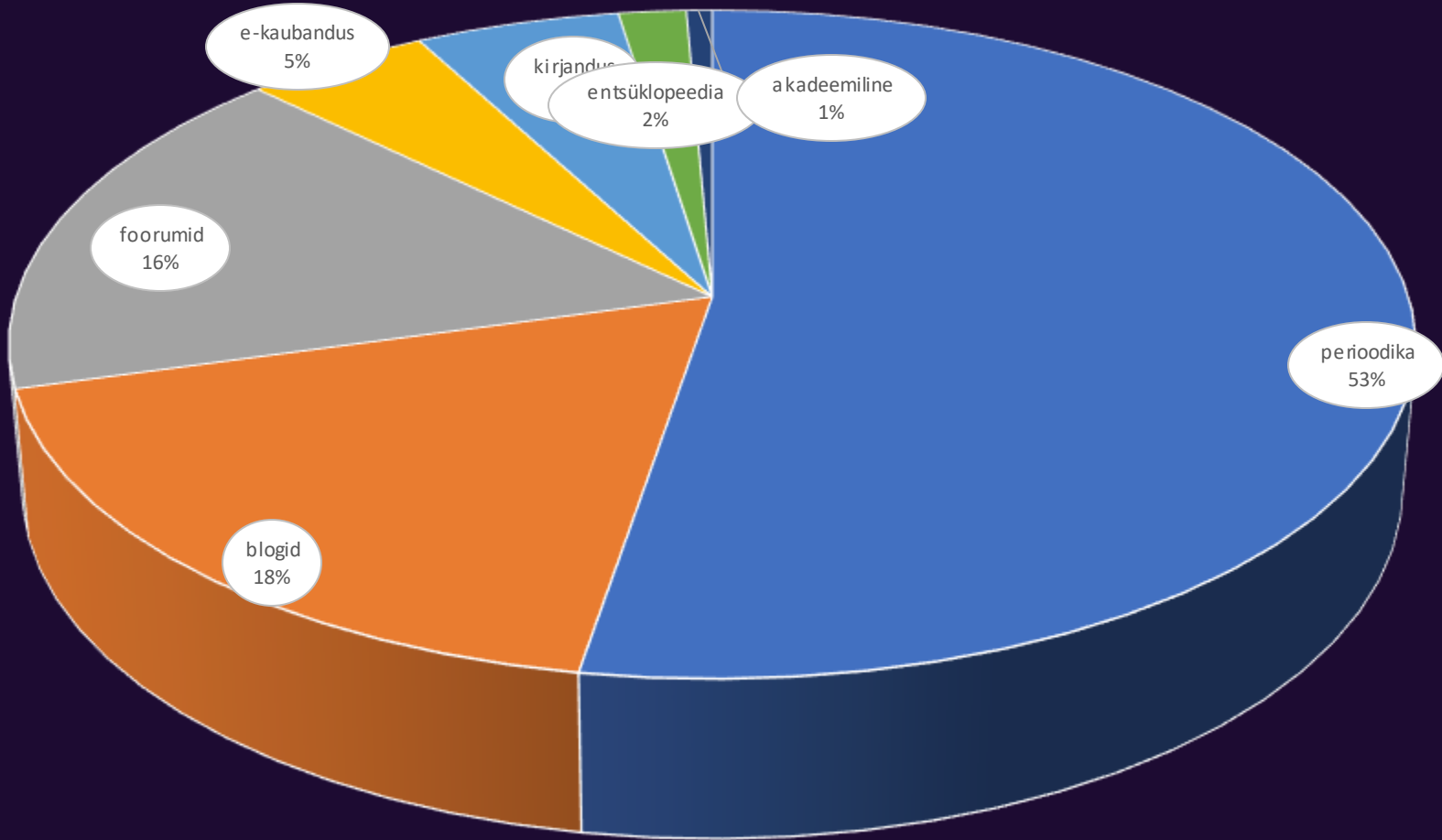
Žanri määrab kirjutamisstiil või avaldamiskoht (nt ilukirjandus, e-kaubandus), (all)teemad (nt sport) võivad olla žanride ülesed (nt sport = sportlase kirjutatud autobiograafia, aga ka spordikaupu müüv e-pood)

Käsitsi klassifitseeritud andmete kasutamine treeningandmetena

Tulemus: žanr on määratud u 79% ja teema u 50% tekstidest

Žanrid

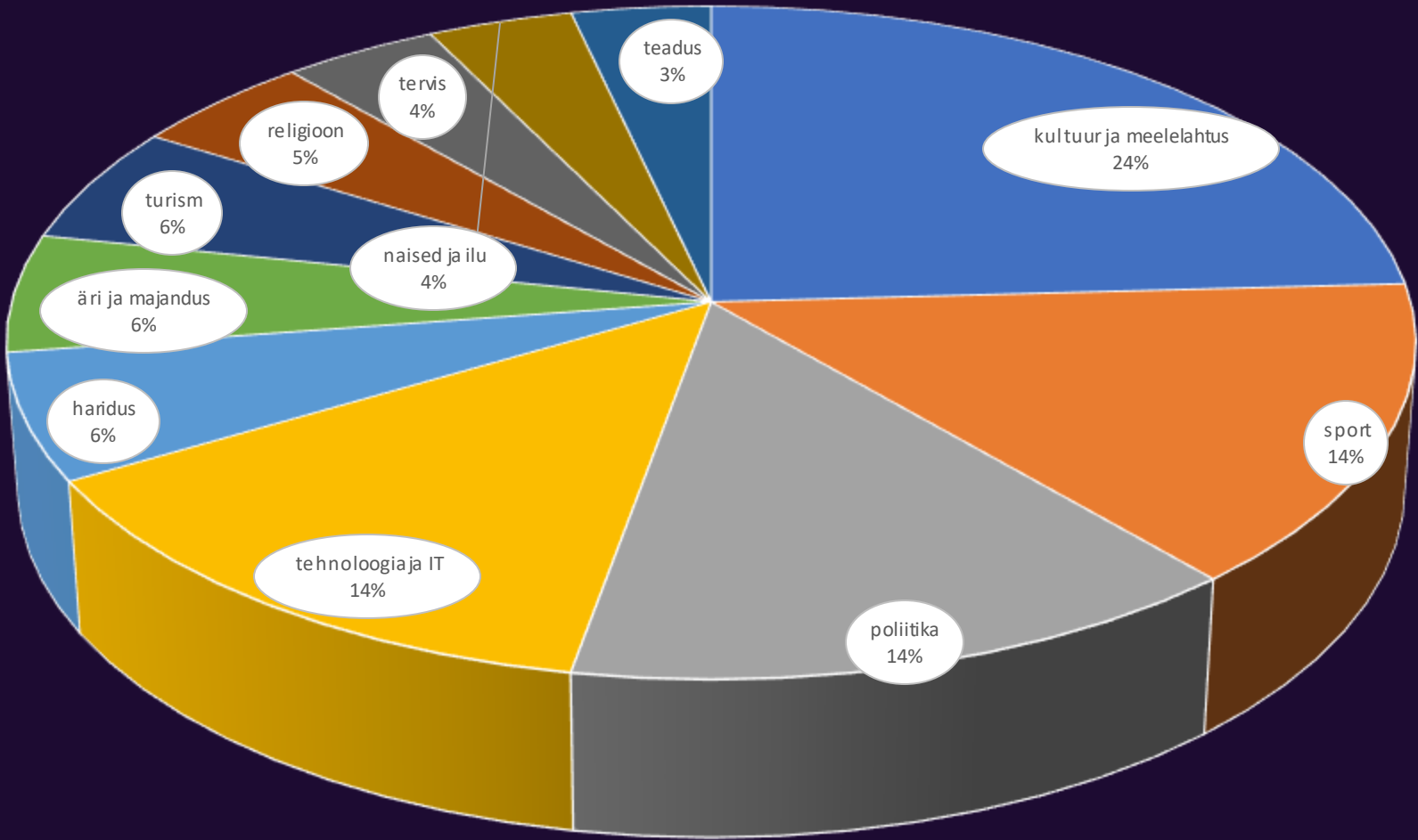
	Tekstisõnu
perioodika	1 574 063 873
blogid	537 993 698
foorumid	497 219 390
(ilu- ja mitteilu)kirjandus	161 255 917
e-kaubandus	145 984 966
entsüklopeedia	53 420 451
akadeemiline	20 329 641



Teemad

	Tekstisõnu
kultuur ja meelelahtus	382 851 434
sport	231 393 259
poliitika ja valitsemine	224 094 079
tehnoloogia ja IT	217 117 346
haridus	96 707 296
äri, rahandus ja majandus	89 853 689
turism	89 325 169
religioon	75 187 695
tervis	63 852 732
naised ja ilu	59 374 774
teadus	57 064 418

Kokku 21 teemat, lisaks: ehitus ja kinnisvara, loodus ja keskkond, õigus ja juura, seks, loomad, pere ja kodu, ajalugu, hasart- ja videomängud, toit ja jook, põllumajandus





Ajamärgistatud uudisvood 2014–2023

Staatiline väljavõte kahest eraldiseisvatest monitorkorpustest

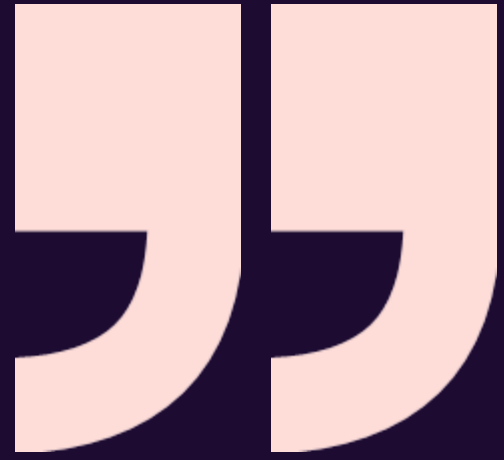
- JSI ajamärgistatud eesti keele uudisvoo korpus (2014–...)
- Estonian Trends (al 2020. a märtsist)

Estonian Trends

- kasutab RSS uudisvoo funktsionaalsust
- tekste kogutakse iga päev, SkE-sse laaditakse u kord kuus
- 2024. a aprillikuu seisuga **218 771 826 sõnet** u 500 veebilehelt

	Attribute value	Token coverage ?
1	pmo.ee	59,736,745
2	err.ee	40,081,999
3	ohtuleht.ee	18,592,449
4	google.com	11,285,859
5	delfi.ee	10,273,230
6	uueduudised.ee	6,212,832
7	minuaeg.com	2,577,072
8	sirp.ee	2,461,510
9	le.ee	2,416,587
10	aripaev.ee	2,398,989
11	lounaeestlane.ee	2,286,454
12	objektiiv.ee	2,054,182
13	geenius.ee	1,903,150
14	goodnews.ee	1,882,573
15	telegram.ee	1,228,988
16	temuki.ee	1,200,075
17	eestikirik.ee	1,177,923
18	valitsus.ee	1,130,075
19	tartu.ee	1,033,810
20	opleht.ee	1,028,139
21	kesknadal.ee	1,014,562
22	riigikogu.ee	1,013,561
23	tv3.ee	921,940
24	piletilevi.ee	854,423
25	nordenbladet.ee	833,838

Korpuse Estonian Trends esimesed 25 veebiallikat



Uued analüüsivõimalused korpuspäringusüsteemis Sketch Engine



Uued analüüsivõimalused

- sõnavisandite grammatika versioon 2.1 (*Sketch Grammar*): kokku 113 grammatilist suhet; süntaksipõhine
- **kollokatsioonide žanrilise ja temaatilise kuuluvuse tuvastus sõnavisandites (*Word Sketch*)**
- **diakrooniline analüüs (*Timeline*)**
- **trendid (*Trends*)**
- **tähendusvihjed (*Word Sense Induction*)**
- terminituvastus (*Terms*)
- sõnavektorid ([Word Embeddings](#))



Zanrid ja (all)teemad sõnavisandites

mängima as verb 1,513,767×					
subjektiga			objektiga		
meeskond	3,751	9.2 ...	roll	5,055	10.8 ...
meeskond mängib			ei mängi rolli		
• concentrated in: Wikipedia 2023 ?			• concentrated in: sports:rally ?		
• concentrated in: sports:football ?			• concentrated in: health ?		
• concentrated in: sports:basketball ?			• concentrated in: gaming & gambling:gambling & casinos ?		
• concentrated in: sports ?			• concentrated in: sex ?		
• concentrated in: gaming & gambling ?			show less		
muusika	2,614	9.1 ...	mäng	4,930	10.3 ...
muusika mängib			mängida mäng		
• concentrated in: culture & entertainment:movies & TV ?			• concentrated in: gaming & gambling:video games ?		
• concentrated in: culture & entertainment ?			• concentrated in: education ?		
• concentrated in: culture & entertainment:music ?			• concentrated in: home, family & children ?		
• concentrated in: culture & entertainment:theatre ?			• concentrated in: technology & IT:computers & IT ?		
• concentrated in: home, family & children ?			• concentrated in: sports:basketball ?		
show less			• concentrated in: culture & entertainment:books & literature ?		
peaosa	1,490	8.6 ...	• concentrated in: gaming & gambling ?		
peaosa mängib			• concentrated in: sports:football ?		
• concentrated in: Wikipedia 2023 ?			• concentrated in: gaming & gambling:gambling & casinos ?		
• concentrated in: Reference Corpus 1990–2008 ?			show less		
• concentrated in: culture & entertainment:books & literature ?			kitarr	2,020	9.6 ...
• concentrated in: culture & entertainment:music ?			mängib kitarri		
• concentrated in: culture & entertainment ?			• concentrated in: Wikipedia 2023 ?		
• concentrated in: culture & entertainment:movies & TV ?			• concentrated in: culture & entertainment:movies & TV ?		
• concentrated in: encyclopedia ?			• concentrated in: culture & entertainment:music ?		
• concentrated in: periodicals:journals ?			• concentrated in: culture & entertainment ?		
			show less		
määrsõnaga					
koos	6,956	9.6 ...	koos		
koos mängida			koos mängida		
• concentrated in: sports:basketball ?			• concentrated in: sports:basketball ?		
• concentrated in: gaming & gambling:video games ?			• concentrated in: gaming & gambling:gambling & casinos ?		
• concentrated in: gaming & gambling:gambling & casinos ?			• concentrated in: sports:football ?		
• concentrated in: sports:football ?			• concentrated in: home, family & children ?		
• concentrated in: home, family & children ?			• concentrated in: pets and animals ?		
• concentrated in: pets and animals ?			• concentrated in: sports ?		
• concentrated in: sports ?			• concentrated in: culture & entertainment:music ?		
• concentrated in: culture & entertainment:music ?			• concentrated in: gaming & gambling ?		
• concentrated in: gaming & gambling ?			• concentrated in: culture & entertainment:theatre ?		
• concentrated in: culture & entertainment:theatre ?			show less		
võõrsil	1,810	9.1 ...	võõrsil		
võõrsil mänginud			võõrsil mänginud		
• concentrated in: Timestamped 2014–2023 ?			• concentrated in: Timestamped 2014–2023 ?		
• concentrated in: sports ?			• concentrated in: sports ?		
• concentrated in: gaming & gambling ?			• concentrated in: gaming & gambling ?		
• concentrated in: sports:football ?			• concentrated in: sports:football ?		
show less			show less		
hästi	6,677	8.5 ...	hästi		
hästi mänginud			hästi mänginud		
• concentrated in: sports:basketball ?			• concentrated in: sports:basketball ?		
• concentrated in: sports ?			• concentrated in: sports ?		
• concentrated in: sports:chess & checkers ?			• concentrated in: sports:chess & checkers ?		
• concentrated in: gaming & gambling:gambling & casinos ?			• concentrated in: gaming & gambling:gambling & casinos ?		
• concentrated in: culture & entertainment:theatre ?			• concentrated in: culture & entertainment:theatre ?		



Trendid (*Trends*)

- Analüüsib keelendi kasutust teatud ajaühiku (kuu, kvartal, aasta) jooksul, võrreldes selle kasutussagedust sama korpusse sees teiste samaväärsete ajavahemike lõikes.
- Tulemus on loend sõnadest, mille kasutussagedus on aja jooksul muutunud:
 - kasutuse suurenedes on trend positiivne (**roheline nool** suunaga üles)
 - kasutuse langedes negatiivne (**punane nool** suunaga alla)

TRENDS

Estonian Trends

Lemma	Trend ↓	Frequency
1 filmisaade	↘	144 ...
2 suurtükimürsk	↗	668 ...
3 ründedroon	↗	488 ...
4 nakatumisjuht	↘	126 ...
5 sõjauuring	↗	935 ...
6 kaudtuleüksus	↗	208 ...
7 nakatunu	↘	14,172 ...
8 ülalolev	↘	97 ...
9 lehesaba	↘	376 ...
10 täiemahuline	↗	3,270 ...
11 föderaaaleelarve	↗	90 ...
12 kommentaaritaotlus	↗	98 ...
13 poliitikaguru	↘	78 ...

Lemma	Trend ↓	Frequency
14 agressioonisõda	↗	1,333 ...
15 rindelõik	↗	823 ...
16 sõjablogija	↗	538 ...
17 sõjaväeülem	↗	114 ...
18 rindejoon	↗	4,235 ...
19 koroonaviirusevastane	↘	180 ...
20 ninaneel	↘	304 ...
21 suuseks	↘	163 ...
22 koroonastatistika	↘	159 ...
23 koroonates	↘	80 ...
24 koroonaaajastu	↘	197 ...
25 kaugmaa	↗	256 ...
26 sõjaväevalitsus	↗	183 ...

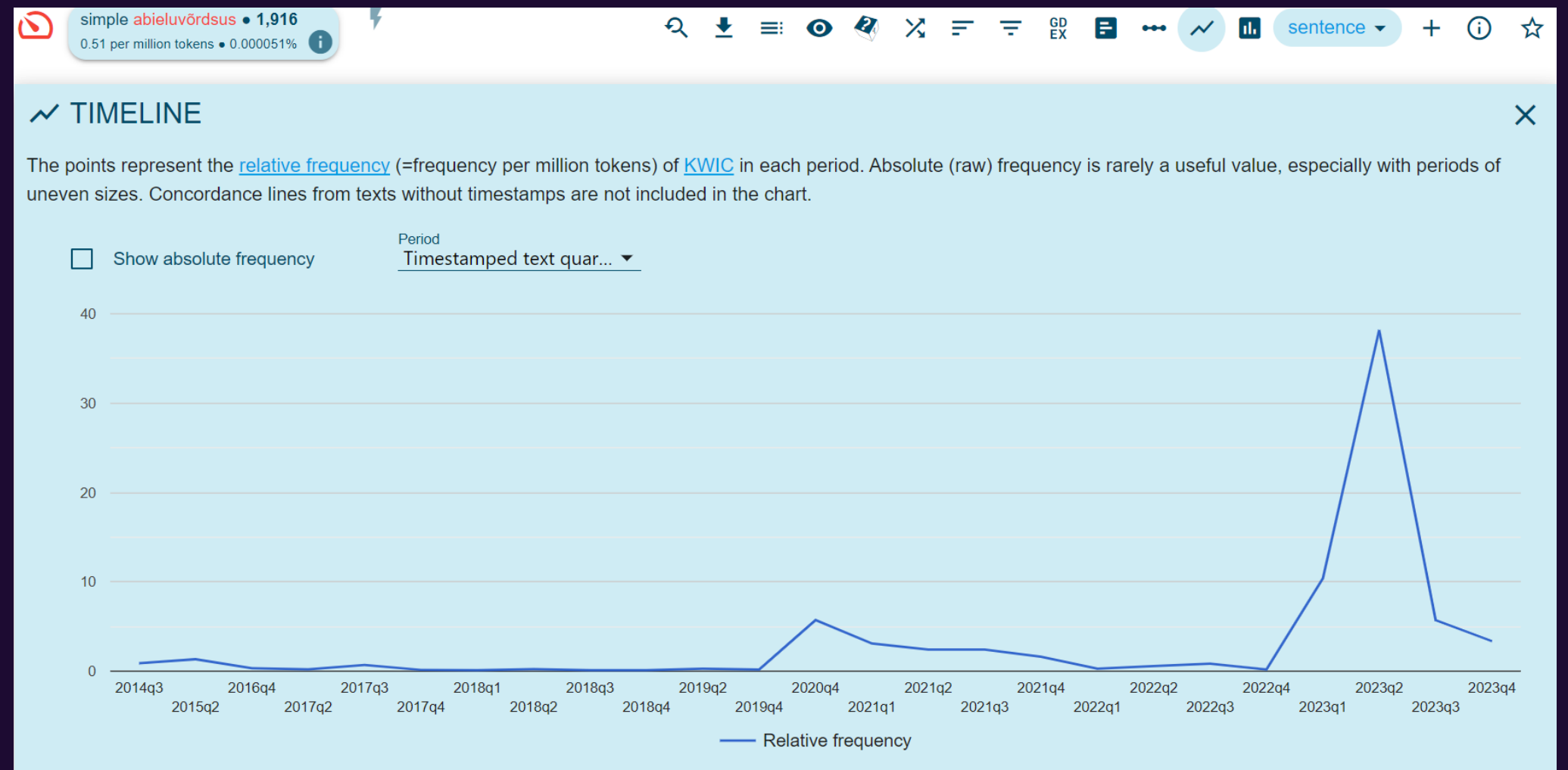
aprill 2024



Ajajoon (*Timeline*): *abieluvõrdsus*

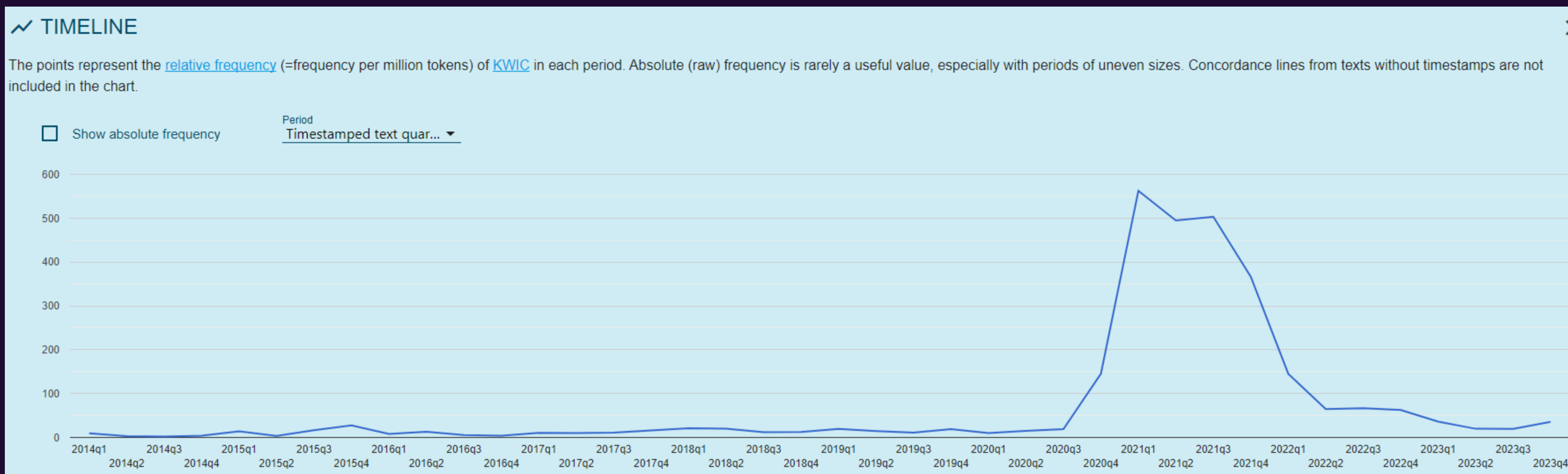
- Näitab sõna sageduse distributsiooni diakroonilise atribuudi (ajamärgise) järgi.
- Pildil** lemma *abieluvõrdsus* ajajoon ENC2023-s – hüppeline kasv 2023. aasta teises kvartalis.*

*2023. a juunis läbis valitsus Riigikogus usaldushääletuse, mis oli seotud abieluvõrdsust kehtestava perekonnaseaduse muutmise ja sellega seondult teiste seaduste muutmise seadusega.





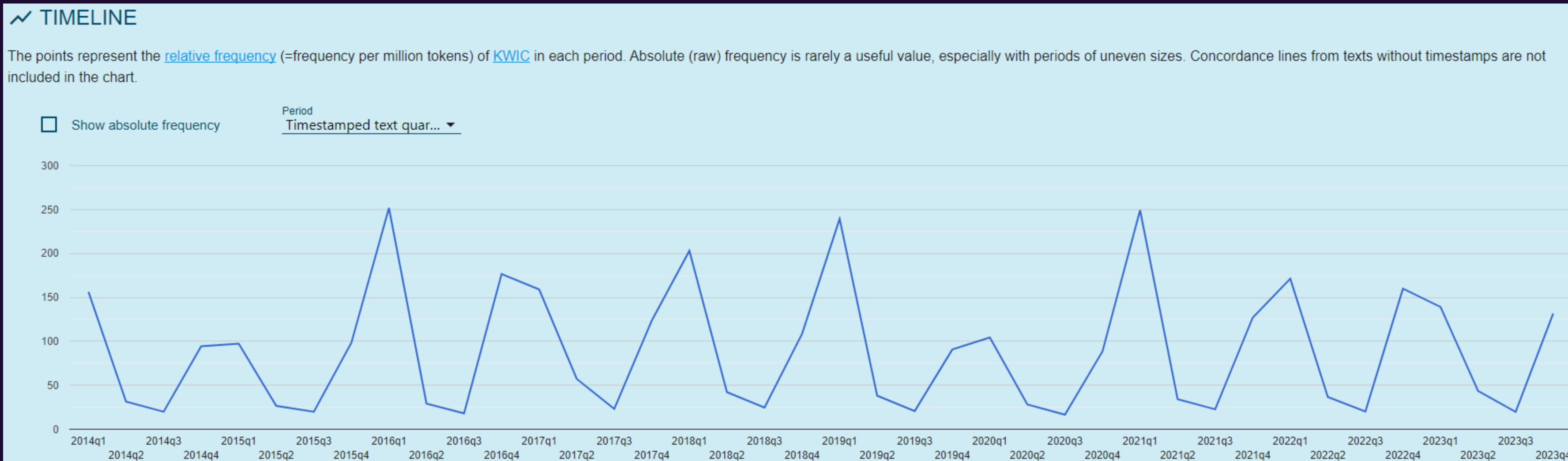
Ajajoon: *vaktsineerimine*



Hüppeline kasv 2021. aasta esimeses kvartalis – Eestis hakati koroonaga vastu vaktsineerima 27.12.2020.

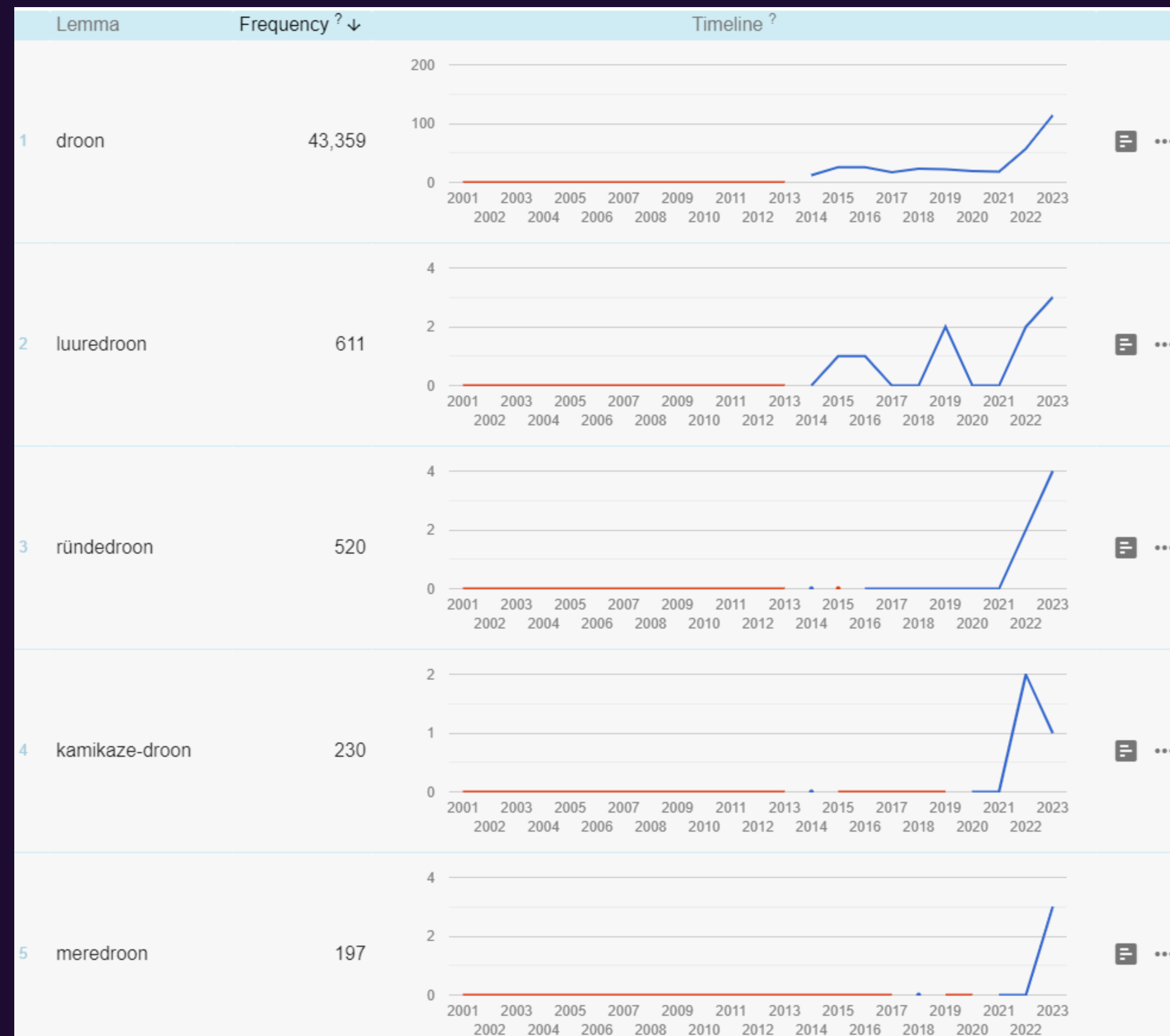


Ajajoon: *lumi*





Ajajoon sõnaloendites (*Wordlist*)





Tähendusvihjed (*Word Sense Induction*)

Automatic identification of word senses

Sketch Engine contains functionality which identifies word senses automatically. This function is included in the Word Sketch tool. Sketch Engine can perform **word sense induction**, not disambiguation. Please see the info box.

The word sense induction tool will categorize the collocations identified by the word sketch into groups corresponding to the different senses of the word. This makes it possible to only display collocations related to a particular sense and investigate the usage of the word when used in this sense only. Each sense is assigned a colour and the related collocations are colour-coded accordingly.

Sense names

The senses are not named, but labelled. The label consists of a number of words which are representative of the word sense. The user must infer the sense name from the labels.

WORD SKETCH card as noun 8,681,658x English Web 2021 (enTenTen21)

SENSES

- ✓ █ deck, discard, shuffle, token
- ✓ █ credit, debit, bank, ATM
- ✓ █ notecard, Cards, souvenir, sticker
- ✓ █ SD stick, SDHC, Firewire

modifiers of "card"	nouns modified by "card"	verbs with "card" as object	verbs with "card" as subject
credit credit card	slot card slot	greet greeting cards	feature card features
gift gift cards	reader card reader	play playing cards	accept credit cards accepted
debit debit card	debt credit card debt	deal cards are dealt	count card counting
graphics graphics card	holder card holders	issue card issued	expire card expires
id ID card	issuer card issuers	accept accept credit cards	carry a card carrying
SD adjective the SD card	payment credit card payments	draw draw a card	depict cards depicting
memory memory card	transaction credit card transactions	insert card is inserted	sort card sorting



Tähendusvihjed: *mängima*

SENSES

✓ | mänguminut, koosseis, väravavaht, mänguliselt —

✓ | hullama, mürama, peitus, põngerjas —

✓ | vastane, meistrisari, võitja, mäng —

✓ | sisustama, lõbus, seltskonnamäng, lustima —

✓ | näitlema, näitlemine, näitleja, anne —

✓ | mängiv, tantsima, roll, näitlema —

✓ | sub, audio, mp3, xbox —

✓ No sense



Tähendusvihjed: *leht*

SENSES

- ✓  link, sait, linkima, kopeerima 
- ✓  toimetuse, artikkel, vastulause, valefakt 
- ✓  hiirekõrv, kastan, vahtrapuu, kuuseoks 
- ✓  2016, 2014, 2015, 2011 
- ✓  risoom, lehevars, õienupp, koor 
- ✓  linkima, veebiaadress, lehekülg, veebileht 
- ✓  link, lehekülg, jooksvalt, Facebook 
- ✓  Other, Selle, -., detail 
- ✓ No sense



Kokkuvõte

- Eesti Keele Instituudis käib süstemaatiline töö korpuste kogumise eesmärgil (veebikorpused, ilukirjandus, teaduskeel, valdkonnakorpused, õppijakeele korpused, õpikute korpused jt)
- EKI keele- ja kõnetehnoloogia osakond tegeleb suurte keelemudelite arendamise eesmärgil eestikeelsete andmete kogumisega, mis võimaldab ühendkorpust tulevikus uuendada sagedamini kui iga kahe aasta tagant

- Jakubíček, Miloš; Kilgariff, Adam; Kovář, Vojtěch; Rychlý, Pavel; Suchomel, Vít 2013.
The TenTen Corpus Family. – 7th International Corpus Linguistics Conference CL 2013, 125–127.
- Koppel, Kristina; Kallas, Jelena 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. Eesti Rakenduslingvistika Ühingu aastaraamat, 18, 207–228.
- Suchomel, Vít 2020. Better Web Corpora For Corpus Linguistics And NLP. Masaryk University, Faculty of Informatics.
- Suchomel, Vít; Kraus, Jan 2021. Website Properties in Relation to the Quality of Text Extracted for Web Corpora. – Proceedings of Recent Advances in Slavonic Natural Language Processing, 167–175.