EESTI
KEELE
INSTITUUT

Estonian National Corpus 2013–2021: the largest collection of Estonian language data

Kristina Koppel, PhD
Senior Computational Lexicographer

Jelena Kallas, PhD
Senior Computational Lexicographer

19th Annual Conference of Applied Linguistics 2022
16.VI.2022

Paper in the 18th Estonian Papers in Applied Linguistics

Koppel, Kristina; Kallas, Jelena (2022). Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 18, 207–228.

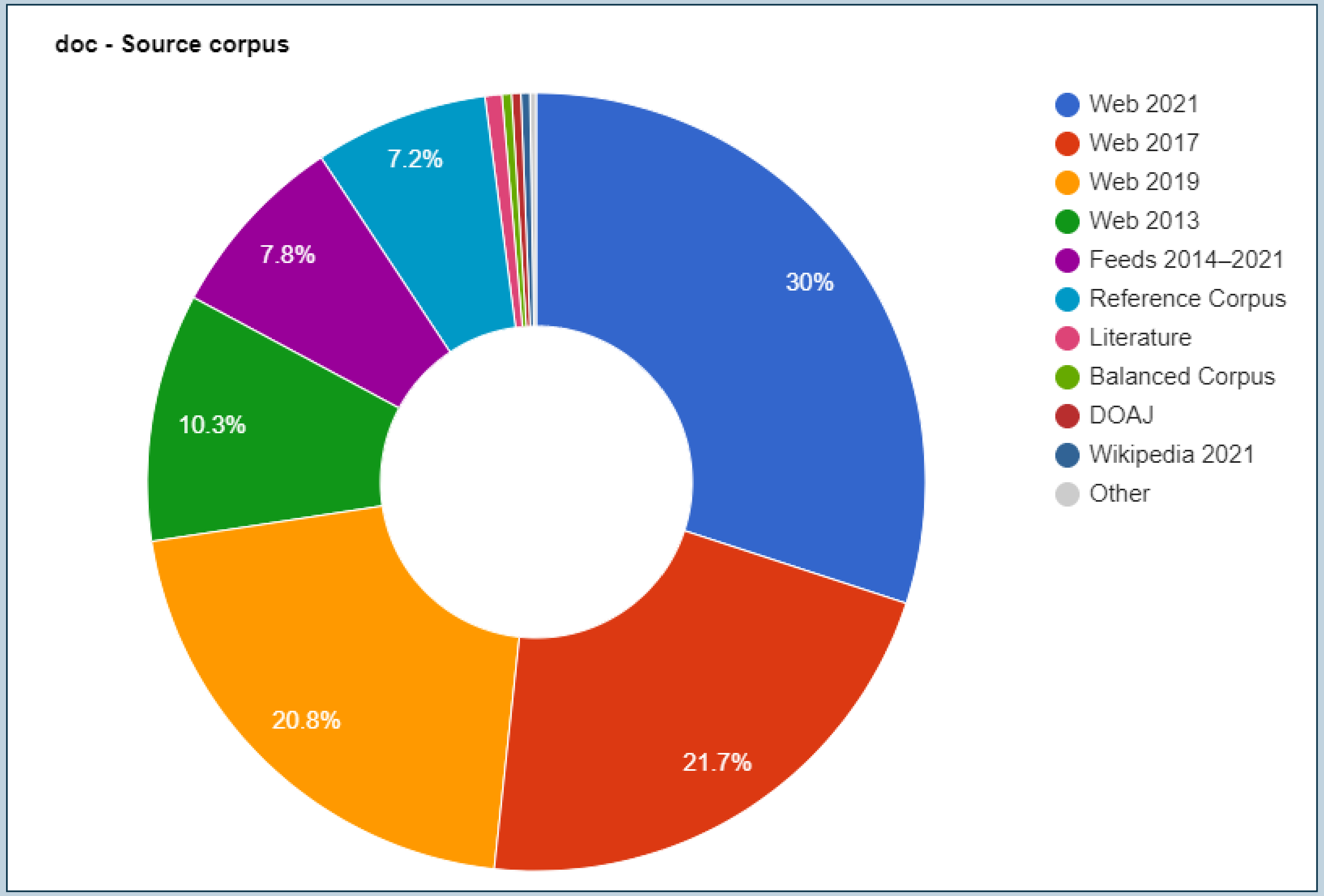
Estonian National Corpus 2013–2021

	Tokens	Words	Sentences	Paragraphs	Documents
ENC 2013	563 mil	464 mil	38 mil	7.5 mil	700 thsnd
ENC 2017	1.3 bil	1.1 bil	88 mil	27 mil	3 mil
ENC 2019	1.8 bil	1.5 bil	120 mil	35 mil	6 mil
ENC 2021	2.9 bil	2.4 bil	197 mil	64 mil	12 mil

EESTI
KEELE
INSTITUUT

Subcorpora of ENC 2021

Subcorpus	Token coverage
Estonian Web 2021	884,525,071
Estonian Web 2017	638,470,413
Estonian Web 2019	613,775,477
Estonian Web 2013	302,402,148
Feeds 2014–2021	230,402,148
Reference Corpus	212,423,989
Literature	20,625,953
Balanced Corpus	11,631,784
DOAJ	11,337,862
Wikipedia 2021	10,941,244
Wikipedia Talk 2017	7,571,854



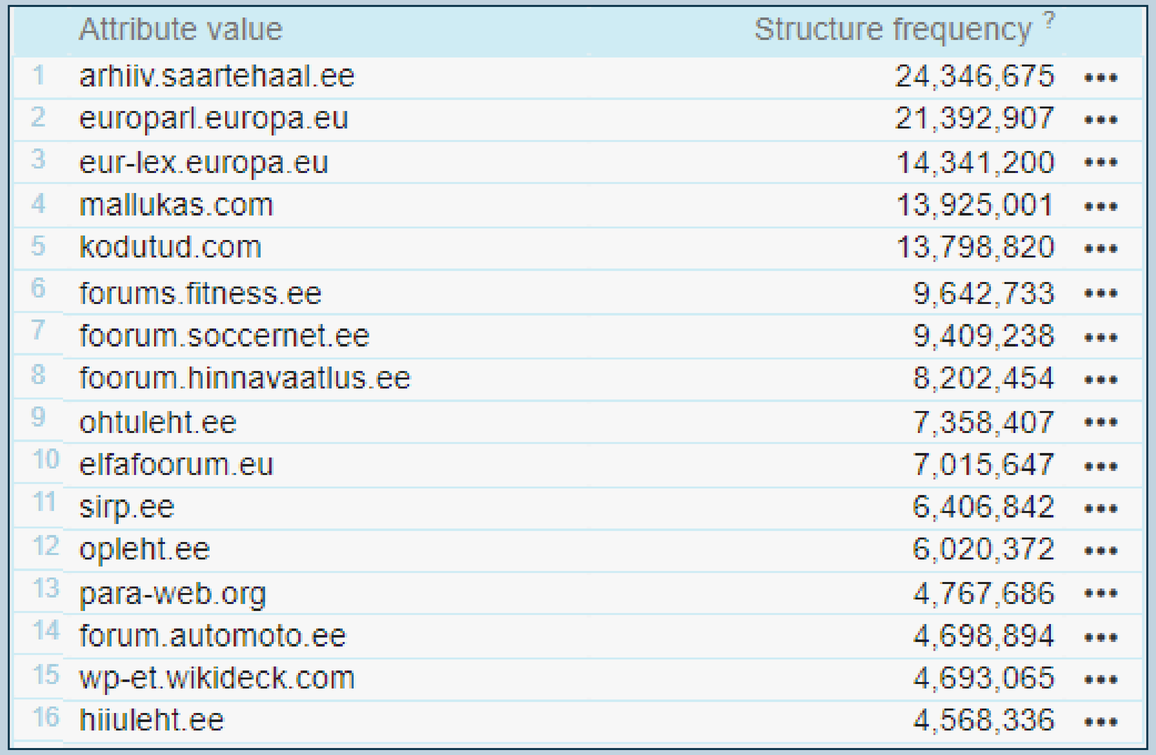
Estonian Web

Web Crawling

1. Crawler looks for text rich resources and avoids websites containing material mostly not suitable for text corpora
2. Starts from trustworthy seed domains
   - In 2019/2021: 851 manually detected URLs, 20,718 from neti.ee
3. Automatic cleaning of crawled data
   - Removing of boilerblates
   - Language detection
   - Removing of (near) duplicates

Manual domain check

1. On the basis of URLs
   - List of 5000 URLs most represented in the corpus
   - Out of 4002 manually checked URLs, 1036 (~26%) were computer generated / machine translated
2. On the basis of Keywords
   - Identifies words that appear more frequently in the focus corpus than in the reference corpus
   - Additional 352 URLs were removed



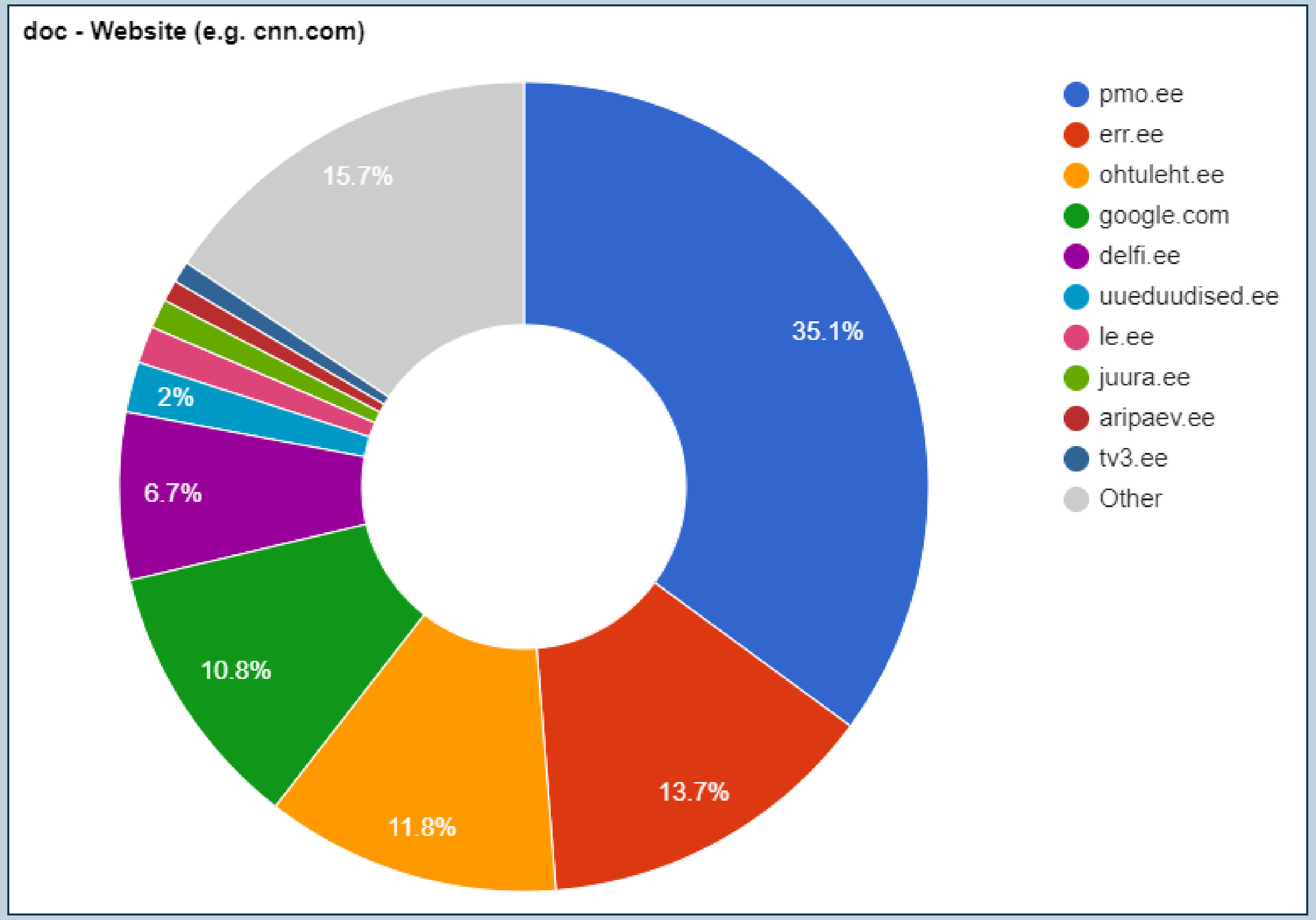
	Attribute value	Structure frequency ?	
1	arhiiv.saartehaal.ee	24,346,675	•••
2	europarl.europa.eu	21,392,907	•••
3	eur-lex.europa.eu	14,341,200	•••
4	mallukas.com	13,925,001	•••
5	kodutud.com	13,798,820	•••
6	forums.fitness.ee	9,642,733	•••
7	foorum.soccernet.ee	9,409,238	•••
8	foorum.hinnavaatlus.ee	8,202,454	•••
9	ohtuleht.ee	7,358,407	•••
10	elfafoorum.eu	7,015,647	•••
11	sirp.ee	6,406,842	•••
12	opleht.ee	6,020,372	•••
13	para-web.org	4,767,686	•••
14	forum.automoto.ee	4,698,894	•••
15	wp-et.wikideck.com	4,693,065	•••
16	hiiuleht.ee	4,568,336	•••

Top web domains in Estonian Web 2021

Feeds 2014–2021 subcorpus

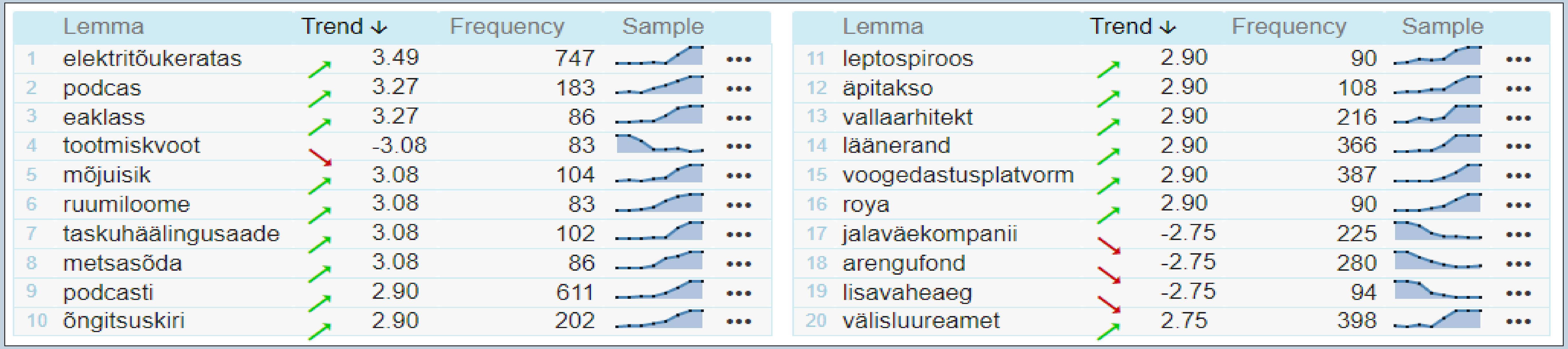
Consists of texts from two monitor corpora as of the end of 2021

1. **Estonian RSS Feeds Corpus** (2020–…)
   - ~700 manually detected seed URLs, 555 active
   - ~86 million tokens
2. **Timestamped JSI web corpus 2014–2020 Estonian**
   - ~80,000 URLs, 524 Estonian web pages
   - ~255 million tokens



Most represented sources in Estonian RSS Feeds Corpus

Trends in Sketch Engine

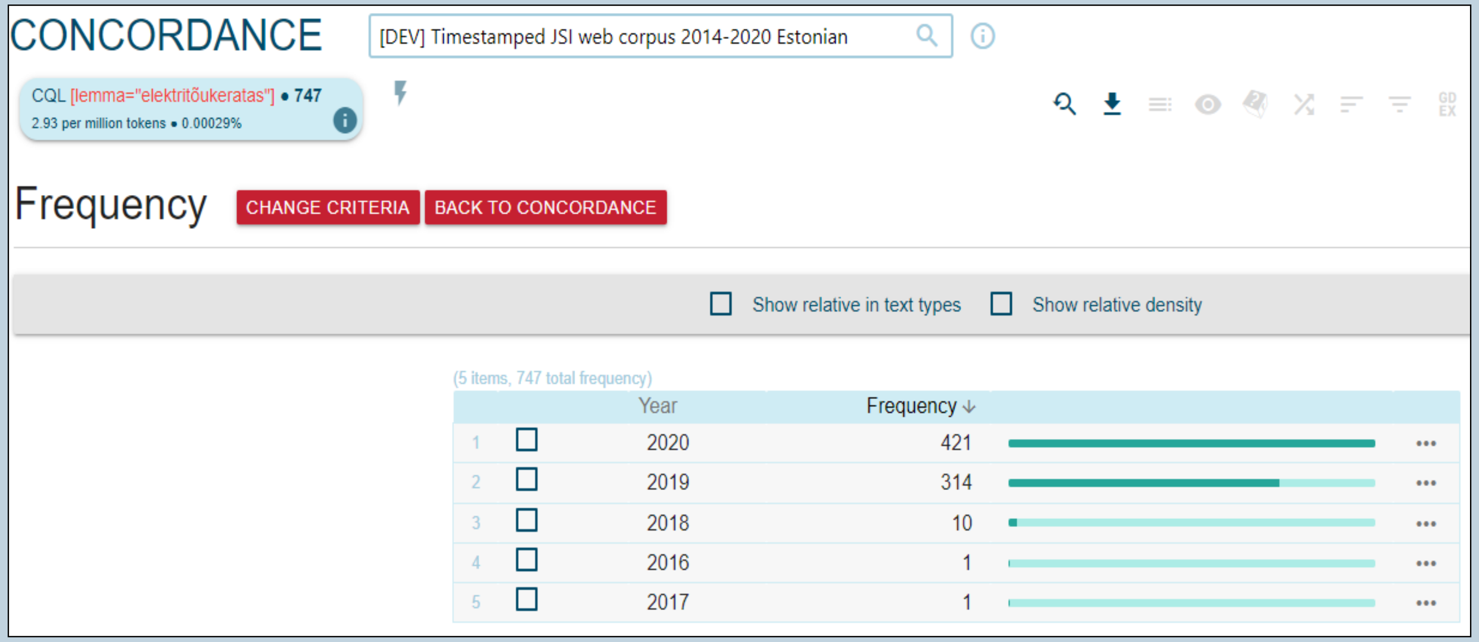


	Lemma	Trend ↓	Frequency	Sample	
1	elektritõukeratas	↗ 3.49	747		•••
2	podcas	↗ 3.27	183		•••
3	eaklass	↗ 3.27	86		•••
4	tootmiskvoot	↘ -3.08	83		•••
5	mõjuisik	↗ 3.08	104		•••
6	ruumiloome	↗ 3.08	83		•••
7	taskuhäälingusaade	↗ 3.08	102		•••
8	metsasõda	↗ 3.08	86		•••
9	podcasti	↗ 2.90	611		•••
10	õngitsuskiri	↗ 2.90	202		•••

	Lemma	Trend ↓	Frequency	Sample	
11	leptospiroos	↗ 2.90	90		•••
12	äpitakso	↗ 2.90	108		•••
13	vallaarhitekt	↗ 2.90	216		•••
14	läänerand	↗ 2.90	366		•••
15	voogedastusplatvorm	↗ 2.90	387		•••
16	roya	↗ 2.90	90		•••
17	jalaväekompanii	↘ -2.75	225		•••
18	arengufond	↘ -2.75	280		•••
19	lisavaheaeg	↘ -2.75	94		•••
20	välisluureamet	↗ 2.75	398		•••

Trends in the Timestamped JSI web corpus

The trends feature analyses the frequency of the use of a word in time by comparing the frequency of use across a series of comparable time periods



CONCORDANCE [DEV] Timestamped JSI web corpus 2014-2020 Estonian

CQL [lemma="elektritõukeratas"] • 747
2.93 per million tokens • 0.00029%

Frequency CHANGE CRITERIA BACK TO CONCORDANCE

□ Show relative in text types □ Show relative density

(5 items, 747 total frequency)

		Year	Frequency ↓	
1	□	2020	421	•••
2	□	2019	314	•••
3	□	2018	10	•••
4	□	2016	1	•••
5	□	2017	1	•••

Distribution of the frequency of lemma *elektritõukeratas* ‘electric scooter’

Reference Corpus and Balanced Corpus

1. Reference Corpus consists of written texts from 1990 until 2008
   - Fiction
   - Periodicals
   - New Media
   - Parliament transcripts
   - PhD dissertations
2. Balanced Corpus
   - Contains equal amount of fiction, journalistic and scientific writing

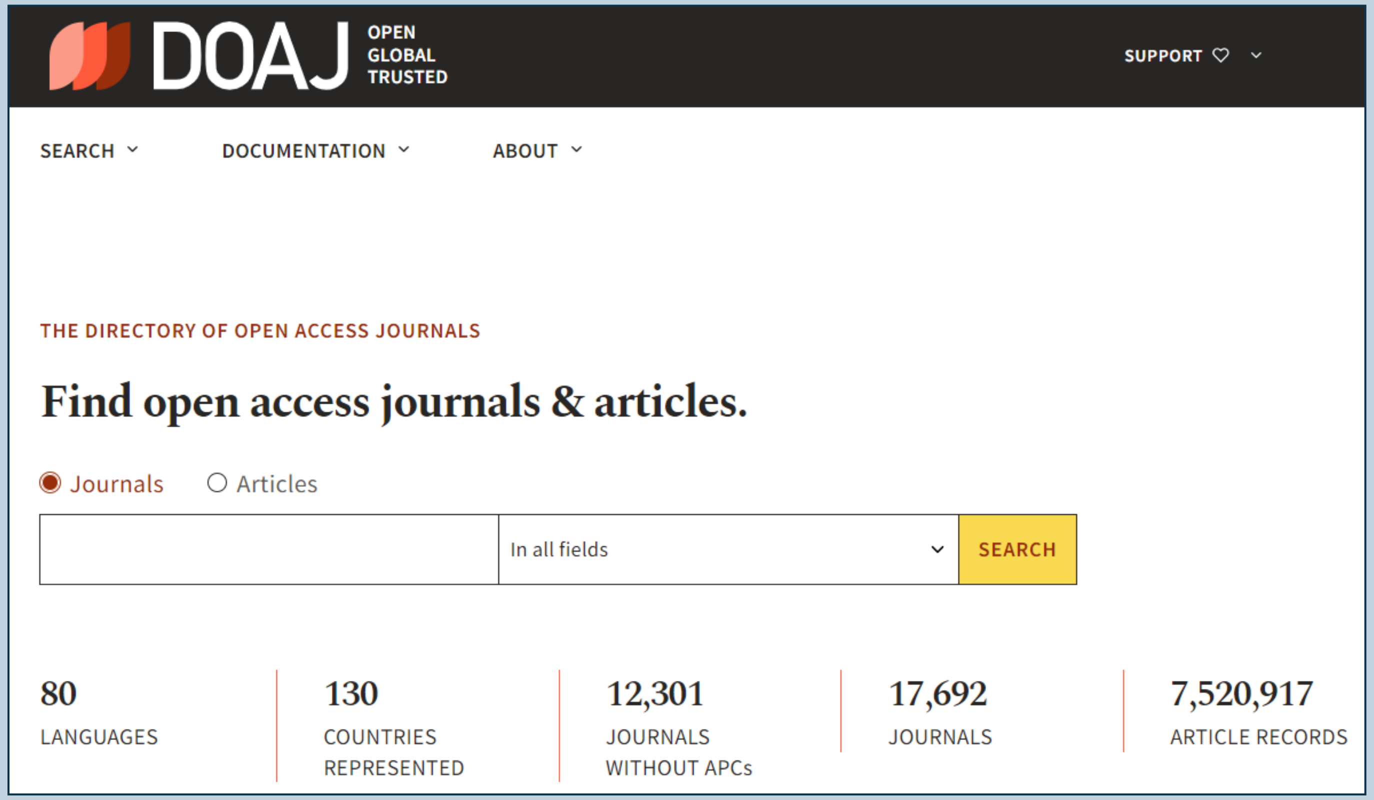
Literature subcorpus

1. 160 books from the Balanced Corpus published in 1990 and onwards
2. 228 new books (original and translations, fiction and non-fiction) published in 2002 and onwards
   - Fiction: Novel, Set of stories, Childrens’ book, Travel writing
   - Non-Fiction: Handbook, Popular Science, Essay, Biography, Cookbook

DOAJ subcorpus

- An online directory that indexes and provides access to 17,500 high quality, open access, peer-reviewed journals
- Journals covering all areas of science, technology, medicine, social sciences, arts and humanities, e.g. Eesti Rakenduslingvistika Ühingu aastaraamat, Methis: Studia Humaniora Estonica, LingVaria, Folklore, Eesti Arst, Eesti Haridusteaduste Ajakiri, Ajalooline Ajakiri, Estonian Journal of Earth Sciences, Eesti Majanduspoliitilised Väitlused ja Agraarteadus

The Directory of Open Access Journals & Articles



DOAJ OPEN GLOBAL TRUSTED

SUPPORT

SEARCH DOCUMENTATION ABOUT

THE DIRECTORY OF OPEN ACCESS JOURNALS

Find open access journals & articles.

● Journals ○ Articles

In all fields

SEARCH

80	130	12,301	17,692	7,520,917
LANGUAGES	COUNTRIES REPRESENTED	JOURNALS WITHOUT APCs	JOURNALS	ARTICLE RECORDS

doaj.org

Wikipedia and Wikipedia Talks

- Wikipedia subcorpus contains texts from main articles
- Wikipedia Talks 2017 subcorpus contains "talk pages" from Wikipedia downloaded in 2017. A talk page in Wikipedia is a separate page with a discussion about the main article where e.g. editors argue about the content of the article

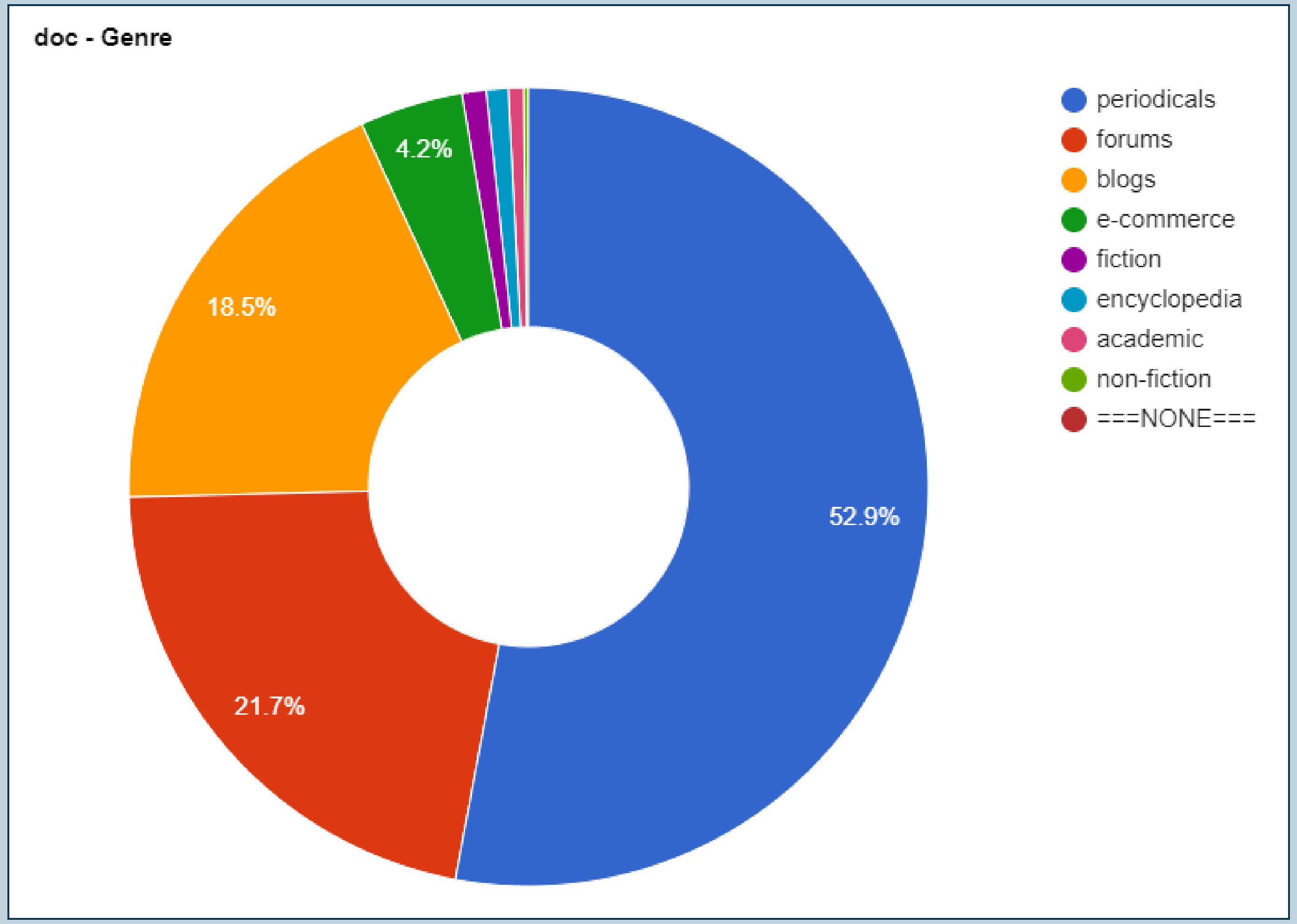
Text classification of web texts

URL-based text classification

1. In parallel to manual domain checking, genres and topics of URLs were recorded
2. Two-level classification was used: genres (broader) and topics (narrower). Genre is determined by the style while topics are determined by words, e.g. a discussion board on characters from Game of Thrones would be Fiction and Discussion at the same time

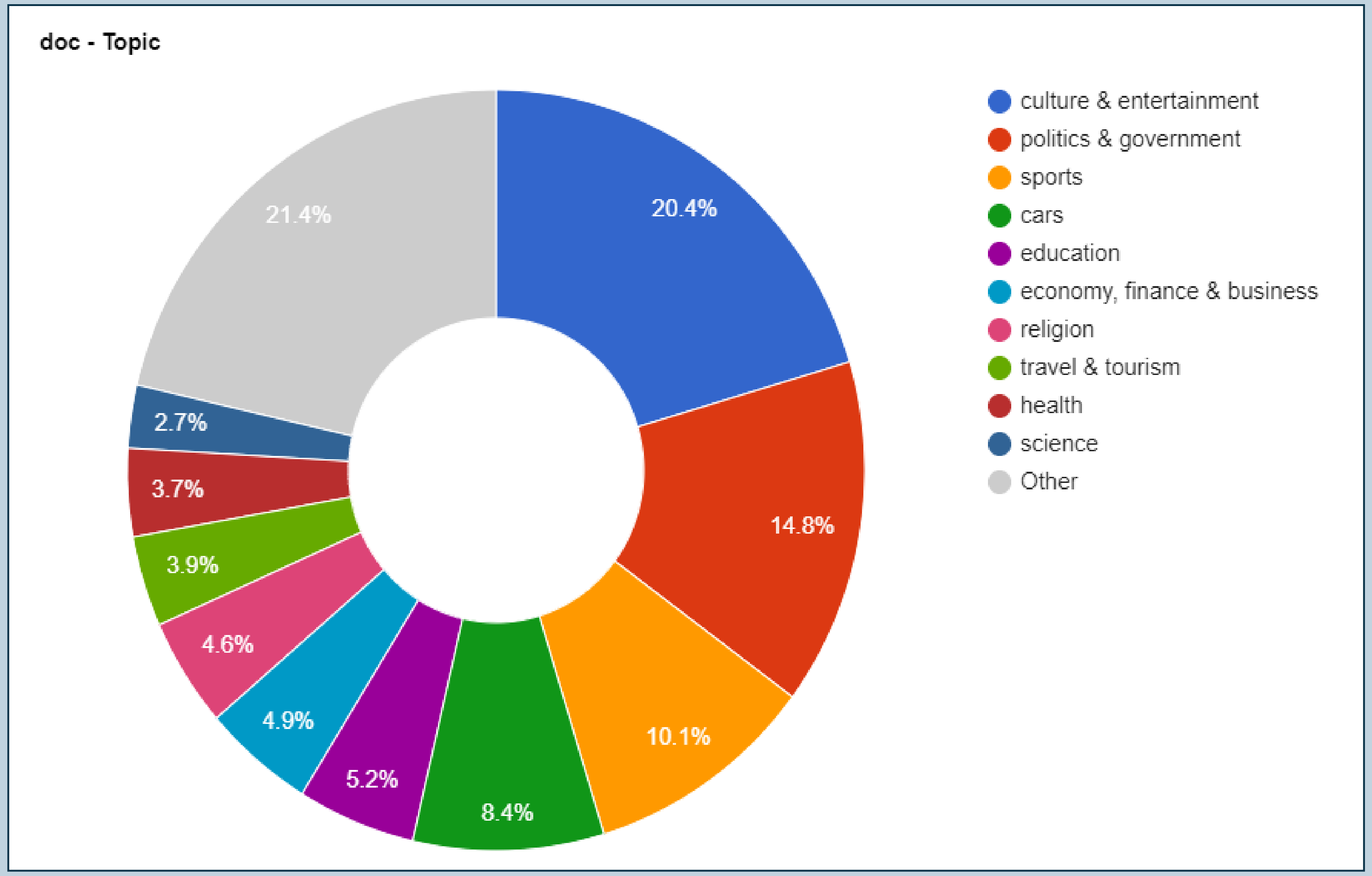
Genres in ENC 2021

Genre	Examples	Number of URLs	Tokens
Blogs	Mallukas, Paljas Porgand, Päevakera, Marimell	761	174,642,782
Encyclopedia	KakuWiki	4	27,205,001
E-commerce	Photopoint, Loverte, Kaup24	180	36,183,138
Forums	Lapsemure, Rahafoorum, Soccernet	121	243,841,952
Periodicals	Õhtuleht, Sirp, Horisont, Anne ja Stiil	208	297,687,046
Sum		**1274**	**794,081,388**



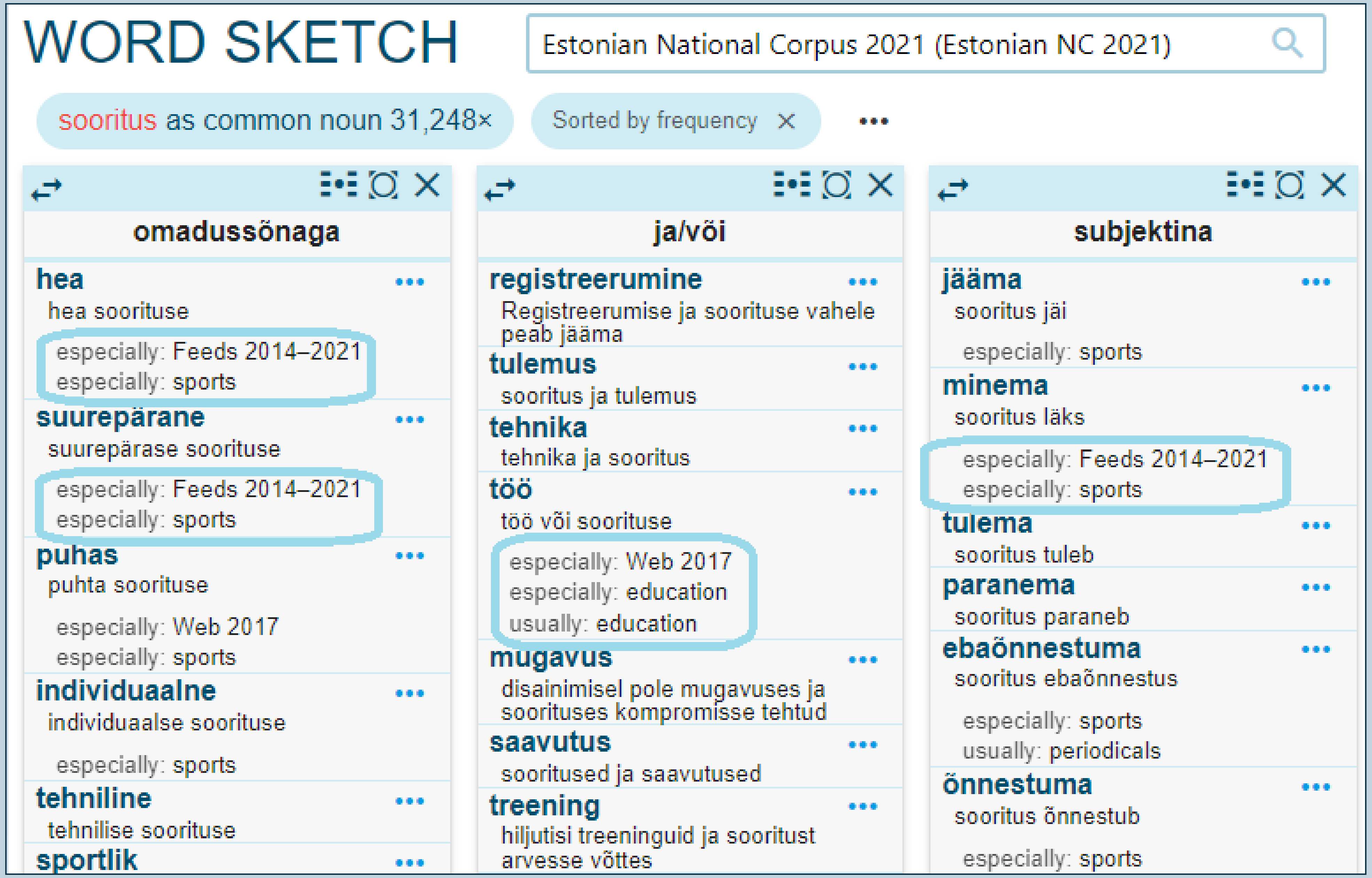
Topics in ENC 2021

Topic*	Examples	Number of URLs	Tokens
Culture & entertainment	Web pages of (art) museums, theatres; blogs and periodicals about reading, music, dance, cinema etc.	241	114,453,884
Education	Õpetajate Leht, web pages of universities and schools	149	42,741,627
Sports	Web pages of sports clubs, periodicals about sports	156	85,893,650
Politics & covernment	Web pages of ministries and political parties, periodicals of political parties (Kesknädal, Uued uudised)	106	84,112,180
Cars	Forums dedicated to different car models (Mazda, BMW)	89	73,003,851
Travel & Tourism	Web pages of travel agencies, travel blogs and forums	78	43,886,494
Food & drinks	Recipes, cooking blogs, cooking magazines	71	13,789,566
	Sum	**1474**	**712,192,346**



* **Altogether 24**, the rest: agriculture; beauty; construction & real estate; economy; finance & business; gambling & casinos; health; history; home, family & children; law & justice; nature & environment; pets and animals; religion; science; sex; technology & IT; video games; women

Genres and topics in Word Sketches



WORD SKETCH

Estonian National Corpus 2021 (Estonian NC 2021)

sooritus as common noun 31,248×

Sorted by frequency ×

omadussõnaga

hea
hea soorituse
especially: Feeds 2014–2021
especially: sports

suurepärane
suurepärase soorituse
especially: Feeds 2014–2021
especially: sports

puhas
puhta soorituse
especially: Web 2017
especially: sports

individuaalne
individuaalse soorituse
especially: sports

tehniline
tehnilise soorituse

sportlik

ja/või

registreerumine
Registreerumise ja soorituse vahele peab jääma

tulemus
sooritus ja tulemus

tehnika
tehnika ja sooritus

töö
töö või soorituse
especially: Web 2017
especially: education
usually: education

mugavus
disainimisel pole mugavuses ja soorituses kompromisse tehtud

saavutus
sooritused ja saavutused

treening
hiljutisi treeninguid ja sooritust arvesse võttes

subjektina

jääma
sooritus jäi
especially: sports

minema
sooritus läks
especially: Feeds 2014–2021
especially: sports

tulema
sooritus tuleb

paranema
sooritus paraneb

ebaõnnestuma
sooritus ebaõnnestus
especially: sports
usually: periodicals

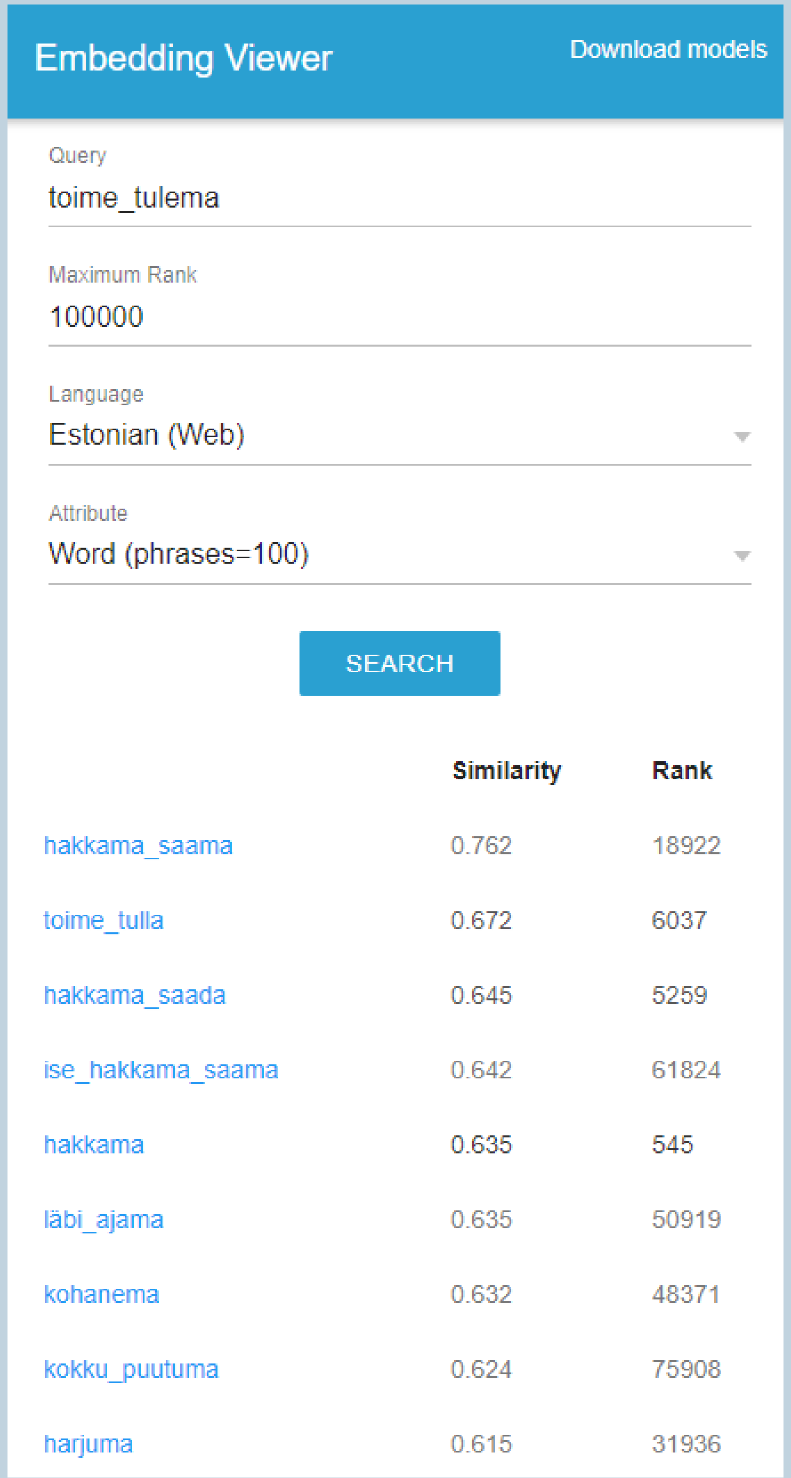
õnnestuma
sooritus õnnestub
especially: sports

Word sketch for lemma *sooritus* ‘performance’

- **"only"** means that >97% of occurrences of this collocation falls into that genre/topic
- **"usually"/"always"** means that >70% (and <97%) of raw ocurrances fall into that genre/topic
- **"especially"** means that the relative frequency within text type X is much higher than the relative frequency in the whole corpus (at least 2x higher)

What else is new?

1. Syntactic annotation
2. SketchGrammar
3. TermGrammar
4. Embeddings with phases



Query: toime_tulema

Maximum Rank: 100000

Language: Estonian (Web)

Attribute: Word (phrases=100)

	Similarity	Rank
hakkama_saama	0.762	18922
toime_tulla	0.672	6037
hakkama_saada	0.645	5259
ise_hakkama_saama	0.642	61824
hakkama	0.635	545
läbi_ajama	0.635	50919
kohanema	0.632	48371
kokku_puutuma	0.624	75908
harjuma	0.615	31936

embeddings.sketchengine.eu

EESTI
KEELE
INSTITUUT

Special thanks! ♡

Lexical Computing Ltd.:

- Vít Suchomel
- Jan Michelfeit
- Jan Kraus
- Marek Blahuš
- Ondřej Herman
- Miloš Jakubíček

Institute of the Estonian Language:

- Helen Kaljumäe
- Madis Jürviste
- Eleri Aedmaa

EstNLTK team:

- Siim Orasmaa
- Sven Laur
- Kadri Muischnek
- Kaili Müürisep

+ Aleksei Kelli

Aitäh!