doi:10.5128/ERYa15.06

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 15, 99–119

EESTI KEELE KUI TEISE KEELE ÕPIKUTE LAUSETE ANALÜÜS JA SELLE RAKENDAMINE ERI KEELEOSKUSTASEMETE SÕNASTIKE NÄITELAUSETE AUTOMAATSEL VALIKUL

Kristina Koppel

Ülevaade. Artikli eesmärk on välja töötada korpuspäringusüsteemi Sketch Engine heade näitelausete tööriista GDEX (*Good Dictionary Example*) eesti mooduli versioonid, mis aitavad korpusest tuvastada eri keeleoskustasemetele vastavaid eri leksikaalse, süntaktilise ja grammatilise keerukusega näitelause kandidaate. Selleks analüüsin eesti keele kui teise keele õpikute lauseid ning teen kindlaks, missugused parameetrid eri keeleoskustasemeid eristavad. Uute eesti mooduli versioonide aluseks on sõnastike näitelausete analüüsi põhjal loodud GDEX-i eesti mooduli versioon 1.4, mille parameetreid vastavalt õpikulausete analüüsi tulemustele kohandan. Uurimistöö tulemusi rakendades saab luua eri keeleoskustasemete õppekorpused, mis sobivad kasutamiseks sõnastikuportaalides (nt Sõnaveeb), keeleõpperakendustes (nt etSkELL) ja muu õppevara loomisel.

Võtmesõnad: korpuslingvistika, korpusleksikograafia, õppeleksikograafia, õppekorpus, eesti keel teise keelena, eesti keel

1. Sissejuhatus

Hea näitelause leidmine korpusest on ajamahukas töö. Tänapäeval on üha enam hakatud eri tüüpi sõnastike koostamisel ja keeleõpperakenduste loomisel kasutama näitelausete automaatset valikut.

Eestis on näitelausete automaatset valikut seni rakendatud "Eesti keele naabersõnade 2019" (eelnevates artiklites: eesti keele kollokatsioonisõnastik) andmebaasi genereerimisel (Kallas jt 2015), Eesti Keele Instituudi portaalis Sõnaveeb[1] ja keeleõppekeskkonnas etSkELL[2] (*Sketch Engine for Estonian Language Learning*). Kõigis neis kasutatakse näitelausete automaatseks tuvastamiseks korpuspäringusüsteemi

[1] https://sonaveeb.ee (1.10.2018).
[2] https://etskell.sketchengine.co.uk/run.cgi (1.10.2018).

Sketch Engine (Kilgarriff jt 2004) tööriista GDEX (*Good Dictionary Example*) (Kilgarriff jt 2008) eesti mooduli versiooni 1.4 (Koppel 2017).

GDEX aitab korpusest automaatselt tuvastada häid näitelauseid. See töötab reeglipõhisel valemil, mis lause komponente hinnates määrab igale korpuslausele skoori ning reastab need skoori alusel paremuse järjekorda. Reeglipõhine valem koos täiendavate parameetritega moodustab GDEX-i konfiguratsiooni. GDEX-i konfiguratsioon sisaldab tervet rida klassifikaatoreid, millest igaüks sisaldab omakorda parameetrit, millele lause peab vastama, ning karistust (ingl *penalty*) või lisapunkte (ingl *bonus*) sellele parameetrile (mitte)vastamise eest.

Klassifikaatorid jagunevad kaheks: tugevateks ja nõrkadeks. Tugevad klassifikaatorid töötavad justkui filtrina, praakides välja tõeliselt ebasobivad laused. Nõrgad klassifikaatorid reastavad sobilikud kandidaadid paremuse järjekorda – need kas vähendavad lause üldskoori ehk karistavad lauset, kui see mingile parameetrile ei vasta, või annavad lausele lisapunkte.

Seni viimase GDEX-i eesti mooduli versiooni 1.4 (GDEX 1.4) töötasin välja naabersõnade sõnastiku (Kallas jt 2015), "Eesti keele sõnaraamatu 2019" (EKS) (Langemets jt 2018) ja "Eesti keele põhisõnavara sõnastiku" (PSV) (2014) näitelausete analüüsile toetudes (Koppel 2017). Kuigi GDEX 1.4 sihtgrupp on B2–C1-tasemel keeleõppija, ei olnud mul selle arendamisel võimalik sobiva andmestiku puudumise tõttu arvestada keeleoskustasemele spetsiifilisi lauseparameetreid. Keeleoskustasemete eristamisel toetun Euroopa keeleõppe raamdokumendile (Raamdokument 2007), mis eristab kolme üldist keeleoskustaset (A ehk algtasemel, B ehk iseseisev ja C ehk vilunud keelekasutaja) ja kuut alltaset (A1 ehk läbimurre, A2 ehk esmane keeleoskus, B1 ehk suhtluslävi, B2 ehk edasijõudnu tase, C1 ehk vaba suhtluse pädevus, C2 ehk haritud emakeelekõneleja tase).

Eri keeleoskustasemetele kohandatud GDEX-i eesti mooduli versioonide loomiseks võtsin aluseks eri tasemetele mõeldud eesti keele kui teise keele õpikute tekstid. Eeldasin, et koostajad on õpikutesse valinud tasemekohased tekstid. Minu eesmärk oli õpikute põhjal välja selgitada need tunnused, mis iseloomustavad eri keeleoskustasemete lauseid ning neile tunnustele toetudes välja arendada GDEX-i eesti mooduli versioonid algtasemel olevale (A-tase), iseseisvale (B-tase) ja vilunud keelekasutajale (C-tase).

2. "Eesti keele A1–C1 õpikute korpus (2018)"

2018. aastal lõime koostöös tarkvarafirmaga Lexical Computing Ltd. "Eesti keele A1–C1 õpikute korpuse (2018)"[3], mis sisaldab eesti keele kui teise keele õpikutest pärit täislauseid. Kokku on korpuses 16 600 lauset, 121 000 sõna ja 151 000 sõnet.[4] "Eesti keele A1–C1 õpikute korpuse (2018)" loomiseks digiteerisime esmalt kaheksa eesti keele kui teise keele õpikut. Õpikute valikul pidasime oluliseks, et need on ükskeelsed, sisaldavad palju tekstilist materjali, ning on välja antud viimase 15 aasta jooksul. Samuti oli oluline, et keeleoskustase oleks selgelt määratletud kas alg-, kesk- või kõrgtasemeks või A1-, A2-, B1-, B2- või C1-tasemeks. C2-tasemele

[3] https://doi.org/10.15155/3-00-0000-0000-0000-071E9L (1.10.2018). Korpus on kättesaadav korpuspäringusüsteemis KORP (korp.keeleressursid.ee/). Korpuse aluseks on "Eesti keele A1–C1 õpikute sisu korpus (2017)", mis sisaldab kõiki õpiku osi – lugemistekste, lünkharjutusi, grammatikaosi, sõnavaraloendeid jms. "Eesti keele A1–C1 õpikute sisu korpus (2017)" valmis koostöös Eesti Keeleressursside Keskusega.

[4] Töörühma liikmed olid Eesti Keele Instituudi arvutileksikograaf Jelena Kallas, vanemtarkvaraarendaja Katrin Tsepelina ja siinkirjutaja.

kui haritud emakeelekõneleja tasemele õppekirjandust ei leidunud, mistõttu see jääb siinses artiklis analüüsist kõrvale.

Nende õpikute puhul, mis olid mõeldud rohkem kui ühele tasemele, aga kus peatükkide raskusaste ei olnud õpikute autorite poolt selgelt eristatud, palusin kolme eksperthindaja abi. Eksperthindajad olid tegevad eesti keele kui teise keele õpetajad,[5] kes määrasid igale õpiku peatükile ühe konkreetse keeleoskustaseme. Palusin neil peatükkide hindamisel keskenduda täistekstidele, kuna need sisaldavad terviklikke lauseid, mis on korpuse loomise seisukohalt õpiku kõige olulisemad üksused. Ma ei andnud ekspertidele ette kindlaid kriteeriume, mille põhjal teksti raskusastet määrata, kuid nad toetusid eelkõige tekstides kasutatud sõnavarale ja lausete grammatilisele keerukusele. Kuna eksperdid määrasid keeleoskustaseme tervele peatükile korraga, mitte igale lausele eraldi, siis said korpuses kõik laused ühe peatüki sees märgitud ühe keeleoskustasemega, nt A2.

Kontrollisin kõik terviklaused enne korpuse lõplikku loomist käsitsi üle. Kustutasin laused, mis ei olnud semantiliselt terviklikud, vajasid konteksti (näited 1–2), olid teksti digiteerimisel valesti tuvastatud (näide 3) või ei sobinud muus mõttes iseseisvana sõnastiku näitelauseks (näide 4).

(1) Millest te seda järeldate?
(2) Aga miks?
(3) Firma võtab tööle juurd soskustega naiste kere vaste rätsepa.
(4) KÜLAS KÄIMA kellel?

Käsitsi puhastatud lausetest genereerisin omakorda viis andmebaasi (A1, A2, B1, B2, C1). Andmebaaside suurused on välja toodud tabelis 1.

Tabel 1. Õpikulausete andmebaaside suurused

Keeleoskustase	Lauseid	Sõnesid
A1	1363	6879
A2	3342	19215
B1	5462	39516
B2	5453	47451
C1	977	9569

Nagu tabelist 1 selgub, sattus korpusesse kõige rohkem B1- ja B2-taseme õpikulauseid. Ilmselt on põhjus selles, et kuna B1- ja B2-tase on kvalitatiivselt väga erinevad (Kitsnik 2018), on nendele alltasemetele välja antud ka rohkem õppematerjali.

3. Õpikulausete analüüs

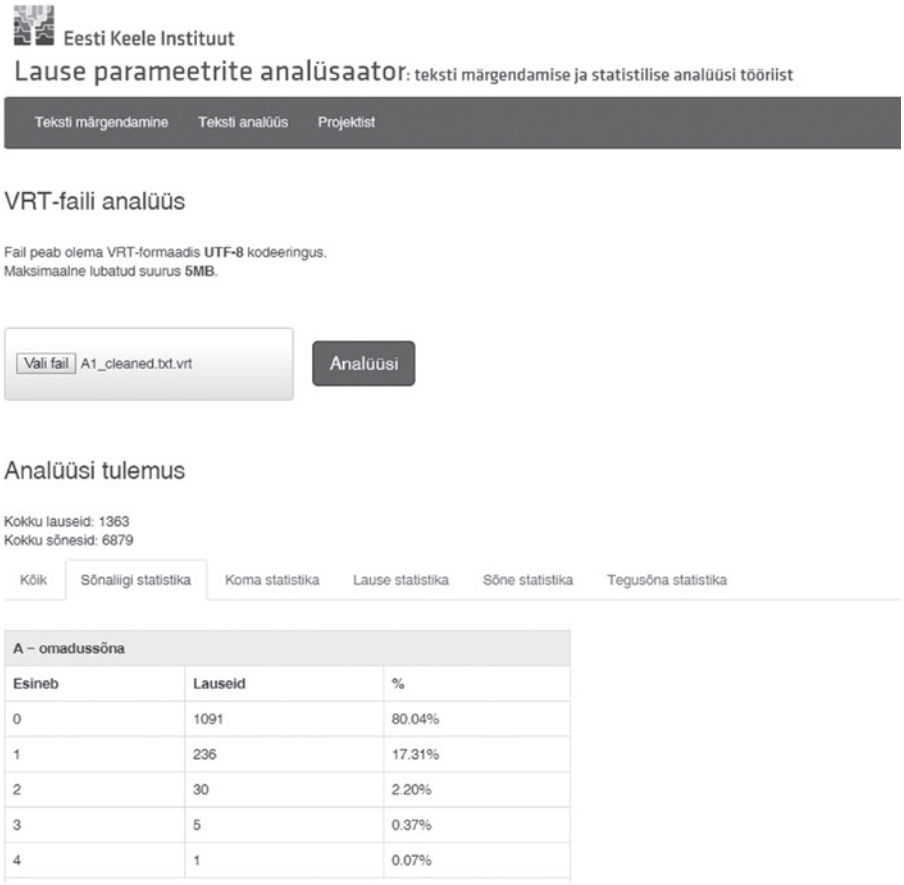
Õpikulausete analüüsimiseks kasutasin programmi "Lause parameetrite analüsaator"[6], mille abil saab teksti morfoloogiliselt märgendada ja analüüsida. Programm kasutab teksti märgendamiseks morfoloogilist analüsaatorit Estnltk[7].

Programmi funktsioonide abil saab statistiliselt analüüsida järgmisi lause tunnuseid:

[5] Tänan Ülle Koppelit, Monika Sooalu ja Marika Kransiveid.
[6] www.eki.ee/keeletase/statistics (1.10.2018). Analüsaatori programmeeris Eesti Keele Instituudi vanemtarkvaraarendaja Katrin Tsepelina.
[7] https://estnltk.github.io/estnltk/1.2/index.html (1.10.2018).

1) sõnaliik – mitu korda esineb lauses teatud sõnaliik (nt asesõna);
2) koma – mitu korda esineb lauses koma;
3) lause – missuguse sõnaliigiga lause algab ja kui pikk on lause sõnedes;
4) sõne – kui pikad on lauses esinevad sõned;
5) tegusõna – täpsem info tegusõna vormide kohta.

Joonisel 1 on näha programmi kasutajaliides koos analüüsi tulemustega.



A – omadussõna		
Esineb	**Lauseid**	**%**
0	1091	80.04%
1	236	17.31%
2	30	2.20%
3	5	0.37%
4	1	0.07%

Joonis 1. Sõnaliigi statistika funktsiooni tulemused programmis "Lause parameetrite analüsaator"

Parameetrite valik toetub GDEX 1.4 arendamiseks tehtud analüüsile, kus leidsin, et otstarbekas on mõõta selliseid parameetreid nagu lause ja sõnede pikkus, lause sõnaliigiline koosseis jmt (Koppel 2017). Õpikulausete puhul aitavad need välja selgitada eri keeleoskustasemete lausete tunnuseid.

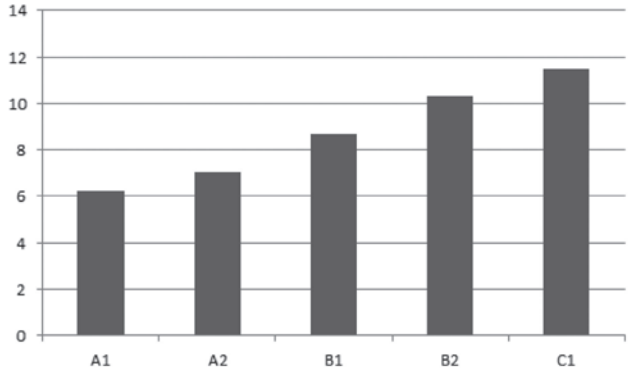
Järgnevalt analüüsin õpikulausete parameetreid alltasemete (A1, A2, B1, B2, C1) kaupa, kuid GDEX-i eesti mooduli versioonid loon üldistele keeleoskustasemetele:

- etBasic-v1 algtasemel keelekasutajale ehk A-tasemele,
- etIndependent-v1 iseseisvale keelekasutajale ehk B-tasemele,
- etProficient-v1 vilunud keelekasutajale ehk C-tasemele.[8]

[8] v1 nimedes tähistab versiooni numbrit ja et eesti keelt (*Estonian*).

3.1. Lause ja sõne pikkus

Kuna keeleõppijaile ei taheta kuvada väga pikki korpuslauseid, määrasin juba GDEX-i eesti mooduli esimeses versioonis lause pikkuseks 4–20 sõnet (Kallas jt 2015). PSV ja EKS-i näitelause on keskmiselt 5–6 sõna[9] pikk (Kallas jt 2015) ning naabersõnade sõnastiku näitelause on keskmiselt 9,8 sõnet pikk (Koppel 2017). Õpikulausete analüüs näitas, et A1-tasemel on lause keskmine pikkus 6,2, A2-tasemel 7, B1-tasemel 8,7, B2-tasemel 10,3 ning C1-tasemel 11,5 sõnet pikk (joonis 2).



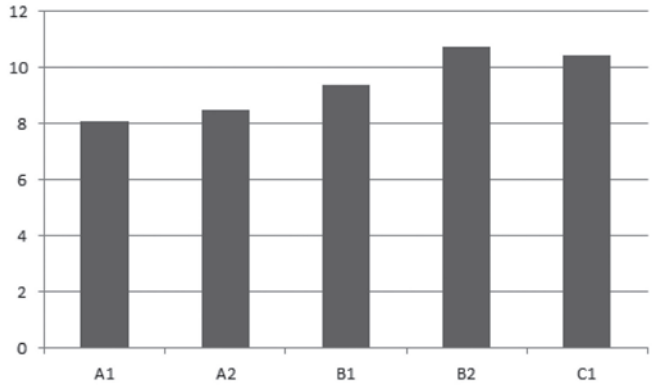
Joonis 2. Õpikulausete keskmine pikkus (sõnedes)

Kõikidel keeleoskustasemetel esines ka lauseid, mis olid pikemad kui 20 sõnet, kuid need olid enamasti üksikud näited. A1-tasemel oli selliseid lauseid kokku 7 (0,5%), A2-tasemel 39 (1%), B1-tasemel 213 (4%), B2-tasemel 328 (6%) ning C1-tasemel 104 (11%). Versioonis etBasic-v1 määrasin lausete pikkuseks 3–14 sõnet, kuna õpikulausetes oli sellest pikemaid lauseid väga vähe: A1-tasemel oli pikemaid kui 13-sõnelisi lauseid vähem kui 1% ning A2-tasemel pikemaid kui 14-sõnelisi lauseid vähem kui 1%. Versioonis etIndependent-v1 määrasin lause pikkuseks 3–18 sõnet, kuna B1- ja B2-tasemel esines pikemaid kui 18 sõnest koosnevaid lauseid vähem kui 1%. Versioonis etProficient-v1 määrasin lause pikkuseks 4–23 sõnet, kuna C1-tasemel oli 23 sõnest pikemaid lauseid vähem kui 1%.

Järgmisena määrasin lause optimaalse vahemiku, arvestades õpikulausete keskmist pikkust. A1- ja A2-taseme laused olid enamasti 4–7-sõnelised, B1-tasemel 4–8-sõnelised, B2-tasemel 4–12-sõnelised ning C1-tasemel 4–14-sõnelised. Versioonis etBasic-v1 määrasin lause optimaalseks vahemikuks 4–7 sõnet, versioonis etIndependent-v1 4–12 sõnet ja versioonis etProficient-v1 6–14 sõnet. Optimaalne vahemik arvestab seda, milline on lause optimaalne pikkus.

Õpikulausetes esinenud sõnede keskmine pikkus on A1-tasemel 8, A2-tasemel 8,5, B1-tasemel 9,4, B2-tasemel 10,7 ning C1-tasemel 10,4 tähemärki (joonis 3). A1-tasemel esines ainult üks sõne (0,01%), mis oli pikem kui 20 tähemärki; A2-tasemel oli neid sõnesid 6 (0,04%); B1-tasemel 24 (0,03%); B2-tasemel 51 (0,09%) ja C1-tasemel 9 (0,09%).

[9] Artiklis Kallas jt 2015 arvestasime sõnastikulausetes esinenud sõnu, mitte sõnesid, mida GDEX lause pikkuse mõõtmisel arvestab.

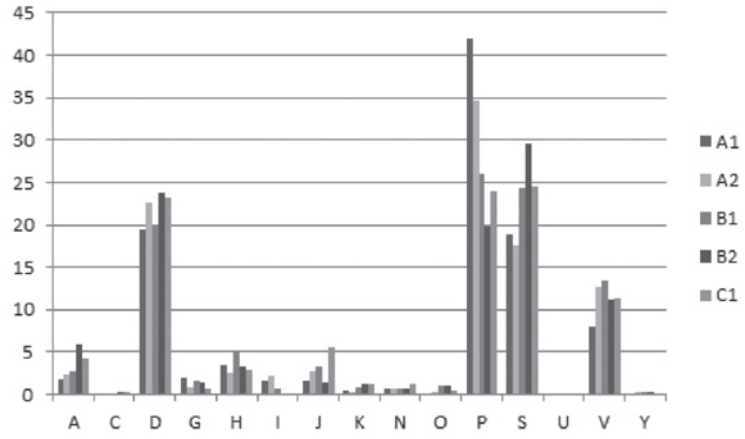


Joonis 3. Õpikulausetes esinevate sõnede keskmine pikkus (tähemärkides)

Inglise, sloveeni, hollandi ja portugali keele GDEX-i moodulid sisaldavad klassifikaatorit, mis karistab sõnesid, mis on pikemad kui teatud arv tähemärke. Inglise keeles on piiriks märgitud 6 tähemärki, sloveeni ja portugali keeles 12 ja hollandi keeles 15 tähemärki. (Kosem jt 2018) GDEX 1.4 sellist klassifikaatorit ei sisaldanud, küll aga on keelatud sõned, mis on pikemad kui 20 tähemärki. Eelpool mainitud keelte eeskujul ja õpikulausetes kasutatud sõnede keskmisi pikkusi arvestades lisasin täiendava klassifikaatori, mis versioonis etBasic-v1 karistab sõnu, mis on pikemad kui 9 tähemärki ja versioonis etIndependent-v1 sõnu, mis on pikemad kui 11 tähemärki.

3.2. Lause esimese sõna sõnaliik

Naabersõnade sõnastiku näitelausete analüüsist selgus, et üle poole lausetest algab nimisõnaga (Koppel 2017). Õpikulausete analüüs tõi aga välja, et lause esimese sõna sõnaliik on tugevas korrelatsioonis sihtgrupi keeleoskustasemega (joonis 4).



Joonis 4. Õpikulausete esimese sõna sõnaliik[10] protsentides

[10] Toetun analüüsis Estnltk märgendusele: A – omadussõna (algvõrre), C – omadussõna (keskvõrre), D – määrsõna, G – käändumatu omadussõna, H – pärisnimi, I – hüüdsõna, J – sidesõna, K – kaassõna, N – põhiarvsõna, O – järgarvsõna, P – asesõna, S – nimisõna, U – omadussõna (ülivõrre), V – tegusõna, X – tegusõna juurde kuuluv sõna (nt *plehku*), Y – lühend.

Õpikulaused algavad A1–B1-tasemel kõige sagedamini asesõnaga. Põhjus on ilmselt selles, et A-tasemel peab õppija oskama võõrkeeles rääkida peamiselt iseendast, enda perest, tuttavatest inimestest ja asjadest (Ilves 2008), B-tasemel peab ta oskama põhjendada oma seisukohti ja plaane (Hausenberg jt 2008), ja õpik annab ette selleks vastava struktuuriga laused (näited 5–7).

(5) **Talle** meeldib süüa teha. (A1)
(6) **Neil** on niisugune traditsioon, et jaanipäeval tulevad kõik sugulased kokku. (A2)
(7) **Ma** pean enne tundi veel natuke õppima. (B1)

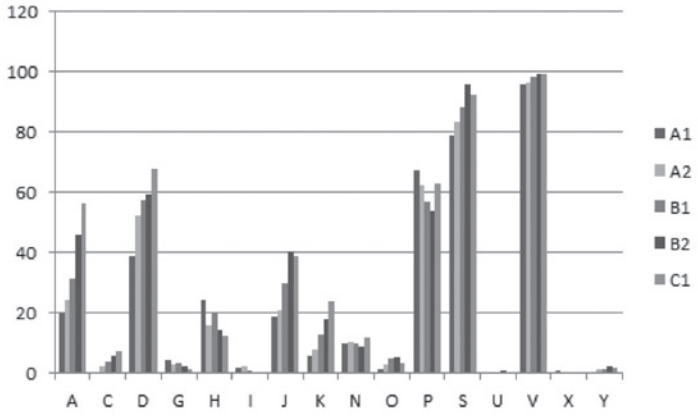
GDEX 1.4 keelab sõnastikulause algusest asesõnad *nemad*, *nad* ja *see*, kuna need viitavad enamasti tagasi eelnevale lausele. Lisaks on sõnastikulause algusest keelatud lühendid, hüüdsõnad, käändumatud omadussõnad ja sidesõnad. (Koppel 2017) Jooniselt 4 selgub, et õpikulaused ei alga üldiselt omadussõna võrdlusastmetega, lühendiga, põhi- ja järgarvsõnaga ega kaassõnaga. Sageli algavad laused hüüdsõna ja sidesõnaga, eriti A1- ja A2-tasemel. Põhjus on õpikulausete spetsiifika, kus dialoogi kõnevoor algab sageli sidesõnaga (näited 8–9).

(8) **Aga** mis me talle kingime? (A1)
(9) **Ja** kuidas teile see töö meeldib? (A2)

GDEX 1.4 ei luba ka sidesõnaga algavaid lauseid (Koppel 2017). Sellised sidesõnaga algavad õpikulaused, nagu näited (8) ja (9), võiks iseseisvana sõnastiku näitelausena esineda küll, kuid kuna ma kustutasin enne õpikute korpuse loomist konteksti vajavad sidesõnaga algavad laused, jätsin ka versioonides etBasic-v1, etIndependent-v1 ja etProficient-v1 alles lause algusest sidesõna keelava klassifikaatori.

3.3. Lause sõnaliigiline koosseis

Naabersõnade sõnastiku näitelausete analüüs näitas, et leksikograafi poolt sõnastiku näitelauseks valitud korpuslauses esines keskmiselt 2 tegusõna, 1 asesõna, 1 määrsõna, 1 pärisnimi, 1 arvsõna ja 1 sidesõna. Analüüsile toetudes karistab GDEX 1.4 lauseid, kus neid sõnaliike esineb rohkem kui eelpool mainitud. (Koppel 2017) Joonisel 5 on esitatud õpikulausete sõnaliigiline koosseis.

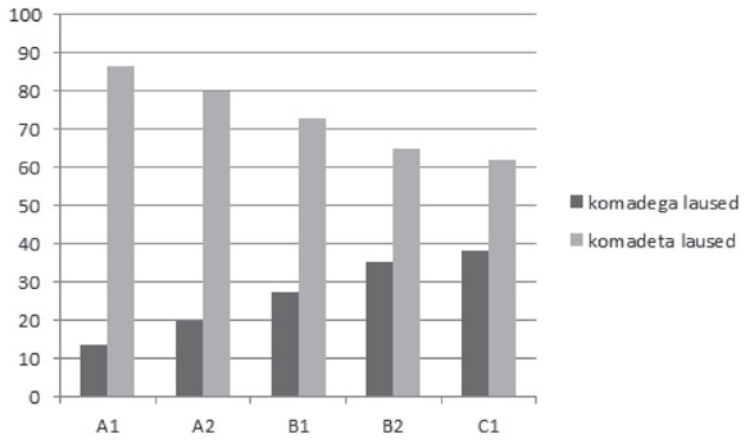


Joonis 5. Sõnaliigi esinemine õpikulausetes (protsentides)

GDEX 1.4 määrab, et lause peab sisaldama tegusõna. Ka õpikulausete analüüs näitab, et pea kõikides lausetes esineb tegusõna (96%-st A1-tasemel kuni 99,6%-ni C1-tasemel). Tegusõna esineb õpikulauses A-tasemel keskmiselt 1,4 korda, B-tasemel 2,1 korda ja C-tasemel 2,3 korda. Peaaegu kõik laused sisaldavad ka nimisõna (79%-st A1-tasemel 96%-ni B2-tasemel, C1-tasemel protsent langeb pisut). Nimisõna esineb õpikulausetes A-tasemel keskmiselt 1,5 korda, B-tasemel 2,8 korda ja C-tasemel 2,9 korda. Ka asesõnu esineb lausetes sageli: A1-, A2- ja C1-tasemel üle 60% lausetest. Joonis 5 näitab veel, et õpikulausetes ei esine eriti hüüdsõnu, lühendeid, tegusõna juurde kuuluvaid sõnu (nt *plehku*) ja omadussõna ülivõrde vorme. Kui GDEX 1.4 määrab, et lause peab kindlasti sisaldama tegusõna, siis versioonidesse etBasic-v1 ja etIndependent-v1 lisasin parameetri, et see peab kindlasti sisaldama ka nimisõna. Versioonis etProficient-v1 ma seda parameetrit ei rakendanud, kuna C-tasemel peaks keeleõppija aru saama ka ainult abstraktsemaid sõnu (nt pronoomenid, adverbid) sisaldavatest lausetest. Samuti lisasin versiooni etBasic-v1 klassifikaatori, mis karistab lauseid, kus esinevad lühendid, tegusõna juurde kuuluvad sõnad ja omadussõna ülivõrde vormid; versiooni etIndependent-v1 lisasin klassifikaatori, mis karistab lauseid, kus esinevad lühendid ja tegusõna juurde kuuluvad sõnad.

3.4. Komade arv lauses

Kallase jt (2015) uurimus näitas, et lihtlausest piisab sõna leksikaalgrammatiliste käitumismallide näitamiseks. Seetõttu on GDEX-i eesti mooduli üks parameetreid olnud komade arv lauses, vältimaks rohkete osalausetega lausete esinemist. Joonisel 6 on näha, mitu koma esineb keskmiselt õpikulausetes.



Joonis 6. Komade esinemine õpikulausetes (protsentides)

Joonis 6 illustreerib, et ka õpikulaused on enamasti komadeta laused. A1-tasemel on komadega lauseid ligikaudu 13% (keskmine koma esinemise arv lauses 0,2), A2-tasemel 20% (keskmine koma esinemise arv lauses 0,3), B1-tasemel 27% (keskmine koma esinemise arv lauses 0,4), B2-tasemel 35% (keskmine koma esinemise arv lauses 0,5) ja C1-tasemel 38% (keskmine koma esinemise arv lauses 0,7).

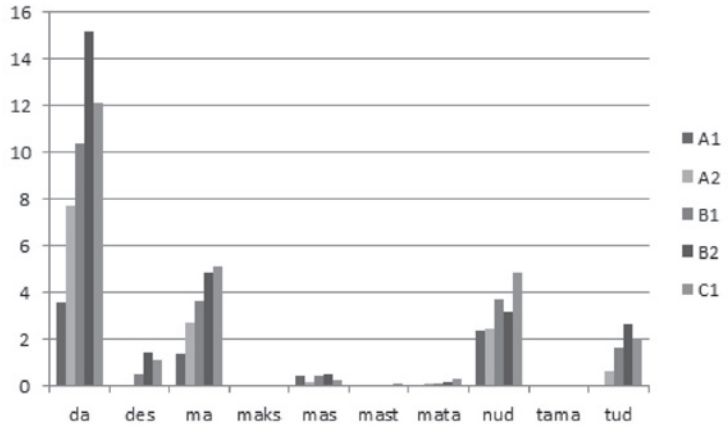
Samas võib lausetes, kus koma esineb, neid olla väga palju (näited 10–14). Põhjus võib olla selles, et õpikute autorid soovivad esitada korraga rohkesti teemakohast informatsiooni (näited 10–11), grammatikateemaga seotud tegusõna vorme (näide 12), teemaga seotud sõnavara (näide 13) või soovivadki esitada autentset, mitte õppeotstarbel koostatud lauset (näide 14).

(10) Tallinna linnaosad on Kesklinn, Mustamäe, Lasnamäe, Õismäe, Kalamaja, Kopli, Kadriorg, Nõmme, Pirita, Kristiine. (A1)
(11) Kihnu, Ruhnu, Muhu, Saaremaa ja Hiiumaa paistavad silma oma kultuuri, keele ja traditsioonidega. (A2)
(12) Edasi liiguti pataljoni territooriumil ringi, kuulati ajateenijate rivilaulu, vaadati arestimaja, köögiviljapeenraid, tutvuti relvade ja vormiriietega, kuulati loengut. (B1)
(13) Abituriendid soovivad kõige enam õppida majandust, õigusteadust, infotehnoloogiat, reklaamindust, ajakirjandust ja turismindust. (B2)
(14) Elu lähebki edasi just selles suunas, nagu iga riigi elu areneb, ja mina usun, et meie inimesed, kes selleks on kutsutud ja seatud, ka kõik selleks teevad, et ta paremaks läheb. (C1)

Õpikulausete analüüsi järel jäi ka versioonides etBasic-v1, etIndependent-v1 ja etProficient-v1 kehtima klassifikaator, mis karistab lauseid, kus esineb rohkem kui üks koma.

3.5. Tegusõna vormid

Kuna GDEX 1.4 sisaldab klassifikaatorit, mis karistab lauseid, kus esinevad *mata-*, *mast-*, *mas-*, *maks-* ja *des*-vormid, otsustasin õpikulausetele toetudes välja selgitada, mis keeleoskustasemel neid ja teisi tegusõna vorme kasutatakse. Lisaks *ma*-tegevusnime käändelistele vormidele ja *des*-vormile tahtsin välja selgitada, millal ilmuvad õpikulausetesse *nud-* ja *tud*-vormid (joonis 7), erinevad kõneviisid (joonised 8–11) ning umbisikuline tegumood (joonis 12).



Joonis 7. *ma-* ja *da*-tegevusnime, *maks-*, *mas-*, *mast-*, *mata-*, *tama-*, *des-*, *nud-*, *tud*-vormide esinemine õpikulausetes (protsentides)

Mare Kitsnik (2018) on uurinud tegusõna vormide kasutust B1- ja B2-tasemel, toetudes keeleõppijate endi loodud kirjalikele tekstidele. Tema sõnul on sagedasemad tegusõna vormid mõlemal keeleoskustasemel peamiselt isikuline tegumood kindla kõneviisi olevikus ja lihtminevikus, tingiva kõneviisi olevik ja tegevusnimed. Kitsniku (2014: 186) uurimuse tulemused õppijakeele kohta ütlevad veel, et võrreldes B1-tasemega suureneb B2-tasemel tegevusnime vormide kasutamise sagedus. Õpikulausetes samasugust muutust esile ei tule – nii *ma*- kui ka *da*-tegevusnime kasutus suureneb keeleoskustasemeti, v.a *da*-tegevusnime kasutus C1-tasemel, kus see hoopis 3% väheneb. Seejuures esineb *da*-tegevusnime kõikide keeleoskustasemete lausetes sagedamini kui *ma*-tegevusnime. Õpikulaused kajastavad üldiselt keeles olemasolevat tendentsi, kuna *da*-tegevusnime esinebki eesti keeles *ma*-tegevusnimest sagedamini ja sellel on ka rohkem süntaktilisi funktsioone (Penjam 2008).

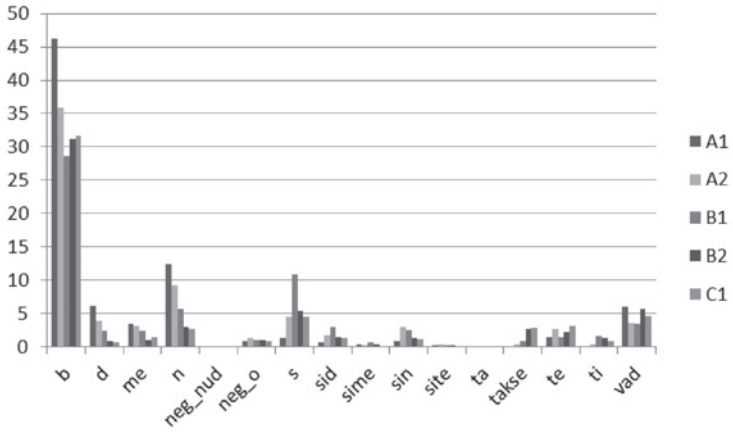
Õpikulausete analüüs näitab, et *ma*-tegevusnime käändelistest vormidest esineb ainult *mas*-vorm (*lugemas*) kõigil keeleoskustasemetel (A1-tasemel 0,5%, A2-tasemel 0,2%, B1-tasemel 0,4%, B2-tasemel 0,5%, C1-tasemel 0,3%), *mata*-vorm (*lugemata*) ilmub õpikulausetesse A2-tasemel, moodustades kõikidest tegusõna vormidest vaid 0,08%. *mast*-vorm (*lugemast*) ilmub õpikulausetesse B1-tasemel, *maks*- (*lugemaks*) ja *tama*-vorm (*loetama*) C1-tasemel, kuid ka nende vormide kasutus ei ole sage – kumbki vormidest esines 0,05%. *ma*-tegevusnime käändelised vormid ja *tama*-vorm moodustavad kõikidest õpikulausetes esinevatest tegusõna vormidest vähem kui 1%.

des-vorm (*lugedes*) ilmub õpikulausetesse B1-tasemel (0,5%), kuid rohkem kasutatakse seda B2- (1,4%) ja C1-taseme lausetes (1,1%).

nud-vorm (*lugenud*) esineb kõikide tasemete õpikulausetes (2,4%-st A1-tasemel kuni 4,8%-ni C1-tasemel), *tud*-vorm (*loetud*) ilmub A2-tasemel, kuid selle kasutus jääb alla 1%. *tud*-vormi kasutussagedus on arvestatav alates B1-tasemest (1,7%, B2-tasemel 2,7%, C1-tasemel 2%).

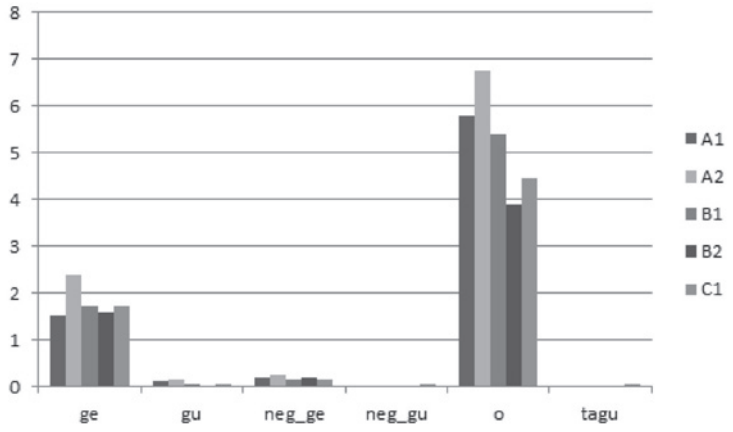
Õpikulausete analüüsile toetudes keelasin versioonis etBasic-v1 lausetest *mast*-, *maks*-, *mata*-, *tama*- ja *des*-vormid ja karistasin *tud*-vormi; versioonis etIndependent-v1 keelasin *maks*- ja *tama*-vormid ning karistasin *mast*- ja *mata*-vorme. Versioonis etProficient-v1 otsustasin *ma*-tegevusnime käändelisi vorme ja *des*-vormi mitte karistada ega keelata.

Kitsniku (2014: 186–187) sõnul esineb B1-tasemel õppijakeeles rohkelt kindla kõneviisi lihtminevikuvorme ja B2-tasemel olevikuvorme, kuid arvab selle põhjuse olevat tekstitüübis, millest tema uurimismaterjal pärines. Keeleoskuse kasvades kindla kõneviisi vormid vähenevad. Õpikulausete analüüs näitab (joonis 8), et kindla kõneviisi 1. ja 3. isiku ainsuse vormide tulemused on õppijakeelega sarnased: 1. isiku oleviku vormide kasutus väheneb õpikulausetes 12,4%-lt A1-tasemel 2,7%-ni C1-tasemel. 3. isiku oleviku vormide kasutus väheneb õpikulausetes 6,2%-lt A1-tasemel 0,7%-ni C1-tasemel. Õpikulausetes esineb kõige sagedamini 2. isiku oleviku vormi (A1-tasemel 46,2%), selle kasutus väheneb kuni B1-tasemeni, kuid tõuseb siis taas 2–3% võrra. Kindla kõneviisi 1., 2. ja 3. isiku mitmuse oleviku vormide kasutuses sellist vähenemist näha ei ole: 2. isiku vormi kasutus keeletaseme arenedes just pigem kasvab (1,4%-lt A1-tasemel 3,1%-ni C1-tasemel). Kindla kõneviisi minevikuvormidest kasutatakse õpikutes rohkem 1., 2. ja 3. pöörde ainsuse vorme.



Joonis 8. Kindla kõneviisi lõputunnused[11] õpikulausetes (protsentides)

Käskiv kõneviis esineb õpikulausetes kõikidel keeleoskustasemetel (joonis 9), kõige sagedamini esineb see oleviku 2. isiku vormis (*loe*). Ainsad vormid, mis ilmuvad õpikulausetesse alles C1-tasemel, on *neg_gu* (*ärgu lugegu*), mis tähistab nii oleviku 3. isiku mitmuse ja ainsuse aktiivi eitavat kõnet kui ka oleviku passiivi eitavat kõnet; ja *tagu*-lõpuga oleviku passiivi jaatav kõne (*loetagu*). Kuigi ka C1-tasemel on *neg_gu*- ja *tagu*-vormide kasutus väga madal, keelasin need vormid vaid versioonidest etBasic-v1 ja etIndependent-v1.



Joonis 9. Käskiva kõneviisi lõputunnused[12] õpikulausetes (protsentides)

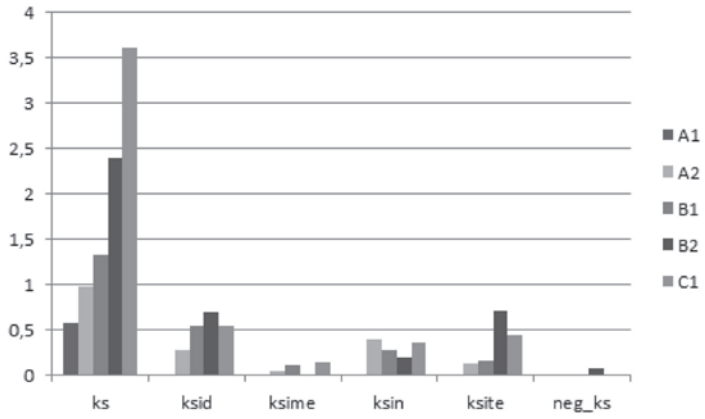
Kitsnik (2018: 69–71) on uurinud ka tingiva kõneviisi kasutust B1- ja B2-taseme õppijate kirjalikes tekstides. Tema uurimusest selgus, et nii B1- kui ka B2-tasemel esineb tingiv kõneviis kõige sagedamini ainsuse 1. ja 3. pöörde vormis, B2-tasemel

[11] Toetun analüüsis Estnltk märgendusele: *b* – olevik 3. isik ainsus (*loeb*); *d* – olevik 2. isik ainsus (*loed*); *me* – olevik 1. isik mitmus (*loeme*); *n* –olevik 1. isik ainsus (*loen*); *neg_nud* – lihtminevik 1., 2. ja 3. isik ainsus/mitmus eitav kõne (*polnud*); *neg_o* – olevik 1., 2. ja 3. isik ainsus/mitmus eitav kõne (*pole*); *s* – lihtminevik 3. isik ainsus (*luges*); *sid* – lihtminevik 2. isik ainsus (*lugesid*); *sime* – lihtminevik 1. isik mitmus (*lugesime*); *sin* – lihtminevik 1. isik ainsus (*lugesin*); *site* – lihtminevik 2. isik mitmus (*lugesite*); *ta* – olevik passiiv eitav kõne (*loeta*); *takse* – olevik passiiv (*loetakse*); *te* – olevik 2. isik mitmus (*loete*); *ti* – lihtminevik passiiv (*loeti*); *vad* – olevik 3. isik mitmus (*loevad*).

[12] Toetun analüüsis Estnltk märgendusele: *ge* – olevik 2. isik mitmus (*lugege*); *gu* – olevik 3. isik ainsus/mitmus (*lugegu*); *neg_gu* – olevik 3. isik ainsus/mitmus (*ärgu lugegu*); *o* – olevik 2. isik ainsus (*loe*); *tagu* – olevik passiiv (*loetagu*).

esineb juba ka üsna palju mitmuse 1. pöörde vormi. Võrreldes B1-tasemega kasutavad B2-tasemel õppijad muid tingiva kõneviisi vorme märgatavalt rohkem (1,9% *vs*. 0,3%). Kitsnik seob tingiva kõneviisi olevikuvormide sageduse kasvu keeleoskuse tõusuga ning ütleb, et just tingiva kõneviisi kasutus on üks neist keelelistest nähtustest, mis B1- ja B2-taseme keeleoskust eristab.

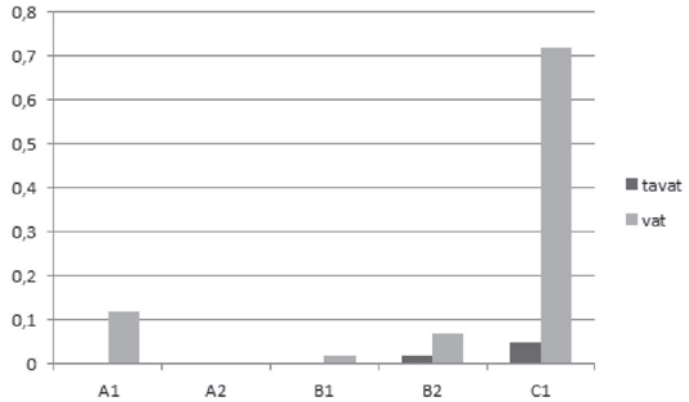
Ka õpikulausete analüüs toetab Kitsniku (2018) väidet: A1-tasemel esineb tingivat kõneviisi (oleviku 1., 2. ja 3. isiku aktiivi jaatav kõne) 0,6%, A2-tasemel 1%, B1-tasemel 1,3%, B2-tasemel 2,4% ja C1-tasemel 3,6% (joonis 10).



Joonis 10. Tingiva kõneviisi lõputunnused[13] õpikulausetes (protsentides)

Lause parameetrite analüsaator näitas, et tingiva kõneviisi *nuks-* (*lugenuks*), *taks-* (*loetaks*) ja *tuks-* (*loetuks*) vorme ei esine ühegi keeleoskustaseme õpikulausetes ning analüüsile toetudes keelasin need vormid versioonides etBasic-v1 ja etIndependent-v1 ning karistasin versioonis etProficient-v1.

Kaudset kõneviisi esineb õpikulausetes vähe (joonis 11).



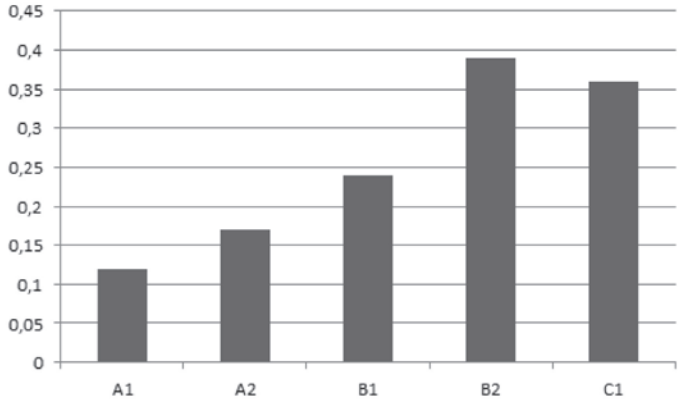
Joonis 11. Kaudse kõneviisi lõputunnused[14] õpikulausetes (protsentides)

[13] Toetun analüüsis Estnltk märgendusele: *ks* – olevik 1., 2., 3. isik ainsus/mitmus (*loeks*); *ksid* – olevik 2. isik ainsus ja 3. isik mitmus (*loeksid*); *ksime* – olevik 1. isik mitmus (*loeksime*); *ksin* – olevik 1. isik ainsus (*loeksin*); *ksite* – olevik 2. isik mitmus (*loeksite*); *neg_ks* – olevik 1., 2. ja 3. isik ainsus/mitmus (*ei loeks*).

[14] Toetun analüüsis Estnltk märgendusele: *tavat* – olevik passiiv (*loetavat*); *vat* – olevik 1., 2. ja 3. isik ainsus/mitmus (*lugevat*); *tuvat* – minevik passiiv (*loetuvat*); *nuvat* – minevik 1., 2., 3. isik ainsus/mitmus (*lugenuvat*); *tavat* – olevik passiiv (*loetavat*).

Oleviku passiivi jaatava kõne vorm *-tavat* (*loetavat*) esineb ainult B2- (0,02%) ja C1-taseme õpikulausetes (0,05%); oleviku 1., 2. ja 3. isiku vorm *-vat* (*lugevat*) esineb A1-tasemel 0,12%, B1-tasemel 0,02%, B2-tasemel 0,07% ja C1-tasemel 0,7%. Kaudse kõneviisi mineviku passiivi jaatava kõne vorm *-tuvat* (*loetuvat*) ja kaudse kõneviisi mineviku 1., 2. ja 3. isiku ainsuse ja mitmuse aktiivi jaatava kõne vorm *-nuvat* (*lugenuvat*) ei esine kordagi. Õpikulausete analüüsi tulemusena lisasin klassifikaatori, mis versioonis etBasic-v1 ja etIndependent-v1 keelab ja versioonis etProficient-v.1 karistab kaudse kõneviisi vorme *-tavat*, *-tuvat*, *-nuvat*. Lisaks keelasin versioonis etBasic-v1 *vat*-vormi, kuna see esines andmebaasis ainult kaks korda (0,1%).

Umbisikulise tegumoe tunnustena käsitlen *takse-* (*elatakse*), *dakse-* (*lauldakse*), *akse-* (*süüakse*), *t-* (*elati*), *d-* (*lauldi*), *ta-* (*elatavat*) ja *da*-vormi (*lauldavat*).



Joonis 12. Umbisikulise tegumoe esinemine õpikulausetes (protsentides)

Jooniselt 12 selgub, et umbisikulise tegumoe esinemine lausetes tõuseb keeletasemeti. C1-tasemel see langeb pisut, kuid languse põhjus võib olla selles, et C1-taseme andmebaasis oli teistega võrreldes palju vähem lauseid. A1-tasemel esineb umbisikulist tegumoodi lauses keskmiselt 0,12 korda, A2-tasemel 0,17, B1-tasemel 0,24, B2-tasemel 0,39 ja C1-tasemel 0,36 korda. Versiooni etBasic-v1 lisasin klassifikaatori, mis keelab umbisikuliste vormide esinemise korpuslausetes, kuna enamik A-taseme õpikulausetest (90%) neid ei kasuta. Versioonis etIndependent-v1 otsustasin neid vorme karistada.

4. GDEX-i eesti mooduli versioonid A-, B- ja C-tasemele

Õpikulausete analüüsi tulemuste põhjal lõin GDEX-i eesti mooduli versioonid etBasic-v1 algtasemel keelekasutajale ehk A-tasemele, etIndependent-v1 iseseisvale keelekasutajale ehk B-tasemele ja etProficient-v1 vilunud keelekasutajale ehk C-tasemele. Tabelis 2 on kokkuvõtvalt välja toodud põhilised keeleoskustasemeid eristavad parameetrid, mis õpikulausete analüüsist välja koorusid, ning mida GDEX-i eesti mooduli uute versioonide arendamisel arvestasin. Võrdluseks on tabelis esitatud sõnastike näitelausete põhjal loodud ja uute versioonide aluseks olnud GDEX 1.4 parameetrid.

Tabel 2. GDEX-i eesti mooduli eri versioonide parameetrid. Märk *x* märgib keelamist ja * karistamist

Parameetrid	GDEX 1.4	etBasic-v1	etIndependent-v1	etProficient-v1
lause pikkus (sõnedes)	4–20	3–14	3–18	4–23
lause optimaalne pikkus (sõnedes)	6–12	4–7	4–12	6–14
sisaldab nimisõna		jah	jah	
pikad sõnad		* (>9 tm)	* (>11 tm)	
ma-infinitiivi käändelised vormid	* (*-mast*, *-maks*, *-mas*)	x (*-mast*, *-maks*, *-mata*, *-tama*)	x (*-maks*, *-tama*)	
des-vorm	*	x		
tingiv kõneviis		x (*-nuks*, *-taks*, *-tuks*)	x (*-nuks*, *-taks*, *-tuks*)	* (*-nuks*, *-taks*)
kaudne kõneviis		x (*tavat-*, *tuvat-*, *nuvat*-vormid)	x (*tavat-*, *tuvat-*, *nuvat*-vormid)	* (*-tuks*, *-tavat*, *-tuvat*, *-nuvat*)
käskiv kõneviis		x (*neg_gu*, *-tagu*)	x (*neg_gu*, *-tagu*)	
umbisikuline tegumood		x (*-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da*)	* (*-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da*)	
tud-vorm		*		
lühendid		*	*	
omadussõna ülivõrre		*		
tegusõna vormi juurde kuuluv sõna (nt *plehku*)		*	*	

Lisaks tabelis 2 välja toodud parameetritele sisaldavad kõik GDEX-i eesti mooduli versioonid alates versioonist 1.4 karistust sõnadele, mille esinemissagedus korpuses on madalam kui 1000, lisaks keelatakse või karistatakse vulgarisme, slängisõnu jmt.

Versiooni etBasic-v1 väljund on esitatud joonisel 13, versiooni etIndependent-v1 väljund joonisel 14 ja versiooni etProficient-v1 väljund joonisel 15.

Rank	Sentence
1	Lõpuks läksin arsti juurde.
2	Tulemused pidin arstile saatma.
3	Arst on ikka vapustav!
4	Arstid on juba üsna nõutud.
5	Palun ärgake, suhelge oma arstidega!
6	Mees läheb arsti juurde ja kurdab oma muret.
7	Röntgen kinnitas arsti kahtlusi.
8	Rahulolev patsient kirjutab:"Väga sõbralik ja tubli arst.
9	Eile käisime arstil Västrikus.
10	Arstide õpetamisel on väga selgelt kaks etappi.

Joonis 13. Versiooni etBasic-v1 väljund lemma *arst* jaoks[15]

Rank	Sentence
1	Mees läheb arsti juurde ja kurdab oma muret.
2	Minulgi täna arstil käidud.
3	Tulemused pidin arstile saatma.
4	Lõpuks läksin arsti juurde.
5	Patsiendid pöörduvad arsti poole väga erinevatel põhjustel ja ka erinevate soovidega.
6	Arstide õpetamisel on väga selgelt kaks etappi.
7	Palun ärgake, suhelge oma arstidega!
8	Arstide probleemi ei ole, aga õdede pakkumine on väga väike.
9	Arsti seisukohalt on töö tema sõnul aga raskemaks läinud: "Pead käima ajaga kaasas.
10	Hommikul komberdasin arsti juurde.

Joonis 14. Versiooni etIndependent-v1 väljund lemma *arst* jaoks

[15] Kasutasin GDEX-i versioonide testimiseks "Eesti keele ühendkorpust 2017" ja programmi GDEX Editor https://gdexed.sketchengine.eu (1.10.2018). Joonisel on esitatud katkend päringu tulemustest.

Rank	Sentence
1	Kaebuse kohaselt eksitab ajakirjanik lugejat ja püüab näidata arste halvas valguses.
2	Arst ütles et sööd liiga palju süsivesikuid.
3	Oleme teinud tihedat koostööd arstide liiduga, sest probleemid on üldisemad.
4	Palun ärgake, suhelge oma arstidega!
5	Otsisin abi kõikvõimalike arstide juurest ja kõikvõimalikke viise katsetades - ei midagi.
6	Arstid oskavad rohkem aidata,kui keegi teine!
7	Arstide probleemi ei ole, aga õdede pakkumine on väga väike.
8	Enne vaktsineerimist tuleb arstile või õele kindlasti mainida, et ootate last.
9	Arsti seisukohalt on töö tema sõnul aga raskemaks läinud: "Pead käima ajaga kaasas.
10	Tulemused kantakse arvutisse, arst hindab neid oma ja kannab kliendile ette.

Joonis 15. Versiooni etProficient-v1 väljund lemma *arst* jaoks

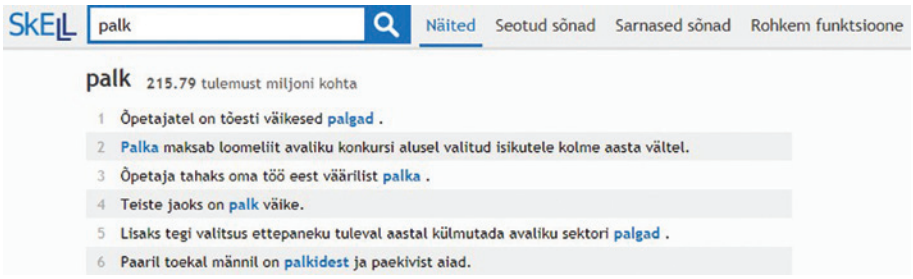
5. Probleemid ja edasiarendused

Kõikide versioonide väljundiks kuvatud lausetes esineb ikka keeleoskustasemele mitte sobivaid sõnu. Näiteks esinevad versiooni etBasic-v1 väljundis (joonis 13) sõnad *rahulolev*, *kinnitama* ja *kurtma* ja etIndependent-v1 väljundis *komberdama* (joonis 14). Väljund paraneks, kui rakendada sõnavarafiltrit, mis valiks korpusest välja ainult sellised laused, mis sisaldavad vaid vastava keeleoskustaseme sõnu. Selleks saab kasutada 2018. aastal loodud sõnavaraloendeid (Kallas, Koppel 2018a, 2018b, 2018c). Samuti tuleks täiendavalt testida sõnade sageduslävesid eri keeleoskustasemetele – kui GDEX 1.4 karistab lauseid, mis sisaldab sõnu, mis esinevad korpuses vähem 1000 korda, siis tõenäoliselt tuleb versioonis etBasic-v1 sõna sagedusläve tõsta.

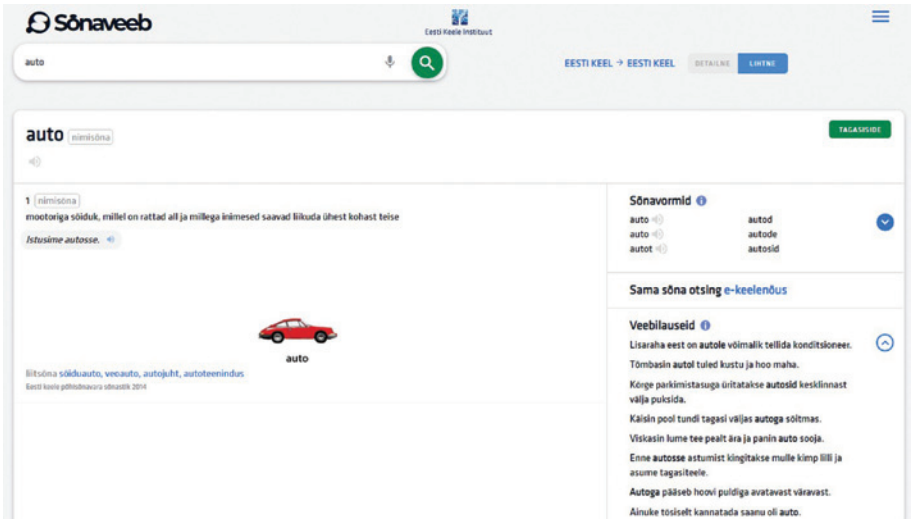
Eri versioonid võimaldavad edaspidi luua eraldi õppekorpused eri keeleoskustasemetele, mida saab rakendada erinevates portaalides nagu etSkELL ja Sõnaveeb, aga ka muu õppevara loomisel.

Mitmed probleemid on ühised nii õpikulausete põhjal loodud versioonidele (etBasic-v1, etIndependent-v1 ja etProficient-v1) kui ka varasemale, sõnastike näitelausete põhjal loodud versioonile GDEX 1.4, mida praegu kasutatakse etSkELL-is (joonis 16) ja Sõnaveebis (joonis 17), ja mille abil päritakse lauseid "Eesti keele õppekorpusest 2018 (etSkELL)"[16] (250 mln sõna).

[16] https://doi.org/10.15155/3-00-0000-0000-0000-07335L (1.10.2018). Korpus valmis koostöös tarkvarafirmaga Lexical Computing Ltd. Tänan korpuse loomise eest tarkvaraarendajat Jan Michelfeiti.



Joonis 16. Lemma *palk* näited keeleõppeportaalis etSkELL



Joonis 17. Lemma *auto* veebilaused sõnastikuportaalis Sõnaveeb

Automaatselt valitud näitelausete kuvamisel on mitmeid kitsaskohti.

1. Lausete valikul ei arvestata otsisõna polüseemiaga. Nt annab päring *leht* vastuseks lauseid, kus sõna esineb eri tähendustes (näited 18–20).

   (18) Leht on avatud ka reklaamidele.
   (19) Lehti ja õisi kogutakse õitsemise ajal.
   (20) Täpsemat infot leiad viisa nõuete lehelt.

2. Lausete valikul ei arvestata otsisõna homonüümsete tähendustega. Nt annab päring *tamm* vastuseks läbisegi lauseid, kus *tamm* on kas puu või vesiehitise tähenduses (näited 21–22).

   (21) Meie kooli õuel kasvasid tammed.
   (22) Kopra loodud märgalade suurus sõltub tammidest ning nende asukohast.

3. Probleemiks on lemmatiseerimise vead, eriti tulevad need esile vormi-homonüümia korral. Nt annab päring *joon* vastuseks lauseid, kus sõna esineb nii nimisõnana (näide 23) kui tegusõna *jooma* kindla kõneviisi oleviku ainsuse 1. pöörde vormina (näide 24).

(23) Joonista peenema pintsliga kaunis joon keset küünt.
(24) Kui tahan, joon õlut.

4. Probleemiks on morfoloogilise märgenduse vead. Nt annab päring *koha* vastuseks lauseid, kus tegemist pole mitte kalaga, vaid nimisõna *koht* omastava käände vormiga (näide 25).

   (25) Meie klassi poisid võitsid II KOHA.

5. Valikusse satuvad kakskeelsetelt lehtedelt kroolitud masintõlkelised laused, mis vastavad küll etteantud parameetritele (sõnad on sagedad ja nende pikkus alla 20 tähemärgi, lause sisaldab verbi ja on maksimaalselt 20 sõnet pikk jmt), kuid ei ole kohati grammatilised (näide 26).

   (26) Õpetused roomakatoliku kirik ei ole kaugeltki selge tõdesid Jumala sõna.

6. Raske on leida lauseid madala sagedusega sõnade jaoks. Nt nimisõna *kalla* ei esine "Eesti keele õppekorpuses 2018 (etSkELL)" ühtegi korda.

Polüseemsete sõnade eri tähendustele sobivate näitelausete leidmist abistaks see, kui korpus oleks ka semantiliselt märgendatud ja sõnaraamat sisaldaks infoüksusena neidsamu semantilisi märgendeid, mida korpuse märgendamisel kasutati. Eesti keele jaoks võiks rakendada Margit Langemetsa (2010) välja töötatud semantilisi tüüpe, mida on kasutatud nt PSV ja EKS-i koostamisel.

Täiendav võimalus, kuidas polüseemia ja vormihomonüümia probleemi lahendada, on arvestada lausete valikul naabersõnade sõnastiku andmebaasis olevate otsisõna kollokaatidega, et väljundis oleks eelkõige sagedasemaid kollokatsioone sisaldavad näitelaused. Nt kui otsisõnal *tamm* on kolm homonüümi ja kasutaja valib neist tähenduse *pais*, kuvatakse esmalt laused kollokatsioonidega *tamm puruneb* ja *tammi ehitama*.

Samuti saab väljundi kvaliteeti parandada, kui päringu koostamisel lähtuda lemposest, mitte lemmast. Lempos sisaldab infot nii lemma kui ka sõnaliigi kohta, nt lemma *noor* jaoks on kaks lempost, kus *noor-s* tähistab kasutust nimisõnana ja *noor-a* kasutust omadussõnana.

Vältimaks automaattõlkeliste lausete sattumist väljundisse, tuleks juba korpuse kroolimisel automaatselt tuvastada ja kõrvale jätta tekstid, kus esineb rohkelt vigase süntaktilise struktuuriga lauseid. Lisaks tuleb uue õppekorpuse loomisel arvesse võtta alliktekstide päritolu – see võimaldaks eelistada nt Eesti Vikipeediast ja perioodikaväljaannetest pärit lauseid blogi- ja foorumipostitustest pärit lausetele.

Madalama sagedusega sõnadele lausete leidmiseks on edaspidi vajalik kombineerida päringuid eri korpustest. Kui sõna ei esine ühes korpuses, siis kuvatakse tulemusi suuremast korpusest.

Kokkuvõtteks võib öelda, et näitelausete automaatne valik ei sõltu üksnes GDEX-iga määratud parameetritest, vaid tulemuse parandamiseks on oluline ka semantilise, süntaktilise ja morfoloogilise info arvestamine, täiendavate filtrite (nt eri keeleoskustaseme sõnavara loendite) rakendamine ja tekstitüüpide arvestamine.

6. Kokkuvõte

Analüüsisin artiklis eri keeleoskustaseme õpikute lausete parameetreid, millele toetudes töötasin välja korpuspäringusüsteemi Sketch Engine heade näitelausete tööriista GDEX eesti mooduli versioonid algtasemel (etBasic-v1), iseseisvale (etIndependent-v1) ja vilunud keelekasutajale (etProficient-v1). Uurimismaterjalina kasutasin "Eesti keele A1–C1 õpikute korpuse (2018)" lausetest loodud viit andmebaasi, mida analüüsisin programmi "Lause parameetrite analüsaator" abil.

Analüüs näitas, et võrreldes seni viimase GDEX-i eesti mooduli versiooniga 1.4 oli vaja kohandada selliseid parameetreid nagu lause ja sõnede pikkus, teatud tegusõna vormide ja eri sõnaliikide esinemine lauses. Nii määrasin versioonis etBasic-v1 lause pikkuseks 3–14 (lause optimaalne vahemik 4–7) sõnet, versioonis etIndependent-v1 3–18 (optimaalne vahemik 4–12) sõnet ja versioonis etProficient-v1 4–23 (optimaalne vahemik 6–14) sõnet. Sõnede maksimaalne pikkus on kõikides mooduli versioonides 20 tähemärki, kuid täiendavalt karistatakse versioonis etBasic-v1 sõnu, mis on pikemad kui 9 tähemärki, ja versioonis etIndependent-v1 sõnu, mis on pikemad kui 11 tähemärki. Versioonides etBasic-v1, etIndependent-v1 ja etProficient-v1 keelatakse või karistatakse tingiva, kaudse ja käskiva kõneviisi, umbisikulise tegumoe ja *ma*-tegevusnime käändeliste vormide esinemist.

Näitelausete automaatse valiku kitsaskohad on sõnade leksikaalne polüseemia, homonüümia, vormihomonüümia, sõnade madal esinemissagedus, vigane lemmatiseerimine ja morfoloogiline märgendus ning masintõlkelised laused. GDEX-i väljundi parandamiseks on vajalik edaspidi arvestada semantilist, süntaktilist ja morfoloogilist infot, täiendavat sõnavarafiltrit ning tekstitüüpe.

GDEX-i eesti mooduli uued versioonid võimaldavad luua eri keeleoskustasemete õppekorpusi, mida saab omakorda kasutada eri tüüpi keeleõpperakendustes ja sõnastikuportaalides, nagu artiklis etSkELL-i ja Sõnaveebi näitel illustreerisin.

Viidatud kirjandus

Hausenberg, Anu-Reet; Ilves, Marju; Kaivapalu, Annekatrin; Kerge, Krista; Kern, Katrin; Kitsnik, Mare; Krall, Ingrid; Rummo, Karin; Rüütmaa, Tiina 2008. Iseseisev keelekasutaja. B1- ja B2-taseme eesti keele oskus ['Independent user: B1- and B2-level proficiency in Estonian']. Tallinn: REKK, Atlex.

Ilves, Marju 2008. Algaja keelekasutaja. A2-taseme eesti keele oskus ['Estonian for beginners: A2-level proficiency in Estonian']. Krista Kerge (Toim.). Tallinn: Eesti Keele Sihtasutus.

Kallas, Jelena; Koppel, Kristina; Tuulik, Maria 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

Kallas, Jelena; Koppel, Kristina 2018a. Eesti keele B1-taseme sõnavara ['Vocabulary lists: B1 Estonian language proficiency level']. Tallinn: Eesti Keele Instituut. http://www.eki.ee/keeletase/lists/B1.pdf (14.2.2019).

Kallas, Jelena; Koppel, Kristina 2018b. Eesti keele A2-taseme sõnavara ['Vocabulary lists: A2 Estonian language proficiency level']. Tallinn: Eesti Keele Instituut. http://www.eki.ee/keeletase/lists/A2.pdf (14.2.2019).

Kallas, Jelena; Koppel, Kristina 2018c. Eesti keele A1-taseme sõnavara ['Vocabulary lists: A1 Estonian language proficiency level']. Tallinn: Eesti Keele Instituut. http://www.eki.ee/keeletase/lists/A1.pdf (14.2.2019).

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

Kilgarriff, Adam; Husák, Miloš; McAdam, Katy; Rundell, Michael; Rychlý, Pavel 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.

Kitsnik, Mare 2014. Verbivormid B1- ja B2-taseme kirjalikus õppijakeeles ['Written learner language verb forms at B1 and B2 levels']. – Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics, 5 (3), 9–35. https://dx.doi.org/10.12697/jeful.2014.5.3.01

Kitsnik, Mare 2018. Iga asi omal ajal: eesti keele B1- ja B2-taseme verbikonstruktsioonid keeleoskuse arengu näitajana ['All in good time: Estonian B1- and B2-level verbal constructions as indicators of the development of language proficiency']. Humanitaarteaduste dissertatsioonid 43. Tallinn: Tallinna Ülikooli Kirjastus.

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

Kosem, Iztok; Koppel, Kristina; Kuhn, Tanara Zingano; Michelfeit, Jan; Tiberius, Carole 2018. Identification and automatic extraction of good dictionary examples: The case(s) of GDEX. – International Journal of Lexicography. https://dx.doi.org/10.1093/ijl/ecy014

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

Langemets, Margit 2010. Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras ['Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources']. Tallinn: Eesti Keele Sihtasutus.

Penjam, Pille 2008. Eesti kirjakeele *da*- ja *ma*-infinitiiviga konstruktsioonid ['The constructions of DA- and MA-infinitives in written Estonian']. Dissertationes philologiae estonicae Universitatis Tartuensis 23. Tartu: Tartu Ülikooli Kirjastus.

PSV = Eesti keele põhisõnavara sõnastik ['Basic Estonian Dictionary']. Jelena Kallas, Mai Tiits, Maria Tuulik (Toim.). Madis Jürviste, Kristina Koppel, Maria Tuulik (Koost.). Tallinn: Eesti Keele Sihtasutus, 2014.

Raamdokument 2007 = Euroopa keeleõppe raamdokument: õppimine, õpetamine, hindamine ['CEFR: Learning, teaching and assessment']. Tartu: Haridus- ja Teadusministeerium, 2007.

Sõnaveeb ['Dictionary portal Wordweb']. https://sonaveeb.ee (14.2.2019).

Võrguviited

Eesti keele naabersõnad 2019 ['The Estonian Collocations Dictionary, ECD']. Jelena Kallas, Kristina Koppel, Maria Tuulik, Geda Paulsen (Toim.). Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee (14.2.2019).

EKS = Eesti keele sõnaraamat 2019 ['The Dictionary of Estonian, DicEst']. Margit Langemets, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll (Toim.). Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee (14.2.2019).

etLex. www.eki.ee/keeletase (1.10.2018).

etSkELL. www.etskell.sketchengine.co.uk (1.10.2018).

GDEX Editor. https://gdexed.sketchengine.eu (1.10.2018).

KORP. https://korp.keeleressursid.ee (1.10.2018).

Lause parameetrite analüsaator: teksti märgendamise ja statistilise analüüsi tööriist ['Analyser of Sentence Parameters']. http://www.eki.ee/keeletase/statistics (1.10.2018).

ANALYSIS OF CEFR-GRADED COURSEBOOK SENTENCES AND THEIR USE FOR AUTOMATIC DETECTION OF GOOD DICTIONARY EXAMPLES

Kristina Koppel
Institute of the Estonian Language

The aim of the study was to develop new Estonian GDEX configurations for A-, B- and C-language proficiency levels. GDEX (Good Dictionary Example) (Kilgarriff et al. 2008) is a software module of the corpus query system Sketch Engine (Kilgarriff et al. 2004), which helps to identify good dictionary example candidates from large corpora.

In order to identify which specific parameters characterise sentences in each proficiency level, full sentences from the Estonian Coursebook Corpus 2018 were analysed using a program called Analyser of Sentence Parameters developed at the Institute of the Estonian Language. The analyser allows to find out how long the sentences and tokens are, what kind of verb forms are used, what syntactic properties the sentences have etc.

The analysis showed that compared to the latest Estonian GDEX configuration 1.4 such parameters as sentence and token length, occurrence of certain verb forms and parts of speech needed to be adjusted. Accordingly, for A-level the sentence length was set to 3–14 tokens (optimal interval 4–7 tokens), for B-level 3–18 tokens (optimal interval 4–12) and for C-level 4–23 tokens (optimal interval 6–14 tokens). A new classifier that penalises tokens longer than 9 characters on A-level and tokens longer than 11 characters on B-level was introduced. On A- and B-levels certain verb forms were penalised or banned from appearing in the sentence.

etSkELL – a corpus tool for Estonian language learning – and the dictionary portal Sõnaveeb (Wordweb) are introduced as possible ways to implement the new GDEX configurations output.

The results of this paper can be applied in compiling corpora and teaching materials for different language proficiency levels.

Keywords: corpus linguistics, corpus lexicography, corpora, learners' corpora, Estonian as a second language, Estonian

Kristina Koppel (Eesti Keele Instituut) on eesti keele naabersõnade sõnastiku töörühma liige ja Tartu Ülikooli doktorant. Põhilised uurimisvaldkonnad: korpuslingvistika, e-leksikograafia.
Roosikrantsi 6, 10119 Tallinn, Estonia
kristina.koppel@eki.ee