

Teksti sisuanalüüsi praktikum: LancsBox X

Analüüsitavad tekstid: https://www.tlu.ee/~kais/Keel_ja_tehnoloogia/2024/sisu/

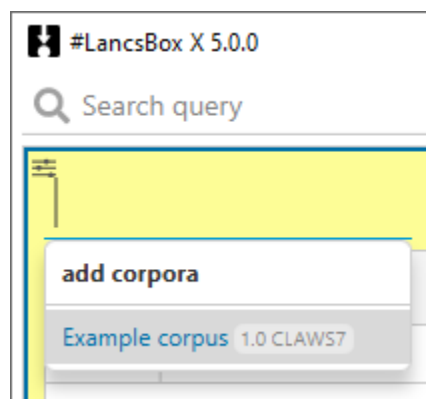
LancsBoxi kasutusjuhend ja videomaterjalid: <https://lancsbox.lancs.ac.uk>

Ettevalmistus: korpuse üleslaadimine ja sellega tutvumine

Praktikumiks koostatud korpus hõlmab kliimamuutuse teemalisi artikleid, mis on ilmunud ERR-i uudisteportaaalis vahemikust oktoober 2023 – september 2024. Otsingu märksõnad olid „kliimamuutus“ ja „kliima soojenemine“. Tulemuste hulgast valiti artiklid, milles kliima on keskne teema. Kokku saadi 44 teksti, mis sisaldavad 31 536 tekstisõna ehk sõnet.

Ülesanne: mõtle, millised teemad võiksid kliimamuutuse kajastuses esile tõusta. Missugused ootused-eeldused enne tekstianalüüsi tekivad?

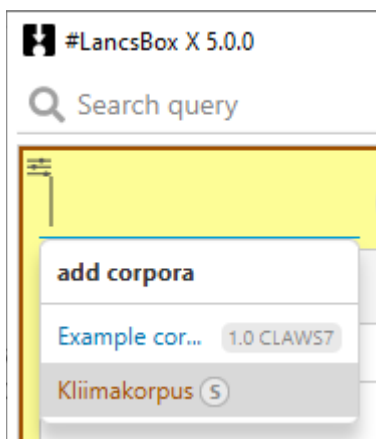
LancsBoxi rakenduses saab analüüsitava korpuse määrata vasakul ülanurgas. Esialgu on valikus vaid ingliskeelne näidiskorpus, mille nimetusel klõpsates avaneb valik *add corpora*.



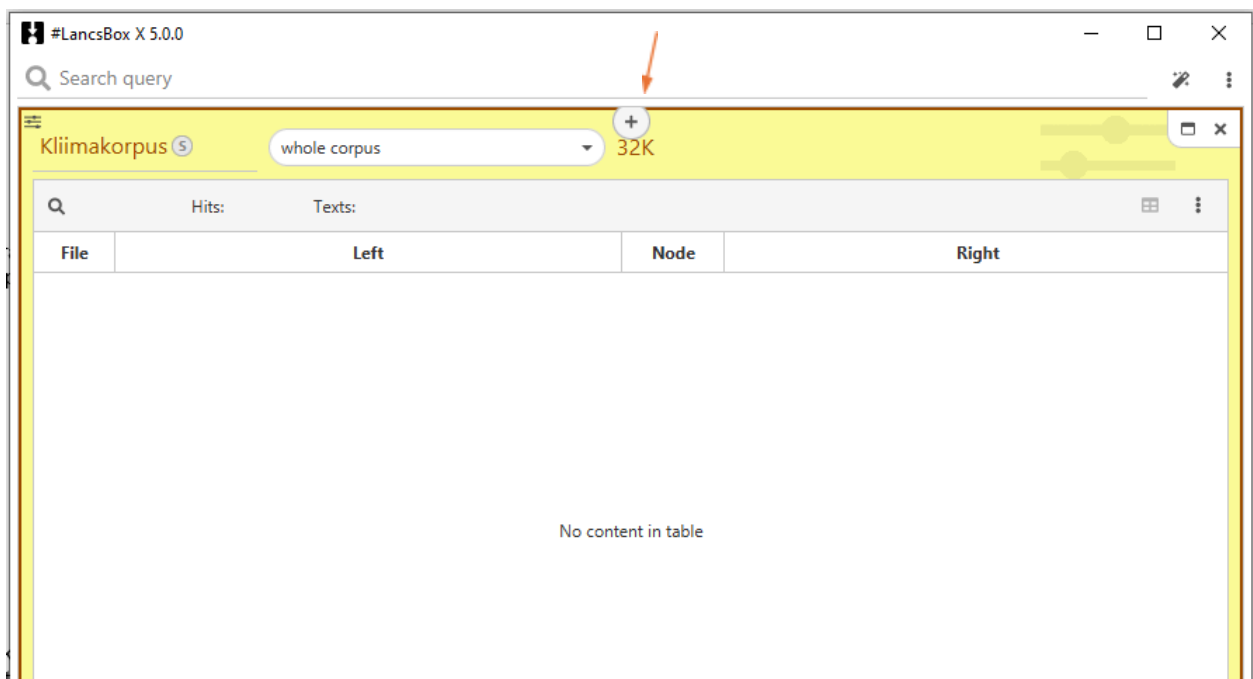
Korpuse lisamise aknas on võimalik muuta aktiivseks mõni sisseehitatud korpus (vaheleht *Corpus hub*), laadida üles omaenda tekstimaterjal (vaheleht *My data*) või koostada korpus automaatselt veebilehtede põhjal (vaheleht *Web*).

Kliimakorpuse sisestamiseks tuleb minna lehele *My data*, anda korpusele nimi, määrata keeleks eesti keel ja laadida üles tekstifailide kaust. Grammatilise märgendamise tugi eesti keelele hetkel puudub.

Nüüd on korpuste menüüsse tekkinud uus valik.

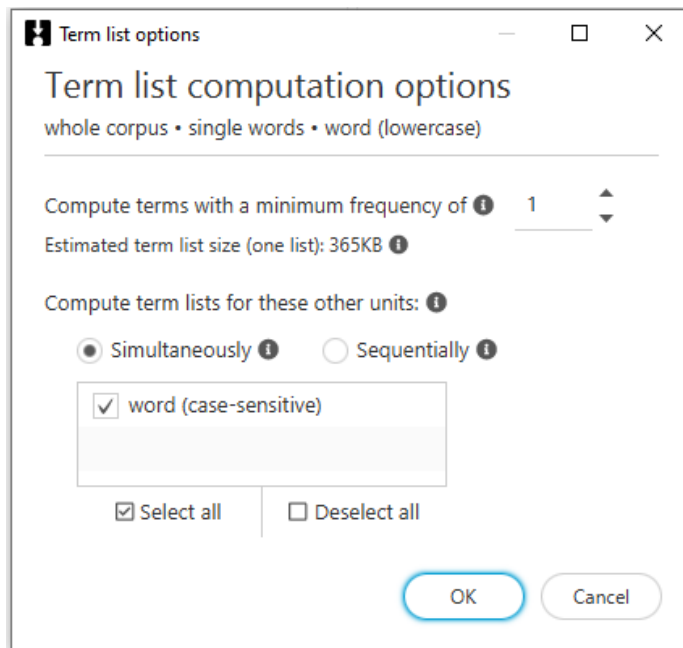


Tekstianalüüsi rakenduste käivitamiseks tuleb klõpsata +-ikoonil, mis ilmub LancsBoxi akna servadesse, kui kursor nende peale lohistada.



Valikus on neli rakendust. **Text** võimaldab sirvida korpuse tekste tervikuna. Näha saab iga teksti mahtu ning sõnavara mitmekesisust MATTR- ja MTLN-indeksi alusel. Topeltklõps failinimel võimaldab teksti lugeda.

Rakendus **Words** kuvab sõnade sagedusloendi. Vaikimisi muudetakse kõik sõnad väiketäheliseks, et lugeda kokku sama sõna kasutused nii lause alguses kui ka lause sees. Paralleelselt võib aga koostada ka loendi, kus suurtähed on alles jäetud. Nii jääb nimede õigekiri korrektseks ja saab teha vahet sõnadel, mille puhul esitäht mõjutab tähendust (nt *Eesti* vs. *eesti*).



Harjutus 1: sõnasagedusloend

1. Kliimakorpusest leiti 10 201 erinevat sõnavormi. Loend on vaikimisi sorditud sõnade absoluutsageduse alusel sagedamast harvemani. Enamik sagedamaid sõnu on n-ö stoppsõnad, mis teksti tähenduse kohta infot ei anna. **Millised on korpuse sagedamad sisusõnad?**
2. Kursoriga sõnale liikudes avaneb eraldi aken kontekstinäidetega. Vali kaks sisusõna, mille kontekste vaatad. **Millist aimu need sõna kasutuse kohta annavad?**
3. Iga tulp saab filtreerida. Jäta alles sõnad, mis esinevad korpuses vähemalt 10 korda. **Kui mitu neid on?**
4. Tulp *Range* näitab, kui suures osas korpuse tekstidest sõna esineb. **Kui palju on sõnu, mida sisaldab vähemalt 50% korpuse tekstidest?**

5. Kuva paremal pool tulpdiaagrammi asemel sõnapilv, klõpsates pilvekese ikoonil. Jäta joonisele alles 100 sagedamat sõna. **Mida saad nende põhjal kliimateemaliste artiklite sisu kohta järeldada?**

Harjutus 2: konkordantsid

Rohkemate sõnakontekstide nägemiseks tuleb kasutada rakendust **KWIC** (*Key Word in Context*). Sõnaloend tasub paralleelselt lahti jätta.

1. Sisesta otsikasti sõna *aasta*. Eelnevate ja järgnevate kontekstide tähestikuliseks sortimiseks klõpsa tulpade pealkirju *Left* ja *Right*. **Kas ja kuidas saab sõna kontekste grupeerida? Mida võib nende põhjal järeldada?**
2. Korda konkordantside analüüsi kahe omavalitud märksõnaga. Võid proovida tärniga otsingut, mis võimaldab leida sarnase algusega sõnu (nt *temperatuur**). Vajadusel vaata pikemat konteksti, lohistades kursori soovitud konkordantsireale. Võid teha seda harjutust koos paarilisega. **Mida huvitavat leidsid?**

Harjutus 3: n-grammid

Rakenduses **Words** saab üksiksõnade sagedusele lisaks vaadata ka sõnajärjendite ehk n-grammide sagedusi.

1. Kahesõnaliste üksuste ehk bigrammide kuvamiseks vali ülemisest rippmenüüst *single words* asemel *2-grams*. **Millised sisukad fraasid leitud väljendite hulgas silma hakkavad?**
2. Vaata ka kolmesõnalisi üksusi ehk trigramme. **Millised on sinu tähelepanekud?**
3. Vali üks bi- või trigramm ja vaatle rakenduse **KWIC** abil selle kontekste. **Mida saad järeldada?**

Harjutus 4: naabersõnad

Rakendus **GraphColl** võimaldab leida soovitud sõnaga sageli koos esinevaid sõnu ehk kollokaate, eestipäraselt naabersõnu.

1. Sisesta otsikasti sõna *temperatuur*. Sorteeeri tulemusi erinevate statistikute (nt LogDice ja MI-skoor) ja koosesinemise sageduse alusel. **Kuidas naabersõnade olulise järjekord sellest muutub?**
2. Vali ka teine märksõna ja sisesta otsinguväljale. Peaksid joonisel nägema kahe otsisõna ühiseid naabersõnu. **Kui on kattuvaid naabersõnu, siis mis need on?**
3. Leia märksõnade *Euroopa* ja *Eesti* jagatud naabersõnad. Selleks, et sõnad joonisel üksteise peale ei kuhjuks, filtreeri välja naabersõnad, mille puhul LogDice'i statistiku

väärtus on alla 10. Vaatle jagatud naabersõnade kontekste koos mõlema märksõnaga.
Kas märkad erinevusi?

Harjutus 5: võtmesõnad

Rakendus **Words** võimaldab lisaks sagedamatele sõnadele leida ka võtmesõnu. Selleks on vaja võrdluskorpust.

Laadi alla British National Corpus (BNC) 2014, valides korpuste menüüst *add corpora*. Korpust saab alla laadida mitmes osas, neist tuleb valida kolmas.

Muuda uus korpus aktiivseks ja tee alamkorpuste rippmenüüs valik *new subcorpus*. Määra tekstiliigiks (*genre*) veebikeel (*elanguage*) ning alamliigiks e-kirjad (*elanguage: email*). Veendu, et valitud alamkorpus oleks aktiivne.

1. Koosta sõnade sagedusloend. Jäta kõik pakutud sõnakujud valituks ja määra vähimaks sageduseks 5. **Kuidas muutub tabel, kui sõnavormide asemel kuvada algvorme (*hw* - *headword*), lemmasid (algvorm koos sõnaliigiga), sõnaliike (*class*) jne?**
2. Leia e-kirjade võtmesõnad esmalt sõnavormide kujul. Klõpsa võtme ikoonil ja vali võrdluskorpuseks BNC2014 veebikoele alamkorpus (*elanguage*). Sõnade vähimaks sageduseks määra samuti 5. **Milliseid tähendusrühmi võib võtmesõnade hulgas esile tuua? Missuguste valdkondade e-kirju tundub korpus sisaldavat?**
3. Koosta võtmesõnade loend ka mingi muu sõnakuju alusel. **Mida saad loendi põhjal järeldada?**
4. Vali üks võtmesõna, mille kontekste vaadelda rakenduses **KWIC**. **Mida said teada sõna kasutuse kohta e-kirjades?**

Harjutus 6: automaatne korpuse koostamine

Korpuse jaoks saab tekste kaevandada, kui liikuda uue korpuse lisamise aknas vahelehele *Web*.

1. Vali huvipakkuv teema, mille kohta leidub Vikipeedias artikkel (nt *Eesti muusika*). Sisesta tekstiotsingu jaoks selle artikli aadress ja määra kaevandamise sügavuseks (*depth*) 1. See tähendab, et korpusesse võetakse tekste ka algsel veebilehel viidatud lehtedelt, aga nendelt lehtedelt edasi ei liiguta. Kui korpus on valmis, siis sirvi seda rakenduses **Text**. **Kuivõrd ootuspärane on valitud artiklite sisu? Kas korpus tundub kasutamiseks sobilik?**
2. Proovi koostada ka ERR-i uudiste korpus, määrates algseks veebiaadressiks <https://www.err.ee> või mõne temaatilise portaali aadressi (nt <https://kultuur.err.ee>). Muuda otsingu parameetreid. Sirvi korpust. **Kas see vajab kasutamiseks eeltöötlust? Kui, siis millist?**

NB! Tekste saad *xml*-formaadis alla laadida korpuste menüüs, klõpsates *Export*.