Eesti Keele Instituut

EÕIK infoseminar: **Sketch Engine**

Kristina Koppel, PhD
EKI vanemarvutileksikograaf
kristina.koppel@eki.ee

Tallinna Ülikool
13.10.2020

Eesti Keele Instituut

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

Eesti Keele Instituut

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 Instituut

Eesti keele korpused (17) Sketch Engine'is

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

Eesti Keele Instituut

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 Instituut

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 Instituut

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 Instituut

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 Keele Instituut

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)

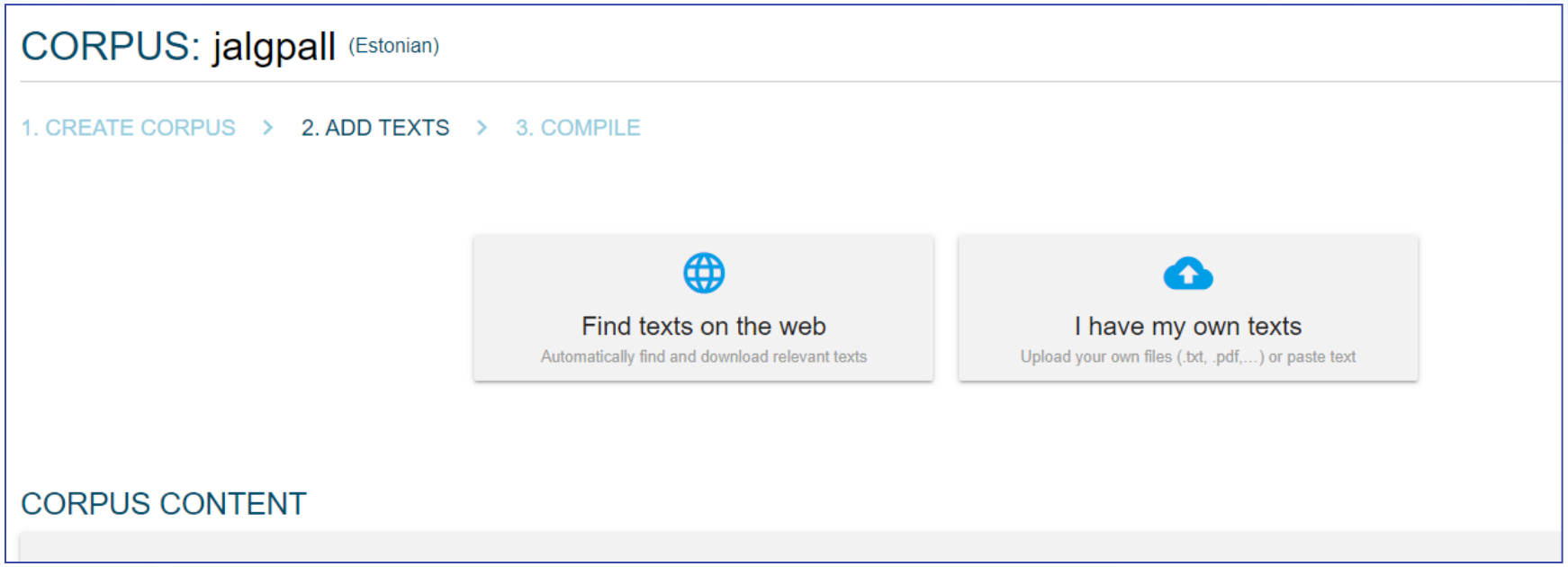
Eesti Keele Instituut

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

Eesti Keele Instituut

Oma korpuse loomine



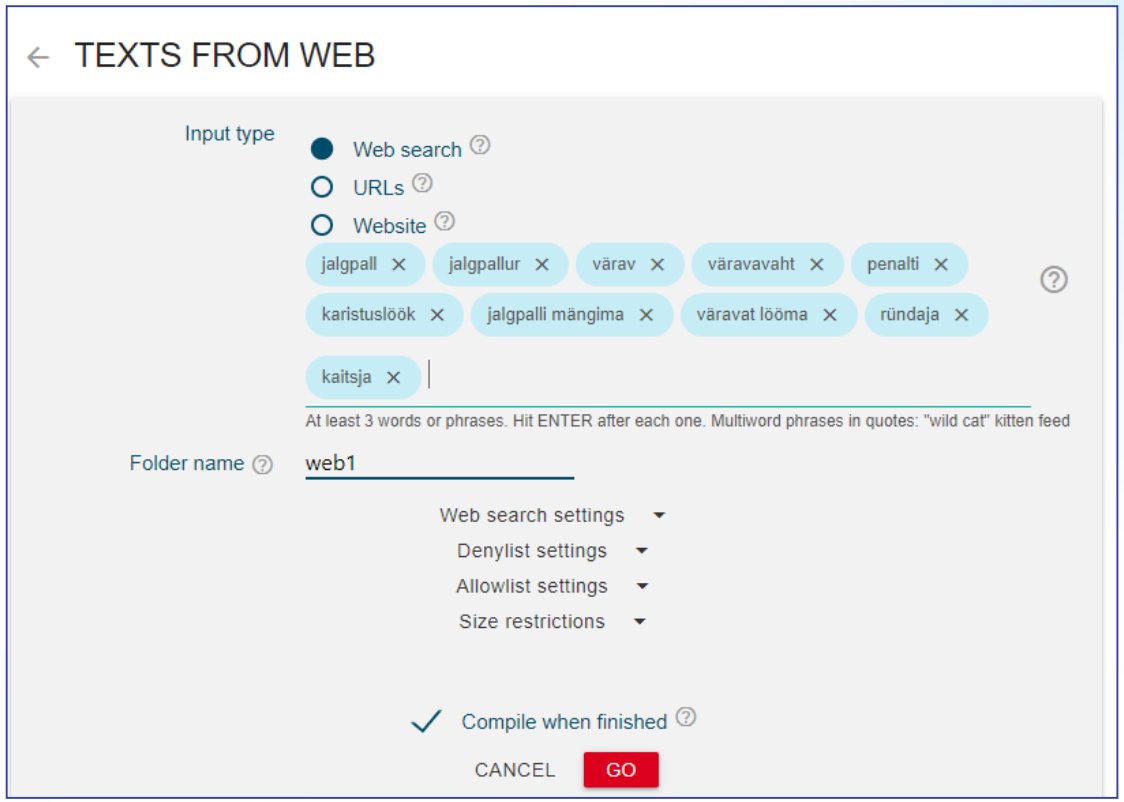
Eesti Keele Instituut

Korpuse loomine veebist

- **3–20 võtmesõna** (*seed word*),

sh mitmesõnalised üksused

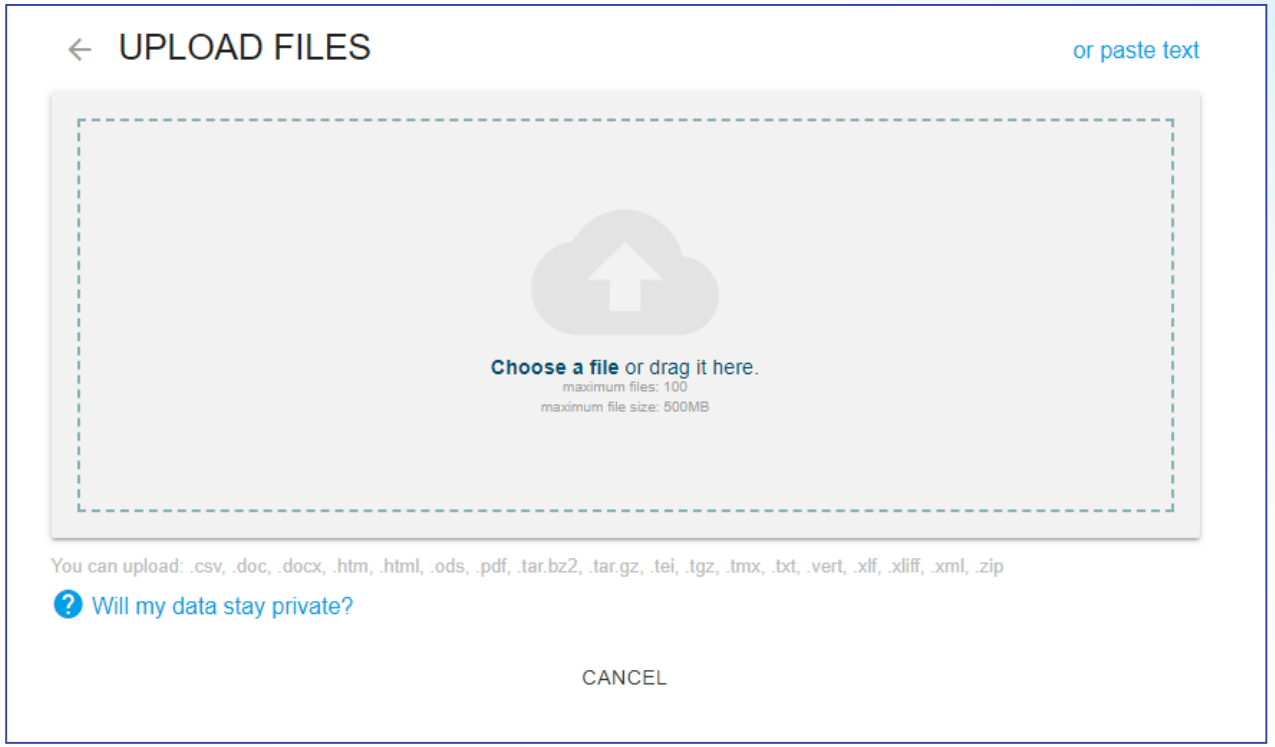
NB!!! eralda Enteri-klahviga, mitmesõnalised peavad olema jutumärkides



Eesti Keele Instituut

Korpuse loomine tekstifailist

- Lemmatiseerija, morfoloogiline analüüs ja ühestamine: EstNLTK



Eesti Keele Instituut

Korpuse märgendamine

- Estnltk – tööriistakast vabatekstide lihtsaks töötlemiseks

  POS tag descriptions – sõnaliigi märgendid

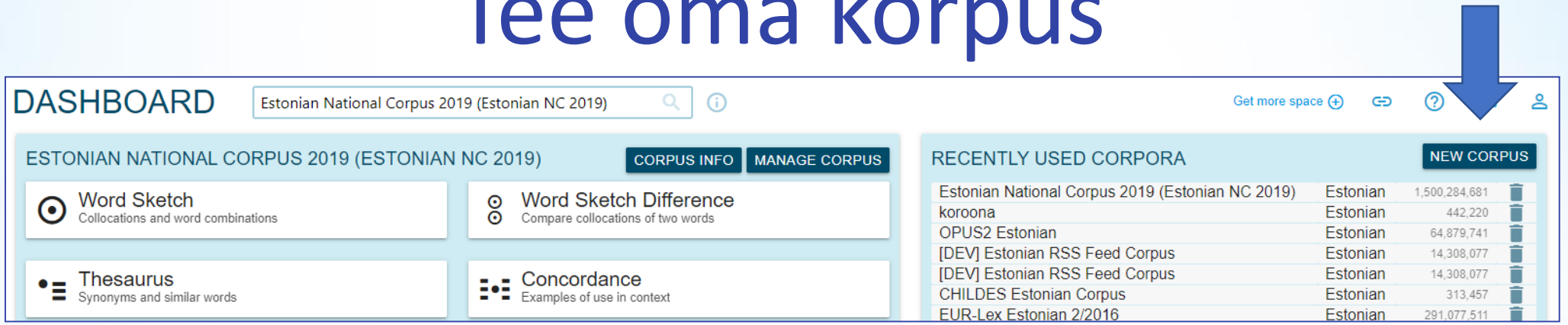
  noun form descriptions – käändsõnade grammatilised kategooriad

  Verb form descriptions – tegusõna grammatilised kategooriad

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

Eesti Keele Instituut

Tee oma korpus



Nt:

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

Eesti Keele Instituut

Konkordants (1)

Konkordants on sõnavorm koos kontekstiga

Eesti Keele Instituut

Konkordantsi metaandmed

Eesti Keele Instituut

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*

Eesti Keele Instituut

Konkordantsiotsing (2)

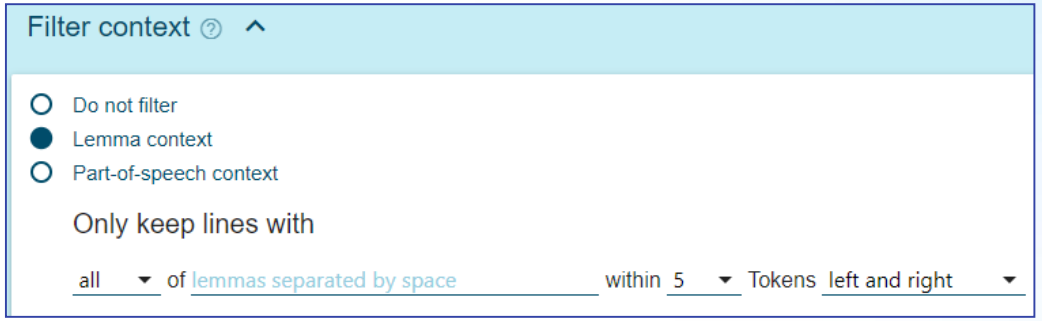
Advanced search:

- **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 *eba*ga algavad sõnad
  - spikker: https://www.sketchengine.eu/guide/regular-expressions/

Eesti Keele Instituut

Konkordantsiotsing (3)

- Otsing (lemma, POS) kontekstis:
  - Otsi *minema*, millele järgnevad nimisõnad on *kool kodu töö*
  - Otsi *viirus*, millele eelnev sõna on adjektiiv



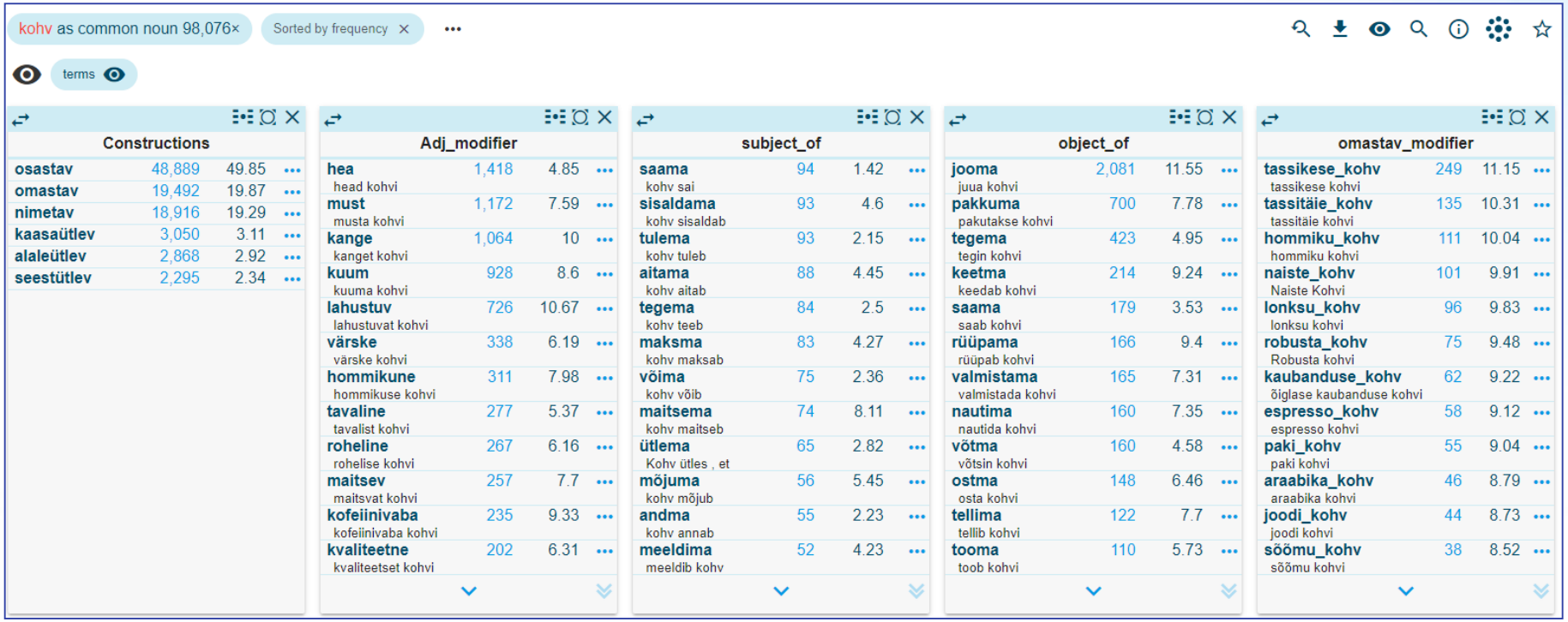
- CQLi (*corpus query language*) otsing, kasutatakse keerukamate leksikaalsete ja grammatiliste struktuuride otsinguks ilma konkreetseid sõnu määramata, nt inglise keele korpuses leiab päring **[word!="a|an|the"][tag="N.*"][tag="V.D"]** kõik nimisõnad, millele ei eelne artikkel, aga millele järgneb mõni minevikuvormis verb

Eesti Keele Instituut

Sõnavisandid

Ühel lehel kuvatav automaatne korpuspõhine kokkuvõte 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*).



Eesti Keele Instituut

Word Sketch Difference

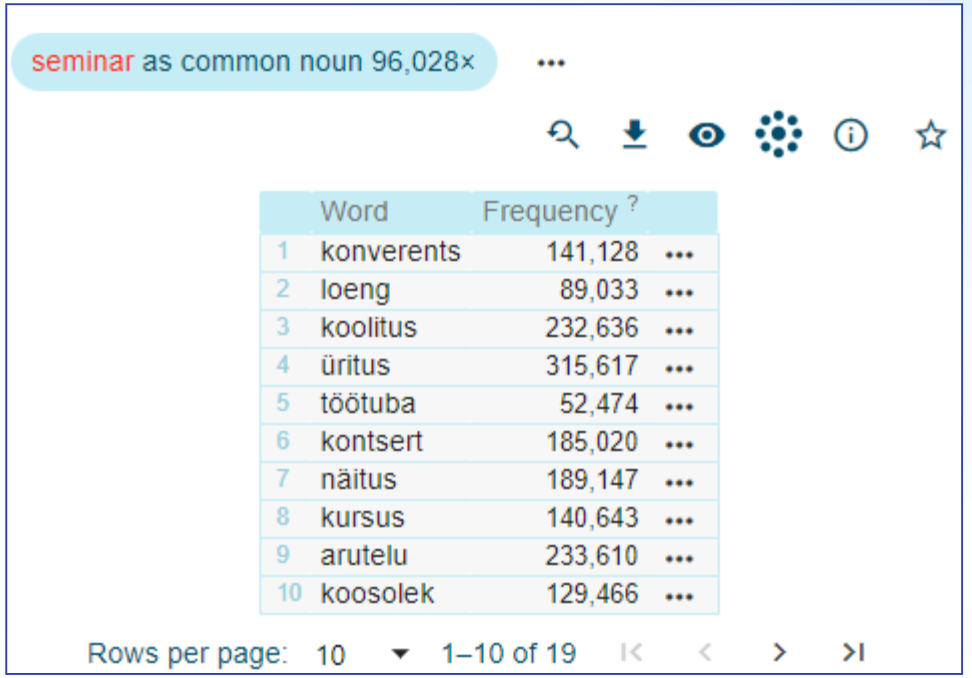
- Võrdleb kollokatsioone
  - **Lemma**: võrdleb kollokaatidele tuginedes kahe erineva lemma kasutust, võrdle nt *prügi* ja *praht*; *kass* ja *koer*
  - **Word forms**: võrdleb kollokaatidele tuginedes sama lemma kahe erineva sõnavormi kasutust, vt nt *kõneleda* ja *kõnelda*
  - **Subcorpora**: võrdleb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

Eesti Keele Instituut

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).



Eesti Keele Instituut

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*)
  - ..

Eesti Keele Instituut

Sagedusloendid

Eesti Keele Instituut

SkELL

https://skell.sketchengine.eu/

Eesti Keele Instituut

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.