doi:10.5128/LV29.03

Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks

KRISTINA KOPPEL
Eesti Keele Instituut

Ülevaade. Artiklis analüüsitakse, kas automaatselt valitud autentsed korpuslaused sobivad Eesti Keele Instituudi leksikograafide ning Tartu ja Tallinna ülikooli eesti keelt teise keelena rääkivate üliõpilaste hinnangul eesti keele B2–C1-keeleoskustaseme õppesõnastiku näitelauseks. Selleks viidi läbi uurimus, kus leksikograafid ja keeleõppijad hindasid nelja tüüpi lauseid: tööriista GDEX ehk Good Dictionary Example eesti mooduli versiooni 1.4 parameetrite järgi head ja halvad korpuslaused, filtreerimata korpuslaused ning leksikograafide koostatud näitelaused.

Artiklis antakse esmalt ülevaade e-leksikograafia hetkeseisust Euroopas ja Eestis, sealhulgas sellest, kui palju autentset korpusmaterjali veebisõnastikes kasutatakse. Seejärel kirjeldatakse hindamisülesande ülesehitust ja läbiviimist ning analüüsitakse tulemusi. Leksikograafide ja keeleõppijate hinnangu põhjal kontrollitakse kolme hüpoteesi: korpuslausete filtreerimine on vajalik, GDEXi eesti mooduli versioon 1.4 suudab korpusest tuvastada optimaalsed näitelausete kandidaadid ning välja filtreerida sobimatud, leksikograafi koostatud näitelaused on head näitelaused.

Võtmesõnad: korpusleksikograafia; õppeleksikograafia; näitelaused; GDEX; eesti keel

1. Sissejuhatus

Viimasel kümnendil on Euroopa korpusleksikograafias toimunud väga kiire areng. Tekstikorpused kasvavad üha suuremaks ja korpuspäringusüsteemid muutuvad aina targemaks, tehes osa tööd juba leksikograafi eest ära. Tavapäraseks on saanud sõnastike täis- ja poolautomaatne koostamine. Praegu genereeritakse poolautomaatselt ligikaudu 31% ja täisautomaatselt umbes 7,5% Euroopa sõnastike andmebaasidest. Sõnaraamatud on kolinud paberilt võrku – umbkaudu pooled (46%) Euroopa sõnastikest avaldatakse ainult veebis. (Kallas jt 2019)

Ka Eesti Keele Instituudis viimastel aastatel valminud sõnaraamatud on ilmunud ainult veebis. Paberil anti 2018. aasta seisuga välja veel vaid "Õigekeelsussõnaraamatut ÕS 2018" (ÕS 2018), "Eesti murrete sõnaraamatu" (EMS) vihikuid ning väikeste murdesõnastike sarja. Samuti on hakatud Eesti Keele Instituudis sõnastikke koostama poolautomaatselt. Esimene selliselt koostatud eesti keele sõnastik on "Eesti keele naabersõnad 2019" (Kallas jt 2015) ning praegu on käsil sünonüümide infokihi poolautomaatne koostamine instituudi uues sõnastikusüsteemis Ekilex (Tavast jt 2018).

Leksikaalse info elektrooniline esitus on kaasa toonud uute meediumide kasutamise sõnastike veebiliidestes. Näiteks on tavapärane veebisõnastikesse integreerida pildi-, video- ja helifaile. Lisaks rakendatakse kõnetuvastust, mille abil saab märksõna sisestada suuliselt, ja kõnesünteesi, mis loeb sünteeshääle abil ette sõnastikuinfot. Viimase viie aasta jooksul on sõnastike sisu loomisel hakatud kasutama ka rahva kaasabi ehk rahvahanget (ingl *crowdsourcing*). Rahvahange seisneb selles, et palutakse (harilikult interneti teel) suurema rahvahulga kaasabi teatud eesmärgi saavutamiseks. Üks tuntumaid rahvahanke teel loodud veebisõnastikke on Urban Dictionary, mille algne eesmärk oli slängisõnadele definitsioone koguda. Kuna rahvahanke rakendamine on leksikograafias suhteliselt uus praktika, suhtutakse sellesse veel üsna ettevaatlikult. Teadaolevalt on Euroopa sõnastikes rahvahanget seni kasutatud sünonüümide, neologismide, sõnaassotsiatsioonide ning

murdetranskriptsioonide kogumiseks. (Kallas jt 2019) Eesti leksikograafias on rahvahanke abil kogutud sõnade assotsiatsioone: "Eesti keele assotsiatsioonisõnastiku" projekti raames koguti umbes 1300 märksõnale ligikaudu 460 000 assotsiatsiooni (Vainik 2018). Ene Vainik ise kasutab rahvahanke asemel terminit rahvateadus (ingl *citizen science*).

Vähesed tänapäeva sõnastikud pakuvad käsitsi valitud või leksikograafi koostatud näitelausetele lisaks automaatselt tuvastatud korpuslauseid. Üks sellistest on näiteks inglise keele sõnaraamatu Longman Dictionary of Contemporary English 5. trükk, kus kasutajatel on võimalus lisaks käsitsi valitud näitelausetele lugeda kuni kümmet autentset korpuslauset. Inglise veebisõnastikus Wordnik kuvatakse kasutajale juba suuremat hulka autentseid korpuslauseid ning ka Google'i automaattõlkel Google Translate on sarnane funktsioon olemas. (Cook jt 2014)

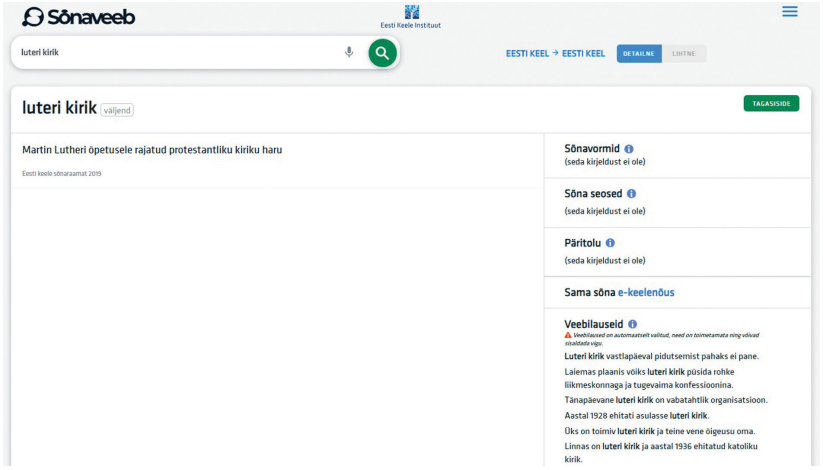
2. Autentsed korpuslaused eesti leksikograafias

Eesti on üks Euroopa riikidest, kus keeleportaalis Sõnaveeb (vt ka Koppel jt 2019b) kuvatakse sõnaraamatu kasutajale autentset korpusinfot veebilausete näol (Koppel jt 2019a). Eriti kasulikud on veebilaused selliste märksõnade juures, kus leksikograafi koostatud näitelaused puuduvad (joonisel 1 pildi all paremas nurgas).

Igasugust autentset korpuslauset aga sõnastiku kasutajale kuvada ei sobi, kuna need võivad olla poolikud (näide 1, otsisõna *maja*), vigased, tundliku sisuga, liiga pikad (näide 2, otsisõna *direktiiv*), sisaldada haruldast sõnavara ja olla keerulise grammatilise või süntaktilise struktuuriga.

(1) 25.10.1687 kohtufoogt oma ülesannetes ja rae erilisel loal loovutab maapealikule (Landtshöffding) ja parunile, Hans Heinrich von Tiesenhausenile **maja** ja kinnistu, mis asub Laial tänaval, surnud Geohard Himseliuse (nr. (etTenTen13)

(2) KOMISJONI OTSUS 96/425/EÜ, milles sätestatakse Mauritaaniast pärit kala- ja akvakultuuritoodete importimise eritingimused 28. juuni 1996 (EMPs kohaldatav tekst) EUROOPA ÜHENDUSTE KOMISJON, võttes arvesse Euroopa Ühenduse asutamislepingut, võttes



Joonis 1. *Eesti Keele Instituudi keeleportaal Sõnaveeb*

arvesse nõukogu 22. juuli 1991. aasta **direktiivi** 91/493/EMÜ (milles sätestatakse kalatoodete tootmise ja turuleviimise tervishoiunõuded, 1 viimati muudetud **direktiiviga** 95/71/EÜ 2), eriti selle artiklit 11, ning arvestades, et : ühenduse eksperdid on teinud Mauritaaniasse kontrollkülastuse, et kontrollida ühendusse saadetavate kalatoodete tootmis-, ladustamis- ja lähetustingimusi; kalatoodete tervishoiukontrolli ja -järelevalvet käsitlevate Mauritaania õigusaktide sätteid võib käsitada samaväärsetena **direktiivi** 91/493/EMÜ sätetega; Mauritaania pädev asutus Ministère des Pêches et de l'Economie Maritime - Centre National de Recherches Océanographiques et des Pêches - Département Valorisation et Inspection Sanitaire (MPEM-CNROP-DVIS) on võimeline tõhusalt kontrollima kehtivate õigusaktide kohaldamist; **direktiivi** 91/493/EMÜ artikli 11 lõike 4 punktis a osutatud veterinaarsertifikaadi saamise kord peab hõlmama samuti näidissertifikaadi määratlust ning sertifikaadi koostamisel kasutatavat keelt (kasutatavaid keeli) ja sellele alla kirjutama volitatud isiku kvalifikatsiooni käsitlevaid miinimumnõudeid; **direktiivi** 91/493/EMÜ artikli 11 lõike 4 punkti b kohaselt tuleks kalatoodete pakenditele kinnitada tähis, kuhu on märgitud kolmanda riigi nimi ja päritoluettevõtte ja külmutuslaeva loanumber; **direktiivi** 91/493/EMÜ artikli 11 lõike 4 punkti c kohaselt

> tuleb koostada heakskiidetud ettevõtete ja/või külmutuslaevade loetelu; kõnealune loetelu tuleb koostada MPEM-CNROP-DVISi poolt komisjonile esitatud teabe põhjal; seepärast peab MPEM-CNROP-DVIS tagama, et täidetakse **direktiivi** 91/493/EMÜ artikli 11 lõike 4 sätteid; MPEM-CNROP-DVIS on andnud ametliku kinnituse **direktiivi** 91/493/EMÜ lisa V peatükis sätestatud eeskirjade järgimise ja kõnealuses **direktiivis** sätestatud ettevõtete ja külmutuslaevade heakskiitmist käsitlevate nõuetega samaväärsete nõuete täitmise kohta; käesoleva otsusega ettenähtud meetmed on kooskõlas alalise veterinaarkomitee arvamusega, ON VASTU VÕTNUD KÄESOLEVA OTSUSE. (NC)

Selleks, et veebisõnastiku kasutaja ei puutuks kokku sellisete lausetega, on vaja korpuslauseid valivale programmile ette anda teatud reeglid, millele toetudes oskab see välja valida optimaalsed näitelause kandidaadid. Eestis on seni korpuslausete filtreerimiseks kasutatud korpuspäringusüsteemi Sketch Engine (Kilgarriff jt 2004) tööriista GDEX ehk Good Dictionary Examples (Kilgarriff jt 2008) eesti mooduli erinevaid versioone (Kallas jt 2015; Koppel & Kallas 2016; Koppel 2017; Kosem jt 2019).

GDEX töötab reeglipõhisel valemil, mis ette määratud tunnuseid arvestades otsib korpusest automaatselt optimaalseid näitelause kandidaate ning reastab need paremuse järjekorda, nii et parimad kandidaadid on nimekirja eesotsas. GDEX on mõeldud eelkõige abimeheks leksikograafile, aidates teda näitelausete valikul. GDEX töötab paremini, kui seda rakendades arvestada lause keelespetsiifilisi parameetreid, nt sõnade ja lause pikkust, märksõna asukohta lauses jmt. Eesti mooduli versioone 1.1–1.4 arendades on seni arvesse võetud eesti sõnastike näitelausete analüüsi tulemusi (Kallas jt 2015; Koppel & Kallas 2016; Koppel 2017; Kosem jt 2019). Versiooni 1.2 kasutati "Eesti keele naabersõnade 2019" andmebaasi genereerimisel (Kallas jt 2015) ning versiooni 1.4 (GDEX 1.4) (Koppel 2017; Kosem jt 2019) abil loodi "Eesti keele õppekorpus 2018 (etSkELL)", mis on keeleõppekeskkonna etSkELL (Sketch Engine for Estonian Language Learning) ning Sõnaveebi autentsete veebilausete allikas. Kõik õppekorpuse laused vastavad GDEX 1.4 poolt

ette määratud hea näitelause parameetritele, näiteks on kõik laused täislaused, minimaalselt 4 ja maksimaalselt 20 sõnet pikad, sisaldavad tegusõna jmt (parameetrite täielikku loetelu vt Koppel 2017: 67). Lisaks on olemas eraldi versioonid eri keeleoskustasemetele (Koppel 2019), kuid neid ei ole seni veel korpuste loomisel rakendatud.

Eesti Keele Instituudi veebisõnastikes ei ole varem autentset korpusmaterjali kuvatud ning kasutajad on oma pikaajalisest kogemusest sõnastike kasutamisel harjunud arvestama sellega, et kogu sõnastikus esitatav info on leksikograafi poolt üle kontrollitud, toimetatud ning seega korrektne. Ka Sõnaveebi kaudu Eesti Keele Instituudi leksikograafidele saadetavas tagasisides on kasutajad osutanud sellele, et mõned veebilaused nende meelest näitelauseks ei sobi.

Autentsete veebilausete kuvamisega on esile kerkinud mitmeid probleeme, mida on ainuüksi reeglipõhisel lähenemisel töötava tööriista abil raske kõrvaldada. Eeskätt tekitavad probleeme polüseemsed sõnad, (vormi)homonüümia, lemmatiseerimise ja morfoloogilise märgenduse vead, madala sagedusega sõnad ja tundliku või sobimatu sisuga (nt masintõlkelised) laused (Koppel 2019; Koppel jt 2019a). Ka teiste keelte GDEXi arendajad on sarnaste probleemidega kokku puutunud. Tanara Z. Kuhn jt (2019) on välja pakkunud meetodi, kuidas veebikorpust puhastada solvavat ja tundlikku keelekasutust sisaldavatest lausetest. Selleks kombineerivad nad reeglipõhist lähenemist masinõppemeetodiga ning rakendavad rahvahanget solvavate ja tundlike sõnade väljaselgitamiseks.

Selleks et välja selgitada, kui hästi GDEX 1.4 heade näitelausete tuvastamisel töötab ning kas autentseid korpuslauseid sobib otse ilma toimetamiseta lõppkasutajale näidata, palusin leksikograafidel ja eesti keelt teise keelena rääkivatel üliõpilastel lauseid hinnata kahes jaos: esimeses hindamisülesandes selgitasin välja üldised hinnangud lausete sobivusele ning jätkuküsitluses küsisin hinnangu põhjendusi. Hindamise tulemused aitavad GDEXi eesti moodulit edasi arendada.

3. Materjali ülevaade ja hindamisülesannete ülesehitus

Lõin esimese hindamisülesande avatud lähtekoodiga platvormis Pybossa, mille abil viiakse läbi lihtsamaid rahvahanke projekte ning analüüsitakse kogutud andmeid. Pybossa võimaldab oma hindamisülesannet ise kujundada, kontrollida osalejate arvu ning hoiustada kogutud andmeid. Esimese hindamisülesande tulemuste analüüsi järel viisin hindajate seas läbi jätkuküsitluse, mille eesmärk oli välja selgitada põhjused, miks nende meelest üks või teine lause sõnastikku ei sobi.

Soovisin lausete hindamisega tõestada kolme hüpoteesi:

1. Korpuslausete filtreerimine on vajalik ning sellel on kaks eesmärki:
   a) abistada leksikograafi näitelausete valikul;
   b) automaatselt kõrvaldada näitamiseks sobimatud laused.
2. Reeglipõhine lähenemine (GDEX) võimaldab tuvastada optimaalsed näitelause kandidaadid ja välja filtreerida sobimatud.
3. Leksikograafi koostatud sõnastiku näitelaused sobivad näitelauseks.

Hüpoteeside tõestamiseks valisin "Eesti keele naabersõnade 2019" kui B2–C1-tasemel keeleõppijale suunatud sõnastiku andmebaasist 40 juhuslikku märksõna – iga sõnaliigi jaoks kümme[1]:

- **tegusõnad**: *tunduma*, *leppima*, *paistma*, *langema*, *koguma*, *kajama*, *ühinema*, *kaitsma*, *erutama*, *kurvastama*;
- **omadussõnad**: *ropp*, *akadeemiline*, *primitiivne*, *ruumiline*, *tundlik*, *väljapaistev*, *mitmeaastane*, *inimtühi*, *atraktiivne*, *varajane*;
- **nimisõnad**: *kontingent*, *rassism*, *käik*, *stseen*, *graafik*, *turnee*, *juubel*, *sõit*, *viin*, *areen*;
- **määrsõnad**: *äkki*, *unarusse*, *julgelt*, *uuesti*, *õnnelikult*, *kangesti*, *natuke*, *tublisti*, *salaja*, *praktiliselt*.

[1] Juhuvalim on võetud SQLi funktsiooniga *random()*.

Seejärel võtsin iga märksõna jaoks juhuvalimi näitelausetest, kuhu kuulus:

- üks korpuslause, mis GDEX 1.4 järgi vastab hea näitelause parameetritele (edaspidi: hea korpuslause);
- üks korpuslause, mis hea näitelause parameetritele ei vasta (edaspidi: halb korpuslause);
- üks filtreerimata korpuslause, mis võis vastata nii hea kui halva näitelause parameetritele;
- üks leksikograafi koostatud sõnastiku näitelause.

Korpuslaused võtsin "Eesti keele ühendkorpusest 2017" ning leksikograafi koostatud näitelaused "Eesti keele sõnaraamatust 2019". Leksikograafi koostatud näitelaused lisasin andmestikku kontrollgrupiks, et näha, kas neid hinnatakse autentsete lausetega võrreldes sama kriitiliselt.

3.1. Esimene hindamisülesanne

Hindamisülesande saatsin tegemiseks kaht tüüpi hindajatele: Eesti Keele Instituudi leksikograafidele ning Tartu ja Tallinna ülikoolis õppivatele eesti keelt teise keelena rääkivatele tudengitele. Osalemise kutse sai kokku seitse leksikograafi – "Eesti keele sõnaraamatu 2019" ja "Eesti keele naabersõnade 2019" sõnastiku koostajad ja toimetajad – ning 31 üliõpilast. Kuna olen GDEXi eesti moodulit arendades silmas pidanud eesti keelt B2–C1-keeleoskustasemel rääkijat, soovisin hindajateks just selle tasemega üliõpilasi, keda aitas leida Tartu Ülikooli eesti keele võõrkeelena dotsent Raili Pool.

Kuna hindajate rühmi oli kaks – leksikograafid ja keeleõppijad –, tegin Pybossas kaks sama sisuga projekti. Selleks, et mõlema projekti tulemused oleksid võrreldavad, määrasin projekte üles ehitades, et kogun kummaski neis igale lausele viie erineva hindaja hinnangu ehk esimeses projektis viie leksikograafi ja teises projektis viie keeleõppija oma. Otsustasin piirduda viie hinnanguga, sest ma ei saanud eeldada, et iga osaleja on motiveeritud ära hindama kogu andmestiku ehk kokku 160 lauset (iga märksõna kohta neli erinevat lauset), mida viis leksikograafi küll

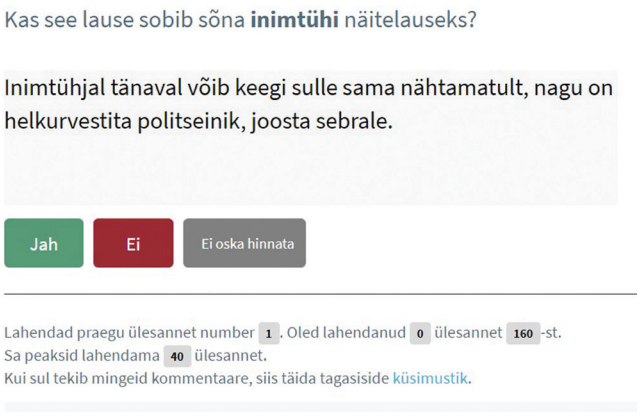
tegid. Kasutajate motiveerimine on teadaolev probleem rahvahanke projektide läbiviimisel (vt nt Leimeister jt 2009; Kaufmann jt 2011), milleks Pybossat kõige sagedamini ka kasutatakse. Sellest probleemist teadlikuna jagasin kogu andmestiku neljaks väiksemaks ülesandeks, kusjuures iga väiksem ülesanne sisaldas kõiki nelja tüüpi lauseid. Kuigi väiksemas ülesandes tuli hinnata 40 lauset, jättis osa keeleõppijatest ülesande täitmise pooleli, kuid nende hinnangud läksid sellegipoolest arvesse. Sellest tulenevalt jõudis hindamisülesandes osaleda üheksa tugengit. Kuivõrd kõik keeleõppijad ei hinnanud kõiki lauseid, võisid hindamistulemusi veidi mõjutada ka individuaalsed erinevused, mille analüüs jääb siinsest artiklist välja.

Keeleõppijatel palusin lauseid hinnata lähtuvalt enda eesti keele oskusest: kas esitatav lause sobib nende keeleoskustasemele vastavasse õppesõnastikku näitelauseks. Vastusevariante oli kolm: *jah*, *ei* ja *ei oska hinnata*. Palusin vastata *jah*, kui lause on nende jaoks kasulik, arusaadav ning kui see näitab, kuidas sõna kasutada. Palusin vastata *ei*, kui lause on nende jaoks keerulise sõnavara ja/või grammatikaga, liiga pikk või liiga lühike või kui miski muu neid lauses segab. Seisukoha puudumisel palusin vastata *ei oska hinnata*.

Leksikograafidel palusin hinnata, kas laused sobiks B2–C1-keeleoskustasemele suunatud õppesõnastikku. Ühtlasi kirjeldasin põgusalt, mida vastava keeleoskustasemega keeleõppijad eesti keeles teha oskavad. Kuna leksikograafid töötavad korpusega ja toimetavad korpuslauseid iga päev, andsin neile neli vastusevarianti – *jah*, *pigem jah*, *ei* ja *ei oska hinnata*. Palusin vastata *jah*, kui nende meelest on lause arusaadav, korrektne, illustreerib sõna tavapärast konteksti, sisaldab tüüpilisi kollokatsioone vmt. Palusin vastata *pigem jah* kui lausel on väikeste mööndustega potentsiaali olla hea näitelause (näiteks kui seda on väga lihtne sobivaks toimetada). Palusin vastata *ei*, kui lause mitte mingil juhul sõnastikku ei sobi ning *ei oska hinnata*, kui neil seisukoht puudub.

Hindaja nägi korraga ühte lauset, millele eelnes küsimus – "Kas see lause sobib sõna X näitelauseks?" Sõna tähenduse selgitust lisatud ei olnud, mis, nagu hiljem selgus, mõjutas tõenäoliselt mõne lause puhul

hindajate arvamust. Eelarvamuste vältimiseks lause allikat hindajale ei kuvatud, st et ei olnud eksplitsiitselt öeldud, kas hinnatav lause pärineb korpusest või sõnastikust (joonis 2).



Joonis 2. *Lause hindamine Pybossa platvormis (keeleõppija vaade)*

Esialgne mõte oli juba esimeses hindamisülesandes pakkuda rohkem vastusevariante (*jah*, *ei*, *ei* (*keeruline sõnavara*), *ei* (*keeruline grammatika*), *ei* (*liiga lühike*), *ei* (*liiga pikk*), *ei* (*semantiliselt tühi*) jmt). Ühelt poolt oleks see aidanud nii võrrelda leksikograafide ja keeleõppijate hinnangut samadele lausetele kui ka välja selgitada, mis neid konkreetsete lausete juures täpsemalt häirib. Teisalt kartsin, et sellise sõnastusega vastusevariandid oleksid tulemusi oluliselt kallutanud, suunates hindajaid lausetesse kriitilisemalt suhtuma. Seetõttu palusin hindajatel oma hinnanguid põhjendada jätkuküsitluse käigus.

3.2. Jätkuküsitlus

Kui esimese hindamisülesande eesmärk oli kvantitatiivselt välja selgitada, kas eri tüüpi laused sobivad leksikograafide ja keeleõppijate hinnangul õppesõnastiku näitelauseks, siis jätkuküsitluse eesmärk oli teada saada põhjuseid, miks leksikograafid ja keeleõppijad häid korpuslauseid

sobimatuks ning halbu korpuslauseid sobivaks pidasid. Seetõttu palusin jätkuküsitluses uuesti hinnata kolme tüüpi lauseid:

1) head korpuslaused, kuid mida enamik hindajaid esimeses hindamisülesandes ei osanud hinnata või ei pidanud sobivaks;
2) halvad korpuslaused, kuid mida enamik hindajaid esimeses hindamisülesandes ei osanud hinnata või pidasid sobivaks;
3) sõnaraamatu laused, mida enamik hindajatest esimeses hindamisülesandes sobivaks ei pidanud.

Jätkuküsitlus ühtegi filtreerimata korpuslauset ei sisaldanud, kuna ei olnud teada, kas need vastasid GDEX 1.4 parameetritele või mitte, ning seega ei olnud nende lausete hinnangute põhjendamine GDEXi eesti mooduli arendamise seisukohalt relevantne.

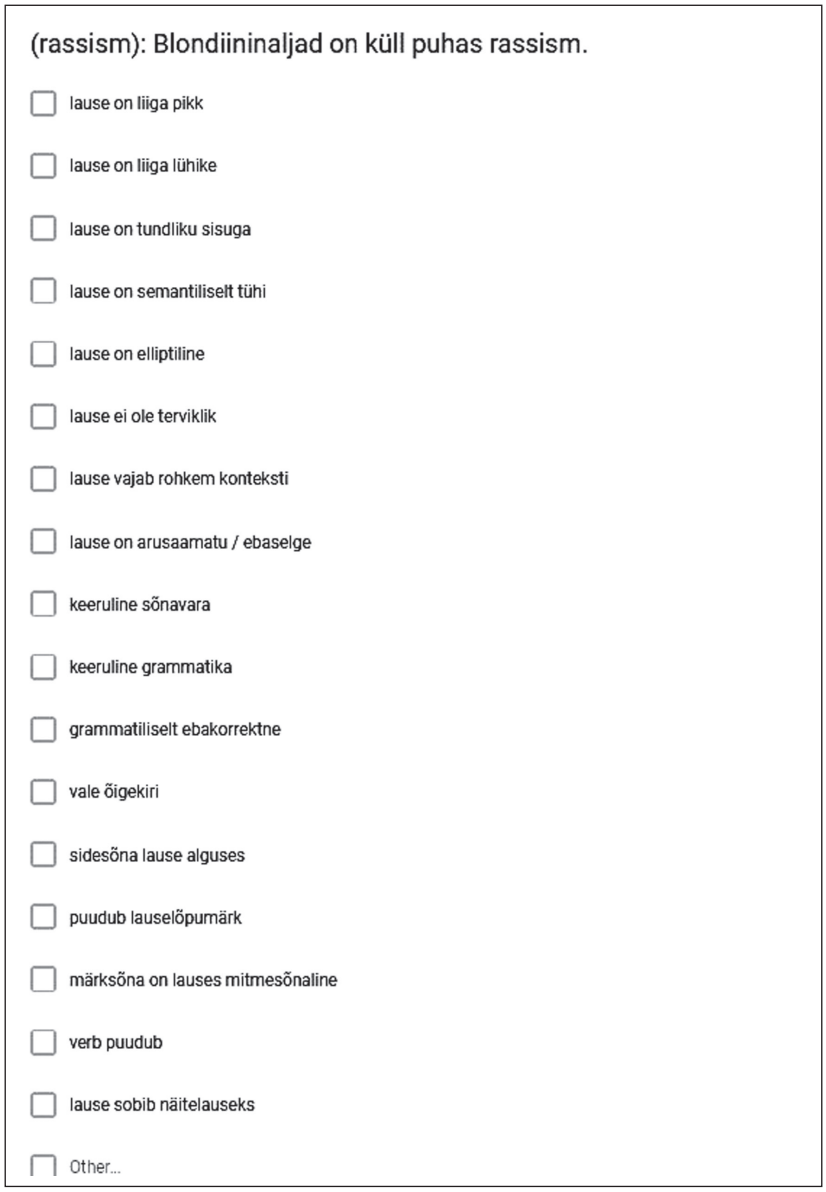
Palve jätkuküsitluses osaleda saatsin esimeses hindamisülesandes osalenud hindajatele (viiele leksikograafile ja üheksale keeleõppijale), kellest vastas viis leksikograafi ja viis keeleõppijat, seega olid ka jätkuküsitluse tulemused võrreldavad.

Jätkuküsitluse maht oli oluliselt väiksem – leksikograafid said uuesti hindamiseks 18 ja keeleõppijad 20 lauset, kusjuures 11 neist kattus. Kasutasin jätkuküsitluseks Google Formsi keskkonda, kus hindajale kuvati märksõna ja näitelause, ning sellele järgnes 17 vastusevarianti, mille sõnastus hindajate tüübiti pisut erines. Näiteks kui leksikograafidel oli üheks vastusevariandiks *alus puudub*, siis keeleõppijatel *lauses puudub tegija*. Ühe vastusevariandina jätsin võimaluse hindajal ise oma otsust põhjendada. Korraga sai valida mitu vastusevarianti (joonis 3).

Etteantud vastusevariandid tuginevad GDEXi eesti mooduli parameetritele ning seda arendades kõige sagedamini ette tulnud probleemidele.

4. Tulemuste analüüs

Analüüsin siinses peatükis esimese hindamisülesande tulemusi ja toon konkreetsete lausete kohta välja jätkuküsitluses toodud hindajate põhjendusi. Analüüsi lihtsustamiseks teisendasin esimeses hindamisülesandes



Joonis 3. *Jätkuküsitlus Google Formsi keskkonnas (leksikograafide vaade)*

antud leksikograafide vastusevariandi *pigem jah* vastuseks *jah*. Juhul kui kaks hindajat hindasid ühe projekti sees lause sobivaks, kolm sobimatuks, teisendasin selle vastusevariandiks *ei*; kui lause hindamisel ei saavutatud konsensust, näiteks kui kahele hindajale lause sobis, kahele mitte ja üks ei osanud hinnata, siis teisendasin selle vastusevariandiks *ei oska hinnata*.

Alapeatükkides 4.1.–4.4. analüüsin esimeses hindamisülesandes antud hinnanguid eri tüüpi lausete sobivusele õppesõnastiku näitelauseks. Joonistel 4–7 on esitatud nii diagramm kui ka tabel, kus on märgitud esimeses hindamisülesandes antud viie leksikograafi ja viie keeleõppija hinnangud lausete sobivuse kohta eraldi, aga ka mõlema hindajarühma (kümne inimese) arvamus kokku. Tabeli viimane rida (*kokku*) joonistel 4–7 ongi saadud kümne vastaja (viie leksikogaafi ja viie keeleõppija) arvamusi kokku liites ning tulemusi teisendades. Kui vastusevariante *jah* ja *ei* anti lausele võrdselt, siis teisendasin selle vastusevariandiks *ei oska hinnata*, muudel juhtudel jäi vastuseks sagedasim vastusevariant.

Esimese hindamisülesande ja jätkuküsitluse tulemuste võrdlus on esitatud joonisel 8 alapeatükis 4.5.

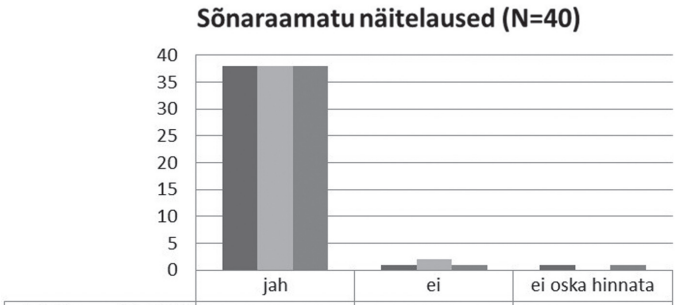
4.1. Sõnaraamatu näitelaused

Sõnaraamatu näitelauseid oli andmestikus kokku 40. Hüpotees oli, et kõik sõnaraamatu näitelaused hinnatakse sobivaks.

Jooniselt 4 on näha, et nii leksikograafide kui ka keeleõppijate arvates sobib näitelauseks 38 lauset ehk 95% lausetest. Kõik leksikograafid ja keeleõppijad hindasid sobivaks 11 lauset (näited 3–13)[2]:

(3) Filmisime üht ja sedasama **stseeni** terve päeva.
(4) Nad on viiskümmend aastat **õnnelikult** koos elanud.
(5) **Paistab**, et tal läheb hästi.
(6) Vaidlus väljus **akadeemilistest** raamidest.
(7) Reibas vanadaam tähistas hiljuti suurt **juubelit**.

[2] Märksõna on näitelauses paksus kirjas.



	jah	ei	ei oska hinnata
leksikograafid (N=5)	38	1	1
keeleõppijad (N=5)	38	2	0
kokku (N=10)	38	1	1

Joonis 4. *Leksikograafide ja keeleõppijate esimeses hindamisülesandes antud hinnang sõnaraamatu näitelausete sobivuse kohta*

(8) Lavastajal on hea visuaalne ja **ruumiline** ettekujutus.
(9) Vanem põlvkond on **areenilt** juba lahkunud.
(10) Ta on viimasel ajal **tublisti** trenni teinud.
(11) Isuäratavad lõhnad **erutavad** meeli.
(12) Büroohooned on õhtuti **inimtühjad**.
(13) Ära **kurvasta**, kõik võib veel hästi minna.

Kuigi ka näide 14 sobib leksikograafide meelest sõnastiku näitelauseks, erinesid nende arvamused selle lause osas kõige rohkem. Neli keeleõppijat viiest pidas seda samuti sobivaks näitelauseks.

(14) **Väljapaistva** arhitektuuriga ehitis.

Tõenäoliselt on põhjus selles, et näites 14 on tegemist korpuses sageli esile tuleva fraasiga, mitte klassikalise sõnaraamatu näitelausega. "Eesti keele sõnaraamatus 2019" on sellised näited tavapärased. Seal esitatakse vahel kasutusnäitena näiteks kollokatsioone, mis võivad omaette üksusena olemas olla ka "Eesti keele naabersõnades 2019", nt *kodune aadress*, *isiklik elu*, *ebavääri kas käitumine*. Ka "Eesti keele sõnaraamatu 2019" autorid (Langemets jt 2018: 951) ise on öelnud, et nende kasutusnäited

on teiste sõnastikega võrreldes teist tüüpi ega pruugi seetõttu automaatselt teistesse sõnastikesse (kakskeelsetesse, õppesõnastikesse) sobida. Samas on "Eesti keele sõnaraamatu 2019" näitelausete eesmärk õppesõnastike näitelausetega sama – toetada seletust ja aidata sõna tähendust paremini mõista. (Langemets jt 2018)

Kõige negatiivsema hinnangu sai leksikograafidelt eelmise näitega sarnane näide 15 – kaks leksikograafi hindasid selle lause sobivaks, kolm mitte. Jätkuküsitluses toodi põhjuseks just seda, et tegemist on fraasiga, lause on elliptiline, liiga lühike ning sealt puudub tegusõna.

(15) **Kangesti** palav ilm.

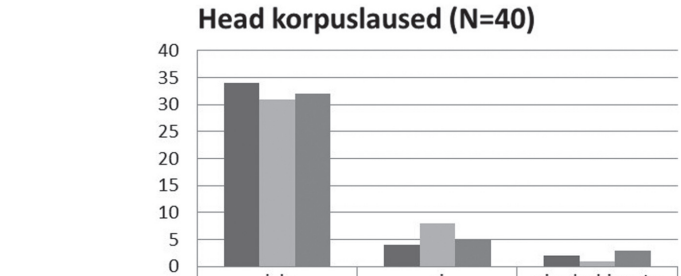
Keeleõppijatele näide 15 aga probleeme ei valmistanud – neli hindajat viiest pidas seda sobivaks, samuti nagu elliptilist näidet 14. Keeleõppijatelt said kõige negatiivsema hinnangu näited 16 ja 17:

(16) Täna on **ropp** tuul!
(17) Blondiininaljad on küll puhas **rassism**.

Neli keeleõppijat viiest pidas näidet 16 sobimatuks ning näidet 17 pidas sobimatuks kolm keeleõppijat. Kolm leksikograafi viiest pidas seevastu näidet 16 sobivaks, kuid näite 17 puhul konsensusele ei jõutud. Jätkuküsitlusest selgus, et keeleõppijate jaoks oli näide 16 kõnekeelne ja tundliku sisuga, samuti vajab see rohkem konteksti. Üks vastajaist leidis, et lause ei näita sõna *ropp* tavapärast kasutust. Näite 17 sisu pidasid kaks leksikograafi ja kaks keeleõppijat tundlikuks ning üks leksikograaf ja kaks keeleõppijat leidsid, et lause sisu on hoopis vale, kuna rassismil sellist tähendust ei ole.

4.2. Head korpuslaused

Korpuslauseid, mis vastasid GDEX 1.4 järgi hea näitelause parameetritele (ehk head korpuslaused), oli andmestikus kokku 40. Hüpotees oli, et GDEX suudab korpusest tuvastada optimaalsed näitelause kandidaadid.



	jah	ei	ei oska hinnata
leksikograafid (N=5)	34	4	2
keeleõppijad (N=5)	31	8	1
kokku (N=10)	32	5	3

Joonis 5. *Leksikograafide ja keeleõppijate esimeses hindamisülesandes antud hinnang heade korpuslausete sobivuse kohta*

Jooniselt 5 selgub, et keeleõppijad suhtuvad autentsete korpuslausete sobivusse kriitilisemalt. Leksikograafid hindasid sobivaks 34 (85%) ning keeleõppijad 31 (77,5%) head korpuslauset. Kokku ühtis leksikograafide ja keeleõppijate arvamus 32 lause puhul ehk 80% lausetest. Kõik leksikograafid ja keeleõppijad hindasid sobivaks neli lauset (näited 18–21):

(18) Talle meeldis **kangesti** juttu ajada.
(19) Küsimuste puhul võta **julgelt** ühendust.
(20) Pühapäeva õhtupoolikul oli restoran täiesti **inimtühi**.
(21) Poes polnud tilkagi **viina** ega veini.

Kõik leksikograafid leidsid ka, et näide 22 sobib näitelauseks.

(22) Kondoom **kaitseb** mõlemat partnerit.

Näidet 22 pidasid sobivaks kolm keeleõppijat, üks mitte ning üks ei osanud hinnata. Kõik keeleõppijad hindasid sobivaks kokku neliteist lauset (lisaks näidetele 18–21 ka näited 23–32):

(23) **Äkki** keegi teab ja mäletab?
(24) Eestlased on ka rahvusvahelisel **areenil** arvestatavad tegijad.
(25) Taevas oli täis eredaid tähti ja lumehelbeid **langes** taevast alla.

(26) Üheks enim poleemikat tekitanud teooriaks on **rassism**.
(27) Küsitluse **käiku** saab jälgida linnavalitsuse koduleheküljelt reaalajas.
(28) Autojuhid ootasid kannatlikult ja aeglustasid **sõitu**.
(29) Elektrijaamade renoveerimistööd peavad **graafiku** järgi algama juba augustis.
(30) Kogu tegevuse jäädvustas **salaja** filmilindile süüdistuse esitanud naabrinaine.
(31) Päike **paistab**, ilm on suviselt soe.
(32) Üritasin lavastada väga suurejoonelist **stseeni**.

Kuigi leksikograafid ei hinnanud sobivaks nelja ja keeleõppijad kaheksat lauset, ei kattunud see arvamus ühegi lause puhul 100%. Kõige negatiivsemad hinnangud said näited 33 ja 34.

(33) Tööreiside tõttu ära jäänud **juubelid** ja teatrietendused.
(34) **Praktiliselt** kaasav ja proovima motiveeriv.

Jätkuküsitluses tõid leksikograafid põhjuseks, et need on elliptilised, tegemist on fraasiga ning lausetes puudub tegusõna. Keeleõppijad tõid välja, et tegemist pole terviklike lausetega ning nad tundsid puudust alusest ja rohkemast kontekstist.

Leksikograafid ei hinnanud sobivaks ka näiteid 35-36.

(35) Väikesed lapsed on selliste kokkupuudete osas eriti **tundlikud**.
(36) Tagavaraks **kogutud** toidu suhtes on neil suurepärane mälu.

Näidet 35 peeti arusaamatuks anafoorse[3] sõna *selliste* tõttu, näites 36 häiris leksikograafe anafoor *neil*, lisaks leiti, et grammatika on kohmakas ning märksõna võiks olla pöördelises vormis. Näidet 35 pidasid sobivaks neli keeleõppijat viiest, näidet 36 kolm keeleõppijat. Keeleõppijad pidasid sobimatuks näiteid 37 ja 38, leksikograafid jäid nende lausete osas eriarvamusele.

(37) Praegu olen hetkes ja **kurvastan**.
(38) Sarve kasv peatus ning naine **leppis** olukorraga.

[3] Anafoor ehk tagasiviide viitab tekstis varem esinenud infole. Eesti keeles toimivad anafoorina harilikult asesõnad.

Näidet 37 pidasid keeleõppijad arusaamatuks, leksikograafid pidasid keeruliseks ka sõnavara (sisaldab väljendit *hetkes olema*). Näidet 38 pidasid mõlemad hindajate grupid arusaamatuks.

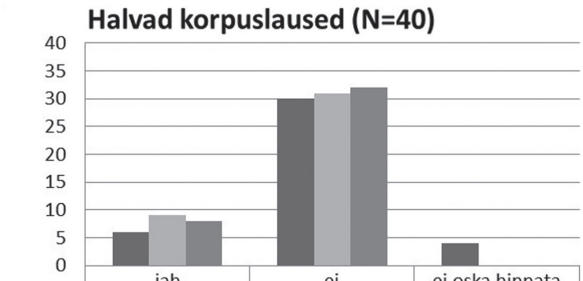
Keeleõppijad ei hinnanud sobivaks ka näiteid 39–41.

(39) Miks peaks mind **erutama** vängete aroomidega autosalong?
(40) Pokkeriga hakkas tegelema **varajases** lapsepõlves.
(41) Tänu sellele on Eesti üks edukamaid Euroopa Liidu kandidaatriike ning sai Praha tippkohtumisel ametliku kutse **ühineda** ka NATO-ga.

Näite 39 kohta tõid keeleõppijad põhjuseks keerulist sõnavara ning samuti tunti puudust kontekstist. Näites 40 tunti puudust alusest. Näidet 41 pidasid keeleõppijad liiga pikaks, samuti tekitas probleeme anafoorne viide *tänu sellele*.

4.3. Halvad korpuslaused

Korpuslauseid, mis GDEX 1.4 järgi hea näitelause parameetritele ei vasta (ehk halvad korpuslaused), oli andmestikus kokku 40. Hüpotees oli, et GDEX suudab välja filtreerida sobimatud näitelause kandidaadid.



	jah	ei	ei oska hinnata
leksikograafid (N=5)	6	30	4
keeleõppijad (N=5)	9	31	0
kokku (N=10)	8	32	0

Joonis 6. *Leksikograafide ja keeleõppijate esimeses hindamisülesandes antud hinnang halbade korpuslausete sobivuse kohta*

Jooniselt 6 on näha, et neljakümnest halvast korpuslausest ei sobi hindajate meelest näitelauseks 32 lauset ehk 80% lausetest. Kõik leksikograafid ja keeleõppijad pidasid sobimatuks kuut lauset (näited 42–47):

(42) siis vedas ju **roppu** moodi - nii palju, kui mina olen neid auhinnakotikesi näinud, siis on need pungil täis spämmi, aga kahjuks on kõik voldikud ja muud ajakirjad taolisel kriitpaberil, et need ei sütti ka bensiini abil...

(43) laupäev, 18. juunikell 10.00 abilinnapea Jane Mets osaleb Rääma Põhikooli lõpuaktusel Endla teatriskell 12.00 abilinnapea Jane Mets osaleb Pärnu Koidula Gümnaasiumi lõpuaktusel Pärnu Kontserdimajaskell 14.00 linnapea Romek Kosenkranius peab tervituskõne rannajalgpalli uue rannaspordi **areeni** avamisüritusel.

(44) Noh aga kui universumit võttagi kui suurt seest tühja toroid,i milles kõik aiva ühtsama ringi keerleb kuigi pildina tundub valgus sirgjooneliselt liikuvat, Kui mind vaimud koolitasid, siis üritasin õppida ka automaatkirjutamist ja joonistamist, automaatjoonistamise kaudu joonistus universumi küsimuse vastus välja suure seest tühja toroidina, milles maa üüratu valgusniitide pundina üha ringe teeb ja miks siis mitte hüpata ühelt niidilt teisele kuid hopp, see on ju teine ajastu ja kuidas sealt samasse ajahetke tagasi saab on küsimuste küsimus, millele need usssiaukude kaudu rändajad ei vasta, kuigi jah iga kord me jookseme minevikku meenutades ajas tagasi ajatusse aega kus ühe jõu sündmused on ühes kohas ja rituaaliga seda jõudu kasutada saad kui oskad või pääsed sinna käima, sõda on selles et neid tunneleid mida pidi käiakse oma minevkus saab võõras jõud kinni panna ja iga kord pead end uuesti läbi kaklema et oma algallika juurde pääseda, nn nõidadel on komme vastaste **käike** kõvasti ja igaveseks kinni toppida, kolage aga mööda gurusid ja nõidu saate korralikult oma jõust lahti ja guru patareiks , omal korralik kogemus olemas selle asjaga.

(45) Kas kaob üleüldse selline rahvakogunemine ja eestlaste ühtsus või muud seda ma ei tea aga sees olid sõnad – see on viimane (kas ümmarguse **juubeli** numbriga vms.)!

(46) Või meenutage mõnda, mille heliriba täidaks midagi sama ägedat, kui Grapsi esitatud sõul-jazz-funk, mille saatel filmi peategelane lõputiitrite eel lendava taldreku trepist üles loivab (üks šefimaid **stseene** eesti kinos, kinnitan teile!

(47) kui esimene **sõit** oleksin tublisti kümnes püsinud, siis teine sõit oleksin pidanud ulmelise 1. koha sõitma.

Esimene hindamisülesanne näitas, et leksikograafid pidasid neljakümnest halvast korpuslausest tegelikult sobivaks kuut (kokku 15%) (näited 48–53) ning keeleõppijad üheksat (kokku 22,5%) lauset (näited 49–57).

(48) **Atraktiivne** naine ei jäta politseinikku sugugi külmaks - peagi leiab too, et suhtub oma kaitsealusesse hoopis teisiti, kui oleks kohane oma ametikohuseid täitvale korravalvurile.
(49) Küsitluste tulemused näitavad, et väga olulisi muudatusi allikate osas, kes peaksid inimesi informeerima, viimase pooleteise aasta jooksul **praktiliselt** ei ole.
(50) See õnnetus lõppes **õnnelikult**, inimesed jäid ellu.
(51) Kuperoosa tekib sagedamini õhukesel, kuival ja **tundlikul** nahal
(52) Neil **paistis** väga huvitav olema.
(53) Sellega me lihtsustaksime muuseas ka kohalike omavalitsuste tööd ja annaksime pensionäridele kindluse, et Riigikogu on nende selja taga ja **kaitseb** nende huve.
(54) **Ruumiline** struktuur
(55) Või **äkki** hoopis selleks, et viia mõtted viletsast majandusseisust eemale ning tekitada uus ühine vaenlane?
(56) Aga olen sellega **kangesti** rahul
(57) Selle asemel **lepi** lapsega kokku internetikasutuse reeglites, mida, kui palju ja millal võib kasutada

Näited 48, 49 ja 54 on GDEX 1.4 parameetrite järgi halvad korpuslaused, kuna lause pikkuseks on määratud minimaalselt 4 ja maksimaalselt 20 sõnet, kuid näites 48 on sõnesid 27, näites 49 on sõnesid 24 ja näites 54 kaks sõnet, lisaks puudub viimasel lauselõpumärk. Näide 50 algab sõnaga *see* ning 57 sõnaga *selle*, mis on GDEX 1.4 järgi lause alguses potentsiaalse anafoorsuse tõttu keelatud, samuti on keelatud sidesõnaga algavad laused (näited 55 ja 56). Näited 51, 54, 56 ja 57 ei ole täislaused (lauselõpumärk puudub) ning GDEX 1.4 parameetrite järgi on need seetõttu halvad näitelause kandidaadid.

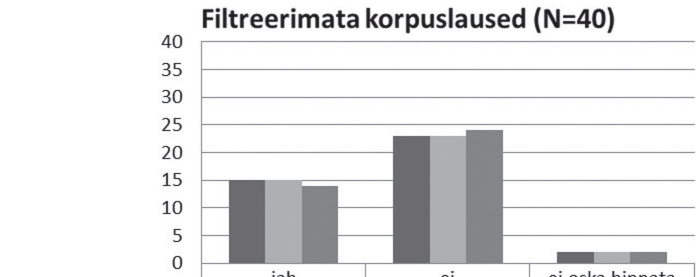
Jätkuküsitluses hindasid leksikograafid sobivaks tegelikult neist vaid kaks – näited 50 ja 52. Näites 48 peeti problemaatiliseks lause pikkust ja

sõnavara; näites 49 samuti pikkust, aga seda peeti ka arusaamatuks. Näites 51 peeti keeruliseks sõnavara, näites 57 pikkust ja anafoorsust (*selle asemel*).

Keeleõppijad pidasid jätkuküsitluse järel sobivaks vaid kolme lauset: näiteid 50–52. Näiteid 49 ja 53 peeti liiga pikaks, samas tunti näites 49 puudust kontekstist, nagu ka näites 57 (ilmselt anafoorse *selle asemel* tõttu). Näidet 54 ei peetud terviklikuks puuduva tegusõna tõttu. Näidet 55 peeti liiga pikaks ning konteksti puudumise tõttu (ilmselt anafoorse sõna *selles* tõttu) ka arusaamatuks. Näite 56 kohta toodi erinevaid põhjendusi, muuhulgas lause pikkust, konteksti puudumist, märksõna ebatavalist kasutust ning kõnekeelsust.

4.4. Filtreerimata korpuslaused

Filtreerimata korpuslauseid oli andmestikus kokku 40. Hüpotees oli, et korpuslausete filtreerimine on vajalik. Filtreerimisel on kaks eesmärki – abistada leksikograafi näitelausete valikul ning näitamiseks sobimatute lausete automaatne kõrvaldamine.



	jah	ei	ei oska hinnata
leksikograafid (N=5)	15	23	2
keeleõppijad (N=5)	15	23	2
kokku (N=10)	14	24	2

Joonis 7. *Leksikograafide ja keeleõppijate esimeses hindamisülesandes antud hinnang filtreerimata korpuslausete sobivuse kohta*

Jooniselt 7 nähtub, et neljakümnest filtreerimata korpuslausest ei sobi leksikograafide ja keeleõppijate meelest sõnastiku näitelauseks kokku 24 lauset ehk 60% lausetest. Näiteks kattus kõikide leksikograafide ja keeleõppijate arvamus, et näide 58 näitelauseks ei sobi, küll aga sobib näide 59.

(58) See kahjustab **tublisti** demokraatlikku otsustamisprotsessi, milles on niikuinii palju nõrkasid punkte.
(59) Eakatel inimestel pole võimalust alustada oma elu **uuesti**.

Hüpotees, et korpuslausete filtreerimine on vajalik, leidis kinnitust – üle poole filtreerimata korpuslausetest hindajate meelest näitelauseks ei sobinud.

4.5. Esimese hindamisülesande ja jätkuküsitluse tulemuste võrdlus

Allpool võrreldakse esimese hindamisülesande käigus saadud vastuste arvu jätkuküsitluse tulemustega. Jätkuküsitluse tulemused näitavad, et vastusevariantide etteandmine mõjutab hindajate arvamust, kuna hinnang lausetele muutus.

Esimene hindamisülesanne näitas, et keeleõppijad ja leksikograafid hindasid 95% "Eesti keele sõnaraamatu 2019" näitelausetest sobivaks. Headest korpuslausetest hinnati sobivaks 80% ja sama palju halbadest korpuslausetest hinnati sobimatuks. Filtreerimata korpuslausetest hinnati sobimatuks 60%.

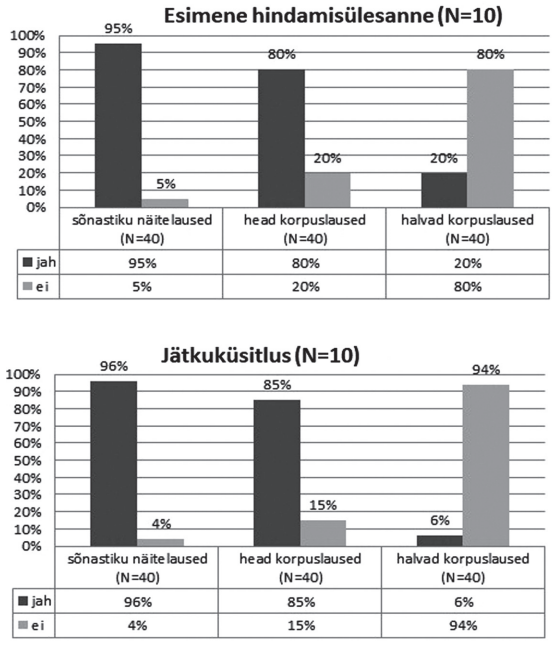
Jätkuküsitluse järel arvas enamik keeleõppijaid, et sõnaraamatu lause, näide 17, tegelikult sobib sõnastiku näitelauseks, kuid näide 16 mitte. Enamik leksikograafe pidas sobimatuks kahte sõnastiku näitelauset (näiteid 15 ja 17). Jätkuküsitluse järel selgus, et 40 sõnaraamatu näitelausest pidasid leksikograafid sobivaks 38 (95%) ja keeleõppijad 39 (97,5%) lauset, ning seega leidis kinnitust hüpotees, et leksikograafi koostatud laused on sobivad näitelaused.

(15) **Kangesti** palav ilm.
(16) Täna on **ropp** tuul!
(17) Blondiininaljad on küll puhas **rassism**.

Jätkuküsitluses muutus keeleõppijate hinnang heade korpuslausete suhtes, leksikograafide oma mitte. Enamik keeleõppijaid hindas sobivaks veel kolm head korpuslauset. Seega pidasid mõlemad hindajate grupid kokku sobivaks 34 lauset (85%), mis GDEX 1.4 järgi vastavad hea näitelause parameetritele. Hüpotees, et GDEX suudab tuvastada optimaalsed näitelause kandidaadid, leidis kinnitust.

Halbade korpuslausete suhtes oldi jätkuküsitluses kriitilisemad kui esimeses hindamisülesandes. Kokku leidsid leksikograafid, et tegelikult ei sobi neljakümnest lausest 38 (95%) ning keeleõppijate arvates 37 (92,5%) halbadest korpuslausetest. Seega leidis kinnitust ka hüpotees, et GDEX suudab välja filtreerida sobimatud (halvad) korpuslaused.

Esimese hindamisülesande ja jätkuküsitluse tulemused on kõrvutavalt välja toodud joonisel 8. Joonisel on mõlema hindajate tüübi (viie leksikograafi ja viie keeleõppija) arvamus kokku liidetud ning tulemus on esitatud absoluutarvude asemel protsentides.



Joonis 8. *Esimese hindamisülesande ja jätkuküsitluse tulemused*

Jätkuküsitluse raames palusin keeleõppijatel vastata ka paarile taustaküsimusele. Kuna taustaküsimustele vastamine oli vabatahtlik, sain vastuseid ainult kolmelt. Kõigi kolme keeleõppija keeleoskustase oli C1. Kõik kolm leidsid, et näitelaused on õppesõnastikes kasulikud ja vajalikud. Küsimusele, mille alusel nad näitelause sobivust hindasid, vastati, et eeskätt pikkuse ja sõnavara põhjal, aga ka selle põhjal, kas lause tundub loomulik ja on arusaadav (ka pikema kontekstita).

Jätkuküsitluses toodi lause sobimatuse põhjusteks kõige sagedamini just anafoorsust ja konteksti puudumist, lause pikkust ja kõnekeelsust. Lisaks toodi mitmel korral välja, et GDEXi väljundisse on sattunud laused, kus märksõna on tegelikult mitmesõnalise märksõna osa.

5. Kokkuvõte

Artiklis analüüsiti, kas autentseted ja toimetamata korpuslaused sobivad leksikograafide ja keeleõppijate hinnangul eesti keele B2–C1-keeleoskustaseme õppesõnastiku näitelauseks. Leksikograafide ja keeleõppijate hinnangu põhjal selgitati välja, kas korpuslausete filtreerimine on enne lõppkasutajale näitamist vajalik ja kas GDEXi eesti mooduli versioon 1.4 suudab korpusest tuvastada optimaalsed näitelausete kandidaadid ning välja filtreerida sobimatud.

Artiklis leiti kinnitus kolmele hüpoteesile: korpuslausete filtreerimine on vajalik, GDEX suudab korpusest tuvastada optimaalsed näitelause kandidaadid ja välja filtreerida sobimatud ning leksikograafi koostatud sõnaraamatu näitelaused on head näitelaused. Esimese hindamisülesande ja jätkuküsitluse tulemusi kokku liites selgus, et leksikograafid peavad kokku sobivaks 95% ja keeleõppijad 97,5% sõnastiku näitelausetest. Mõlemad hindajate grupid hindasid sobivaks 85% GDEX 1.4 järgi hea näitelause parameetritele vastavatest korpuslausetest ning sobimatuks 94% GDEX 1.4 järgi halva näitelause parameetritele vastavatest korpuslausetest. Lausete sobimatuse puhul toodi kõige sagadamini välja anafoorsust, konteksti puudumist, lause pikkust ja kõnekeelsust. Järelikult tuleb GDEXi eesti mooduli versiooni edasi arendades senisest

veelgi suuremat tähelepanu pöörata anafooride esinemisele lauses. Samuti tasub täiendavalt testida lause optimaalset pikkust – kuigi pikki lauseid peeti sageli liiga pikaks, kippus lühematel lausetel olema vähem konteksti (samale järeldusele jõudis ka Kuhn 2017: 265).

Jätkuküsitluse tulemused näitasid, et hindajate põhjendused lausete sobimatuse osas pigem erinesid kui ühtisid. See näitab, et nii palju kui on erinevaid hindajaid, on ka erinevaid hinnanguid sellele, mis teeb lausest hea näitelause.

Tulevikus saaks siinses artiklis kirjeldatud hindamismeetodit rakendada ka Eesti Keele Instituudi keeleportaalis Sõnaveeb kuvatavate veebilausete näitamisel. Sellisel juhul saaksid Sõnaveebi kasutajad nende jaoks sobimatuid veebilauseid maha hääletada ning kui lause on teatud arvu negatiivseid hinnanguid kogunud, õpiks masin selliseid lauseid kasutajale edaspidi enam mitte näitama.

Kirjandus

Cook, Paul, Michael Rundell, Jay Han Lau, Timothy Baldwin 2014. Applying a word-sense induction system to the automatic extraction of diverse dictionary examples. – Proceedings of the XVI EURALEX International Congress, 319–328.

Kallas, Jelena, Svetla Koeva, Iztok Kosem, Margit Langemets, Carole Tiberius 2019. Lexicographic Practices in Europe: A Survey of User Needs. ELEXIS – European Lexicographic Infrastructure. https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (8.4.2019).

Kallas, Jelena, Kristina Koppel, Maria Tuulik 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel ['New possibilities in corpus lexicography based on the example of the Estonian Collocations Dictionary']. – Eesti Rakenduslingvistika Ühingu aastaraamat 11, 75–94. https://dx.doi.org/10.5128/ERYa11.05

Kaufmann, Nicolas, Thimo Schulze, Daniel Veit 2011. More than fun and money. Worker motivation in crowdsourcing – A study on Mechanical Turk. – Proceedings of the 17th Americas Conference on Information Systems, 1–11.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý 2008. GDEX: Automatically finding good dictionary examples in a corpus. – E. Bernal, J. DeCesaris (Eds.). Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425–432.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smr, David Tugwell 2004. The Sketch Engine. – G. Williams, S. Vessier (Eds.). Proceedings of the 11th EURALEX International Congress. Lorient: Université de Bretagne Sud, 105–115.

Koppel, Kristina 2017. Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks [‘Automatic detection of good dictionary examples in Estonian learner’s dictionaries’]. – Eesti Rakenduslingvistika Ühingu aastaraamat 13, 53–71. https://dx.doi.org/10.5128/ERYa13.04

Koppel, Kristina 2019. Eesti keele kui teise keele õpikute lausete analüüs ja selle rakendamine eri keeleoskustasemete sõnastike näitelausete automaatsel valikul [‘Analysis of CEFR-graded coursebook sentences and their use for automatic detection of good dictionary examples’]. – Eesti Rakenduslingvistika Ühingu aastaraamat 15, 99–119. https://dx.doi.org/10.5128/ERYa15.06

Koppel, Kristina, Jelena Kallas 2016. Õppijasõbralik korpuslause: automaatse valiku võimalusi [‘User-friendly corpus sentence: Parameters for automatic selection’]. – Lähivõrdlusi. Lähivertailuja 26, 222–250. https://dx.doi.org/10.5128/LV26.07

Koppel, Kristina, Maria Khokhlova, Jelena Kallas, Vít Baisa, Vít Suchomel, Jan Michelfeit 2019a. SkELL corpora as a part of the language portal Sõnaveeb: Problems and perspectives. – Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, Jose Pedro Ferreria, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius (Eds.). Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of eLex 2019 conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 519–536.

Koppel, Kristina, Arvi Tavast, Margit Langemets, Jelena Kallas 2019b. Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. – Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, Jose Pedro Ferreria, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius (Eds.). Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of eLex 2019 conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 434–452.

Kosem, Iztok, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit, Carole Tiberius 2019. Identification and automatic extraction of good dictionary examples: The case(s) of GDEX. – International Journal of Lexicography 32 (2), 119–137. https://dx.doi.org/10.1093/ijl/ecy014

Kuhn, Tanara Zingano 2017. A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students. PhD thesis. Universidade de Lisboa.

Kuhn, Tanara Zingano, Peter Dekker, Branislava Šandrih, Rina Zviel-Girshin 2019. Crowdsourcing corpus cleaning for language learning – an approach proposal. – Posterettekanne. enetCollect 3th annual meeting, Lisbon, 14-15 March. https://dx.doi.org/10.13140/RG.2.2.31326.48964

Langemets, Margit, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll 2018. Eesti keel uues kuues: Eesti keele sõnaraamat 2018 ['Estonian lexis revisited. The Dictionary of Estonian 2018']. – Keel ja Kirjandus 12, 942–958.

Leimeister, Jan Marco, Michael Huber, Ulrich Bretschneider, Helmut Krcmar 2009. Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition. – Journal of Management Information Systems 26, 197–224.

Tavast, Arvi, Margit Langemets, Jelena Kallas, Kristina Koppel 2018. Unified Data Modelling for presenting lexical data: The case of EKILEX. – Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (Ed.). Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018. Ljubljana: Ljubljana University Press, Faculty of Arts, 749–761.

Vainik, Ene 2018. Compiling the Dictionary of Word Associations in Estonian: From scratch to the database. – Eesti Rakenduslingvistika Ühingu aastaraamat 14, 229–245. https://doi.org/10.5128/ERYa14.14

ÕS 2018 = Eesti õigekeelsussõnaraamat ÕS 2018 ['Dictionary of Standard Estonian 2018']. Maire Raadik, Tiiu Erelt, Tiina Leemets, Sirje Mäearu (Toim.). Tallinn: EKSA.

Võrguviited

Eesti keele assotsiatsioonisõnastik. Eesti Keele Instituut. http://www.eki.ee/dict/assotsiatsioonid/assotsiatsioonid.html (12.5.2019).

Eesti keele naabersõnad 2019 ['The Estonian Collocations Dictionary 2019']. Jelena Kallas, Kristina Koppel, Maria Tuulik, Geda Paulsen (Toim. & Koost.). Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee (12.5.2019).

Eesti keele sõnaraamat 2019 [‘The Dictionary of Estonian 2019’]. Margit Langemets, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll (Toim. & Koost.). Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee (12.5.2019).
Eesti keele õppekorpus 2018 (etSkELL) [‘Estonian Corpus for Learners 2018’]. https://doi.org/10.15155/3-00-0000-0000-0000-07335l
Eesti keele ühendkorpus 2017 [‘Estonian National Corpus 2017’]. https://doi.org/10.15155/3-00-0000-0000-0000-071e7l (12.5.2019).
Ekilex. Online dictionary system. https://ekilex.eki.ee (12.5.2019).
EMS = Eesti murrete sõnaraamat [‘The Dictionary of Estonian Dialects’]. Eesti Keele Instituut. http://www.eki.ee/dict/ems/ (15.5.2019).
etSkELL. etskell.sketchengine.co.uk (12.5.2019).
Google Translate. https://translate.google.com (12.5.2019).
Longman Dictionary of Contemporary English. http://ldoce.longmandictionaries-online.com/main/Home.html (12.5.2019).
Pybossa. https://pybossa.com (12.5.2019).
Sõnaveeb. Eesti Keele Instituut. https://sonaveeb.ee (12.5.2019).
Urban Dictionary. https://www.urbandictionary.com (12.5.2019).
Wordnik. https://www.wordnik.com/ (12.5.2019).

Kristina Koppel
Eesti Keele Instituut
Roosikrantsi 6, 10119 Tallinn, Estonia
kristina.koppel@eki.ee

Suitability of automatically selected example sentences for learners' dictionaries as tested on lexicographers and language learners

KRISTINA KOPPEL
Institute of the Estonian Language

This paper reports on an assessment task carried out among students of Tallinn University and the University of Tartu, who speak Estonian at B2-C1 proficiency level, and among lexicographers working at the Institute of the Estonian Language. The purpose of the task was to determine whether, according to the above two types of annotators, authentic and unedited corpus sentences would be suitable as example sentences for learners' dictionaries on B2-C1 level. The results of the assessment task were also to help evaluate the output of version 1.4 of the Estonian module of GDEX (GDEX 1.4) used to choose and display web sentences in the Institute's new language portal Sõnaveeb. GDEX (Good Dictionary Example) is a function of the corpus query system Sketch Engine, designed to find optimal example sentence candidates from large corpora.

The results of the assessment task confirmed three hypotheses: 1) Before displaying authentic corpus sentences to end-users, a filtering of corpus sentences is necessary; 2) GDEX 1.4 can identify good example candidates from corpora and filter out inapropriate candidates; 3) example sentences compiled by lexicographers are suitable example sentences. Both types of annotators considered as many as 96% of the dictionary examples to be suitable example sentences and 85% of corpus sentences chosen as good examples by GDEX 1.4. Only 6% of the sentences that were discarded by GDEX 1.4 were considered as suitable, meaning that 94% of the bad candidates had been filtered out successfully. As for unfiltered corpus sentences, 60% of those were considered unsuitable. When asking for the annotators' reasons for considering a sentence unsuitable, the most common arguments were that the sentences include anaphora and hence need more context, or that the sentences are colloquial, too long or too short.

Keywords: corpus lexicography; learners' lexicography; example sentences; GDEX; Estonian