



Eesti Keele Instituut

# EÕIK infoseminar: Sketch Engine

Kristina Koppel, PhD  
EKI vanemarvutileksikograaf  
[kristina.koppel@eki.ee](mailto:kristina.koppel@eki.ee)

Tallinna Ülikool  
13.10.2020



# Põhimõisted

**korpus** (*corpus*) – kirjalike või suuliste tekstide elektrooniline kogu

**korpuspäringusüsteem** (*Corpus Query System*) – tarkvara, mis võimaldab korpusandmete mitmekülgset analüüsi, nt [Sketch Engine](#), [KORP](#), [Keeleveeb](#)

**sõne** (*token*) – tekstis esinev sõna või selle muutevorm (hrl tekstikorpuse suuruse mõõtühikuna)

**lemma** – sõna või väljendi algvorm

**lempos** – sõna algvorm (*lemma*) + sõnaliik (*POS*) (nt *noor-s*, *noor-a*; *kerge-a*, *kerge-s*)

## Sketch Engine'is:

- 527 korpus
- 95 keelt
- suurim (Timestamped JSI web corpus 2014-2020 English) 50 miljardit sõna



# Korpus lingvistilise info allikana

- sagedusinfo
- grammatiline kasutusinfo (nt kas ainsuses või mitmuses, eitavas/jaatavas kõnes, teatud käändes)
- neologismid ja diakrooniline analüüs
- sõna tähendus(ed)
- kollokatsioonid ehk naabersõnad
- leksikaal-semantilised suhted (sarnastes kontekstides esinevad sõnad, sh sünonüümid), nt [Embedding Viewer](#) (vektorsemantika)
- näitelaused
- tõlkevasted (paralleelkorpuste alusel)
- definitsioonid
- terminid (valdkonnakorpuste alusel)
- mitmesõnalised märksõnad



**sõnastike automaatne koostamine**



# Eesti keele korpused (17) Sketch Engine'is

<b>Estonian</b>	<b>Estonian</b> National Corpus 2019 (Estonian NC 2019)	1,500,284,681
<b>Estonian</b>	<b>Estonian</b> National Corpus 2017 (Estonian NC 2017)	1,107,584,469
<b>Estonian</b>	<b>Estonian</b> National Corpus 2013 (Estonian NC 2013)	463,827,780
<b>Estonian</b>	EUR-Lex <b>Estonian</b> 2/2016	291,077,511
<b>Estonian</b>	Corpus of <b>Estonian</b> Web sentences 2020	280,961,465
<b>Estonian</b>	<b>Estonian</b> Corpus for Learners 2020 (etSkELL)	280,572,215
<b>Estonian</b>	<b>Estonian</b> Web 2013 (etTenTen13)	260,559,829
<b>Estonian</b>	[DEV] <b>Estonian</b> Corpus for Learners 2018 (etSkELL)	248,203,200
<b>Estonian</b>	<b>Estonian</b> Reference corpus 1990-2008 (EstonianRC)	203,267,951
<b>Estonian</b>	OPUS2 <b>Estonian</b>	64,879,741
<b>Estonian</b>	DGT, <b>Estonian</b>	34,155,488
<b>Estonian</b>	EUR-Lex judgments <b>Estonian</b> 12/2016	15,029,608
<b>Estonian</b>	[DEV] <b>Estonian</b> RSS Feed Corpus	14,308,077
<b>Estonian</b>	EUROPARL7, <b>Estonian</b>	11,171,727
<b>Estonian</b>	Open Access Journals (DOAJ - <b>Estonian</b> single)	6,040,679
<b>Estonian</b>	CHILDES <b>Estonian</b> Corpus	313,457
<b>Estonian</b>	<b>Estonian</b> coursebook corpus 2018	121,114



# Eesti keele ühendkorpuste sari

- Estonian National Corpus (Estonian NC)
- Iga ~2a tagant uus versioon
- Sisaldab:
  - eesti keele koondkorpus, sh tasakaalus korpus
  - veebikorpus (internetist alla laetud eestikeelsed veebilehed)
    - **Spiderling** is a web spider for linguistics. It can crawl text-rich parts of the web and collect a lot of data suitable for text corpora
    - **JusText** is a HTML boilerplate removal tool. It can strip navigation links, headers, footers, etc. From HTML pages and leave just regular text containing full sentences
    - **Chared** is a tool for detecting the character encoding of a text in a known language. It contains models for a wide range of languages
    - **Onion** (ONe Instance ONly) is a de-duplicator for large collections of texts. It can measure the similarity of paragraphs or whole documents and drop duplicate ones based on the threshold you set
    - **wiki2corpus** is a script which downloads Wikipedia articles (for a given language) and outputs them in the form of prevertical which can be further processed by other corpus tools



# Eesti keele ühendkorpus 2013

Estonian National Corpus 2013 (**563 mln sõnet**)

- eesti keele koondkorpus (1990–2008) (250 mln sõnet), sh tasakaalus korpus (koondkorpuse tasakaalustatud alamhulk) (15 mln sõnet)
- eesti veebikorpus 2013 (330 mln sõnet)



# Eesti keele ühendkorpus 2017

## Estonian National Corpus 2017 (1,3 mld sõnet)

- koondkorpus, sh tasakaalus korpus
- veebikorpus 2013 (323 mln sõnet)
- veebikorpus 2017 (764 mln sõnet)
- eesti Wikipedia (38 mln sõnet)



# Eesti keele ühendkorpus 2019

## Estonian National Corpus 2019 (1,8 mld sõnet)

- eesti keele koondkorpus, sh tasakaalus korpus
- eesti veebikorpused 2013 (303 mln), 2017 (639 mln), 2019 (615 mln) (.ee-domeenid, blogid, foorumid, haridus, ilukirjandus, toit, tervis, ajakirjad, uudised, religioon, teadus, seks, ühiskond, sport)
- eesti Wikipedia 2017 (32 mln sõnet) ja 2019 (8 mln sõnet)
- DOAJ (avatud lähtekoodiga eestikeelsed teadusajakirjad) (8 mln sõnet)





# Eesti uudisvoogude korpus 2020

- Estonian RSS Feed Corpus 2020
- (Usaldusväärsete) veebilehtede uudisvoogude eestikeelsed tekstid
- Täineb jooksvalt aasta vältel
- Suurus **17,5 mln sõnet** (09.10.2020 seisuga, kogutud alates veebruarist 2020)



# Õppe- ja õppijakeele korpused

- [Eesti keele õppekorpus 2018 \(etSkELL\)](#) / Estonian Corpus for Learners 2018 (etSkELL)
  - Aluseks eesti keele ühendkorpus 2017
  - Sisaldab ainult “häid näitelauseid”
- Eesti keele õppekorpus 2020 (etSkELL) / Estonian Corpus for Learners 2020 (etSkELL)
  - Aluseks eesti keele ühendkorpus 2019
  - Sisaldab ainult “häid näitelauseid”
- [Eesti keele õpikute korpus 2018](#) / Estonian Coursebook Corpus 2018
  - Korpus sisaldab A1, A2, B1, B2 ja C1 keeleoskustasemega täiskasvanud eesti keele õppijatele suunatud õpikute tekstidest eraldatud terviklauseid
  - Kokku kaheksast õpikust
- Lapsekeelekorpus CHILDES
  - sisaldab emakeelsete laste ja nende vanemate spontaanse kõne lindistusi



# Oma korpuse loomine

CORPUS: jalgpall (Estonian)

1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE



Find texts on the web

Automatically find and download relevant texts



I have my own texts

Upload your own files (.txt, .pdf,...) or paste text

CORPUS CONTENT



# Korpuse loomine veebist

- **3–20 võtmesõna** (*seed word*),  
sh mitmesõnalised üksused

NB!!! eralda Enteri-klahviga,  
mitmesõnalised peavad olema  
jutumärkides

← TEXTS FROM WEB

Input type

- Web search <sup>?</sup>
- URLs <sup>?</sup>
- Website <sup>?</sup>

jalgpall × jalgpallur × värav × väravavaht × penalti × <sup>?</sup>

karistuslõök × jalgpalli mängima × väravat lööma × ründaja ×

kaitsja × |

At least 3 words or phrases. Hit ENTER after each one. Multiword phrases in quotes: "wild cat" kitten feed

Folder name <sup>?</sup> web1

- Web search settings ▾
- Denylist settings ▾
- Allowlist settings ▾
- Size restrictions ▾

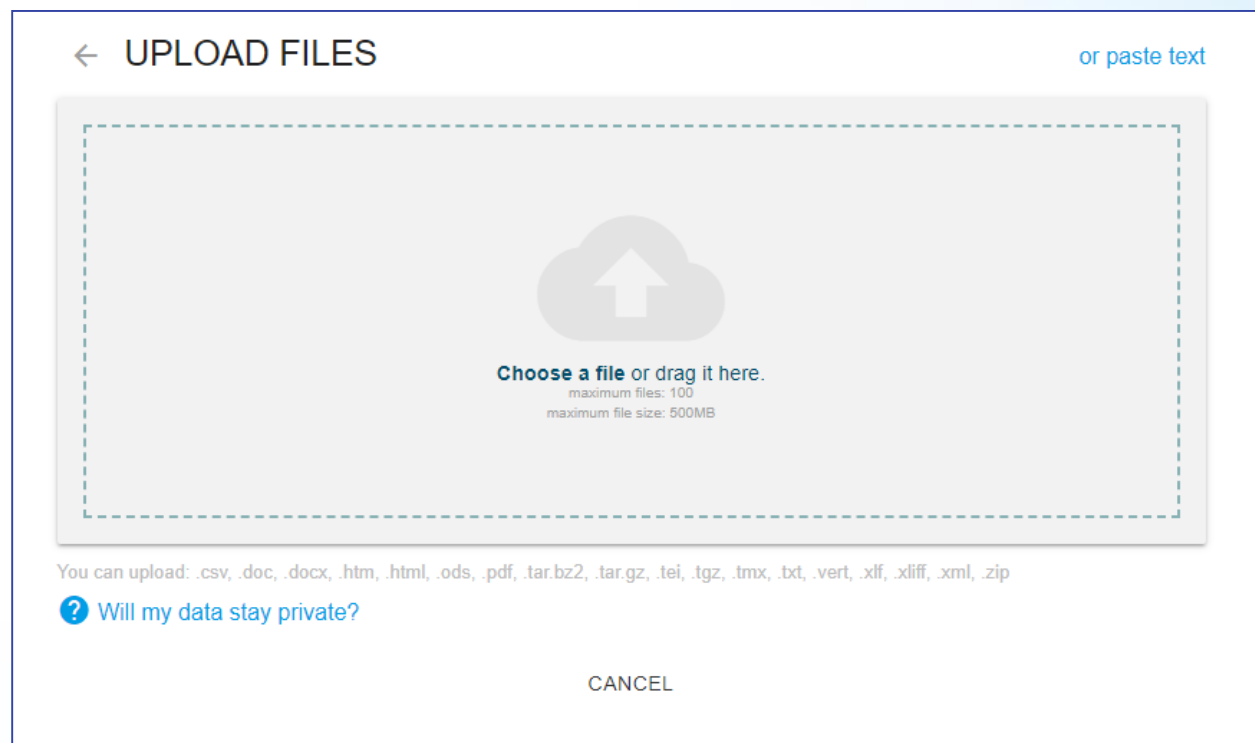
Compile when finished <sup>?</sup>

CANCEL GO



# Korpuse loomine tekstifailist

- Lemmatiseerija, morfoloogiline analüüs ja ühestamine: EstNLTK





# Korpuse märgendamine

- [Estnltk](#) – tööriistakast vabatekstide lihtsaks töötlemiseks

[POS tag descriptions](#) – sõnaliigi märgendid

[noun form descriptions](#) – käandsõnade grammatilised kategooriad

[Verb form descriptions](#) – tegusõna grammatilised kategooriad

- morfoloogiline analüüs ja süntees
- sõnade lemmatiseerimine
- (osa)lausestamine
- ajaväljendite tuvastamine
- nimeüksuste tuvastamine
- verbiahelate tuvastamine
- Eesti Wordneti liidestamine



# Tee oma korpus

DASHBOARD Estonian National Corpus 2019 (Estonian NC 2019)

ESTONIAN NATIONAL CORPUS 2019 (ESTONIAN NC 2019) **CORPUS INFO** **MANAGE CORPUS**

**Word Sketch**  
Collocations and word combinations

**Word Sketch Difference**  
Compare collocations of two words

**Thesaurus**  
Synonyms and similar words

**Concordance**  
Examples of use in context

RECENTLY USED CORPORA **NEW CORPUS**

Estonian National Corpus 2019 (Estonian NC 2019)	Estonian	1,500,284,681
koroona	Estonian	442,220
OPUS2 Estonian	Estonian	64,879,741
[DEV] Estonian RSS Feed Corpus	Estonian	14,308,077
[DEV] Estonian RSS Feed Corpus	Estonian	14,308,077
CHILDES Estonian Corpus	Estonian	313,457
EUR-Lex Estonian 2/2016	Estonian	291,077,511

Nt:

- jalgpall, värav, väravavaht, jalgpallur, ründaja, kaitsja, penalti, karistuslööök, „väravat lööma“, „jalgpalli mängima“
- koroona, koroonakriis, koroonaviirus, nakkuskordaja, desoaine, eneseisolatsioon, covid-19, karantiin, nakatunu, riskipiirkond, eriolukord, epideemia, lähikontakt, riskirühm, liikumispiirang, kaitsemask, kodukontor, distantstõpe, Wuhan



# Konkordants (1)

Konkordants on sõnavorm koos kontekstiga

**CONCORDANCE** [DEV] Estonian RSS Feed Corpus Get more space ↻ ? ! 👤

simple **koroona** 1,770 (101.09 per million)

Details Left context KWIC Right context


	Left context	KWIC	Right context
41	<a href="#">google.com</a> jis. </s></s> KLÕPS   Rahvasaadik Marko Mihkelson mõnuleb päikese käes ja trotsib	<b>koroonat</b>	: rohkem ollakse mures hiljutise Sahara liivatormi tagajärgede pärast </s></s> Riigiko
42	<a href="#">google.com</a> ara liivatormi tagajärgede pärast </s></s> Riigikogu liige Marko Mihkelson ei lasknud	<b>koroona</b>	levikul ennast heidutada ning otsustas siiski suve nautima minna. </s></s> Mihkelson
43	<a href="#">google.com</a> </s></s> Mihkelson jagas sotsiaalmeedias pilti päikeseliselt Tenerifelt, kus tema sõnul	<b>koroona</b>	mingeid märke ei näita. </s></s> "Tenerifel ollakse rohkem mures hiljutise Sahara liiva
44	<a href="#">ohtuleht.ee</a> > Terviseameti pressiesindaja Eike Kingsepp kinnitas kolmapäeval kell 15.50, et teine	<b>koroonasse</b>	nakatunud eestlane on tartlane. </s></s> Kolmapäevase seisuga on Eestis kinnitatud
45	<a href="#">google.com</a> : Riigil on alati õigus. </s></s> L. Glikman. </s></s> 2016. </s></s> Tallinn Music Week	<b>koroona</b>	tõttu ära ei jää: kahe tuvastatud nakatunuga Eesti ei ole praegu riskipiirkond </s></s>
46	<a href="#">google.com</a> viseameti ekspert Ester Õpik: hetkel ei ole teada, kui palju Eestisse tulnud inimesi oli	<b>koroonasse</b>	nakatunuga ühel lennul </s></s> Eile kinnitati Eestis teine koroonaviirusesse nakatun






# Konkordantsi metaandmed

## Display and count metadata

Select the metadata to be displayed in the concordance. Click  to calculate statistics.

Display above lines ?  Shorten to 15 characters

ime kaasa oma rahva suurteo teadvustamisele ja väärtustamisele. </s><s> **KOROONA** JA EUROVISION: paar riiki ei osale tähtsal kohtumisel, Tomi Rahula kinnita  

Website (e.g. cnn.com)   (1)

<input type="checkbox"/> Token number	4907834	
<input type="checkbox"/> Document number	19333	
<input type="checkbox"/> doc.downloaded	2020-03-06	
<input type="checkbox"/> doc.feed	<a href="https://www.ohtuleht.ee/rss">https://www.ohtuleht.ee/rss</a>	
<input type="checkbox"/> doc.fetched	2020-03-06	
<input type="checkbox"/> doc.published	2020-03-06	
<input type="checkbox"/> doc.wordcount	219	
<input type="checkbox"/> p.heading	1	
<input type="checkbox"/> Top-level domain (e.g. com)	ee	
<input type="checkbox"/> uri	<a href="https://elu.ohtuleht.ee/994377/koroona-ja-eurovision-lauluvoistluse-boss-sai-reisikeelu-rootsi-ja-iisrael-ei-osale-tahtsal-kohtumisel">https://elu.ohtuleht.ee/994377/koroona-ja-eurovision-lauluvoistluse-boss-sai-reisikeelu-rootsi-ja-iisrael-ei-osale-tahtsal-kohtumisel</a>	
<input type="checkbox"/> Web domain (e.g. news.blogs.cnn.com)	elu.ohtuleht.ee	
<input checked="" type="checkbox"/> Website (e.g. cnn.com)	ohtuleht.ee	



# Konkordantsiotsing (1)

**Basic search:** leiab nii lemmad kui ka sõnavormid, mis vastavad otsisõnale

1. Otsi lemmat *lammas, toime tulema*
2. Otsi sõnavormi *lammasteni, tuli toime*



# Konkordantsiotsing (2)

## Advanced search:

- **Simple:** leiab nii sõnavormid kui lemmad, mis vastavad otsisõnale
- **Lemma:** (sõna algvorm) leiab otsisõna kõikides vormides
- **Phrase:** leiab mitmesõnalise fraasi (kirjutatud kujul), nt *tulin toime, ei liha ega kala*
- **Word:** leiab konkreetse sõnavormi (kirjutatud kujul), nt *ngo*
- **Character:** leiab sõned, mis sisaldavad konkreetset tähemärki või tähemärgi jada, nt !
- Saab kasutada regulaaravaldisi, vt nt lemma otsing:
  - \* – suvaline sümbol (suvaline arv sümboleid, 0–...), nt *meh\*aanika* leiab nii *mehaanika* kui ka *mehhaanika*
  - . – suvaline üks sümbol, nt *.ala* leiab nii *pala, sala, tala, kala*; *.lu* leiab *elu, ilu, õlu, ..*
  - *eba.\** – leiab kõik *ebaga* algavad sõnad
  - spikker: <https://www.sketchengine.eu/guide/regular-expressions/>



# Konkordantsiotsing (3)

- Otsing (lemma, POS) kontekstis:
  - Otsi *minema*, millele järgnevad nimisõnad on *kool kodu töö*
  - Otsi *viirus*, millele eelnev sõna on adjektiiv
- CQLi (*corpus query language*) otsing, kasutatakse keerukamate leksikaalsete ja grammatiliste struktuuride otsinguks ilma konkreetseid sõnu määramata, nt inglise keele korpuses leiab päring **[word!="a|an|the"]**[tag="N.\*"]**[tag="V.D"]** kõik nimisõnad, millele ei eelne artikkel, aga millele järgneb mõni minevikuvormis verb

Filter context ? ^

Do not filter

Lemma context

Part-of-speech context

Only keep lines with

all ▼ of lemmas separated by space within 5 ▼ Tokens left and right ▼



# Sõnavisandid

Ühel lehel kuvatav automaatne korpuspõhine kokkuvõtte sõna grammatilisest ja kollokatsioonilisest käitumisest.

**Kollokatsioon** on sisusõnade tähenduslikud ja statistiliselt esilduvad kombinatsioonid teiste leksikaalsete ja grammatiliste üksustega, nt *must kohv*, *kange kohv*, *lahustuv kohv*.  
Grupeeritud grammatiliste suhete kaupa (nt *adj\_modifier*, *subject\_of*, *object\_of*).

kohv as common noun 98,076x Sorted by frequency x

terms

Constructions	Adj_modifier	subject_of	object_of	omastav_modifier
<b>osastav</b> 48,889 49.85 ...	<b>hea</b> 1,418 4.85 ... head kohvi	<b>saama</b> 94 1.42 ... kohv sai	<b>jooma</b> 2,081 11.55 ... juua kohvi	<b>tassikese_kohv</b> 249 11.15 ... tassikese kohvi
<b>omastav</b> 19,492 19.87 ...	<b>must</b> 1,172 7.59 ... musta kohvi	<b>sisaldama</b> 93 4.6 ... kohv sisaldab	<b>pakkuma</b> 700 7.78 ... pakutakse kohvi	<b>tassitäie_kohv</b> 135 10.31 ... tassitäie kohvi
<b>nimetav</b> 18,916 19.29 ...	<b>kange</b> 1,064 10 ... kanget kohvi	<b>tulema</b> 93 2.15 ... kohv tuleb	<b>tegema</b> 423 4.95 ... tegin kohvi	<b>hommiku_kohv</b> 111 10.04 ... hommiku kohvi
<b>kaasaütlev</b> 3,050 3.11 ...	<b>kuum</b> 928 8.6 ... kuuma kohvi	<b>aitama</b> 88 4.45 ... kohv aitab	<b>keetma</b> 214 9.24 ... keedab kohvi	<b>naiste_kohv</b> 101 9.91 ... Naiste Kohvi
<b>alaleütlev</b> 2,868 2.92 ...	<b>lahustuv</b> 726 10.67 ... lahustuvat kohvi	<b>tegema</b> 84 2.5 ... kohv teeb	<b>saama</b> 179 3.53 ... saab kohvi	<b>lonksu_kohv</b> 96 9.83 ... lonksu kohvi
<b>seestütlev</b> 2,295 2.34 ...	<b>värske</b> 338 6.19 ... värske kohvi	<b>maksma</b> 83 4.27 ... kohv maksab	<b>rüüpama</b> 166 9.4 ... rüüpad kohvi	<b>robusta_kohv</b> 75 9.48 ... Robusta kohvi
	<b>hommikune</b> 311 7.98 ... hommikuse kohvi	<b>võima</b> 75 2.36 ... kohv võib	<b>valmistama</b> 165 7.31 ... valmistada kohvi	<b>kaubanduse_kohv</b> 62 9.22 ... õiglase kaubanduse kohvi
	<b>tavaline</b> 277 5.37 ... tavalist kohvi	<b>maitsema</b> 74 8.11 ... kohv maitseb	<b>nautima</b> 160 7.35 ... nautida kohvi	<b>espresso_kohv</b> 58 9.12 ... espresso kohvi
	<b>roheline</b> 267 6.16 ... rohelise kohvi	<b>ütleva</b> 65 2.82 ... Kohv ütles , et	<b>võtma</b> 160 4.58 ... võtsin kohvi	<b>paki_kohv</b> 55 9.04 ... paki kohvi
	<b>maitsev</b> 257 7.7 ... maitsvat kohvi	<b>mõjuma</b> 56 5.45 ... kohv mõjub	<b>ostma</b> 148 6.46 ... osta kohvi	<b>araabika_kohv</b> 46 8.79 ... araabika kohvi
	<b>kofeiinivaba</b> 235 9.33 ... kofeiinivaba kohvi	<b>andma</b> 55 2.23 ... kohv annab	<b>tellima</b> 122 7.7 ... tellib kohvi	<b>joodi_kohv</b> 44 8.73 ... joodi kohvi
	<b>kvaliteetne</b> 202 6.31 ... kvaliteetset kohvi	<b>meeldima</b> 52 4.23 ... meeldib kohvi	<b>tooma</b> 110 5.73 ... toob kohvi	<b>sõõmu_kohv</b> 38 8.52 ... sõõmu kohvi



# Word Sketch Difference

- Võrdleb kollokatsioone
  - **Lemma:** võrdleb kollokaatidele tuginedes kahe erineva lemma kasutust, võrdle nt *prügi* ja *praht*; *kass* ja *koer*
  - **Word forms:** võrdleb kollokaatidele tuginedes sama lemma kahe erineva sõnavormi kasutust, vt nt *kõneleda* ja *kõnelda*
  - **Subcorpora:** võrdleb kollokaatidele tuginedes sama lemma kasutust erinevates allkorpustes, vrd nt *muna* allkorpustes *health* ja *food*



# Tesaurus

Automaatselt genereeritud pilv sõnadest (sh sünonüümid), mis kuuluvad samasse kategooriasse/semantilisse välja.

Pilves esitatud sõnad esinevad valitud korpuses sarnastes kontekstides (nt jagavad samu kollokaate).

seminar as common noun 96,028x ...

🔍 ⬇️ 👁️ ⚙️ ⓘ ☆

	Word	Frequency ?
1	konverents	141,128 ...
2	loeng	89,033 ...
3	koolitus	232,636 ...
4	üritus	315,617 ...
5	töötuba	52,474 ...
6	kontsert	185,020 ...
7	näitus	189,147 ...
8	kursus	140,643 ...
9	arutelu	233,610 ...
10	koosolek	129,466 ...

Rows per page: 10 ▼ 1–10 of 19 ⏪ ⏩ ⏴ ⏵



# N-grammid

- N-grammid ehk mitmesõnalised üksused (*multi word expressions*)
- Tuvastab sõnede järjendid
- N – number
  - 2grams (kahesõnaline, nt *ei ole, kui ka, mis on*)
  - 3grams (kolmesõnaline, nt *ei ole võimalik, ma ei tea, ei saa aru*)
  - 4grams (neljasõnaline, nt *ma ei saa aru, mul on hea meel*)
  - ..





# Sagedusloendid



# SkELL

<https://skell.sketchengine.eu/>



# Kirjandus

Kallas, Jelena (2013). Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. (Doktoritöö, Tallinna Ülikool). Tallinn: Tallinna Ülikool.

Kallas, Jelena, Kristina Koppel, Maria Tuulik (2015). Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. Eesti Rakenduslingvistika Ühingu aastaraamat, 11, 75–94.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smr, David Tugwell (2004). The Sketch Engine. – Geoffrey Williams, Sandra Vessier (Eds), Proceedings of the XI EURALEX International Congress. Lorient, France: Université de Bretagne Sud, 105–115.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014). The Sketch Engine: ten years on. – Lexicography, 1(1), 7–36.

Koppel, Kristina (2020). Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. (Doktoritöö, Tartu Ülikool). Tartu: Tartu Ülikooli Kirjastus.