DISSERTATIONES
LINGUISTICAE
UNIVERSITATIS
TARTUENSIS
38

KRISTINA KOPPEL

Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
38

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
38

KRISTINA KOPPEL

Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele

TARTU ÜLIKOOL
kirjastus

Tartu Ülikooli humanitaarteaduste ja kunstide valdkond, eesti ja üldkeeleteaduse instituut

Väitekirja on filosoofiadoktori kraadi saamiseks kaitsmisele suunanud Tartu Ülikooli eesti ja üldkeeleteaduse instituudi nõukogu otsusega 9. jaanuaril 2020.

Juhendajad: dotsent Raili Pool (Tartu Ülikool)
dr Jelena Kallas (Eesti Keele Instituut)

Oponent: professor Annekatrin Kaivapalu (Ida-Soome Ülikool)

Kaitsmine toimub 23. märtsil 2020 kell 16.15 Tartu Ülikooli senati saalis.

Doktoritöö valmimist on toetanud keeleteaduse, filosoofia ja semiootika doktorikool, rahastanud Euroopa Regionaalarengu Fond (Tartu Ülikooli ASTRA projekt PER ASPERA, Eesti Keele Instituudi projekt EKI-ASTRA), ISCH COSTi projekt IS10305 „Euroopa elektroonilise leksikograafia võrgustik" ja Horisont 2020 programmi projekt 731015 „European Lexicographic Infrastructure".

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

ISSN 1406-5657
ISBN 978-9949-03-300-3 (trükk)
ISBN 978-9949-03-301-0 (pdf)

Tartu Ülikooli Kirjastus
www.tyk.ee

SISUKORD

EESSÕNA 7

PUBLIKATSIOONIDE LOEND 8

KESKSED MÕISTED JA LÜHENDID 9

1. TEMAATILINE ÜLEVAADE 11
   1.1. Tänapäeva leksikograafia arengusuunad 11
   1.2. Uurimisobjekt, -meetod ja analüüsimaterjal 14
   1.3. Töö eesmärgid 16
   1.4. Ülevaade väitekirja publikatsioonidest 16

2. KORPUSLAUSE NÄITELAUSE ALLIKANA 18
   2.1. Sõnastiku näitelause tüübid, funktsioonid ja valiku põhimõtted 18
   2.2. Sõnastiku näitelause tunnused 21
   2.3. Korpuslause keeleõppes 23

3. NÄITELAUSETE AUTOMAATSE TUVASTAMISE MEETODID 25
   3.1. Masinõppemeetod, reeglipõhine lähenemine ja kombineeritud meetod 25
   3.2. Reeglipõhine tööriist Good Dictionary Examples ehk GDEX 27
   3.3. GDEXi eri keelte moodulid 28

4. UURIMISTULEMUSED: GDEXi EESTI KEELE MOODUL 33
   4.1. GDEXi eesti keele mooduli versioonid 33
      4.1.1. GDEX 1.2 33
      4.1.2. GDEX 1.3 34
      4.1.3. GDEX 1.4 35
      4.1.4. GDEXi versioonid eri keeleoskustasemetele 38
   4.2. GDEXi eesti keele mooduli eri versioonide parameetrid 39
      4.2.1. Must ja hall nimekiri 40
      4.2.2. Lause alguses keelatud sõnad ja sõnapaarid 42
      4.2.3. Tegusõnavormid 43
   4.3. GDEX 1.4 väljundi evalveerimine 43
   4.4. GDEXi rakendamine 47
      4.4.1. GDEXi eesti keele moodul Sketch Engine'is 48
      4.4.2. „Eesti keele õppekorpus 2018 (etSkELL)" 48
      4.4.3. etSkELL ehk Sketch Engine for Estonian Language Learning 49
         4.4.3.1. Näited 49
         4.4.3.2. Naabersõnad 50
         4.4.3.3. Sarnased sõnad 53
      4.4.4. Keeleportaal Sõnaveeb 54

5. PROBLEEMID JA EDASIARENDUSED 56
   5.1. Korpuse sisu ja maht 56
   5.2. Märgendamise kvaliteet 57
      5.2.1. Lemmatiseerimise ja morfoloogilise märgenduse vead 58

5.2.2. Lausestamine 58
5.2.3. Mitmesõnalised üksused 59
5.2.4. Leksikon 59
5.2.5. Trükivead 60
5.3. Grammatiline mitmesus 61
5.4. Semantiline mitmesus 62
5.5. Edasiarendused 63
5.5.1. Täiendavad klassifikaatorid 63
5.5.2. Eri sihtgruppidele kohandatud konfiguratsioonid ja uued õppekorpused 63
5.5.3. Leksikaalne filter 64
5.5.4. API sätted 64
5.5.5. Õppekorpuse kvaliteedi parandamine kasutajate abil 65
6. KOKKUVÕTE 66
SUMMARY: Corpus-based automatic detection of example sentences for dictionaries for Estonian learners 69
KIRJANDUS 79
LISA 1. Koondtabel GDEXi eesti keele mooduli eri versioonide parameetritest 90
LISA 2. GDEX 1.2 konfiguratsioonifail 95
LISA 3. GDEX 1.3 konfiguratsioonifail 97
LISA 4. GDEX 1.4 konfiguratsioonifail 99
LISA 5. etBasic-v1 konfiguratsioonifail 102
LISA 6. etIndependent-v1 konfiguratsioonifail 105
LISA 7. etProficient-v1 konfiguratsioonifail 108
LISA 8. Must nimekiri 111
LISA 9. Hall nimekiri 112
PUBLIKATSIOONID 113
ELULOOKIRJELDUS 246
CURRICULUM VITAE 247

EESSÕNA

Olen sõnaraamatutööga tegelenud alates 2009. aastast, mil magistrantuuri järel Eesti Keele Instituuti tööle sattusin. Siis ei uskunud, et tahan veel kunagi astuda doktorantuuri, kuigi Silvi Vare mind sellele mõtlema suunas, iga kord kui „Eesti keele sõnaperede" andmebaasi toimetades esile kerkinud sõnamoodustuse küsimuste üle arutlesime. Tõelise tõuke andsid liitumine eesti keele õppesõnastike töörühmaga, kellega 2013. aastal korraldasime Tallinnas rahvusvahelist e-leksikograafia teemalist konverentsi „eLex 2013: Electronic Lexicography in the 21st Century", kus sai ühtlasi alguse minu koostöö kolleegidega mujalt Euroopast; ning 2014. aastal alanud „Eesti keele naabersõnade 2019" sõnastiku projekt, mille raames sain ülesandeks teha prooviuuringu, selgitamaks välja eesti keele sõnastike näitelauseid iseloomustavad tunnused.

Minu kõige suurem tänu läheb minu juhendajatele Jelena Kallasele ja Raili Poolile, kes mind selle viie aasta jooksul õigel kursil hoidsid. Aitäh, Jelena, et värbasid mind oma töörühma liikmeks, palusid appi konverentsi korraldama, olid kõhkluseta nõus võtma mind oma esimeseks doktorandiks ja et leidsid ka kõige kiirematel perioodidel aega lugeda minu tupikusse jooksnud tööd ning anda väljapäästvaid suuniseid. Sinu lennukus ja töökus on olnud mulle tõeliseks inspiratsiooniks! Aitäh, Raili, et motiveerisid mind oma optimismiga iga poolelioleva kirjatüki lugemise järel nendega jätkama. Sinu asjatundlikkus eesti keele kui teise keele õpetamise alal on minu jaoks hindamatu väärtusega.

Olen tänulik oma retsensentidele Annekatrin Kaivapalule ja Ulla Vanhatalole konstruktiivsete parandusettepanekute ning Maria-Maren Linkgreimile töö keelelise ja tehnilise toimetamise eest. Tahan tänada oma toredaid kolleege ning kaasdoktorante Eesti Keele Instituudist – ilma teie toetuseta ei oleks see olnud võimalik. Eriline tänu kuulub Margit Langemetsale, kes andis alati minu kirjutistele väärtuslikku tagasisidet. Olen tänulik reedeklubile ja eeskätt Helle Metslangile. Meie regulaarsed kohtumised aitasid väitekirjaga järjekindlalt edasi liikuda. Tänan Iztok Kosemi inspireerivate vestluste eest erinevatel konverentsidel ja töötubades ning Euroopa erinevais jäätisekohvikuis – ilma temata olnuks see väitekiri hoopis mõnel muul teemal. Tänuga mõtlen Arvi Tavastile, kes andis eneseusku ja teotahet, kui katuspeatüki kirjutamise jõud hakkas raugema. Aastate jooksul on mind tehnilistes küsimustes aidanud Ülle Viks, Indrek Hein, Katrin Tsepelina, Arvi Tavast, Jan Michelfeit, Jaka Čibej ja Cyprian Laskowski – suur aitäh abi eest!

Lõpetuseks siiras tänu minu perekonnale, sugulastele ja sõpradele, kelle julgustuse ja toeta ei oleks mul olnud jaksu jõuda oma tööga võiduka lõpuni. Olen igavesti tänulik oma emale, kes on mind kõigis minu tegemistes toetanud ning mulle õpetanud, et suure töö ja tahtmisega on võimalik saavutada suuri asju, aga et ka puhkama peab. Ja kui võimalik, siis ilusas kohas ja hästi kaua. Siiras tänu sugulastele, eriti Anule ja tema perele Tartu öömaja, lugematu arvu kohvitasside ning koogiviilude eest. Ülle, Robert, Kristo, Helen, Jaava, Liis, Ilmar, Rael, Maria, Chris, Sander, Arvi, Jesper ja Katrin – suur aitäh, et hoidsite minu töö- ja puhkusreisidel Süsi ja Pipi elus!

Tallinnas 15. jaanuaril 2020.

PUBLIKATSIOONIDE LOEND

[P1] **Koppel, Kristina, Jelena Kallas** (2016). Õppijasõbralik korpuslause: automaatse valiku võimalusi. Lähivõrdlusi. Lähivertailuja, 26, 222–250.

[P2] **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.

[P3] **Koppel, Kristina** (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks. Eesti Rakenduslingvistika Ühingu aastaraamat, 13, 53–71.

[P4] **Koppel, Kristina** (2019). Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks. Lähivõrdlusi. Lähivertailuja, 29, 84–112.

[P5] **Koppel, Kristina** (2019). Eesti keele kui teise keele õpikute lausete analüüs ja selle rakendamine eri keeleoskustasemete sõnastike näitelausete automaatsel valikul. Eesti Rakenduslingvistika Ühingu aastaraamat, 15, 99–119.

KESKSED MÕISTED JA LÜHENDID

anafoor (*anaphora*) – (tagasi)viide tekstis varem esinenud infole. Eesti keeles käituvad anafoorina tavaliselt asesõnad, siinses töös käsitletakse anafoorina nt ka deiksiseid ja konnektiivlaiendeid.

API ehk **programmiliides** (*application programming interface*) – protokoll, mille abil üks programm teise käest veebi kaudu andmeid pärib.

etSkELL ehk Sketch Engine for Estonian Language Learning – automaatne keeleõppekeskkond eesti keele jaoks.

GDEX ehk Good Dictionary Examples – korpuspäringusüsteemi Sketch Engine integreeritud tööriist, mis tuvastab näitelauseks sobivaid korpuslauseid.

GDEXi eesti keele moodul – eesti keele jaoks loodud GDEXi konfiguratsioonid, mis arvestavad keelespetsiifilisi parameetreid.

GDEXi konfiguratsioonifail (*GDEX configuration file*) – reeglipõhine valem koos klassifikaatoritega, mis kombineerib parameetritele (mitte)vastamise eest antud skoorid ühtseks üldskooriks (*GDEX score*).

hall nimekiri (*greylist*) – nimekiri sõnadest, mille eest saab lause karistada, nt kõnekeelsed sõnad.

kaal (*weight*) – klassifikaatori olulisuse määr. Aitab vahet teha sellel, millised klassifikaatorid on lause kvaliteedi määramisel olulisemad kui teised. Mida olulisem klassifikaator, seda suurem kaal.

karistama (*penalize*) – lause üldskoori vähendama, kui lause ei vasta teatud nõrga klassifikaatori alla liigituvale parameetrile.

klassifikaator (*classifier*) – algoritm, mis analüüsib sisendiks oleva andmestiku vastavust etteantud parameetritele ning määrab kindlaks selle sobivuse (nt kui lause ei lõpe lauselõpumärgiga, siis see ei sobi näitelauseks). Jagunevad tugevateks ja nõrkadeks.

- **tugevad klassifikaatorid** (*hard classifiers*) – parameetrid, millele hea näitelause peab alati vastama, nt peab tegemist olema täislausega.
- **nõrgad klassifikaatorid** (*soft classifiers*) – parameetrid, mis lause skoori vähem mõjutavad, nt teatud elementide (komade, asesõnade vm) arv lauses.

kollokatsioon (*collocation*) – sisusõnade tähenduslikud ja statistiliselt esilduvad kombinatsioonid teiste leksikaalsete ja grammatiliste üksustega (nt *päike paistab*, *kange kohv*, *pakast trotsima*). Omakeelse sünonüümina kasutatakse siinses töös mõistet *naabersõnad*.

korpus (*corpus*) – suur elektrooniline tekstikogu, mis on reeglina otsitav korpuspäringusüsteemi kaudu.

korpuslause (*corpus sentence*) – autentsetest tekstidest koosneva korpuse lause.

korpuspäringusüsteem (*corpus query system*) – tarkvara, mis võimaldab korpuse mitmekülgset analüüsi.

kroolimine (*crawling*) – veebilehtede süstemaatiline sirvimine ja tekstide kogumine spetsiaalse veebiroboti (*crawler*) abil.

lemma (*lemma*) – sõna või väljendi algvorm.

lempos (*lempos*) – kombinatsioon lemmast ja sõnaliigist (POS). Nt omadussõna *noor* lempos on *noor-a*, kus ühetäheline lühend *a* tähistab adjektiivi ehk omadussõna; nimisõna *noor* lempos on *noor-s*, kus ühetäheline lühend *s* tähistab substantiivi ehk nimisõna.

must nimekiri (*blacklist*) – nimekiri sõnadest, mis on lauses keelatud, nt vulgarismid.

naabersõnad – vt kollokatsioon.

näitelause (*example*) – sõnastikus näitena toodud lause.

parameeter (*parameter*) – tunnus, mille alusel programm näitelauseid valib.

pikim tüüpiline kontekst (*longest commonest match*) – kahesõnalise kollokatsiooni laienemine mitmesõnaliseks üksuseks.

SkELL ehk Sketch Engine for Language Learning – automaatne keeleõppekeskkond inglise keele jaoks.

skoor (*GDEX score*) – lausele määratud punktisumma, mis moodustub individuaalsete parameetrite kombineeritud summast.

sõnaartikkel (*dictionary entry*) – sõnaraamatu märksõna koos juurdekuuluva infoga.

sõnastikusüsteem (*dictionary writing system*) – haldussüsteem, mis võimaldab sõnastikke koostada, toimetada ja küljendada, veebis avaldada; teha lihtsaid ja keerulisi struktuuripõhiseid päringuid ning päringutulemusi sortida.

sõnavisand (*word sketch*) – ühel lehel kuvatav automaatne korpuspõhine kokkuvõte sõna grammatilisest ja kollokatiivsest käitumisest.

sõne (*token*) – tekstisõna.

TBL-meetod (*tickbox lexicography*) – sõnaartikli komponentide ühekaupa valimine korpusandmetest ja nende automaatne ülekandmine sõnastikusüsteemi.

teine kollokaat (*second collocate*) ehk **kollokatsiooni kollokaat** – esilduv sõna, mis esineb lauses sageli koos kollokatsiooniga. Nt kollokatsiooni *kontsaga kingad* teine kollokaat võib olla *kõrge* või *madal* (*kõrge/madala kontsaga kingad*).

õppekorpus (*pedagogical corpus*) – korpus, mis on loodud pedagoogilistel eesmärkidel ja sobib kasutada nii keele õpetamisel kui ka õppimisel. Õppekorpus ei sisalda võõrkeele õppijate sihtkeelseid kirjalikke tekste.

õppesõnastik (*pedagogical dictionary*, *learners' dictionary*) – sõnastik, mille sihtgrupp on keeleõppija.

ühekliki sõnaraamat (*one-click dictionary*) – sõnastiku kõigi infoüksuste korpuspõhine täisautomaatne genereerimine eeldefineeritud parameetrite alusel.

1. TEMAATILINE ÜLEVAADE

Väitekiri koosneb sissejuhatavast osast ja viiest publikatsioonist, mis on avaldatud aastatel 2016–2019. Väitekirja keskmes on parameetrid, millele toetub näitelausete automaatne tuvastamine eesti õppesõnastike jaoks.

Sissejuhatava osa esimeses peatükis annan ülevaate tänapäeva leksikograafia arengusuundadest Eestis ja Euroopas. Teises peatükis kirjeldan sõnastiku näitelause tüüpe, funktsioone ja valiku põhimõtteid ning toon välja hea näitelause tunnuseid; samuti annan ülevaate korpuslausete kasutusvõimalustest (õppe)leksikograafias ja keeleõppes üldisemalt. Kolmandas peatükis tutvustan eri keelte näitel meetodeid, mida on näitelausete automaatseks tuvastamiseks kasutatud. Tähelepanu on reeglipõhisel valemil töötaval tööriistal Good Dictionary Examples ehk GDEX, mida on seni eesti leksikograafias näitelausete automaatseks tuvastamiseks kasutatud. Neljandas peatükis kirjeldan GDEXi eesti keele mooduli eri versioone ning spetsiifilisi lause parameetreid, mis eesti keele sõnastike näitelauseid ja eesti keele õpikute lauseid (õpikulauseid) iseloomustavad. Keskendun versioonile GDEX 1.4 ning analüüsin selle väljundi evalveerimiseks läbi viidud hindamisülesande tulemusi. Tutvustan GDEX 1.4 abil genereeritud „Eesti keele õppekorpust 2018 (etSkELL)“ ning selle rakendamisvõimalusi keeleõppekeskkonna etSkELL ja keeleportaali Sõnaveeb näitel. Viiendas peatükis keskendun kitsaskohtadele, mis tulevad esile autentsete lausete kuvamisega lõppkasutajale, ning pakun võimalikke lahendusi. Kuuendas peatükis toon välja töö põhitulemused ja teen kokkuvõtte. Lisades on esitatud koondtabel väitekirja raames loodud kuue GDEXi eesti keele mooduli versiooni parameetritest, nende konfiguratsioonifailid ning must ja hall nimekiri.

1.1. Tänapäeva leksikograafia arengusuunad

Euroopas liigutakse aina enam traditsioonilisest leksikograafiast e-leksikograafia suunas [P2]. Traditsiooniline leksikograafia tähendab, et sõnastikke koostatakse lihtsamates tekstitöötlusprogrammides (nt Microsoft Wordis), sõnaartiklite koostamisel toetutakse sedelitele ning sõnastikke avaldatakse ainult paberil. E-leksikograafia tähendab, et sõnastikke koostatakse sõnastikusüsteemides tekstikorpuste põhjal ning avaldatakse peamiselt veebis. Tänapäeval räägitakse digipõhistest sõnastikest (*born-digital*), mis on loodud spetsiaalselt elektroonilise meediumi jaoks ning mis pakuvad uuenduslikke võimalusi leksikaalse info organiseerimiseks ja esitamiseks. (Kallas, Koeva jt 2019)

E-leksikograafiline töö eeldab tekstikorpuse, korpuspäringusüsteemi ja sõnastikusüsteemi olemasolu. Korpused on mahukad elektroonilised tekstikogud, mis on koostatud keeleteaduse, arvutilingvistika ja leksikograafia vajadusi silmas pidades. Korpusi töödeldakse spetsiaalse tarkvaraga ehk korpuspäringusüsteemiga, mis võimaldab korpusandmete mitmekülgset analüüsi. Eesti Keele Instituudis kasutatakse alates 2011. aastast põhiliselt korpuspäringusüsteemi

Sketch Engine (vt ka Kilgarriff, Rychlý jt 2004, Kilgarriff, Baisa jt 2014), mis on laialdaselt kasutusel ka Euroopa leksikograafide seas (Kallas, Koeva jt 2019). 2019. aasta seisuga on Sketch Engine'is kokku 15 eri tüüpi eesti keele korpust, millest suurim „Eesti keele ühendkorpus 2017" sisaldab umbes 1,3 miljardit sõnet. Ühendkorpus koosneb „Eesti keele koondkorpusest" (250 mln sõnet), sealhulgas tasakaalus korpusest (15 mln sõnet), ning eesti veebikorpustest[1] „Estonian Web 2013" (233 mln sõnet) ja „Estonian Web 2017" (763 mln sõnet). Vähemal määral kasutatakse instituudis ka korpuspäringusüsteemi KORP, kus on 2019. aasta seisuga erinevaid eesti korpusi kokku enam kui 850 miljoni sõne mahus.

Seoses mobiilse interneti tulekuga on inimesed võrguga pidevalt ühendatud. See on ka viimase kümne aasta jooksul leksikograafias kaasa toonud arusaamise, et pabersõnastikud kuuluvad pigem minevikku (Krek 2019: 115, Langemets, Tiits jt 2018). Euroopas avaldatakse praegu peaaegu pooled (46%) sõnastikest ainult elektrooniliselt (Kallas, Koeva 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" vihikuid ning väikeste murdesõnastike sarja. Senini on enamik eesti keele sõnastike veebiversioone olnud pabersõnastike täpsed koopiad. 2014. aastal ilmunud „Eesti keele põhisõnavara sõnastik" (Kallas, Tuulik 2011, Kallas, Tuulik, Langemets 2014) oli Eestis esimene, mille veebiversioonis kasutati paberkandjal esitatud infole lisaks kõnesünteesi, helifaile, hüperlinke ja navigeeritavaid pilte (Kallas, Langemets jt 2019). Väitekirja kirjutamise ajal Eesti Keele Instituudis arendamisel oleva sõnastiku- ja terminibaasisüsteemiga Ekilex (Tavast jt 2018)[2] liigutakse eri sõnakogude veebiversioonidest üheainsa (agregeeritud) leksikograafilise andmebaasi suunas, mis tavakasutajale avaldub keeleportaalis Sõnaveeb (Koppel, Tavast jt 2019).

E-leksikograafia areng on leksikograafia uue haruna kaasa toonud korpusleksikograafia (*corpus lexicography*) ning selle kitsama valdkonna automaatse leksikograafia (*automated lexicography*). Siinne väitekiri kuulubki nii korpusleksikograafia kui ka automaatse leksikograafia valdkonda. Korpusleksikograafia (Kilgarriff, Rychlý jt 2004) on interdistsiplinaarne lingvistika valdkond, mille eesmärk on luua meetodeid, mis võimaldavad analüüsida suurte tekstikorpuste põhjal sõnastike koostamiseks vajalikke andmeid. Automaatne leksikograafia (Gantar jt 2016, Kallas, Koeva jt 2019) keskendub meetoditele, mis võimaldavad korpusest leksikograafiliste infoüksuste ekstraheerimist ning nende alusel leksikograafiliste andmebaaside (pool)automaatset genereerimist.

---

[1] Veebikorpuse loomiseks kogutakse kokku kõik veebis leiduvad eestikeelsed tekstid, st kroolitakse eestikeelset veebi. Kroolimine (*crawling*) käib spetsiaalse tarkvara SpiderLing (Pomikalek, Suchomel 2012) abil. Kroolimise järel kodeeritakse tekst UTF-8 formaati, korpus puhastatakse ning sealt eemaldatakse duplikaadid. Eestikeelset veebi kroolib Eesti Keele Instituut koostöös tarkvarafirmaga Lexical Computing Ltd.

[2] Kuni 2019. aastani koostati sõnastikke Eesti Keele Instituudi sõnastikusüsteemis EELex (Langemets jt 2006, Järviste jt 2011). EELexis jõuti alates 2000ndate keskpaigast koostada umbes 70 sõnastiku- ja terminibaasi.

Korpusleksikograafia arengus oleme Eestis jõudnud tasemeni, kus korpuspäringusüsteemi Sketch Engine abil on võimalik suuri keelelisi andmeid väga kiiresti analüüsida, kuna lisaks statistilistele meetoditele on kasutusele võetud ka reeglipõhine lähenemine (Kilgarriff, Husák jt 2008). Sketch Engine võimaldab automaatselt koostada sagedusloendeid (Kilgarriff 2010a), leida statistilisi kollokaate, genereerida sõnavisandeid (*word sketch*) (Kilgarriff, Rychlý jt 2004: 105, Kilgarriff, Kovář jt 2010) ehk kokkuvõtteid sõna süntaktilisest ja kollokatiivsest käitumisest, koostada tesaurust (Rychlý, Kilgarriff 2007), leida definitsioone (Kovář jt 2016) ja termineid (*term extraction)* (Jakubíček jt 2014), aga tuvastada ka sõnastiku näitelauseks sobivaid korpuslauseid. Viimase jaoks on Sketch Engine'isse integreeritud reeglipõhisel valemil töötav tööriist Good Dictionary Examples ehk GDEX (Kilgarriff, Husák jt 2008). Eesti keele korpuspõhise analüüsi tarbeks on välja töötatud spetsiaalne moodul sõnavisandite (Kallas 2013) ja terminite (Kallas, Suchomel, Khokhlova 2017) jaoks. Eesti keele näitelausete tuvastamise moodul (GDEXi eesti keele moodul) on loodud käesoleva väitekirja raames.

Sketch Engine'it kasutatakse ka sõnastike andmebaaside (pool)automaatseks genereerimiseks. See eeldab vastava programmi olemasolu, mille abil ekstraheeritakse leksikograafilised üksused automaatselt tekstikorpusest sõnastikusüsteemi, kus leksikograaf neid edasi toimetab. 2018. aastal viidi leksikograafide ja leksikograafiaga tegelevate institutsioonide seas läbi üleeuroopaline küsitlus, milles osalenud 159 leksikograafist vastas sõnaartikli üksuste automaatse ekstraheerimise küsimustele 89. Küsitluse tulemusena selgus, et neist 20,8% tuvastab ja ekstraheerib tekstikorpusest automaatselt märksõnade loendeid, 12,7% kollokatsioone, 11,3% sagedusinfot, 8% mitmesõnalisi märksõnu, 7,5% näitelauseid, 6,1% sõnakuju variante, 4,7% süntaktilisi malle, 3,8% neologisme, 3,8% leksikaalsemantilisi suhteid, 4,4% infot domeeni kohta, 3,8% mitmekeelseid andmeid paralleelkorpustest, 3,3% definitsioone ja 2,4% helinäiteid kõnekorpusest. (Kallas, Koeva jt 2019) Siinse töö fookuses on näitelausete automaattuvastuse probleemistik.

Euroopas genereeritakse poolautomaatselt ligikaudu 31% ja täisautomaatselt umbes 7,5% sõnastike andmebaasidest (Kallas, Koeva jt 2019). Sõnastiku andmebaasi täisautomaatne genereerimine tähendab, et kõik sõnastiku üksused (märksõnastik, kollokatsioonid, näitelaused jm) ekstraheeritakse automaatselt korpuspäringusüsteemist sõnastikusüsteemi. Andmebaasi genereerimisele järgneb üldjuhul järeltoimetamine, mille käigus leksikograaf automaatselt loodud sisu kontrollib ja puhastab (n-ö poolautomaatne koostamine). Täisautomaatse koostamise korral korpusest genereeritud sisu ei toimetata. Selliselt on loodud näiteks SkELLi keeleõppekeskkondade[3] sari (Baisa, Suchomel 2014, Koppel, Kallas jt 2019). Euroopas on koostatud mitmeid sõnastikke, mille andmebaasi automaatsele genereerimisele on järgnenud järeltoimetamise faas: nt sloveeni leksikaalne andmebaas

[3] Seda tüüpi keskkondade jaoks ei ole juurdunud nimetust veel välja kujunenud ning viitan siinses töös SkELL-tüüpi liidesele kui *automaatsele keeleõppekeskkonnale*, kuigi see ei sisalda harjutusi, ei võimalda hinnata keeleoskustaset, saata tagasisidet jmt.

Slovene Lexical Database (Kosem, Gantar, Krek 2013), sloveeni keele tesaurus Sopomenke 1.0 (Krek jt 2017, Holdt jt 2018), sloveeni kollokatsioonisõnaraamat „Kolokacijski Slovar Sodobne Slovenščine“ (Kosem, Krek jt 2018), inglise keele leksikaalne andmebaas DANTE (Kilgarriff 2010b), „Macmillan Collocations Dictionary for Learners of English“ (MCD 2010) ja suur hollandi keele sõnaraamat „Algemeen Nederlands Woordenboek (ANW)“ (Tiberius, Schoonheim 2016).

Eestis on poolautomaatselt koostatud „Läti-eesti sõnaraamat“ (2015) ja „Eesti keele naabersõnad 2019“. Esimest korda rakendati Eestis näitelausete automaatset tuvastust 2014. aastal „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi täisautomaatsel genereerimisel (Kallas, Koppel, Tuulik 2015). Alates 2019. aastast koostatakse Eesti Keele Instituudis eri allikate alusel täisautomaatselt sünonüümide andmebaasi. Samuti kuvatakse 2019. aastast alates keeleportaalis Sõnaveeb automaatselt tuvastatud näitelauseid (portaalis nimetusega *veebilaused*) juba ka otse sõnastiku kasutajatele. Sõnaveebis ilmneb veebilausete kasulikkus eriti siis, kui leksikograafi koostatud näitelaused puuduvad (vt lähemalt ptk 4.4.4).

Näitelausete automaatne tuvastamine on eesti leksikograafias suhteliselt uus praktika. Sellist tüüpi uurimistöö järele tekkis suur vajadus seoses automaatse leksikograafia saabumisega Eestisse. Siinkirjutajale teadaolevalt on Eestis uuritud vaid mees- ja naissoo kujutamist õigekeelsussõnaraamatu „Eesti õigekeelsussõnaraamat ÕS 2013“ näidetes (Raadik 2016), kuid automaatse leksikograafia eesmärkidel sõnastiku näitelauseid seni uuritud ei ole.

1.2. Uurimisobjekt, -meetod ja analüüsimaterjal

Väitekirja uurimisobjekt on tunnused ehk parameetrid, mis iseloomustavad eesti (õppe)sõnastike näitelauseid ja eesti keele kui teise keele õpikute lauseid ning mille alusel saab korpusest automaatselt tuvastada sõnastiku näitelauseks sobivaid korpuslauseid. Näitelausena tuvastatav korpuslause peab vastama ortograafilistele nõuetele: algama suure tähega ja lõppema lauselõpumärgiga (Erelt, Metslang 2017: 87). Korpus, eriti selline, mis sisaldab ohtralt veebitekste ja kust näitelause kandidaate automaatselt tuvastatakse, sisaldab ka palju teist tüüpi, n-ö mitteortograafilisi lauseid. Enamasti on need pärit blogi- või foorumipostitustest, kus sageli ei järgita õigekirjareegleid, või laused, mille piiride automaatse tuvastamisega ei ole (osa)lausestaja saanud hakkama.

Näitelausete automaatseks tuvastamiseks on Eesti Keele Instituudis seni kasutatud korpuspäringusüsteemi Sketch Engine integreeritud tööriista Good Dictionary Examples ehk GDEX. GDEXi keskmes on universaalne reeglipõhine valem, mida paremate tulemuste saavutamiseks täiendatakse keelespetsiifiliste parameetritega. GDEX valib nende eeldefineeritud parameetrite abil kõikidest korpuslausetest välja vaid need, mis oma struktuuri ja sisu poolest leksikograafiliseks analüüsiks kõige paremini sobivad. Olen arendanud GDEXi eesti keele moodulit reeglipõhist lähenemist kasutades, kuid parameetrite peenhäälestamiseks

osaliselt kasutanud ka masinõppe elemente (meetoditest lähemalt ptk 3.1) ning loonud analüüsi tulemusi arvestades eesti keele moodulile kuus erinevat versiooni (Kallas, Koppel, Tuulik 2015, [P1], [P3], [P5]).

Peamise analüüsimaterjalina olen kasutanud Eesti Keele Instituudis koostatud sõnastikke: „Eesti keele sõnaraamat 2019“, „Eesti keele põhisõnavara sõnaraamat“ (2014), „Eesti keele naabersõnad 2019“ (eelnevalt: eesti keele kollokatsioonisõnastik, vt Kallas, Koppel, Tuulik 2015); samuti „Eesti keele A1–C1 õpikute korpust 2018“.[4] Väitekirja artiklid keskenduvad sõnastike näitelausete ja eesti keele õpikulausete parameetritele, mille alusel õppesõnastikku sobivaid näitelauseid automaatselt tuvastada. Õppesõnastike sihtgrupina on siinse väitekirja raames peamiselt silmas peetud B2–C1-keeleoskustasemel eesti keele kui teise keele valdajaid[5], kuna GDEXi eesti keele mooduli arendamine on käinud paralleelselt nendele suunatud „Eesti keele naabersõnade 2019“ sõnastiku koostamisega. Sõnastike näitelausete ja õpikulausete analüüsi põhjal olen kindlaks teinud hea näitelause n-ö kuldstandardi ehk välja selgitanud, millised parameetrid nimetatud sõnastike näitelauseid ja õpikulauseid iseloomustavad. Teise analüüsimaterjalina olen kasutanud eesti keele korpuseid („Eesti keele ühendkorpus 2013“ (563 mln sõnet), „Eesti keele ühendkorpus 2017“ (1,3 mld sõnet)), mille peal olen GDEXi eesti keele mooduli eri versioone testinud. Versioonide testimiseks on Sketch Engine'i külge loodud spetsiaalne kasutajaliides GDEX Editor (vt ptk 4.1.3), mille abil saab kahte erinevat versiooni omavahel võrrelda ning lihtsasti välja selgitada, millised on lausete kvaliteeti kõige rohkem mõjutavad parameetrid.

GDEXi eesti keele moodulit on seni rakendatud kahe õppekorpuse loomiseks: versiooni 1.3 kasutati „Eesti keele ühendkorpuse 2013“ põhjal õppekorpuse „EstonianNC GDEX“[6] loomiseks [P1] ja versiooni 1.4 [P3] kasutati „Eesti keele ühendkorpuse 2017“ põhjal „Eesti keele õppekorpuse 2018 (etSkELL)“ (294 mln sõnet) loomiseks. Õppekorpuse all pean siinses töös silmas korpust, mis on loodud pedagoogilistel eesmärkidel ning mida sobib kasutada nii keele õpetamisel kui ka õppimisel. „EstonianNC GDEX“ ja „Eesti keele õppekorpus 2018 (etSkELL)“ on esimesed spetsiaalsed autentseid lauseid sisaldavad eesti keele õppekorpused. „Eesti keele õppekorpus 2018 (etSkELL)“ on omakorda allikaks automaatselt loodud keeleõppekeskkonnale etSkELL ehk Sketch Engine for Estonian Language Learning ja veebilausetele keeleportaalis Sõnaveeb. Töö edasiarendusena on plaan luua õppekorpused ka eri keeleoskustasemetele. Selleks olen välja töötanud GDEXi eesti keele mooduli versioonid üldistele keeleoskustasemetele ehk A-, B- ja C-tasemele [P5].

Kuna GDEXi eesti keele mooduli versiooni 1.4 [P3] on rakendatud õppekorpuse loomisel, mida omakorda on kasutatud keeleõppekeskkonnas etSkELL

---

[4] Õpikute korpus sisaldab eesti keele kui teise keele õpikutest pärinevaid täislauseid (loe lähemalt [P5]).

[5] Keeleoskustasemete eristamisel toetutakse Euroopa keeleõppe raamdokumendile (Raamdokument 2007), mis eristabki kolme üldist keeleoskustaset (A-, B- ja C-tase) ning kuut alltaset (A1-, A2-, B1-, B2-, C1-, C2-tase).

[6] Korpus on kättesaadav Sketch Engine'i arhiivis.

ja keeleportaalis Sõnaveeb, olen selle väljundit evalveerinud (lähemalt ptk 4.3). Evalveerimiseks kasutasin avatud lähtekoodiga platvormi Pybossa, mida tavaliselt kasutatakse eri tüüpi rahvahanke (*crowdsourcing*) projektide läbiviimiseks.

1.3. Töö eesmärgid

Väitekirja eesmärgid võib tinglikult jagada teoreetilisteks ja rakenduslikeks. Teoreetiline eesmärk on saada ülevaade sõnastiku hea näitelause tunnustest; rakenduslik eesmärk on eesti keele jaoks välja töötada meetod, mis korpusest näitelauseid automaatselt tuvastab. Täpsemad eesmärgid on järgmised.

- Saada ülevaade hea näitelause tunnustest nii traditsioonilise teoreetilise leksikograafia kui ka korpusleksikograafia seisukohalt.
- Välja selgitada hea näitelause formaalsed parameetrid eesti keelele, võttes aluseks eri eesti keele sõnastike näitelausete ja eesti keele kui teise keele õpikute lausete analüüsi tulemused.
- Anda ülevaade meetoditest, mida kasutatakse näitelausete automaatseks tuvastamiseks.
- Luua GDEXi eesti keele mooduli eri keeleoskustasemetele suunatud versioonid, mis hea näitelause formaalseid parameetreid arvestades tuvastavad automaatselt korpusest sobivad näitelause kandidaadid.
- Luua GDEXi eesti keele mooduli versiooni 1.4 abil õppekorpus, mille sihtgrupp on eesti keelt B2–C1-oskustasemel valdaja.
- Evalveerida GDEXi eesti keele mooduli versiooni 1.4 väljundit, mille laused sobivad eesti keele B2–C1-oskustasemele.
- Vaadelda korpuslausete kasutusvõimalusi (õppe)leksikograafias.
- Rakendada loodud õppekorpust keeleõppekeskkonnas etSkELL ja Eesti Keele Instituudi keeleportaalis Sõnaveeb.

1.4. Ülevaade väitekirja publikatsioonidest

Artikli „Õppijasõbralik korpuslause: automaatse valiku võimalusi“ [P1] kaasautor on Jelena Kallas (Eesti Keele Instituut). Olen kirjutanud artikli põhiosa, sissejuhatus ja kokkuvõte on valminud ühiselt. Artikkel annab ülevaate korpuslausete kasutusvõimalustest õppeleksikograafias ning esitleb nüüdisaegset keeleõppeportaali SkELL ehk Sketch Engine for Language Learning, mis kasutab ainult GDEXi abil välja valitud korpuslauseid. [P1] tutvustab GDEXi eesti keele mooduli versiooni 1.3 ning analüüsib selle abil loodud esimese autentseid lauseid sisaldava õppekorpuse „EstonianNC GDEX“ lauseid.

Artikkel „Identification and automatic extraction of good dictionary examples: the case(s) of GDEX“ [P2] on kirjutatud koos Iztok Kosemi (Ljubljana ülikool), Jan Michelfeiti (Lexical Computing Ltd.), Carole Tiberiuse (Hollandi Keele Instituut) ja Tanara Zingano Kuhniga (Coimbra ülikool). Olen artikli GDEXi eesti keele mooduli peatüki autor, samuti olen panustanud artikli sissejuhatuse ja

kokkuvõtte kirjutamisse. Artikkel arutleb hea näitelause tunnuste üle ning annab ülevaate erinevatest leksikograafia- ja keeleõppeprojektidest, kus kasutatakse automaatselt tuvastatud näitelauseid. [P2] kirjeldab tööriista GDEX arendamise ajalugu ja tööpõhimõtteid ning nelja erineva keele (sloveeni, hollandi, eesti, portugali) GDEXi mooduleid ning nende keelespetsiifilisi parameetreid.

Artiklis „Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks“ [P3] analüüsin „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi näitelauseid. Andmebaasi ekstraheeriti GDEXi versiooniga 1.2 iga kollokatsiooni kohta viis korpuslauset, mille seast valis leksikograaf toimetamise käigus välja ühe. Koostasin naabersõnade sõnastiku näitelausetest artikli tarbeks omakorda kaks andmebaasi, millest ühte kuulusid nn head näitelaused ja teise nn halvad näitelaused. Heade näitelausete andmebaasi liigitusid need laused, mille sõnastiku koostajad olid viie ekstraheeritud korpuslause seast kollokatsiooni näitelauseks valinud; halbade näitelausete andmebaasi liigitusid ülejäänud, valituks mitteosutunud laused. Andmebaaside analüüsi tulemustele toetudes lõin GDEXi eesti keele mooduli versiooni 1.4, mille abil tuvastatud lauseid kasutab automaatne keeleõppekeskkond etSkELL ning keeleportaal Sõnaveeb (2019. aasta seisuga).

Artiklis „Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks“ [P4] kirjeldan GDEX 1.4 väljundi evalveerimise tulemusi. Hindajateks olid Eesti Keele Instituudis töötavad leksikograafid ja Tartu ning Tallinna Ülikooli eesti keelt kui teist keelt B2–C1-oskustasemel valdavad üliõpilased. Evalveerimise viisin läbi kahe ülesande käigus: esimese hindamisülesande eesmärk oli välja selgitada, kui suur hulk GDEX 1.4 poolt valitud korpuslauseid hinnatakse sobivaks näitelause kandidaadiks ning kui suur hulk GDEX 1.4 poolt kõrvale jäetud (välja filtreeritud) korpuslauseid hinnatakse sobimatuks näitelause kandidaadiks. Esimesele hindamisülesandele järgnenud jätkuküsitluse raames kogusin hindajate põhjendusi sellele, miks nad teatud lauseid sobivaks või sobimatuks hindasid. Evalveerimine näitas, et 85% GDEX 1.4 abil välja valitud näitelausetest hinnati sobivateks ning 94% GDEX 1.4 poolt kõrvale jäetud (välja filtreeritud) lausetest hinnati sobimatuteks näideteks.

Artiklis „Eesti keele kui teise keele õpikute lausete analüüs ja selle rakendamine eri keeleoskustasemete sõnastike näitelausete automaatsel valikul“ [P5] analüüsin eesti keele kui teise keele õpikute lauseparameetreid ning kirjeldan analüüsi tulemusena üldistele keeleoskustasemetele (A-, B- ja C-tasemele) loodud GDEXi eesti keele mooduli versioone. Kuigi olen GDEXi eesti keele moodulit algusest peale arendanud eesti keele kui teise keele õppijaid silmas pidades (eelkõige B2–C1-taset), siis ei olnud varasemalt sobiva andmestiku puudumise tõttu võimalik arvestada keeleoskustasemetele spetsiifilisi lauseparameetreid. 2018. aastal loodud „Eesti keele A1–C1 õpikute korpus (2018)“ võimaldas esmakordselt välja selgitada, millised need on. Selleks analüüsin eri keeleoskustaseme õpikulauseid Eesti Keele Instituudis loodud teksti märgendamise ja statistilise analüüsi tööriista „Lause parameetrite analüsaator“ abil. Samuti kirjeldan artiklis SkELLi-sarja kuuluvat eesti keelele loodud keeleõppekeskkonda etSkELL, mis kasutab spetsiaalset GDEX 1.4 abil loodud õppekorpust.

2. KORPUSLAUSE NÄITELAUSE ALLIKANA

Selles peatükis kirjeldan esmalt sõnastiku näitelausete tüüpe ja nende funktsioone. Seejärel seletan, mis otstarvet näitelause sõnaartiklis täidab ning annan ülevaate hea näitelause tunnustest traditsioonilise keeleteaduse ja korpusleksikograafia ning automaatse leksikograafia seisukohalt. Samuti kirjeldan eri keelte näitel, kuidas korpuslauseid on kasutatud (õppe)leksikograafias ning keeleõppes üldisemalt.

2.1. Sõnastiku näitelause tüübid, funktsioonid ja valiku põhimõtted

Lause on keelelise suhtluse põhiüksus (Erelt, Metslang 2017: 53). Tüüpiline lause sisaldab finiitset ehk pöördelist tegusõnavormi ning selle juurde kuuluvaid fraase. Lause tähistab mingit situatsiooni ning täidab erinevaid funktsioone (nt semantilisi, pragmaatilisi), milleks kasutab leksikaalseid ja grammatilisi vahendeid.

Sõnastiku näitelause tüüpide eristamisel toetun kahele rahvusvaheliselt tunnustatud akadeemilisele leksikograafiaalasele tervikkäsitlusele: Oxford University Pressi raamatule „The Oxford Guide to Practical Lexicography“ (Atkins, Rundell 2008) ja Cambridge University Pressi raamatule „A Handbook of Lexicography. The Theory and Practice of Dictionary-Making“ (Svensén 2009). Näitelause võib olla kahte tüüpi: selline, mis illustreerib midagi, mida on mujal sõnaartiklis juba mainitud (nt kirjeldab mingi grammatilise vormi kasutust); või selline, mis lisab sõnaartiklisse uut informatsiooni, öeldes kasutajale midagi, mida ei ole mujal sõnaartiklis esitatud (Atkins, Rundell 2008: 225). Päritolu poolest võib näitelause olla kas autentne (*authentic example*) või tehislik (*invented example*) ehk leksikograafi koostatud. Autentne lause võib omakorda olla kas täiesti autentne või autentse lause toimetatud versioon (*adapted example*) (Svensén 2009: 283, Atkins, Rundell 2008: 225). Kui varasemalt kasutati autentsete lausete allikana sõnasedelitest koosnevaid kartoteeke, siis tänapäeva leksikograafias on autentse lause allikas mahukas tekstikorpus, mistõttu kasutan siinses töös sünonüümselt mõisteid *korpuslause* ja *autentne lause*.

Selle üle, kas korpuslause sobib sõnastiku näitelauseks, on arutletud alates 1980ndate lõpust, kui ilmus esimene täielikult korpuspõhine inglise keele sõnastik COBUILD. Sellele järgnes äge debatt – kas näitelaused peaksid olema leksikograafide poolt välja mõeldud või peaksid need olema pärit autentsetest tekstidest. Hulk autoreid (vt nt Rundell 1998: 334–335, Kilgarriff 2013, Simpson 2003: 269, Svensén 2009: 284) on väljendanud seisukohta, et sõnastikes tuleb eelistada autentseid lauseid leksikograafi poolt koostatud lausetele. Varem, kui sõnastikke avaldati vaid paberil, eelistati ruumipuuduse tõttu tehislikke näiteid, kuna leksikograaf mõtles sageli välja sellise lause, mis täitis samaaegselt mitut ülesannet (Cowie 1978: 131). Eriti eelistati tehislikke lauseid õppesõnastikes (Atkins, Rundell 2008: 455–456). Gwyneth Fox (1987: 138–144, viidatud Atkins, Rundell

2008 kaudu) seadis kahtluse alla emakeelse kõneleja (leksikograafi) oskuse produtseerida loomulikke näitelauseid. Foxi arvates on tehislikud näitelaused sageli liiga isoleeritud ja iseseisvad, kuna ühte lausesse üritatakse mahutada võimalikult palju informatsiooni. Fox oli seisukohal, et kui korpuses leiduvat päris keelekasutust kasutatakse sõna kasutusmustrite väljaselgitamiseks, siis oleks kummaline näitelauseid tehislikult välja mõelda, selle asemel et korpusest autentseid lauseid võtta. John Sinclair (1987) peab tehislikke lauseid pigem definitsiooni osaks. Tema sõnul ei saa (sõna)kasutust välja mõelda, vaid seda saab ainult (korpuse põhjal) registreerida.

Näitelause on sõnaartikli oluline üksus, mis peab täitma selget eesmärki ning lisama sõnaartiklile väärtust. Traditsiooniliselt koostatakse sõnastikke üht sihtgruppi silmas pidades ja kogu sõnastiku sisu peaks vastama selle sihtgrupi vajadustele. Ka see, milline on hea näitelause, sõltub (sõnastiku)projekti tüübist ja sihtgrupist. Emakeelsele kasutajale suunatud ükskeelses sõnastikus illustreerivad näitelaused sõnakasutust ja täiendavad sageli definitsiooni. Hästi valitud näitelause aitab vahet teha polüseemsete sõnade erinevatel tähendustel. (Atkins, Rundell 2008: 454, 461) Keeleõppijale suunatud sõnastikes peaksid näitelaused illustreerima nii sõna kasutust tavapärases kontekstis kui ka sõna süntaktilist ja kollokatiivset käitumist, registrit jmt (Atkins, Rundell 2008, Zöfgen 1986, Harras 1989, Laufer 1992). Isegi sellise lihtsa sõna nagu *televiisor* (*television*) kohta on keeleõppijal oluline teada, et seda saab *sisse lülitada* (*turn on*) ja *välja lülitada* (*turn off*) ja et seda *vaadatakse* (*watch*), mitte ei *nähta* (*see* või *look at*) (Fox 1987: 137, viidatud Atkins, Rundell 2008 kaudu). Keeleõppijad otsivad sõnastikest sageli näitelauseid lootuses leida sealt selline näide, mis on sarnane lausele, mida ta on varem kuulnud või lugenud; või mis kinnitaks, et teatud sõna võib just niimoodi kasutada. Seetõttu peaksid õppesõnastike näitelaused olema just sellised, millistega keeleõppijad tõenäoliselt igapäevaelus kokku puutuvad, ehk autentsed. Lisaks on näitelausete abil võimalik õppijale edasi anda grammatilist infot, näiteks et teatud sõna kasutatakse tavaliselt ainult ainsuses või mitmuses. (Bowker 2010: 164)

Atkins ja Rundell (2008: 37, 225, 454) osutavad, et õppesõnastikes on näitelaused vajalikumad kui emakeelsetele kasutajatele suunatud ükskeelsetes sõnastikes, mistõttu on tavaks anda neis näitelauseid rohkem, aitamaks paigutada tundmatut sõna keeleõppija passiivsesse (ja aktiivsesse) sõnavarasse.[7] Kuigi definitsioon ja näitelause peaksid ideaalis olema iseseisvad üksused, saab paljudel juhtudel sõna tähendus selgeks alles siis, kui näitelauset lugeda, ning vahel võibki sõnaartikkel ilma näitelauseta jääda arusaamatuks. Ana Frankenberg-Garcia uurimused (2012, 2014) on näidanud, et kohati on uue sõna mõistmisel kasutusnäidetest rohkem abi kui definitsioonist, ning kuna näitelause jääb sageli õppija ainsaks kokkupuuteks uue sõnavara ja grammatikaga, siis soovitab ta ühe sõna või kasutusmustri kohta esitada näitelauseid rohkem kui ühe. Samal arvamusel on ka teised autorid. Robert Lew'i ja Arleta Adamska-Sałaciaki (2015) sõnul on

[7] Passiivsest sõnavarast saab inimene aru, aga ise ei kasuta. Aktiivset sõnavara inimene tunneb ja oskab kasutada. (Richards, Schmidt 2013)

kasulik just ükskeelsetesse õppesõnastikesse lisada illustratiivseid näitelauseid (*illustrative quotation*), mille eesmärk on kirjeldada definitsiooni, kuna võõrkeelseid definitsioone on keeleõppijal sageli raske mõista. Sean Michael Burke'i (2003: 247) väitel on ka sõnastike emakeelsetele kasutajatele kasulik, kui näitelauseid on mitu, kuna need aitavad mõista sageli abstraktseks jäävaid definitsioone. John Simpsoni (2003: 268–269) järgi võib isegi kogenud leksikograafil vahel olla raske ilma näitelauseta eristada kahte sarnast definitsiooni. Ka Bo Svensén (2009: 284) leiab, et isegi kui autentsed laused on tehislikest näidetest enamasti palju pikemad, tuleks sõna erinevate kasutusmustrite ilmestamiseks esitada rohkem (autentseid) näitelauseid. Ka eesti keele kui teise keele õppijate hinnangul võiks sõnastik sisaldada rohkem näitelauseid, mis aitaksid illustreerida sõna erinevaid kasutusviise (Teral 2015: 122).

Keeleõppijate seas on tehtud mitmeid eksperimente (Hubbard jt 1986, Cobb 1997, Baicheng 2009, Tolmachev, Kurohashi 2017) näitelausete kasulikkuse kohta: tulemused on näidanud, et uute sõnade õppimisel ja nende meelde jätmisel on näitelausetest väga suur abi. Zhang Baichengi (2009) eksperiment näitas, et keeleõppijale jääb uus sõna paremini meelde siis, kui ta peab õpitavale sõnale ise näitelauseid välja mõtlema, ning halvemini siis, kui ta kohtub uue sõnavaraga õpetaja poolt valitud näitelausetes. Arseny Tolmachevi ja Sadao Kurohashi (2017) eksperiment näitas, et tuleb kasuks, kui keeleõppijatele kuvatavad näitelaused on semantiliselt, leksikaalselt ja grammatiliselt võimalikult mitmekesised, nii et need illustreeriksid sõna igat (all)tähendust. Chieh-Yang Huangi ja Lun-Wei Ku (2016) vaatluse tulemused näitasid, et keeleõppijad on võimelised näitelausete abil kahe semantiliselt lähedase sõna (lähisünonüümi) kasutuskontekste võrreldes keelt kaudselt õppima.[8]

Kakskeelsete õppesõnastike kohta arvamused lahknevad. Taku Kaneta (2011) sõnul kakskeelsed sõnaartiklid tingimata näitelauseid ei vaja, kuna nende peamine eesmärk on aidata lähte- ja sihtkeele märksõnu dekodeerida. Jorge Lázaro jt (2017) eksperiment näitas jällegi, et kui lähte- ja sihtkeelel on erinev lingvistiline süsteem, siis peaksid kakskeelsed sõnastikud sisaldama just sihtkeele näitelauseid, kuna need aitavad paremini mõista sõna tähendust. Atkinsi ja Rundelli (2008: 506) sõnul täiendavad näitelaused aktiivse suunitlusega kakskeelses sõnastikus tõlgetega edasi antud infot ning nende eesmärk on aidata lähtekeele valdajatel valida õigeid sihtkeele vasteid ning neid korrektselt kasutada. Ka Mike Hannay (2003: 151) leiab, et kakskeelsed õppesõnastikud peaksid sisaldama autentseid näitelauseid. Adamska-Sałaciak (2013: 226–227) on jällegi seisukohal, et see, kas näitelaused on autentsed või tehislikud, oleneb sõnastiku kasutaja keeleoskustasemest. Kuna kakskeelseid sõnastikke kasutavad erineva keeleoskustasemega inimesed, ei saa näitelaused Adamska-Sałaciaki sõnul alati olla täielikult autentsed, vaid neid tuleb toimetada. Tüüpiliselt ei otsi kakskeelse sõnastiku kasutaja infot mitte

[8] Eristatakse kahte tüüpi õppimist: otsest (*explicit learning*) ja kaudset õppimist (*implicit learning*). Otsene õppimine toimub teadlikult (hõlmab nt grammatika õppimist ja vigade parandust). Kaudne õppimine toimub enesele teadvustamata (nt keelekeskkonnas viibimise teel). (DeKeyser 2008: 314, 321)

oma emakeele kohta, vaid sihtkeele kohta, ja seega peaksid sihtkeele näitelaused olema kohased võõrkeelse kasutaja keeleoskustasemele.

Vahekokkuvõtteks võib öelda, et igal sõnastikul on oma põhimõtted, mille järgi näitelauseid valitakse või koostatakse. Üht tüüpi, harilikult passiivsetes sõnastikes kasutatakse näitelauseid ainult märksõna ja selle alltähenduste illustreerimiseks; teist tüüpi, harilikult aktiivsetes sõnastikes pakub näitelause aga hoopis grammatilist tuge. Kui sõnastiku sihtgrupp on emakeelne kasutaja, võivad näitelaused sisaldada haruldast sõnavara ning olla keerulise grammatilise struktuuriga. Kui sõnastiku sihtgrupp on keeleõppija, peaksid laused olema tasemekohased, näiteks peaksid alg- või kesktasemele suunatud näitelaused olema lühemad, sisaldama sagedasemat sõnavara ega tohiks olla grammatiliselt keerukad.

Kuigi Atkinsi ja Rundelli (2008: 457–458) järgi on ideaalne näitelause võetud otse korpusest toimetamata kujul, on isegi tänapäevastest üle miljardi sõna suurustest korpustest raske leida sellist näidet, mis toimetamata kujul kõigile hea näitelause tunnustele vastaks. Tavaliselt sobib näitelauseks mingi korpuslause osa, selle keskne tuumik (nt 4–6 sõna), mis illustreerib kõige paremini sõna tüüpilist kasutuskonteksti. Sageli on autentseid lauseid vaja enne (õppe)sõnastikku lisamist vähemal või rohkemal määral toimetada. Kõige harilikumad redigeerimise strateegiad ongi korpuslause lühendamine (nt osalause väljajätt), segava pärisnime või pronoomeni muutmine ja keerulise sõnavara lihtsustamine. Autentsete lausete toimetamine on Atkinsi ja Rundelli (2008) sõnul õigustatud küll juhul, kui tegemist on mitteemakeelsele sihtgrupile suunatud sõnastikuga. Õppesõnastikes eelistatakse sageli just lühikesi lauseid, aga kuna nende eesmärk on toetada teksti loomist, siis soovitatakse samas, et laused sisaldaksid rohkelt konteksti. Autentset lauset lühendades kaotab see oma loomulikkuse ning ilma piisava kontekstita pole laused piisavalt informatiivsed. Aga isegi juhul, kui leksikograaf on otsustanud korpusest leitud lauset lühendada või muul moel toimetada, tuleb need korpusest esmalt üles leida. (Loe lähemalt [P2].)

2.2. Sõnastiku näitelause tunnused

Leksikograafias kirjeldatakse head sõnastiku näitelauset kõige sagedamini kui loomulikku, tüüpilist, informatiivset ja arusaadavat (Harras 1989, Atkins, Rundell 2008, Kilgarriff, Husák jt 2008: 426). Tüüpiline näitelause sisaldab sõna sagedasi ja levinud süntaktilisi ja kollokatiivseid kasutusmustreid. Näitelause aitab avada sõna tähendust, kusjuures oluline on, et selles kajastatav informatsioon ei satuks konflikti definitsioonis öelduga. Eriti oluline on see õppesõnastikes, kus keeleõppija peab esmalt töötlema definitsiooni kaudu saadud infot ning näitelause ei tohiks seda infot ümber lükata. Näiteks ei tohiks ingliskeelset fraasi *common cold*, mis on defineeritud kui ‘tavaline külmetus, mida inimesed sageli põevad’, illustreerida vasturäägiva näitelausega *A common cold could kill her* ‘Tavaline külmetus võib ta tappa’. Loomulikkus on pigem intuitiivne kui objektiivne mõõde, mida siiski on võimalik sõnastikus tagada, kui järgida nt kolligatsiooni: sõna kalduvust esineda lauses teatud grammatilises mustris, näiteks kindlas ajas, arvus,

kõneviisis vmt. Loomulikkusele aitab kaasa ka see, kui lause ei sisalda idiolekte ja on võimalikult üldkeelne. Samuti peaks loomulik näitelause järgima ainult ühte registrit, nii ei tohiks kõnekeelne lause sisaldada ametliku keelekasutuse sõnu. Korpuslaused kipuvad sageli pakkuma vähem konteksti kui tarvis, või on vastupidi üle koormatud deiktiliste ja anafoorsete viidetega inimestele või asjadele lausest väljaspool. Informatiivsuse tagamisel on oluline leida tasakaal liiga lühikese (konteksti puudumise) ja liiga pika (liigse kontekstiga) lause vahel. Piisav hulk konteksti aitab kaasa ka lause loomulikkusele. Arusaadav lause ei sisaldada keerulist sõnavara ega tarindeid, segavaid või keerulisi nimesid. (Atkins, Rundell 2008: 459–461)

Ka korpuslause saab vastata eelnevalt kirjeldatud hea näitelause tunnustele. Korpuslausete loomulikkuse tagab see, et need on autentsed ehk pärinevad reaalsest keelekasutusest. Tüüpilisuse tagab see, kui korpuslause illustreerib sõna enamlevinud kasutust konteksti, süntaksi, fraseoloogia jms kohalt. Informatiivsuse tagab see, kui korpuslause on iseseisev ning selle sisu on arusaadav ka ilma laiema kontekstita. Arusaadavuse tagab see, kui korpuslaused ei ole liiga pikad, ei sisalda keerulisi süntaktilisi mustreid ega haruldast või erialast sõnavara. [P2] Korpuslause informatiivsust saab programm mõõta lause pikkusele toetudes. Kui see on liiga lühike, võib lause mõistmiseks kontekstist puudu jääda; kui see on liiga pikk, siis peab lausest arusaamiseks tegema palju tööd. Lisaks on väga pika lause struktuur ja sõnavara tõenäoliselt keerukam. Tüüpilisust aitab tagada see, kui korpuslauseid välja valiv programm eelistab lauseid, mis sisaldavad sagedasi kollokatsioone või süntaktilisi mustreid.

Ka eesti leksikograafias toetutakse juba üsna pikka aega näitelausete valikul korpuse andmetele. Erandiks võib pidada 2014. aastal ilmunud „Eesti keele põhisõnavara sõnastikku“, mis on küll korpuspõhine, kuid selle näitelaused on koostanud leksikograafid, kasutades korpuses sageli esinevaid kollokatsioone. Kogu sõnastikus (definitsioonides, näitelausetes, kollokatsioonides, õppekommentaarides) kasutatav sõnavara on piiratud märksõnastikus oleva 5000 sõnaga. „Eesti keele põhisõnavara sõnastiku“ näitelausete eesmärk on muuhulgas aidata sõna tähendusi paremini mõista. (Kallas, Koppel, Tuulik 2014) „Eesti keele sõnaraamatu 2019“ näitelausete eesmärk on toetada definitsiooni, kuid selle näitelaused ei pruugi vahetult sobida teist tüüpi sõnastikesse, näiteks õppesõnastikesse. Näidete valikul on toetutud süntagmaatilistele funktsioonidele (konstruktsioonidele, rektsioonidele, kollokatsioonidele), peale selle on püütud vältida hinnangulisust ning säilitada neutraalsust. Kasutusnäitena on sageli kasutatud ka lühemaid fraase ja kollokatsioone. (Langemets, Tiits jt 2018: 950–951) „Eesti keele naabersõnade 2019“ sõnastiku (Kallas, Koppel, Tuulik 2015) näitelaused on suuremalt jaolt täiesti autentsed. Teatud juhtudel on korpuslauseid lihtsustatud ja lühendatud, kuid tehislikke lauseid sisaldab see minimaalselt.

2.3. Korpuslause keeleõppes

1990ndatel tehti mitmeid eksperimente, kus selgus, et korpuspõhine ehk korpuste abil toimuv õpe on õpilastele meelepärasem kui õpikuid ja grammatikaid kasutavad traditsioonilised meetodid (Johns 1991). Korpuse ainesel põhinev õppimine stimuleerib õpilasi, esitab neile suuremaid väljakutseid, tekitab uudishimu, mõjub motiveerivalt ning on efektiivne viis grammatikaga tutvuda ja sõnavara suurendada (Leech 1997, Aston 1997, Dodd 1997, Gavioli 2005). Ka hilisemad uurimused (Frankenberg-Garcia 2012, 2014) on toetanud korpuste kasutamist keeleõppes. Õppijad teevad korpusmaterjaliga töötades keele kohta ise järeldusi. Laused, mis sisaldavad vihjeid kontekstile, toetavad uute sõnade tähenduse mõistmist, ning kollokatsioone ja süntaktilisi mustreid sisaldavad laused aitavad parandada vigu, mida teist keelt õppides tüüpiliselt tehakse.

Siiski ei saa eeldada, et keeleõppija oskab iseseisvalt korpusest leida üles vajamineva info – tavaline sõnaotsing korpusest võib olla töömahukas, tuua kaasa müra ega pruugi õiget vastust anda (Kilgarriff, Husák jt 2008, Kilgarriff 2009, Kilgarriff, Marcowitz jt 2015). Konkordantside[9] lugemine on edasijõudnud lingvistiline oskus ja keeleõppijate enamikule liiga raske ülesanne. Pealegi on konkordantside lugemise peamine eesmärk (üles noppida kõige tavalisemad kasutusmustrid, milles märksõna esineb) juba iseenesest abstraktne ja keeruline ülesanne. Vaid edasijõudnud ja tugevalt motiveeritud õppijad võivad konkordantside lugemisest kasu saada. Seevastu Atkins ja Rundell (2008: 457) on öelnud, et keeleõppes puututaksegi kokku igat tüüpi keelekasutuse näidetega, ka sellistega, mis on ebaloomulikud.

Kilgarriff, Marcowitz jt (2015) on välja pakkunud kaks viisi, kuidas korpusi keeleõppijatele tutvustada. Esimene on need n-ö sõnastikuks maskeerida. Selleks tuleb korpusmaterjal esitada sellisena, nagu see oleks sõnastikuinfo. Korpused ja sõnaraamatud on mõlemad keeleressursid, mis paiknevad sama skaala erinevates otstes. Korpused ei kirjelda keelt, vaid näitavad, kuidas seda päriselt kasutatakse. Need pakuvad keelelist toormaterjali – sedasama, millega leksikograafid iga päev töötavad, mida analüüsivad, filtreerivad, sorteerivad ja kust vajalikku infot välja valivad. Keeleõppijaid tuleks õpetada korpusi kasutama, täpselt nagu neid on õpetatud kasutama sõnastikke. Keeleõppijale esitatava korpusmaterjali mahtu tuleks aga piirata, filtreerida ja süstematiseerida. Samuti aitaks korpuste kasutamist keeleõppijate seas populariseerida see, kui korpusmaterjali keeleõppijale lihtsustatud või piiratud kujul esitada, nagu on tehtud näiteks SkELL keeleõppekeskkondade sarjas. Korpustest on keeleõppijatele abi ka siis, kui sõnastikus esitatavatest näitelausetest ei piisa sõna kõikide kasutusmustrite illustreerimiseks. Samuti saavad keeleõppijad korpuslauseid kasutada mallina teksti loomisel. Teine võimalus korpusi keeleõppijatele lähemale tuua on tutvustada neid kui interneti otsinguportaale. Kindlasti kasutavad ka keeleõppijad keeleküsimuste lahendamisel internetti, sealhulgas otsingumootoritesse integreeritud tõlketeenuseid ja -sõnastikke (nt Google Translate, Bing Microsoft Translation). Kui interneti otsinguportaalist abi otsides peab keeleõppija suures infomüras iseseisvalt navigeerima

[9] Konkordants on sõnavorm koos kontekstiga (McEnery, Hardie 2012: 241).

ning oskama sealt kasulikku infot ise üles noppida, siis korpuspäringusüsteem teeb selle töö mõnes mõttes ära, kuvades kasutajale vastused juba süstematiseeritud kujul (nt täislausena, kollokatsioonina, tesaurusena).

Eestis ei ole korpuste kasutust keeleõppes eraldi uuritud. Arvutipõhist ehk arvutite vahendusel ja arvutitega korraldatud eesti keele õpet on üldisemalt uurinud Maarika Teral (2015), kelle uurimuses selgus, et arvutipõhine õpe on õppijate meelest otstarbekas ja tulemuslik, kuid võib samas olla ka ajamahukam kui kontakttunnid.

Siinne väitekiri, mille teema sündis „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi täisautomaatse genereerimisega, keskendub näitelausete automaatsele tuvastamisele ja on Eestis esimene katse tuua korpuslaused otse (õppe)sõnastiku kasutajateni. Enne korpuslausete kuvamist lõppkasutajatele on neid aga tarvis filtreerida, kõrvaldamaks näitamiseks sobimatud (poolikud, ülipikad, vigased jms) laused. Selleks tuleb näitelauseid valivale programmile ette anda reeglid, millele toetudes oskaks see välja pakkuda kõige paremad näitelause kandidaadid ning välja filtreerida sobimatud. Korpuslausete filtreerimine on eriti oluline siis, kui sihtgrupp on keeleõppija (nagu seda on naabersõnade sõnastiku kasutaja), kuna keeleõppijale ei sobi kuvada liiga pikki, grammatiliselt ebakorrektseid ega leksikaalselt keerulisi lauseid. Järgmises peatükis tutvustan teiste keelte näitel erinevaid meetodeid, mida näitelausete automaatseks tuvastamiseks kasutatakse.

3. NÄITELAUSETE AUTOMAATSE TUVASTAMISE MEETODID

Viimase aastakümne jooksul on korpusleksikograafias väga palju keskendutud näitelausete automaatsele tuvastamisele. Uurimused on näidanud, et näitelausete automaattuvastus vähendab oluliselt leksikograafide ajakulu sõnaartikli koostamisel (Kosem, Gantar, Krek 2013, Kosem, Husák, McCarthy 2011). Selles peatükis tutvustangi meetodeid, mida on näitelausete automaatseks tuvastamiseks kasutatud: masinõppemeetodit, reeglipõhist lähenemist ja nende kahe kombinatsiooni ehk kombineeritud meetodit. Keskendun reeglipõhisele lähenemisele, mida kirjeldan korpuspäringusüsteemi Sketch Engine integreeritud tööriista Good Dictionary Examples ehk GDEX näitel. Samuti tutvustan GDEXi inglise, sloveeni, hollandi, portugali, vene, jaapani ja soome keele moodulit.

3.1. Masinõppemeetod, reeglipõhine lähenemine ja kombineeritud meetod

Masinõppe keskmes on algoritm, mis õpib empiirilistele andmetele toetudes otsuseid tegema ning nende põhjal tundmatu andmestiku kohta midagi ennustama (Witten jt 2016). Näiteks kui sisendiks on andmestik *X*, siis õpib funktsioon *f(X)* selle andmestiku põhjal prognoosima väljundiks andmestikku *Y* (Pilán 2018: 41). Andmestik, mida masinõppes tüüpiliselt kasutatakse, jaguneb treening- ja testandmestikuks. Treeningandmestikku kasutab masinõppe algoritm õppimiseks, näiteks võivad selleks olla leksikograafi poolt valitud näitelaused. Testandmestikku kasutatakse algoritmi resultatiivsuse mõõtmiseks, näiteks saab automaatselt tuvastatud näitelausete kvaliteeti võrrelda leksikograafi poolt valitud näitelausetega. Masinõppes kasutatakse tavapäraselt kolme tüüpi lähenemist: juhendatud (*supervised learning*), juhendamata (*unsupervised learning*) (Hastie jt 2009) ja pooljuhendatud (*semi-supervised learning*) (Søgaard 2013) õppimist. Juhendatud õppimise meetodiga antakse algoritmile „õiged vastused“ ette (nt leksikograafi valitud näitelaused), st programmile öeldakse, kuidas midagi teha ehk milliseid lauseid korpusest otsida. Juhendamata õppimise lähenemises õpib algoritm suurest hulgast andmetest (nt juhuslikest korpuslausetest) ise tuvastama neid lause tunnuseid, mille alusel näitelauseid valida. Pooljuhendatud õppimine toetub nii käsitsi valitud andmestikule kui ka iseõppimisele. (Pilán 2018: 41–42) Reeglipõhine lähenemine eeldab teatud eeldefineeritud parameetreid, mille alusel see meetod töötab. Näiteks saab programmile öelda, et hea näitelause on täislause ning maksimaalselt 10 sõnet pikk – nii valib programm korpusest automaatselt välja just nendele parameetritele vastavad korpuslaused. Kombineeritud meetod ühendab reeglipõhist lähenemist masinõppe algoritmidega.

Nii reeglipõhist lähenemist kui ka kombineeritud meetodit kasutanud uuringud on näidanud, et kombineeritud meetodiga on näitelausete kvaliteet parem (Didakowski jt 2012, Lemnitzer jt 2015). Nikola Ljubešić ja Mario Peronja (2015)

kasutasid heade näitelausete ekstraheerimiseks masinõppemeetodit ning saavutasid väga hea tulemuse (90%-suurune saagis kolme esimese näitelause kandidaadi pealt). Masinõppemeetodit ja kombineeritud meetodit on kasutatud nii näitelausete automaatseks tuvastamiseks lemmadele kui ka lemma erinevatele alltähendustele. Nii näiteks klasterdab Beto Boullosa jt (2017) poolt arendatud masinõppel põhinev süsteem korpuslauseid automaatselt tähendusjaotuste järgi (toetudes lausetes esinevatele sarnastele teemadele) ning võtab arvesse ka kasutajate tagasisidet klasterdamise täpsuse kohta. Paul Cooki jt (2014) kombineeritud meetodil põhinev mudel otsib märksõna jaoks võimalikult mitmekesise kasutusmustriga lauseid.

Automaatselt tuvastatud näitelauseid kasutatakse leksikograafias tüüpiliselt kolmel viisil (Kosem, Husák, McCarthy 2011).

1. **Korpuspäringusüsteemis**. Leksikograafile pakutakse korpuspäringusüsteemis nimekiri korpuslausetest, mille seast ta valib välja sobivaima ning kopeerib sõnastiku näitelauseks (nt *tickbox lexicography*-meetod, loe lähemalt ptk 3.3).
2. **Sõnastikusüsteemis**. Teatud arv korpuslauseid (nt kümme) ekstraheeritakse korpusest automaatselt spetsiaalse programmi abil sõnastikusüsteemi, kus leksikograaf neid edasi toimetab, näiteks valib ekstraheeritud korpuslausete seast välja ühe, mis kõige paremini sõnastiku näitelauseks sobib. Eestis kasutati sellist lähenemist „Eesti keele naabersõnade 2019“ sõnastiku koostamisel (loe lähemalt ptk 4.1.1).
3. **Sõnastikuportaali osana**. Korpuslauseid kuvatakse otse sõnastiku lõppkasutajale. Sellisel juhul on kõik laused täiesti autentsed ja toimetamata ehk leksikograafi poolt üle kontrollimata. Eestis kasutatakse sellist lähenemist näiteks keeleportaalis Sõnaveeb (loe lähemalt ptk 4.4.4).

Ent näitelausete automaatset tuvastust ei kasutata mitte ainult leksikograafias, vaid ka keeleõppes ning keeleõpperakenduste loomisel. Ildikó Pilán jt (2013) on kombineeritud meetodit kasutades välja töötanud süsteemi HitEx, mis leiab korpusest automaatselt sellised laused, mis sobivad keeleõppe erinevat tüüpi harjutustesse. Sarnast kombineeritud meetodit on kasutanud ka Chieh-Yang Huang ja Lun-Wei Ku (2016), kelle loodud süsteem GiveMeExample valib korpusest automaatselt näitelaused sõnade rühmale, mille vahelistest erinevustest on keeleõppijal raske aru saada (*confusing words*, sinna alla kuuluvad nt ka lähisünonüümid). Nende mudel õpib lausete klasterdamise teel ära iga sõna kõige tüüpilisemad kasutusmustrid ning valib iga sõna jaoks välja just sellise näitelause, mis kõige paremini näitab selle sõna kasutust. Arseny Tolmachev ja Sadao Kurohashi (2017) on masinõppemeetodiga loonud sõnasedelite (*flashcard*) süsteemi, kus õppijale sõna korrates näidatakse uut korpuslauset. Nende süsteem tagab, et õppijale kuvatavad laused oleksid süntaktiliselt võimalikult mitmekesised (erineva argumentstruktuuriga) ning illustreeriksid märksõna erinevaid kasutusmustreid. Anneliis Halling (2016) on reeglipõhist lähenemist kasutades loonud õppeprogrammi, mis ilukirjanduskorpuse lauseid kasutades genereerib harjutusi eesti keele käänete õppimiseks.

Kuna korpuses on halbu näitelause kandidaate palju rohkem kui häid näitelause kandidaate, tasub parameetrid välja selgitada mõlema jaoks – nii headele

kui ka halbadele. Sageli suudavad leksikograafid palju paremini kirjeldada just neid lause omadusi, mida nad näitelause juures halvaks peavad, kui neid lause omadusi, mis heal näitelausel olema peavad. [P2] Siinse väitekirja keskmes on reeglipõhisel valemil töötav tööriist Good Dictionary Examples ehk GDEX ning selle eesti keele mooduli erinevad versioonid, mis arvestavad eesti keele spetsiifilisi lause parameetreid (Kallas, Koppel, Tuulik 2015, [P1], [P3], [P5]). GDEXi eesti keele moodulit on arendatud reeglipõhist lähenemist kasutades, kuid parameetrite häälestamiseks on osaliselt kasutatud ka masinõppe elemente: klassifikaatorite väärtuste optimeerimiseks ja neile kaalu määramiseks on võrdlevalt analüüsitud nn heade ja halbade näitelausete andmebaase (loe lähemalt [P3]).

3.2. Reeglipõhine tööriist Good Dictionary Examples ehk GDEX

Good Dictionary Examples ehk GDEX on korpuspäringusüsteemi Sketch Engine integreeritud tööriist, mis teatud eeldefineeritud parameetrite abil analüüsib korpuslauseid ning reastab need paremuse järjekorda. GDEXi loomise algne eesmärk oli eelkõige aidata arvutil n-ö eeltööd teha ja vähendada leksikograafide ajakulu näitelausete valimisel korpusest (loe lähemalt Kilgarriff, Husák jt 2008), kuid hiljem hakati GDEXit rakendama ka laiemalt, võttes peale keeleteadlaste ja leksikograafide arvesse ka keeleõppija vajadusi (loe lähemalt Baisa, Suchomel 2014, Koppel, Kallas jt 2019).

Lihtsustatult öeldes töötab GDEX justkui filtrina, praakides välja tõeliselt ebasobivad korpuslaused ning reastades kõik ülejäänud näitelause kandidaadid paremuse järjekorda. GDEXi keskmes on reeglipõhine valem, mis hindab etteantud parameetrite alusel korpuslause komponente ja määrab igale lausele skoori (*GDEX score*), mille alusel neid kasutajale järjestatakse. Skoor jääb 0 ja 1 vahele – mida kõrgem skoor, seda sobivam näitelause kandidaat. Skoori väärtus sõltub lause omadusi mõõtvatest klassifikaatoritest, mis omakorda jagunevad kaheks: tugevateks (*hard classifiers*) ja nõrkadeks (*soft classifiers*). GDEXi eesti keele moodulis moodustavad tugevad ja nõrgad klassifikaatorid kumbki lause üldskoorist 50% ehk annavad kumbki maksimaalselt kokku 0,5 punkti (0,5 + 0,5 = 1). Tugevate klassifikaatorite abil tuvastatakse kõik sobimatud näitelause kandidaadid, nõrgad reastavad ülejäänud näitelause kandidaadid paremuse järjekorda. Nõrgad klassifikaatorid kas vähendavad lause üldskoori ehk karistavad (*penalize*) lauset, kui see mingile etteantud parameetrile ei vasta (mis tähendab, et lause liigub kandidaatide nimekirjas allapoole), või annavad lausele lisapunkte (mis tähendab, et lause liigub kandidaatide nimekirjas ülespoole) (loe lähemalt [P3]).

Klassifikaatorid sisaldavad eeldefineeritud leksikaalseid ja süntaktilisi parameetreid (nt lause ja sõna pikkus, sõnade sagedus korpuses, märksõna asukoht lauses, märksõna kordumine), mis on masina abil mõõdetavateks tunnusteks tõlgendatud. Reeglipõhine valem koos klassifikaatorite ja täiendavate parameet-

ritega moodustavad GDEXi konfiguratsioonifaili (joonis 1), mis sisaldab kahte tasandit: kohustuslikku valemit (*formula*) ja valikulisi muutujaid (*variables*).

```
formula: >
    (50 * is_whole_sentence() * blacklist(words, illegal_chars) * blacklist(lemmas, parsnips)
    + 50 * optimal_interval(length, 10, 14)
    * greylist(words, rare_chars, 0.1)
    * greylist(tags, pronouns, 0.1)
    ) / 100
variables:
    illegal_chars: ([<|\]\[>/\\^@])
    rare_chars: ([A-Z0-9'.,!?)(;:-])
    pronouns: PRON.*
    parsnips: ^(tory, whisky, jesus, cowgirl, meth, commie, bacon)$
```

Joonis 1. GDEXi konfiguratsioonifail

Sketch Engine'is on olemas ka universaalne ehk keelest sõltumatu GDEXi konfiguratsioon. Oma olemuselt on see inglise keele konfiguratsiooni lihtsustatud versioon, mis on kavandatud sobima teistele keeltele. See sisaldab kolme tugevat klassifikaatorit (tegemist peab olema täislausega, teatud tähemärgid on keelatud (joonisel 1 *illegal_chars*), sõnele on määratud minimaalne esinemissagedus korpuses) ning kolme nõrka klassifikaatorit (lause optimaalne pikkus, karistus harvadele sõnadele ja märkidele (joonisel 1 *rare_chars*)). (Srdanović, Kosem 2016) Universaalne on ka Jaccardi sarnasuse indeksi[10] (*Jaccard similarity index*) kasutamine, mis tagab, et väljundis kuvatavad laused ei korduks, vaid oleksid võimalikult mitmekesised [P2].

Järgnevalt annan lühiülevaate GDEXi inglise, sloveeni, hollandi, portugali, soome, vene ja jaapani keele moodulitest.

3.3. GDEXi eri keelte moodulid

GDEX loodi algselt inglise keele sõnastike koostamiseks TBL-meetodiga (*tickbox lexicography*) (Kilgarriff, Kovář, Rychlý 2010). TBL-meetod seisneb selles, et kõigepealt kuvatakse leksikograafile sõnavisand, mis on üheleheline automaatne korpuspõhine kokkuvõte sõna grammatilisest ja kollokatiivsest käitumisest, kus ta märgib ükshaaval konkreetse lekseemi jaoks sobivad kollokatsioonid ja GDEXi poolt pakutud näitelaused, mis kantakse seejärel automaatselt sõnastikusüsteemi. Kõige olulisemad parameetrid inglise keele näitelausete valikul on lause pikkus ja sõnade sagedus korpuses. (Kilgarriff, Husák jt 2008) Inglise GDEXi konfiguratsiooni sloveeni keele peal rakendades selgus, et paremate tulemuste saavutamiseks on vaja arvesse võtta keelespetsiifilisi parameetreid

[10] Varasemates versioonides kasutati lausete mitmekesisuse tagamiseks Levenshteini distantsi [P2].

(Kosem, Husák, McCarthy 2011, Kosem, Gantar, Krek 2013). Inglise ja sloveeni keele konfiguratsioonid on olnud lähtepunktiks paljude keelte, näiteks hollandi (Tiberius, Kinable 2015), portugali (Kuhn 2017), jaapani (Srdanović, Kosem 2016), vene (Koppel, Kallas jt 2019), soome (Langemets, Heinonen jt 2017) ja ka eesti [P3] moodulitele.

GDEXi sloveeni keele mooduli keelespetsiifiliste parameetrite väljaselgitamiseks analüüsiti masinõppemeetodiga leksikograafide poolt käsitsi välja valitud sloveeni leksikaalse andmebaasi näitelauseid. Kõige olulisemateks parameetriteks osutusid lause pikkus, märksõna suhteline asukoht ning märksõna korduse keelamine. GDEXi sloveeni keele mooduli esimest versiooni kasutati sloveeni leksikaalse andmebaasi (Gantar, Krek 2011) koostamiseks TBL-meetodiga. Töö teise versiooniga (Kosem, Gantar, Krek 2013) nõudis klassifikaatorite peenhäälestust, kuna see loodi spetsiaalselt leksikaalse info (grammatiliste suhete, kollokatsioonide, näitelausete) automaatseks ekstraheerimiseks. Olulisim täiendus oli teise kollokaadi (*second collocate*) ehk kollokatsiooni kollokaadi klassifikaatori lisamine. Samuti loodi eri sõnaliikidele eri konfiguratsioonid, mis arvestasid märksõna asukohta lauses. Näiteks selgus, et kui märksõna on tegusõna, siis heades näitelausetes ei esine see tavaliselt lause esimeses pooles.

GDEXi hollandi keele mooduli esimene versioon loodi 2015. aastal sõnaraamatu „Algemeen Nederlands Woordenboek“ (ANW) jaoks. Klassifikaatorite häälestamiseks analüüsiti selleks hetkeks ANW andmebaasi käsitsi valitud näitelauseid. Kuna ANW on emakeelsele kõnelejale suunatud akadeemiline sõnaraamat, kus näitelaused illustreerivad nii definitsioone kui ka kinnistunud väljendeid, siis võivad need rohkema konteksti andmise eesmärgil olla detailsed ja keerulised, koosnedes kohati isegi rohkem kui ühest lausest. Keelespetsiifiliste parameetrite väljaselgitamiseks analüüsiti definitsioone illustreerivaid lauseid ning seda tehti üksikute lausete tasandil. Analüüsi tulemusena otsustati karistada näiteks ilma pöördelise tegusõnavormita lauseid, asesõnu sisaldavaid lauseid ja lauseid, milles on sõnu, mis on pikemad kui 15 tähemärki. (Tiberius, Kinable 2015, [P2])

GDEXi portugali keele moodulil on akadeemiline sihtgrupp ehk portugali keelt emakeelena valdavad Portugali üliõpilased. Keelespetsiifiliste parameetrite väljaselgitamiseks analüüsiti lauseid, mis pärinesid Brasiilia ja Portugali akadeemilisi ajakirju sisaldavast tasakaalus korpusest (Kuhn, Ferreira 2016). Portugali keele GDEXi moodul andis paremaid tulemusi, kui igale lauses esinevale lemmale ja sõnele seati sagedusläve miinimum (lemma sagedusläveks määrati 500, sõne sagedusläveks 50). Samuti parandas väljundit teise kollokaadi klassifikaator, mis annab iga lauses esineva kollokatsiooni kollokaadi eest lisapunkti. Võrreldes sloveeni mooduliga on portugali moodulis teise kollokaadi klassifikaatorile määratud suurem kaal ja kõrgemad lisapunktid. (Kuhn 2017)

GDEXi vene keele mooduli esimest versiooni (Apresjan jt 2016) kasutati korpuse ruSkELL 1.5 loomiseks samanimelise keeleõppekeskkonna ruSkELL jaoks. ruSkELL 1.5 väljund sisaldas nii ülipikki (kuni 150 sõna) kui ka ühesõnalisi lauseid, selle laused ei alanud sageli suure tähega ning sisaldasid ebasobivat sõna-

vara (Koppel, Kallas jt 2019). Uue versiooni 1.6[11] põhjal loodi uus korpus ruSkELL 1.6. Vene keele moodulis oli oluline keelata näiteks ladina tähtedes kirjutatud sõnad ning sellised kirillitsas kirjutatud sõnad, mis on tegelikult hoopis ukraina või valgevene keele sõnad. Korpust ruSkELL 1.6 kasutatakse ka Eesti Keele Instituudi keeleportaalis Sõnaveeb venekeelsete veebilausete allikana. (Koppel, Kallas jt 2019)

GDEXi jaapani keele moodulit luues peeti silmas kahte eesmärki: üldine leksikograafiline vajadus näitelauseid korpusest tuvastada ja GDEXi poolt valitud lausete kasutamine keeleõppe eesmärgil (sh õppesõnastike koostamisel). Kõige olulisemad lause parameetrid olid pikkus (8–30 sõnet, optimaalne pikkus 10–25 sõnet); ladina tähtede, teatud sümbolite ja jaapani keele spetsiifiliste märkide karistamine; teine kollokaat; hüüumärgi keelamine sõna lõpust. Kuna jaapani sõnad koosnevad enamasti kuni neljast tähemärgist, karistatakse sõnu, mis on pikemad kui seitse tähemärki. Keeleõppe eesmärgil on jaapani keele jaoks loodud GDEXi mooduli versioonid ka viiele eri oskustasemele. Need erinevad teineteisest peamiselt selle poolest, et karistatakse teatud sõnu ja lemmasid, mis sellele konkreetsele tasemele ei kuulu, ning antakse lisapunkte teatud hulga sõnade eest, mis sellele konkreetsele tasemele kuuluvad. (Srdanović, Kosem 2016)

GDEXi soome keele moodul loodi eesmärgiga leida näitelauseid eesti-soome veebisõnastiku (Langemets, Heinonen jt 2017)[12] jaoks, kuid praktikas seda ei kasutatud. Soome mooduli lähtekohaks oli eesti keele mooduli versioon 1.4 [P3]. Soome keeles tekitab probleeme kõnekeel, mida veebist kroolitud korpus ohtralt sisaldab. Seetõttu on sagedasemad kõnekeelsed sõnad (nt sm *oon*, *oot*, *mä*, *sä*, *mulle*, *sulla*, *niiku*, *kans*) lausest keelatud.

Tabel 1 pärineb artiklist [P2] ning illustreerib inglise, sloveeni, hollandi, eesti[13] ja portugali keele moodulite parameetreid. Tabelist nähtub, milliseid parameetreid kasutatakse kõigi viie keele moodulites ning millised on keelespetsiifilised. Tugevate klassifikaatorite alla liigituvad parameetrid on märgitud paksus, nõrkade alla liigituvad parameetrid tavalises kirjas.

[11] Versiooni autor on Peterburi Riikliku Ülikooli teadur Maria Khokhlova (loe lähemalt Koppel, Kallas jt 2019).

[12] Soome mooduli autor on Soome Kodumaa Keelte Keskuse (KOTUS) teadur Tarja Heinonen.

[13] Tabelis 1 toodud GDEXi eesti keele mooduli parameetrid on versioonist 1.4 (loe lähemalt ptk 4.1.3).

Tabel 1. GDEXi eri keelte moodulites kasutatavate parameetrite võrdlus [P2]. + märgib parameetri olemasolu, Ø puudumist

parameeter	**inglise**	**sloveeni**	**hollandi**	**eesti**	**portugali**
täislause	+	+	+	+	+
must nimekiri: keelatud tähemärgid	+	+	+	+	+
must nimekiri: tundlikud sõnad	Ø	Ø	+	Ø	Ø
spämm	+	+	Ø	+	+
sõnede miinimumsagedus	Ø	3	Ø	5	50
märksõna kordus	Ø	+	+	+	+
lause minimaalne ja maksimaalne pikkus (sõnedes)	6–28	7–60	Ø	4–20	7–30
karistus sarnastele lausetele	+	+	+	+	+
lause optimaalne pikkus (sõnedes)	8–12	15–40	10–25	6–12	10–30
teine kollokaat	Ø	+	+	+	+
märksõna suhteline asukoht	+	ainult tegu-sõnadele	Ø	Ø	Ø
karistus pikkadele sõnadele	pikemad kui 6 tähemärki	pikemad kui 12 tähemärki	pikemad kui 15 tähemärki	Ø	pikemad kui 12 tähemärki
karistus harvadele tähemärkidele	+	+	Ø	+	+
karistus suurtähtedele	+	+	Ø	+	Ø
karistus sõnedele, mis sisaldavad sümboleid	Ø	+	Ø	+	+
karistus pärisnimedele	Ø	+	Ø	+	Ø
karistus asesõnadele	+	+	+	+	Ø
lause alguses keelatud sõnad	Ø	+	Ø	+	Ø
lause alguses keelatud sõnapaarid	Ø	+	Ø	+	Ø
lause alguses keelatud sõnaliigid	Ø	Ø	Ø	+	Ø

parameeter	inglise	sloveeni	hollandi	eesti	portugali
karistus harvadele sõnadele	sagedus korpuses väiksem kui 1 miljoni kohta	sagedus korpuses väiksem kui 1000	Ø	sagedus korpuses väiksem kui 1000	sagedus korpuses väiksem kui 500
karistus komadele	Ø	3 või rohkem	Ø	2 või rohkem	3 või rohkem
karistus lausetele, kus puudub finiitne ehk pöördeline tegusõnavorm	Ø	Ø	+	infiniitsed ehk käände-lised vormid	Ø
lisapunkt lausetele, mis pärinevad teatud allkorpusest	Ø	Ø	Ø	+	Ø
karistus rohkem kui kaks korda järjestikku esinemise eest	Ø	Ø	Ø	Ø	+

4. UURIMISTULEMUSED: GDEXi EESTI KEELE MOODUL

Siinses peatükis keskendun GDEXi eesti keele mooduli arendamise ajaloole, selle eri versioonidele (GDEX 1.2, GDEX 1.3, GDEX 1.4, etBasic-v1, etIndependent-v1 ja etProficient-v1) ning neis kasutatud eesti keele spetsiifilistele parameetritele. Ka eesti keele peal on testitud Sketch Engine'i universaalset GDEXi konfiguratsiooni (versioon 1.1), kuid korpuslausete ekstraheerimiseks kasutati versiooni 1.2 (Kallas, Koppel, Tuulik 2015), kuhu oli juba lisatud eesti keele spetsiifilisi parameetreid.

4.1. GDEXi eesti keele mooduli versioonid

4.1.1. GDEX 1.2

Eesti leksikograafias kasutati esmakordselt näitelausete automaatset tuvastamist 2014. aastal, kui Eesti Keele Instituudis genereeriti täisautomaatselt „Eesti keele naabersõnade 2019" sõnastiku andmebaas. Naabersõnad ehk kollokatsioonid on sisusõnade tähenduslikud ja statistiliselt esilduvad kombinatsioonid teiste leksikaalsete ja grammatiliste üksustega (nt *päike paistab*, *kange kohv*, *pakast trotsima*). (Kallas, Koppel, Tuulik 2015) Keeleõppe sisukohalt pakub info kollokatsioonide kohta huvi peamiselt edasijõudnutele, kellel on juba paremad teadmised õpitava keele struktuurist (Lew 2004: 23).

„Eesti keele naabersõnad 2019" on aktiivse[14] suunitlusega sõnastik, mis pakub keelele omast leksikat teksti loomiseks. Sõnastiku sihtgrupp on edasijõudnud ja vilunud eesti keele õppijad (B2–C1-keeleoskustase), mis tähendab, et ka GDEXi eesti keele moodulit arendades olen peaasjalikult silmas pidanud just sellel tasemel eesti keele valdajat. Sõnastiku andmebaasi automaatseks genereerimiseks[15] kasutati korpuspäringusüsteemi Sketch Engine sõnaloendi ja sõnavisandite funktsiooni ning heade näitelausete tuvastamise tööriista GDEX, mille eesti keele moodul tuli genereerimise eel alles luua.

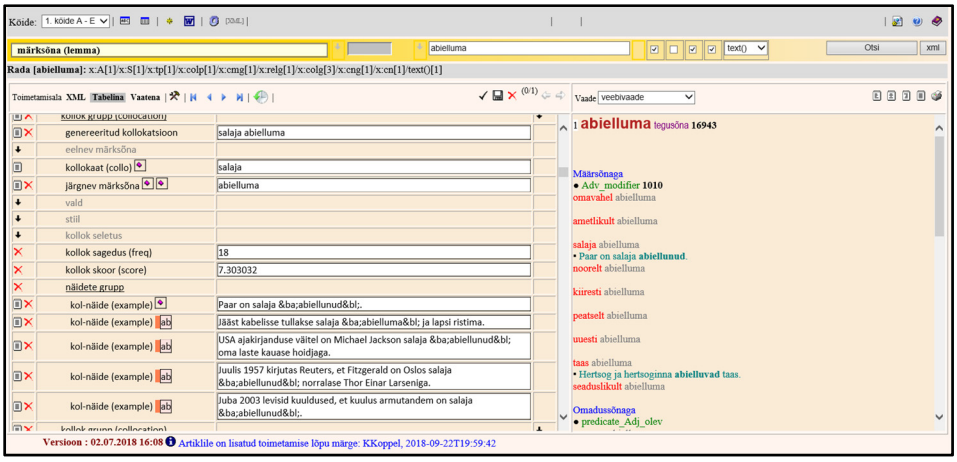
Eesti keele spetsiifiliste lause parameetrite väljaselgitamiseks analüüsisin tol hetkel koostamisel oleva „Eesti keele sõnaraamatu 2019" ja 2014. aastal ilmunud „Eesti keele põhisõnavara sõnastiku" näitelauseid: sõnade arvu lauses, lause keskmist pikkust (sõnades), sõna keskmist pikkust (tähemärkides) ning ka kõrvallausetega lausete osakaalu. Analüüsi tulemused näitasid, et laused on sõnastikes küllaltki lühikesed (keskmiselt 4–7 sõna), sõnade keskmine pikkus on 5–7 tähemärki ning kõrvallausete osakaal on üsna väike. Analüüsi tulemustele toetudes

[14] Aktiivse suunitlusega sõnastikud toetavad kasutajat teksti loomisel (kirjutamisel), passiivse suunitlusega sõnastikud teksti mõistmisel (lugemisel) (Tavast, Taukar 2013).

[15] Andmebaas genereeriti „Eesti keele ühendkorpuse 2013" põhjal, hiljem sõnastiku toimetamise faasis toetuti juba „Eesti keele ühendkorpusele 2017". Kokku ekstraheeriti 10 939 märksõna, 493 971 kollokatsiooni ja 2 469 855 näitelauset.

lõin GDEXi mooduli versiooni 1.2 (Kallas, Koppel, Tuulik 2015), mille kõik parameetrid on koondtabelina kirjas lisas 1 (vt ka konfiguratsioonifaili lisas 2).

GDEX 1.2 abil ekstraheeriti iga kollokatsiooni jaoks „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi viis korpuslauset (Kallas, Koppel, Tuulik 2015). Leksikograafi ülesanne sõnastiku koostamisel oli muuhulgas nende viie ekstraheeritud korpuslause seast välja valida üks – selline, mis kollokatsiooni kasutust kõige paremini illustreerib. Joonisel 2 on näha „Eesti keele naabersõnade 2019“ andmebaasi kuvatõmmist Eesti Keele Instituudi veebipõhises sõnastikusüsteemis EELex (Langemets, Loopmann, Viks 2006, Järviste jt 2011). Vasakul pool on toimetamisala, kus leksikograaf saab sõnastiku sisu töödelda – kollokatsioone kustutada, lisada, parandada ning näitelauseid valida. Vasakul pool on näha ka kollokatsiooni *salaja abielluma* jaoks GDEX 1.2 abil ekstraheeritud viit korpuslauset, mille seast leksikograaf on näitelauseks välja valinud ühe (*Paar on salaja abiellunud.*), mis on nähtaval ka paremal pool olevas veebivaates.



Joonis 2. Märksõna *abielluma* „Eesti keele naabersõnade 2019“ sõnastiku andmebaasis

Kuigi iga kollokatsiooni jaoks ekstraheeriti viis korpuslauset, oli sageli (eriti madala sagedusega kollokatsioonide puhul) nende seast ühe autentse lause välja valimine keeruline ülesanne, kuna sekka oli sattunud ka sobimatuid või toimetamist vajavaid kandidaate. Näiteks esines palju tegusõnata ning rohkete anafooridega lauseid. See tõi kaasa vajaduse GDEXi eesti keele moodulit edasi arendada.

4.1.2. GDEX 1.3

Versiooni 1.3 lisasin klassifikaatori, mis tuvastab korpusest ainult tegusõna sisaldavaid lauseid. Kontekstisidususe vältimiseks keelasin lause alguses sõnad *näiteks*, *kui*, *ühesõnaga*, *seejärel*, *nagu* ning karistasin lauset asemäärsõnade *siin*, *siia*, *siit*, *sinna*, *seal* ja *siis* esinemise eest. GDEX 1.3 on täpsemalt kirjeldatud

artiklis [P1], selle kõik parameetrid on toodud koondtabelina lisas 1 ning konfiguratsioonifail lisas 3.

GDEX 1.3 abil lõi Eesti Keele Instituut koostöös tarkvarafirmaga Lexical Computing Ltd. korpuse „EstonianNC GDEX“, mis oli esimene katse luua autentseid lauseid sisaldav õppekorpus eesti keele õppijatele. Korpuse analüüs näitas, et lause alguses keelatud sõnade ja lauses esinevate anafooride nimekirja tuleb täiendada. Samuti olid korpuses sageli esil pärisnimed, numbrid ja madala sagedusega sõnad. Neid probleeme üritasin lahendada järgmise versiooniga 1.4.

4.1.3. GDEX 1.4

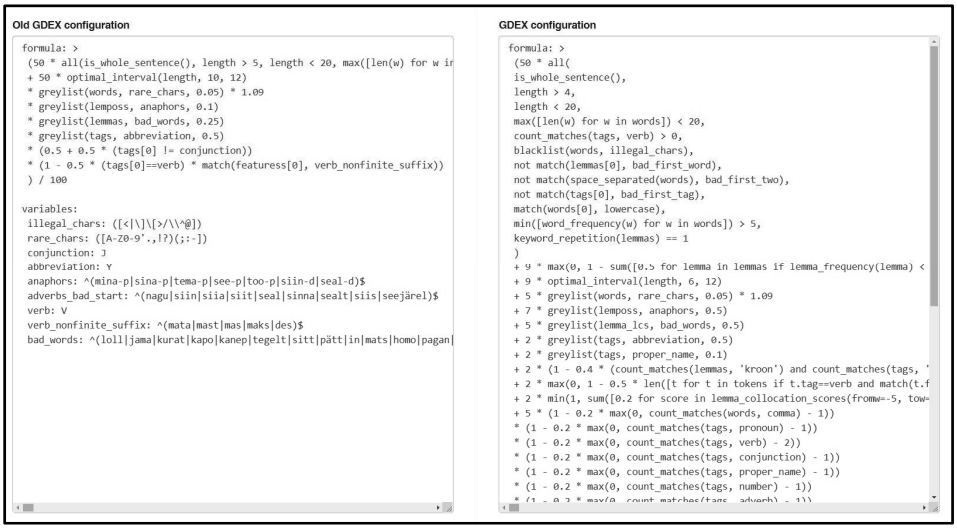
GDEX 1.4 loomiseks analüüsisin tol hetkel veel koostamisel olnud[16] „Eesti keele naabersõnade 2019“ sõnastiku näitelauseid. Lõin neist kaks andmebaasi, mida nimetasin tinglikult heade näitelausete andmebaasiks ja halbade näitelausete andmebaasiks. Nagu eespool mainitud, ekstraheeriti „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi iga kollokatsiooni jaoks viis korpuslauset, millest leksikograaf pidi välja valima ühe. Heade näitelausete andmebaasi läksidki need laused, mille leksikograaf oli algsetest GDEX 1.2 abil ekstraheeritud korpuslausetest välja valinud. Kokku oli heade näitelausete andmebaasis 44 038 lauset. Laused, mis valituks ei osutunud, liikusid halbade lausete andmebaasi: kokku 128 239 lauset. Need andmebaasid ei peegeldanud küll korpuse tegelikku sisu, kuna nii heade kui ka halbade näitelausete andmebaas sisaldas juba eeldefineeritud parameetritele (GDEX 1.2) vastavaid korpuslauseid. Samuti ei olnud kõik heade näitelausete andmebaasi laused täiesti autentsed, kuna mitmeid olid leksikograafid redigeerinud. Kindlasti sattus hulka ka tehislikke ehk leksikograafi koostatud näitelauseid, kuid nende osakaal kogu andmebaasis on väga väike. Ka kõik halbade näitelausete andmebaasi kuuluvad laused ei olnud tingimata halvad, lihtsalt leksikograaf pidi ühe viiest ekstraheeritud lausest valima. Sellele vaatamata olid sellise sisuga andmebaasid analüüsiks sobivad, kuna artikli [P3] peamine eesmärk oli välja selgitada näitelauseks valitud korpuslauseid iseloomustavad parameetrid.

Selleks, et heade ja halbade näitelausete andmebaase oleks võimalik analüüsida, tuli need esmalt morfoloogiliselt märgendada. Seejärel määrasin kindlaks, missuguseid parameetreid soovin lähemalt uurida (lause pikkus, märksõna asukoht lauses, lause esimese sõna sõnaliik jmt), ning viisin läbi andmebaaside statistilise analüüsi.

GDEX 1.4 parameetrite testimiseks ja klassifikaatorite peenhäälestamiseks kasutasin spetsiaalset kasutajaliidest GDEX Editor[17], mille Lexical Computing Ltd. tarkvaraarendaja Jan Michelfeit lõi GDEX 1.4 arendamisega paralleelselt ISCH COST programmi IS10305 European Network of e-Lexicography lühiajalise teadusmissiooni käigus Ljubljana ülikoolis dr Iztok Kosemi juhendamisel.

[16] Selleks ajaks oli „Eesti keele naabersõnade 2019“ sõnastikku koostatud umbkaudu 1,5 aastat.
[17] GDEX Editori kasutamiseks tuleb Sketch Engine’isse esmalt sisse logida.

Kui varasemalt oli klassifikaatorite häälestamine ja testimine GDEXi mooduli arendajatele olnud üsna tülikas ülesanne – selleks pidi konfiguratsioonifaile korduvalt alla laadima, toimetama ning taas üles laadima –, siis GDEX Editor võimaldab parameetreid mugavalt häälestada veebipõhiselt ning kõik konfiguratsioonis tehtud muudatused on liideses kohe nähtavad. Korraga saab võrrelda kahte erinevat GDEXi konfiguratsiooni (joonis 3).



Joonis 3. Kahe konfiguratsiooni võrdlus GDEX Editoris

Võrdluse tulemusena kuvatakse nimekiri lausetest, millel on kaks erinevat kandidaatide järjekorda ja kaks erinevat GDEXi skoori, mis on arvutatud kummagi konfiguratsiooni järgi. Tulemusi saab sortida nii GDEXi skoori kui ka näitelause kandidaatide loendi numbri järgi (joonis 4).

GDEX Editoris on võimalik testida sedagi, kuidas iga individuaalne klassifikaator lause üldskoori mõjutab. See lihtsustab oluliselt GDEXi moodulite arendajate tööd versioonide testimisel, klassifikaatorite peenhäälestamisel ja konfiguratsioonifailide kirjutamisel.

GDEX 1.4 parameetrite testimiseks valisin „Eesti keele naabersõnade 2019“ sõnastiku andmebaasist välja 40 märksõna (10 nimisõna, 10 omadussõna, 10 määrsõna, 10 tegusõna) erinevatest sagedusklassidest: kõrge (>5000 ja rohkem), keskmine (1000–5000) ja madal sagedus (<1000). Valisin iga märksõna kohta välja ühe grammatilise suhte (nt Adj_modifier, modifies, object, V_modifies), mille alt valisin omakorda välja kolm kollokatsiooni (nt *pidulikult tähistama*, *pidulikult avama*, *pidulikult lõpetama*), mille näitelauseid GDEX Editoris analüüsisin.

Old rank	Rank	Sentence	Old score	Score	Flag
16	1	Planeedi pinna temperatuur arvatakse olevat 700 kraadi.	0.78	0.98	
10	2	Teadlased on teistelt planeetidelt vee leidmisest huvitatud, sest vett peetakse elu tekkeks vältimatuks.	0.85	0.95	
31	3	Igaüks meie planeedil on maailmaime.	0.59	0.93	
8	4	Te peaksite tõsiselt uurima, mis siin planeedil toimub.	0.87	0.92	
20	5	Diana avastab, et Maa on lähim planeet, kus Külalised oma liiki jätkata saavad.	0.75	0.91	
19	6	Jumal on siin planeedil meelevalla andnud inimestele.	0.76	0.90	
11	7	Näiteks liiguvad lisaks tähtedele taevas planeedid, millest mõnigi on oma heleduselt tähtedest üle.	0.85	0.90	
18	8	Astroloogia on kõige vanem teadus siin planeedil.	0.76	0.89	
9	9	Pluuto oli kuni 2006. aastani Päikesesüsteemi üheksas planeet.	0.85	0.89	
4	10	Planeedid omavad üldjuhul ühte juhtorganit, mis allub Keskusele.	0.94	0.89	
2	11	Aastakümneid on harjutud nägema Pluutot just Päikesesüsteemi üheksanda planeedina.	0.94	0.88	
7	12	Seni troonib planeedi hinnalisima eralennuki kohal jätkuvalt Gulfstream.	0.87	0.88	
12	13	Film on Greeri kõige hilisem pingutus paljastada meie planeedi külaliste olemasolu võimalikku kinnimätsimist.	0.85	0.87	
25	14	Üksköik kuidas me seda planeeti ka ei kohtleks, 25 liiki on homseks meie seast kadunud!	0.69	0.87	
14	15	Unusta natukeseks ajaks kõik see halb, mis sa planeedile teinud oled.	0.84	0.86	

Joonis 4. Märksõna *planeet* kahe erineva konfiguratsiooni valitud näitelausete võrdlus GDEX Editoris

Vaatlesin selliseid parameetreid nagu lausete ja sõnede pikkus, lause sõnaliigiline koosseis, märksõna asukoht lauses, märksõna kordumine lauses, teatud elementide (koma, asesõnade, pärisnimede jmt) arv lauses jm. Lisaks võimaldas halbade näitelausete andmebaas välja selgitada sõnad ja sõnapaarid, mis lause alguses esinedes kipuvad olema anafoorsed (vt ptk 4.2.2). Analüüsisin ka heade ja halbade näitelausete esimese sõna sõnaliiki ning selgus, et suurem osa lausetest algavad nimisõnaga, millele järgnesid omadussõnaga, asesõnaga ja harvemini tegusõnaga algavad laused. Lause esimese sõna sõnaliigi kategooria võib viidata keeltevahelistele erinevustele ning samuti anda märku keeleregistrist. Näiteks algavad ka portugali üldkeele laused tavaliselt nimisõnaga, millele järgnevad tegusõnaga algavad laused. Akadeemilises portugali keeles algavad laused kõige sagedamini just tegusõnaga. Asesõnaga ja omadussõnaga algavaid lauseid esineb portugali keele kummaski registris harva. (Kuhn 2017: 326–327)

Esmakordselt määrasin versioonis 1.4 nõrkadele klassifikaatoritele kaalu, mis näitab klassifikaatori olulisuse määra. Kuna lause optimaalne pikkus ja sõnade sagedus korpuses mõjutasid GDEX 1.4 väljundit kõige rohkem, määrasin neile ka kõige suurema kaalu. GDEX 1.4 on kirjeldatud põhjalikult artiklis [P3]. Selle kõik parameetrid on toodud koondtabelina lisas 1 ning konfiguratsioonifail lisas 4.

Kuna GDEX 1.4 abil loodi „Eesti keele õppekorpus 2018 (etSkELL)“, millele omakorda toetuvad keeleõppekeskkond etSkELL ja veebilaused keeleportaalis Sõnaveeb, olen selle versiooni väljundit ka evalveerinud (loe lähemalt ptk 4.3 ja [P4]).

4.1.4. GDEXi versioonid eri keeleoskustasemetele

Kuigi olen GDEXi eesti keele moodulit arendanud „Eesti keele naabersõnade 2019“ koostamisega paralleelselt ehk eelkõige eesti keele B2–C1-oskustaset silmas pidades, siis ei olnud varasemalt sobiva andmestiku puudumise tõttu võimalik arvestada keeleoskustasemele spetsiifilisi lause parameetreid. Samal põhjusel ei olnud varem võimalik luua õppekorpusi eri keeleoskustasemega sihtgruppidele. 2018. aastal loodud „Eesti keele A1–C1 õpikute korpus (2018)“, kus iga lause on märgendatud keeleoskustasemega A1, A2, B1, B2 või C1, võimaldas eri keeleoskustaseme lauseid iseloomustavad parameetrid välja selgitada.

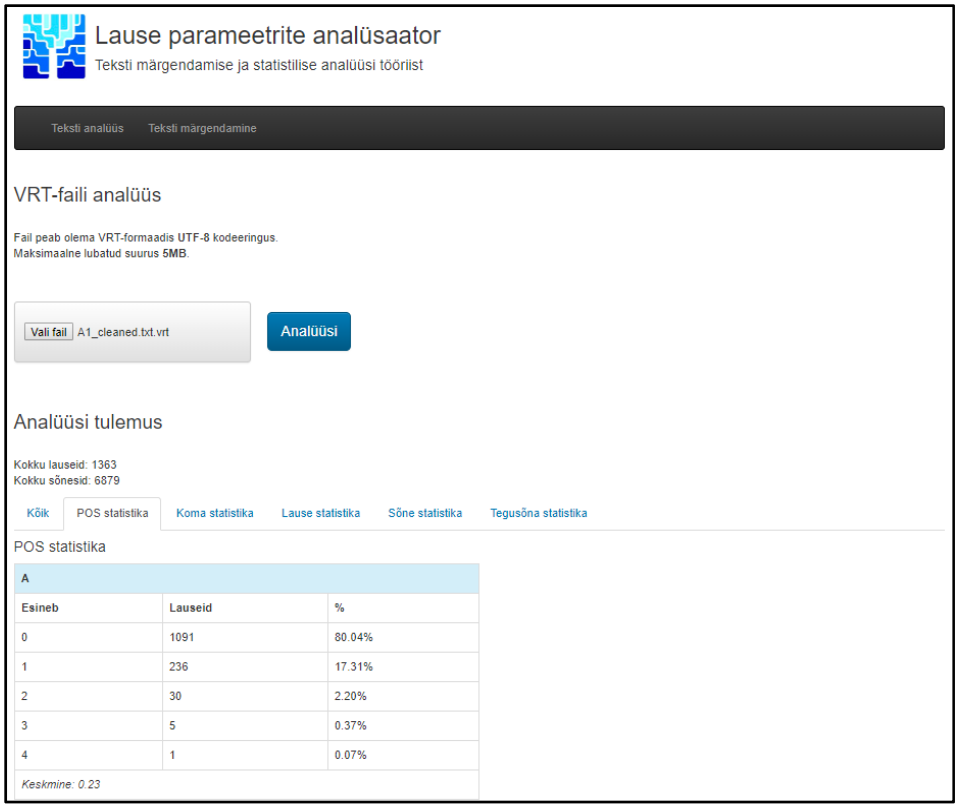
Selleks, et kokku sobitada GDEXi eesti keele mooduli konfiguratsioonid konkreetsete keeleoskustasemetega, lõin õpikute korpuse lausetest (õpikulausetest) viis andmebaasi (tabel 2). Kuna andmebaaside suurused ei olnud võrdsed, analüüsisin lausete parameetreid küll alltasemete kaupa (A1, A2, B1, B2, C1), kuid GDEXi eesti keele mooduli versioonid lõin üldistele keeleoskustasemetele: versiooni etBasic-v1 A-tasemele, versiooni etIndependent-v1 B-tasemele ja versiooni etProficient-v1 C-tasemele.

Tabel 2. Õpikulausete andmebaaside suurused

keeleoskustase	lauseid	sõnesid
A1	1363	6879
A2	3342	19215
B1	5462	39516
B2	5453	47451
C1	977	9569

Õpikulauseid analüüsisin Eesti Keele Instituudis loodud teksti märgendamise ja statistilise analüüsi tööriista „Lause parameetrite analüsaator“[18] abil (joonis 5). Analüsaator võimaldab mõõta lause ja sõnede pikkust, lause sõnaliigilist koosseisu, komade arvu lauses, lause esimest sõnaliiki ja tegusõnavormide esinemist lauses. Parameetrite valik toetub GDEX 1.4 arendamiseks tehtud analüüsile [P3] ning ka versioonide etBasic-v1, etIndependent-v1 ja etProficient-v1 konfiguratsioonide aluseks võtsin versiooni 1.4. Versioone on lähemalt kirjeldatud artiklis [P5] ning nende kõik parameetrid on toodud ka lisa 1 koondtabelis (konfiguratsioonifailid on esitatud lisades 5, 6 ja 7).

[18] Programmi autor on Eesti Keele Instituudi vanemtarkvaraarendaja Katrin Tsepelina.



Joonis 5. Programmi „Lause parameetrite analüsaator“ kasutajaliides

Eraldi korpuste loomine A-, B- ja C-keeleoskustasemetele ja nende väljundi evalveerimine kuulub edasiarenduste hulka.

4.2. GDEXi eesti keele mooduli eri versioonide parameetrid

Selles alapeatükis kirjeldan täpsemalt mõningaid tähtsamaid GDEXi eesti keele mooduli parameetreid, sealhulgas halli ja musta nimekirja, lause alguses keelatud sõnu ja sõnapaare ning karistada saavaid tegusõnavorme. Osa lause parameetreid on olnud kasutusel kõigis eelnevalt kirjeldatud GDEXi eesti keele mooduli versioonides. Need on järgmised:

- lause algab suure tähega ja lõpeb lauselõpumärgiga;
- lauses ei esine sõnu, mille sagedus korpuses on madalam kui 5;
- sõnad ei sisalda sümboleid;
- halli nimekirja (vt ptk 4.2.1) kuuluvad sõnad saavad karistada;
- lause ei alga sidesõnaga.

Eri versioonide kõik parameetrid on välja toodud lisa 1 koondtabelis.

4.2.1. Must ja hall nimekiri

„Eesti keele ühendkorpus 2013“ (563 mln sõnet), kust „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi jaoks GDEXi abil näitelauseid ekstraheeriti, sisaldas suuremas osas veebitekste. Kuna veebikorpused sisaldavad palju netikeelt, slängi, vulgarisme, (teadlikult) valesti kirjutatud sõnu jmt (vt näiteid 1–6), mis võivad olla tundlikud või solvavad ega pruugi seetõttu sobida õppesõnastiku näitelausetesse, rakendasin juba GDEXi eesti keele mooduli esimeses versioonis 1.2 nn halli nimekirja (*greylist*). Halli nimekirja kasutuselevõtt oli tingitud kahest eesmärgist, mis GDEXi eesti keele moodulit arendades algusest peale kaasas on käinud: selle sihtgrupp on „Eesti keele naabersõnade 2019“ sõnastiku kasutajad (B2–C1-keeleoskustasemel) ning luua selle abil keeleõppijatele kasutamiseks sobivaid õppekorpusi.

(1) For **fuck** sake, milleks seda vaja oli. (Estonian Web 2013)

(2) **Siuke** väike poiss, minust lühem, aga hull kiire naaskel. (Estonian Web 2013)

(3) **Kuradi jobukari** mõtlesin ma, aga noh las ta jääb ja jätkasime liikumist kui järsku ühel hetkel olidki meie ees vastutegevus, kes meilt konkreetselt talonge nõudsid. (Estonian Web 2013)

(4) Ole õnnelik, et sa pole veel tina saanud **kuradi värdjas**. (Estonian Web 2017)

(5) Anna kõigile **russidele** kodakondsus, tee ainukeseks riigikeeleks vene keel ja ühine venemaaga, siis saabub rahu. (Estonian Web 2013)

(6) **Türa** ma ütlen teil on järelikult **pask** arvuti kui video käima ei lähe! (Estonian Web 2017)

Halli nimekirja aluseks oli OÜ Filosofti[19] nimekiri sõnadest, mida eesti keele speller ei tohi valesti kirjutatud või tundmatute sõnade asenduseks pakkuda. Täiendasin nimekirja „Eesti keele seletava sõnaraamatu“ (2009), „Eesti keele sõnaraamatu 2019“ ja „Eesti keele sõnaperede“ (Vare 2012) stiilimärgenditele (vulgaarne, halvustav, kõnekeelne, släng) toetudes. Samuti olen nimekirja lisanud interneti akronüüme (nt *omg*, *wtf*, *lol*, *irw*), võõrkeelseid sõimusõnu (nt *fuck*, *pohui*) ja nende mugandatud variante (nt *fakk*, *pohh*) ning kirjakeele normist erinevalt kirjutatud sõnu (nt *shantazheerima*, *päälegi*). [P3]

Kui lauses esines ükskõik milline halli nimekirja kuuluv sõna, sai see lause võrdselt karistada, olenemata sellest, kas tegemist oli vulgaarse (nt *puts*) või kõnekeelse sõnaga (nt *ment*). Sellesse kuulus 2017. aasta seisuga kokku 476 sõna. Enne „Eesti keele õppekorpuse 2018 (etSkELL)“ loomist jagasin halli nimekirja kuuluvad sõnad kaheks: halli ja musta (*blacklist*) nimekirja. Halli nimekirja liigitasin sõnad, mille eest saab lause karistada; musta nimekirja sõnad, mis on lausest keelatud. Hall ja must nimekiri on mõistena kasutusel ka teiste keelte

[19] Autor tänab Heiki-Jaan Kaalepit sõnaloendi eest.

konfiguratsioonides ning need võivad tähistada üldisemaid üksusi (sõnu, märke, sõnaliike jne), mis saavad karistada või on keelatud. Portugali konfiguratsioonis on näiteks eesti konfiguratsiooni mõistes halli nimekirja kuuluvad sõnad nimetatud tabusõnadeks (*taboo words*). Musta ja halli nimekirja kuuluvad sõnad ei esine eesti õppesõnastikes märksõnadena mitmel (pedagoogilisel) põhjusel: madal sagedus, registriline kuuluvus (madalkeelsus, släng), ebavajalikkus igapäevases elus jmt.

Musta nimekirja liigitasin vulgarismid (nt *munn*), halvustavad sõnad (nt *pede*), võõrkeelsed sõimusõnad (nt *bljat, fuck*), erinevaid rahvusi halvustavad sõnad (nt *tibla*, *murjam*, *negru*), veebilausetes sageli esinevad võõrkeelsed sõnad (nt *awesome*, *story*) jmt. Must nimekiri liigitub tugevate klassifikaatorite alla, mis tähendab, et GDEX filtreerib automaatselt välja kõik laused, milles esineb kasvõi üks musta nimekirja kuuluv sõna. Musta nimekirja kuulub väitekirja kirjutamise seisuga 267 sõna[20] (vt lisa 8).

Halli nimekirja kuuluvad halvustavad ja tundlikud sõnad (nt *bimbo*, *idikas*, *pedofiil*, *grupikas*), slängisõnad (nt *muti*), kirjakeele normist erinevalt kirjutatud sõnad (nt *zhest*, *öigus*), kõne- ja netikeelsed sõnad (nt *burks*, *bemar*) jmt. Tulenevalt geopeituse logidest, mida veebikorpused „Estonian Web 2013“ ja „Estonian Web 2017“ massiliselt sisaldavad, lisasin halli nimekirja veel geopeitusele omased sõnad *aare*, *mugu*, *leid* ja *peidukoht* (vt näiteid 7–11).

(7) Igal juhul olime lõpuks õiges kohas ja avanes **aare**. (Estonian Web 2017)

(8) Pika otsimise peale paljastus õnneks **aare** ka. (Estonian Web 2017)

(9) Ilmselt vihmase ilma tõttu ühtki **mugu** ei kohanudki. (Estonian Web 2017)

(10) Õiges punktis matkarajalt kõrvale astuda ja kiire **leid**. (Estonian Web 2013)

(11) Tõstsin **peidukoha** raskusastet ühe punkti võrra. (Estonian Web 2013)

Halli nimekirja kuulub väitekirja kirjutamise ajal 451 sõna (vt lisa 9). Hall nimekiri liigitub nõrkade klassifikaatorite alla, mis tähendab, et GDEX karistab lauset iga halli nimekirja kuuluva sõna eest. Halli nimekirja rakendamine aitab tagada, et nimekirja kuuluvaid sõnu sisaldavad näitelaused ei satu GDEXi poolt pakutud kandidaatide nimekirja etteotsa, aga ei tähenda, et need sõnad GDEXi väljundis üldse ei esine.[21]

[20] Must ja hall nimekiri ei ole lõplikud ning sõnu võib edaspidi veelgi lisanduda.

[21] Sama kehtib ka teiste nõrkade klassifikaatorite alla liigituvate parameetrite kohta. Näiteks on GDEXi eesti keele moodulis määratud karistus lausetele, kus esinevad teatud märgid, nt jutumärgid, kuid see ei tähenda, et jutumärkidega laused ei satu valikusse üldsegi – need ei ole lihtsalt GDEXi poolt pakutavate näitelause kandidaatide seas esimeste hulgas.

4.2.2. Lause alguses keelatud sõnad ja sõnapaarid

Vähendamaks kontekstisidusate lausete sattumist väljundisse, olen alates GDEXi eesti keele mooduli versioonist 1.3 [P1] rakendanud klassifikaatorit, mis keelab lause alguses teatud sõnade esinemise. Need sõnad on oma olemuselt anafoorsed ehk viitavad seosele lausest välja ning nendest aru saamiseks on vaja laiemat konteksti (näited 12–13). Eesti keeles käituvad anafoorina harilikult asesõnad, kuid siinses väitekirjas käsitlen anafoore laiemalt, liigitades nende alla näiteks ka deiksised ja konnektiivlaiendid.

(12) **Seejärel** konvektiivne aktiivsus vaibus. (Estonian Web 2013)

(13) **Näiteks** kui sõlmida fiktiivne laenuleping vms. (Estonian Web 2013)

GDEX 1.3 nimekirjas oli anafoorseid sõnu viis (*näiteks*, *kui*, *ühesõnaga*, *seejärel*, *nagu*), GDEX 1.4 nimekirjas 62: *aga*, *ega*, *ehk*, *esiteks*, *hoolimata*, *ikka*, *iseasi*, *jah*, *ju*, *just*, *järelikult*, *järgnevalt*, *ka*, *lihtsalt*, *muidu*, *nad*, *nagu*, *nemad*, *niisiis*, *niisugune*, *nimelt*, *no*, *noh*, *nõnda*, *näiteks*, *ometi*, *pealegi*, *pigem*, *põhjuseks*, *samamoodi*, *samas*, *samuti*, *seal*, *sealjuures*, *see*, *see-eest*, *seega*, *seejuures*, *seejärel*, *seepeale*, *seepärast*, *seetõttu*, *seevastu*, *sellegipoolest*, *sellekohaselt*, *sellepärast*, *selletõttu*, *seniks*, *sestap*, *siin*, *siis*, *säärane*, *tagajärjeks*, *teiseks*, *teisisõnu*, *tere*, *too*, *vastupidi*, *või*, *võrdluseks*, *ühesõnaga*, *ülejäänud*.

Väitekirja kirjutamise ajal olen nimekirja oluliselt täiendanud ning lisanud 40 sõna: *mistõttu*, *misjärel*, *misläbi*, *selle-eest*, *sellevastu*, *nõndasamuti*, *nõndasama*, *nõndamoodi*, *nõndaviisi*, *samalaadne*, *samasugune*, *samaviisi*, *sealtkandist*, *sealtkaudu*, *sealtmaalt*, *sealtpeale*, *sealtpoolt*, *sealtsaadik*, *sealtsamast*, *sedamoodi*, *sedapidi*, *sedasi*, *sedaviisi*, *sedakaudu*, *seekõrval*, *seepoolest*, *seeläbi*, *sellepoolest*, *seevõrd*, *seevõrra*, *sellegipärast*, *seetarvis*, *sellejagu*, *sellekohane*, *sellevõrra*, *selliselt*, *sinna*, *sinnajuurde*, *sinnani*, *teisalt*.

GDEX 1.4 nimekirjas on 70 sõnapaari, mis on lause alguses potentsiaalse anafoorsuse tõttu keelatud (näited 14–15): *Ainult et*, *Ainult nii*, *Ehk siis*, *Ehk teisisõnu*, *Eriti juhul*, *Eriti just*, *Eriti kui*, *Eriti siis*, *Eriti veel*, *Isegi siis*, *Just need*, *Just nii*, *Just see*, *Just seetõttu*, *Kuid juhtumisi*, *Küll aga*, *Lisaks sellele*, *Muidugi eeldusel*, *Muidugi ka*, *Nii et*, *Nii nagu*, *Nüüd aga*, *Peale seda*, *Peale selle*, *Sama asi*, *Sama kehtib*, *Samal ajal*, *Samal põhjusel*, *Samal viisil*, *Seda enam*, *Seda eriti*, *Seda kõike*, *See omakorda*, *See tähendab*, *Selleks ajaks*, *Selleks on*, *Selleks peab*, *Selleks pead*, *Selleks peaks*, *Selleks peame*, *Selleks peavad*, *Sellele vaatamata*, *Selles mõttes*, *Selles osas*, *Selles valguses*, *Sellest hoolimata*, *Sellest johtuvalt*, *Sellest lähtudes*, *Sellest lähtuvalt*, *Sellest omakorda*, *Sellest tulenevalt*, *Sellisel juhul*, *Sellisel moel*, *Sellisel puhul*, *Ses mõttes*, *Teisel juhul*, *Teisel korral*, *Teiselt poolt*, *Teisest küljest*, *Teisiti öeldes*, *Teiste* sõnadega, *Umbes nagu*, *Vaatamata sellele*, *Vastasel juhul*, *Vastasel korral*, *Veel enam*, *Veelgi enam*, *Viimasel juhul*, *Välja arvatud*, *Ühelt poolt*.

(14) **Sellele vaatamata** hakkasid umbes 1993. aastast asjad viltu kiskuma. (Estonian Web 2013)

(15) **Teiselt poolt** tagab sõna, et sõltumata oludest toimub asjaajamine eesti keeles. (Estonian Web 2013)

Väitekirja kirjutamise ajal olen nimekirja täiendanud kümne sõnapaariga: *Ikka selleks*, *Ikka seetõttu*, *Ikka seda*, *Ilmselt et*, *Ilmselt isegi*, *Ilmselt seetõttu*, *Ilmselt ka*, *Ilmselt kuna*, *Muudkui et*, *Seni aga*.[22]

Samuti on alates versioonist GDEX 1.2 lause alguses keelatud teatud sõnaliigid, nt sidesõnad (loe lähemalt [P3]: 59–61, [P5]: 104–105; vt ka koondtabelit lisas 1).

4.2.3. Tegusõnavormid

Lausete grammatilise lihtsuse huvides said alates GDEXi eesti keele mooduli esimesest versioonist 1.2 näitelause kandidaadid karistada, kui seal esinesid teatud käändelised tegusõnavormid (*-mata*, *-mast*, *-mas*, *-maks*, *-des*).[23] See otsus põhines introspektiivsel analüüsil, mis osutas, et kantseliitlikumale ja ametlikumale registrile on iseloomulik käändeliste tegusõnavormide ületarvitus. 2018. aastal läbi viidud õpikulausete analüüs aitas aga empiiriliselt välja selgitada, missugused tegusõnavormid on tasemekohased vastavatele keeleoskustasemetele.

Õpikulausete analüüsi tulemus näitas, et *des*-vorm; *ma*-tegevusnime käändelised vormid *-mast*, *-maks*, *-mata*, *-tama*; tingiva kõneviisi vormid *-nuks*, *-taks*, *-tuks*; kaudse kõneviisi vormid *-tavat*, *-tuvat*, *-nuvat*; käskiva kõneviisi vormid *neg_gu* ja *-tagu* ja umbisikulise tegumoe vormid *-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da* A-taseme lausetes ei esine, mistõttu keelati need versioonis etBasic-v1. Karistada saab *tud*-vorm. B-tasemele suunatud versioonis etIndependent-v1 on keelatud *ma*-tegevusnime vormid *-maks* ja *-tama*; tingiva kõneviisi vormid *-nuks*, *-taks*, *-tuks*; kaudse kõneviisi vormid *-tavat*, *-tuvat*, *-nuvat* ja käskiva kõneviisi vormid *neg_gu* ja *-tagu*. Umbisikulise tegumoe vormid *-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da* saavad karistada. C-tasemele suunatud versioonis etProficient-v1 saavad karistada tingiva kõneviisi vormid *-nuks* ja *-taks* ning kaudse kõneviisi vormid *-tuks*, *-tavat*, *-tuvat* ja *-nuvat*.

4.3. GDEX 1.4 väljundi evalveerimine

Kuna GDEX 1.4 abil loodi „Eesti keele õppekorpus 2018 (etSkELL)“, mis on omakorda allikaks keeleõppekeskkonnale etSkELL ja veebilausetele keeleportaalis Sõnaveeb, olen selle versiooni väljundit evalveerinud. Kuna Eesti Keele Instituudis on lähitulevikus plaanis luua õppekorpused eri keeleoskustasemega sihtgruppidele, olen loonud GDEXi mooduli versioonid ka üldistele keeleoskustasemetele (A-, B- ja C-tasemele), kuid nende versioonide evalveerimine ja nende versioonide alusel õppekorpuste loomine jääb edaspidiseks ülesandeks.

[22] Uus nimekiri lause alguses keelatud sõnadest ja sõnapaaridest läheb kasutusele järgmises GDEXi versioonis, mille abil luuakse uus õppekorpus.

[23] Tegusõnavormide eristamisel toetun eesti keele tekstianalüsaatori EstNLTK märgendusele.

Keskendun artiklis [P4] GDEX 1.4 väljundi evalveerimisele. Selleks palusin Eesti Keele Instituudis töötavatel leksikograafidel ning Tartu ja Tallinna Ülikoolis eesti keelt B2–C1-oskustasemel valdavatel üliõpilastel täita hindamisülesanne, millele järgnes jätkuküsitlus.[24] Hindamisülesande lõin avatud lähtekoodiga platvormis Pybossa, mida kasutatakse rahvahanke projektide läbiviimiseks ning kogutud andmete analüüsimiseks. Pybossa võimaldab oma hindamisülesannet ise kujundada, kontrollida osalejate arvu ning hoiustada kogutud andmeid. Hindamisülesande eesmärk oli välja selgitada, kas:

1. korpuslausete filtreerimine on vajalik;
2. GDEX suudab tuvastada sobivaid näitelause kandidaate ja välja filtreerida sobimatuid;
3. leksikograafi koostatud sõnastiku näitelaused on hindajate arvates sobivad näitelaused.

Selleks, et hinnata, kas GDEX suudab tuvastada sobivaid näitelause kandidaate ja välja filtreerida sobimatud, lisasin hindamisülesandesse korpuslauseid, mis vastasid GDEX 1.4 järgi hea näitelause parameetritele, ning korpuslauseid, mis GDEX 1.4 järgi hea näitelause parameetritele ei vastanud. Selleks, et hinnata, kas korpuslausete filtreerimine on üleüldse vajalik, lisasin hindamisülesandesse ka korpuslauseid, mille kohta ei olnud teada, kas need vastasid hea näitelause parameetritele või mitte. Leksikograafi koostatud näitelaused lisasin andmestikku kontrollgrupiks, et näha, kuidas neid korpuslausetega võrreldes hinnatakse. Hindamisülesande laiem eesmärk oli evalveerida GDEX 1.4 tulemusi.

Hindamiseks vajaminevate märksõnade juhuvalim võeti automaatselt Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteemi Ekilex andmebaasist SQLi funktsiooniga *random()*.[25] SQL (*Structured Query Language*) on päringukeel, mida kasutatakse relatsiooniliste andmebaasidega suhtlemiseks. Päring juhuvalimi saamiseks oli järgmine:

```
select f.value
from word w
 left join paradigm p on p.word_id = w.id
 left join form f on f.paradigm_id = p.id
 left join lexeme l on l.word_id = w.id
where f.mode = 'WORD'
and  l.dataset_code = 'kol'
and  exists (select lp.pos_code
      from lexeme_pos lp
      where lp.lexeme_id = l.id
      and  lp.pos_code = 's')
order by random() limit 10;
```

[24] Vastava keeleoskustasemega üliõpilasi aitas leida Tartu Ülikooli eesti keele võõrkeelena dotsent Raili Pool. Autor tänab Railit abi eest.

[25] Autor tänab Arvi Tavastit päringu koostamise eest.

Päringu eeltingimus oli, et Ekilexis sisalduv märksõna kuulub ka „Eesti keele naabersõnade 2019“ kui B2–C1-keeleoskustasemele suunatud sõnastiku andmebaasi. Kokku valiti 40 märksõna:

- 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*.

Seejärel võeti iga märksõna jaoks juhuvalim näitelausetest[26], kuhu kuulus kokku 160 lauset (iga märksõna jaoks neli erinevat tüüpi lauset):

- üks korpuslause, mis GDEX 1.4 järgi vastab hea näitelause parameetritele;
- üks korpuslause, mis GDEX 1.4 järgi hea näitelause parameetritele ei vasta;
- üks filtreerimata korpuslause, mis võis vastata nii hea kui ka halva näitelause parameetritele;
- üks leksikograafi koostatud näitelause „Eesti keele sõnaraamatust 2019“.

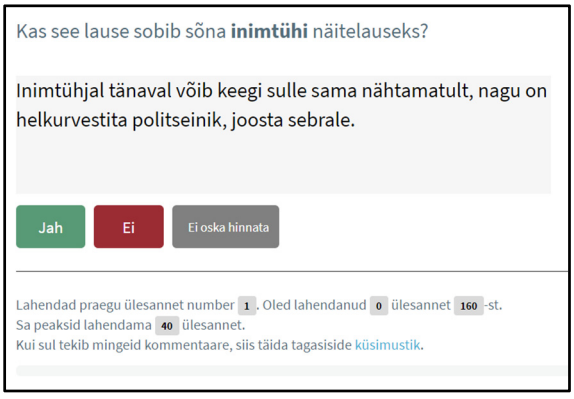
Kuna hindajate rühmi oli kaks – leksikograafid ja keeleõppijad –, tegin Pybossas kaks sama sisuga projekti. Neis mõlemas sai ühte lauset hinnata viis erinevat hindajat – kui üks lause kogus ühe projekti sees juba viie erineva leksikograafi või viie erineva keeleõppija hinnangu, siis seda järgmisele hindajale enam ei kuvatud. Kokku said kutse hindamises osaleda seitse leksikograafi ja 31 keeleõppijat, kellest osales viis leksikograafi ja üheksa keeleõppijat. Kutse saanud leksikograafide seas oli kolm „Eesti keele naabersõnade 2019“ ja neli „Eesti keele sõnaraamatu 2019“ koostajat, kes tol hetkel neidsamu sõnastikke aktiivselt koostasid.

Otsustasin ühe projekti sees piirduda viie hinnanguga ühele lausele, kuna ei saanud eeldada, et iga osaleja on valmis hindama kogu andmestikku ehk kokku 160 lauset (mida viis leksikograafi küll tegid). Kasutajate motiveerimine on teadaolev probleem rahvahanke projektide läbiviimisel (vt nt Leimeister jt 2009, Kaufmann jt 2011).[27] Sellest probleemist teadlikuna jagasin andmestiku neljaks väiksemaks ülesandeks, kusjuures iga väiksem ülesanne sisaldas kõiki nelja tüüpi lauseid. Kuigi väiksemas ülesandes esitati korraga hindamiseks vaid 40 lauset, jättis osa keeleõppijaid ka selle tegemise pooleli, kuid antud hinnangud läksid sellegipoolest arvesse. Hindamisülesandes osales kokku üheksa tudengit.

[26] Korpuslausete juhuvalimi päringu tegi „Eesti keele ühendkorpusest 2017“ Lexical Computing Ltd. tarkvaraarendaja Jan Michelfeit.

[27] Rahvahankes motiveeritakse inimesi osalema nt rahalise tasuga, kinkekaartidega vmt. Samuti kasutatakse ülesannete mängustamist, nt pannakse rahvahanke projektis osalejad punkte koguma ja omavahel võistlema.

Hindajale näidati ühte lauset korraga. Ülesande läbiviimise lihtsustamiseks andsin hindajale ette väga üldised vastusevariandid. Märksõna definitsiooni ega lause allikat lisatud ei olnud, seega ei olnud leksikograafid teadlikud sellest, et hinnatavate lausete hulka on kontrollgrupiks lisatud ka nende endi koostatud või valitud näitelauseid. Joonisel 6 olev lause on GDEX 1.4 parameetritele vastav korpuslause.



Joonis 6. Lause hindamine Pybossa platvormis (keeleõppija vaade)

Hindamisülesande tulemuste analüüs näitas, et sobivaks hinnati 95% sõnaraamatu näitelausetest, 80% GDEX 1.4 parameetritele vastavatest korpuslausetest, 40% filtreerimata korpuslausetest ning 20% GDEX 1.4 parameetritele mittevastavatest korpuslausetest. Kuna 5% sõnastiku näitelausetest, 20% GDEX 1.4 parameetritele vastavatest ja 60% filtreerimata korpuslausetest hinnati sobimatuks; samuti hinnati sobivaks 20% GDEX 1.4 parameetritele mittevastavatest korpuslausetest, tahtsin kuulda hindajate endi põhjendusi nende lausete sobivuse või mittesobivuse kohta. Selleks viisin keskkonnas Google Forms läbi jätkuküsitluse, mis oli oluliselt väiksema mahuga: leksikograafid said uuesti hindamiseks 18 lauset ning keeleõppijad 20 lauset. Jätkuküsitluses esitasin hindajale need laused koos 18 erineva vastusevariandiga (nt *lause on liiga pikk*, *alus/tegija puudub*, *vajab rohkem konteksti* jms), mille hulgast sai valida mitu vastust, seal-hulgas oma otsust põhjendada.

Jätkuküsitluse tulemused näitasid esiteks, et vastusevariantide etteandmine mõjutab hindajate arvamust, kuna hinnang lausetele muutus. Teiseks näitasid tulemused, et hindajate lausete sobimatuse põhjendused pigem erinesid kui ühtisid. Lausete sobimatuse puhul toodi kõige sagedamini välja anafoorsust, konteksti puudumist, lause pikkust ja kõnekeelsust. Järelikult tuleb GDEXi eesti keele 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 vähe konteksti (samale järeldusele jõudis ka Kuhn 2017: 265)

Jätkuküsitluse järel hinnati sobivaks pea kõik (96%) leksikograafi koostatud näitelaused. GDEX 1.4 parameetritele vastavatest korpuslausetest hinnati sobivaks

85% ning koguni 94% GDEX 1.4 parameetritele mittevastavatest korpuslausetest hinnati sobimatuks. Tulemused näitavad, et korpuslausete filtreerimine on vajalik ning et GDEX töötab edukalt korpusest sobivate näitelause kandidaatide tuvastamisel ja sobimatute korpuslausete välja filtreerimisel.

Eespool kirjeldatud hindamisülesanne oli katse kombineerida väitekirja tulemuste hindamist rahvahankega. Ülesande saanuks üles ehitada ka teisiti, näiteks kasutada andmestikuna konkreetse märksõna kõiki GDEX 1.4 parameetritele vastavaid lauseid või teatud hulka erinevate märksõnade GDEX 1.4 parameetritele vastavatest lausetest. Samuti oleksid tulemused veenvamad, kui informantide hulk olnuks suurem ning hindajate rühmi rohkem, nt võinuks kaasata ka eesti keele kui teise keele õpetajaid, kellel on hea arusaam erinevatel keeleoskustasemetel olevatest õppijatest. Samuti ei kogutud evalveerimisülesande käigus personaalset infot (vanus, sugu, sünni- ja elukoht jmt) hindajate kohta, mistõttu ei olnud võimalik analüüsida nende individuaalseid erinevusi, mis võisid ka hindamistulemusi mõjutada. Minu esimene kogemus rahvahanke meetodil läbi viidud katsega näitas, et esiteks on keeruline leida vajalikul keeleoskustasemel olevaid informante ning teiseks neid osalema motiveerida. Väitekirja kirjutamise hetkel kuulun COSTi keeleõppe ja rahvahanke ühendamise Euroopa võrgustikku enetCollect, mille raames püütakse muuhulgas välja töötada masinõppe algoritmi, mis suudaks tuvastada tundliku sisuga laused ning need korpusest kõrvaldada. Selle projekti raames on plaanis küsitleda ka suuremat eesti informantide hulka (osaleda saavad ka emakeelsed) ning selle tulemusel loodud algoritmi rakendada uue korpuse loomisel.

4.4. GDEXi rakendamine

GDEXi eesti keele mooduli versioone on kasutatud „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi täisautomaatsel genereerimisel ja autentseid lauseid sisaldavate õppekorpuste („EstonianNC GDEX“ ja „Eesti keele õppekorpus 2018 (etSkELL)“) loomisel. „Eesti keele õppekorpus 2018 (etSkELL)“ on omakorda allikaks automaatsele keeleõppekeskkonnale etSkELL ehk Sketch Engine for Estonian Language Learning ja veebilausetele keeleportaalis Sõnaveeb.

Kuna olen välja töötanud ka GDEXi versioonid eri keeleoskustasemetele, saab tasemekohased õppekorpused tulevikus luua ka A-, B- ja C-tasemele. Õppekorpuste loomise eesmärgil tuleb rakendada ka leksikaalset filtrit, milleks saab kasutada nt 2018. aastal valminud sõnaloendeid (Kallas, Koppel 2018a, 2018b, 2018c, loe lähemalt ptk 5.5). Leksikaalne filter aitab tagada, et õppekorpus sisaldab ainult selliseid lauseid, mis vastavad hea näitelause parameetritele ning sisaldavad vastavale keeleoskustasemele kohast sõnavara.

4.4.1. GDEXi eesti keele moodul Sketch Engine'is

Korpuspäringusüsteem Sketch Engine kasutab 2019. aasta seisuga eesti korpuste lausete filtreerimiseks vaikimisi GDEXi eesti keele mooduli versiooni 1.4, seega saavad seda kasutada kõik Sketch Engine'i kasutajad. Seoses Horisont 2020 projektiga ELEXIS arendatakse Sketch Engine'i kõrval paralleelselt sõnastikusüsteemi Lexonomy (Měchura 2017), mis võimaldab sõnastikke koostada ühekliki sõnaraamatu (*OneClick Dictionary*) meetodil. Lexonomy rakendab Sketch Engine'i sõnastike automaatse koostamisega seotud funktsioone: märksõna loendite genereerimist, sõnaliikide, kasutusmärgendite, näitelausete, kollokatsioonide, sünonüümide ja tesauruse, definitsioonide ja/või tõlkevastete automaatset tuvastamist. Tuvastatud üksused ekstraheeritakse ja kantakse automaatselt sõnastikusüsteemi üle. Sketch Engine'i ja Lexonomi loojad leiavad, et leksikograafid ei peaks kulutama oma aega tööle, mida masinad on juba võimelised nende eest ära tegema, ning usuvad, et sõnastike andmebaaside automaatne loomine ja nende järeltoimetamine on nii ajalise kui ka rahalise ressursi seisukohalt kõige säästlikum viis sõnastikke koostada. Lexonomy süsteemis sõnastiku andmebaasi järeltoimetades saavad leksikograafid paralleelselt Sketch Engine'iga ühenduses olla. Ka eesti keele jaoks saab ühekliki sõnaraamatu meetodit rakendada ning sellisel juhul valib sõnastiku koostaja näitelauseid GDEXi 1.4 poolt pakutud kandidaatide seast.

4.4.2. „Eesti keele õppekorpus 2018 (etSkELL)"

2018. aastal lõi Eesti Keele Instituut koostöös tarkvarafirmaga Lexical Computing Ltd. „Eesti keele õppekorpuse 2018 (etSkELL)" (300 mln sõnet). Õppekorpus sisaldab GDEX 1.4 abil „Eesti keele ühendkorpusest (2017)" (1,3 mld sõnet) välja valitud lauseid, mis tähendab, et kõik õppekorpuse laused vastavad GDEX 1.4 poolt ette määratud hea näitelause parameetritele. Õppekorpuse loomise protsess oli kaheosaline. Kõigepealt filtreeriti tugevate klassifikaatorite abil välja kõik need „Eesti ühendkorpuse (2017)" laused, mis oma parameetrite poolest näitelauseks ei sobinud. Järelejäänud lausetele määrati nõrkade klassifikaatorite abil lõplik GDEXi skoor. Samuti on õppekorpusesse lisatud „Eesti keele A1–C1 õpikute korpuse (2018)" (121 000 sõna) laused. Õppekorpus on otsitav korpuspäringusüsteemides Sketch Engine ja KORP.

Erinevalt teist tüüpi korpustest ei sisalda SkELLi-sarja korpused terviklikke dokumente (eeldusel, et keeleõppijad neid ei vaja), vaid iseseisvaid korpuslauseid, mille skoor on teatud numbrist kõrgem (eesti keele moodulis peab skoor olema kõrgem kui 0,5). Kuna SkELL-tüüpi korpused ei sisalda terviktekste, on lausetevaheline kontekst puudu. Samas on lausepäring korpusest oluliselt kiirem, kuna kõik laused on juba hinnatud ning neid sorditakse ja kuvatakse kasutajale vaikimisi skoori alusel ehk paremuse järgi. (Koppel, Kallas jt 2019)

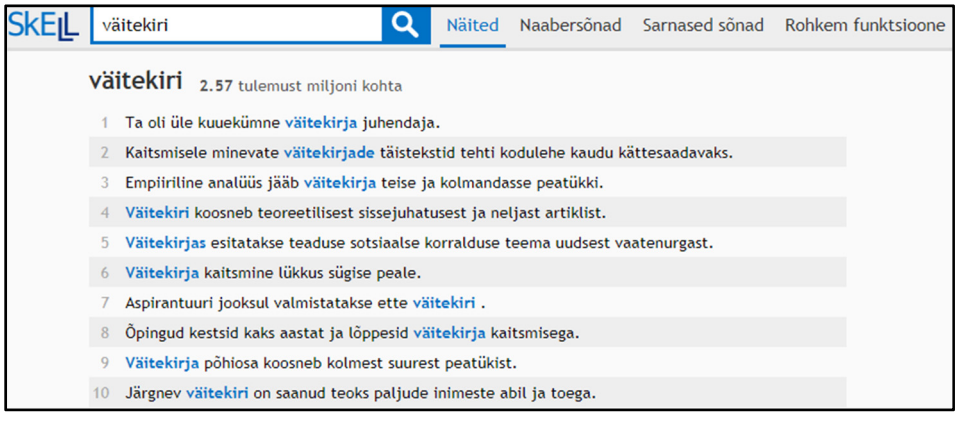
4.4.3. etSkELL ehk Sketch Engine for Estonian Language Learning

SkELL ehk Sketch Engine for Language Learning (Baisa, Suchomel 2014) on tuntuim keeleõppekeskkondade sari, kus kasutatakse GDEXi abil korpusest automaatselt tuvastatud lauseid. See on oma olemuselt korpuspäringusüsteemi Sketch Engine lihtsustatud versioon, mille abil saab kasutada selle olulisemaid funktsioone: lugeda näitelauseid (näited), vaadata naabersõnu ja sarnaseid sõnu (tesaurus). SkELLi keeleõppekeskkonna saab luua kõikidele keeltele, mille jaoks on olemas morfoloogiliselt märgendatud korpus(ed), süntagmaatiliste suhete tuvastamiseks koostatud sõnavisandite (Kilgarriff, Kovář jt 2010) grammatika (eesti keele kohta loe lähemalt Kallas 2013) ja välja töötatud korpuslausete valiku parameetrid ehk GDEXi konfiguratsioon. Seni on SkELLi keeleõppekeskkonnad loodud inglise, vene, saksa, itaalia, tšehhi ja eesti keele jaoks, loomisel on ka SkELLi korpus ja veebiliides portugali keele jaoks (Kuhn jt 2019).

SkELLi eelis on veebiliidese tehniline lihtsus, läbipaistev metakeel ja väljastatud info piiratud maht. SkELLi veebiliidese abil näeb keeleõppija n-ö valmis tükke, mis aitavad suurendada nii passiivset kui ka aktiivset sõnavara (Kilgarriff, Marcowitz jt 2015). Keeleõppijad ise on SkELLi suurimaks eeliseks pidanud vaba juurdepääsu autentsele keelematerjalile, mille abil nad saavad just neid huvitavaid leksikaalseid ja grammatilisi kombinatsioone ise uurida (Hirata, Hirata 2018).

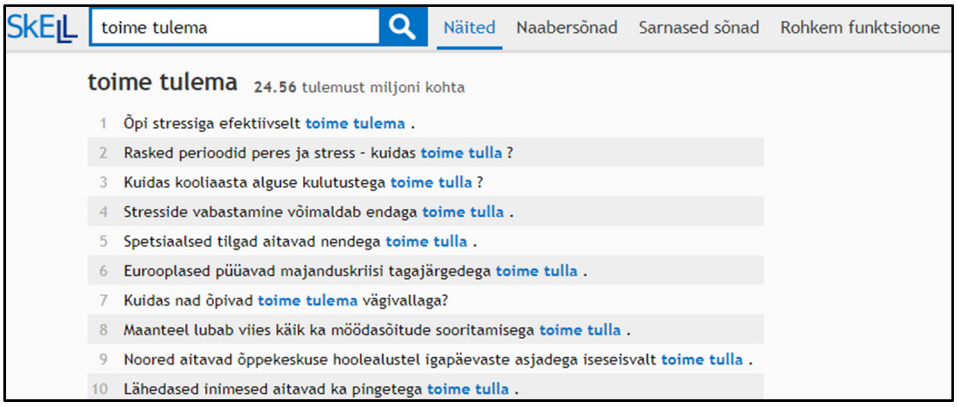
4.4.3.1. Näited

Näidete abil näeb keeleõppija otsitavat sõna või fraasi nende loomulikus ümbruses. etSkELLi päringuaknasse saab kirjutada nii lemma ehk sõna algvormi kui ka kindla sõnavormi. Lemma kuvatakse päringus erinevates sõnavormides (joonis 7), sõnavorm ainult selles konkreetses vormis, milles seda otsiti. Näitelaused kuvatakse GDEXi skoori alusel, st kõige paremad näitelause kandidaadid on nimekirja eesotsas.



Joonis 7. Lemma *väitekiri* eri sõnavormides näitelaused

Otsida saab ka mitmesõnalisi üksusi (nt verbiühend *toime tulema*), kuid sel juhul tuleb silmas pidada, et liides otsib sõnu täpselt niipidi, nagu need on päringuaknasse kirjutatud (vrd *toime tulema* (joonis 8) ja *tulema toime* (joonis 9)).

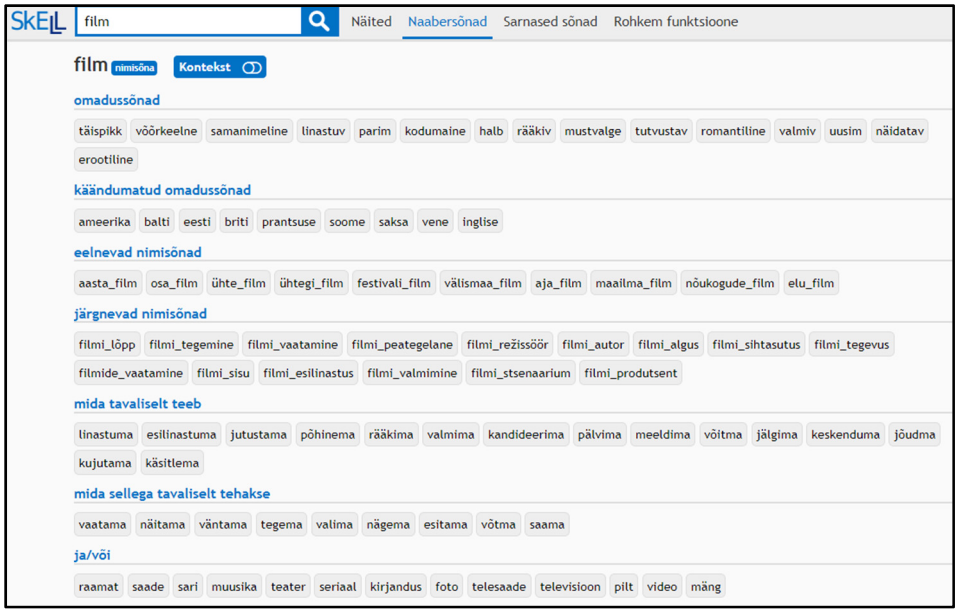


Joonis 8. Mitmesõnalise üksuse *toime tulema* näitelaused

Joonis 9. Mitmesõnalise üksuse *tulema toime* näitelaused

4.4.3.2. Naabersõnad

etSkELLi naabersõnade sakis kuvatakse sõnavisandeid, kust keeleõppija näeb märksõna tüüpilisemaid kollokaate. Sõnavisandid kuvatakse ainult sisusõnadele: nimisõnadele, omadussõnadele, tegusõnadele ja määrsõnadele. Kollokaadid on reastatud esilduvuse (*salience*) ehk sõnadevahelise seose tugevuse järgi ning on jaotatud gruppidesse vastavalt sellele, mis sõnaliiki märksõna kollokaat kuulub. Nimisõnal on kokku seitse gruppi: omadussõna + nimisõna, käändumatu omadussõna + nimisõna, eelnev nimisõna + nimisõna, nimisõna + järgnev nimisõna, tegusõna + nimisõna aluse funktsioonis, tegusõna + nimisõna sihitise funktsioonis, ja/või suhe (joonis 10).

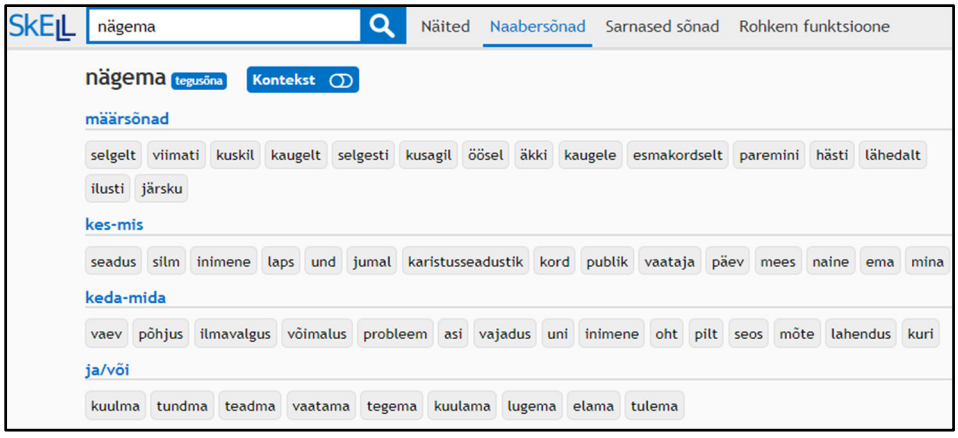


Joonis 10. Nimisõna *film* kollokaadid

Omadussõnal on kokku kolm gruppi: omadussõna + nimisõna, määrsõna + omadussõna, ja/või suhe (joonis 11).

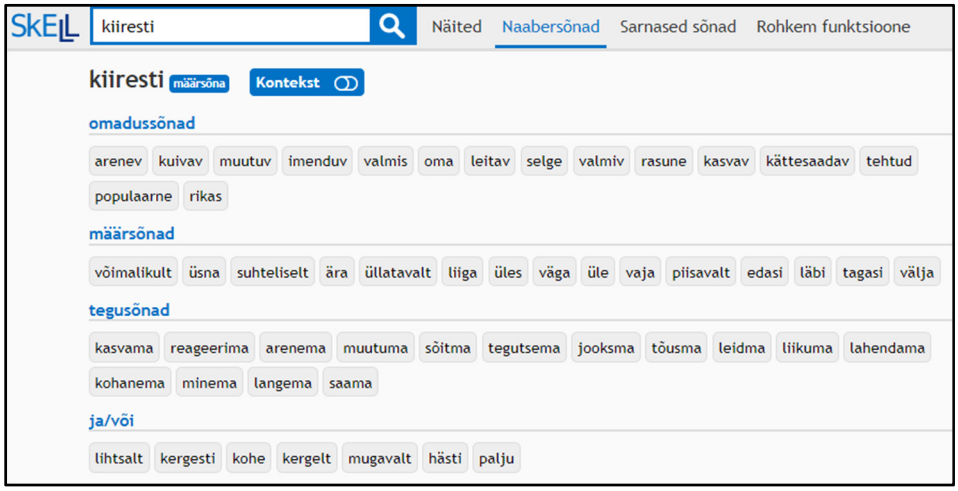
Joonis 11. Omadussõna *ilus* kollokaadid

Tegusõnal on kokku neli gruppi: määrsõna + tegusõna, nimisõna aluse funktsioonis + tegusõna, nimisõna sihitise funktsioonis + tegusõna, ja/või suhe (joonis 12).



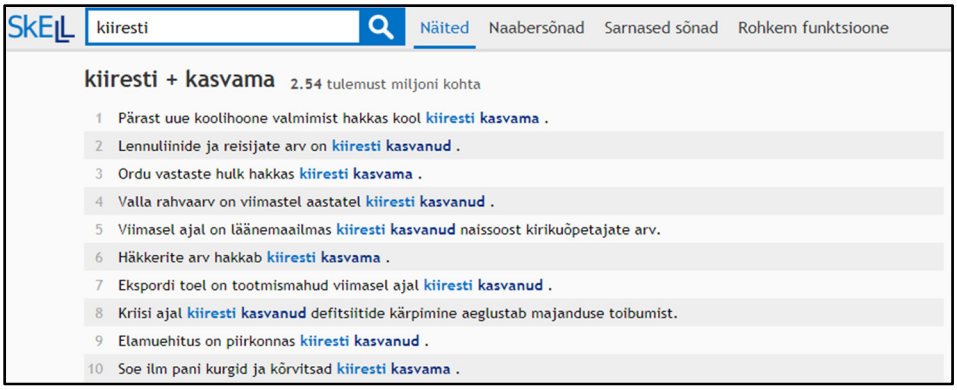
Joonis 12. Tegusõna *nägema* kollokaadid

Määrsõnadel on kokku neli gruppi: määrsõna + omadussõna, määrsõna + määrsõna, määrsõna + tegusõna, ja/või suhe (joonis 13).



Joonis 13. Määrsõna *kiiresti* kollokaadid

Jooniselt 13 on näha, et määrsõna *kiiresti* kõige esilduvamad tegusõna-kollokaadid on *kasvama*, *reageerima* ja *arenema*. Kollokaadile klikkides kuvatakse kollokatsiooni näitelaused (joonis 14).



Joonis 14. Kollokatsiooni *kiiresti kasvama* näitelaused

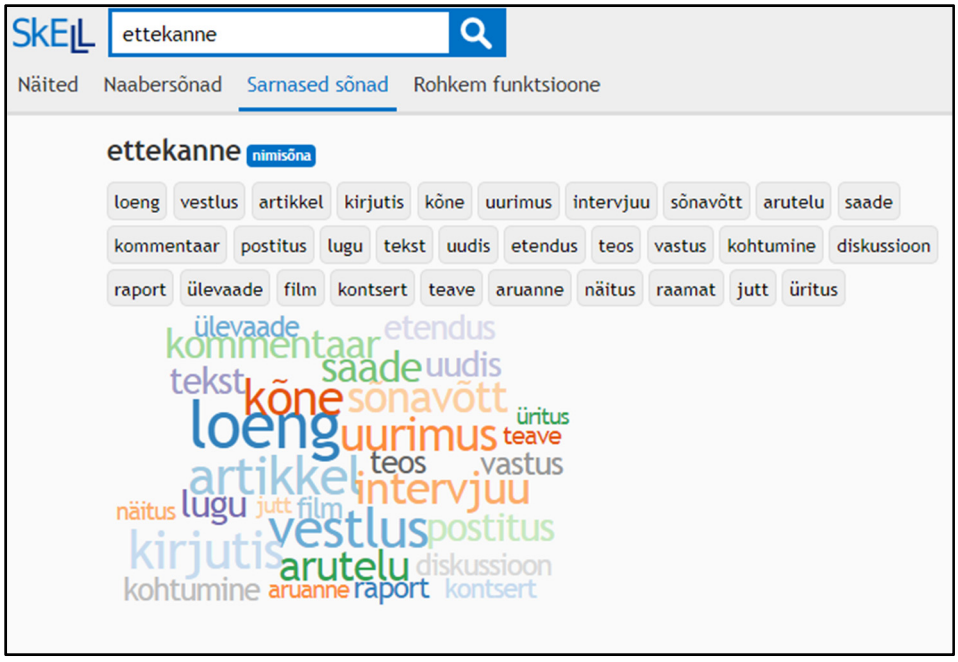
Naabersõnade sakis on võimalik vaadata ka kollokatsiooni pikimat tüüpilist konteksti (*longest-commonest match*) (Kilgarriff, Baisa jt 2015). See leitakse korpusest algoritmiga, mis toetudes kahesõnalistele kollokatsioonidele leiab üles kõige esilduvamad mitmesõnalised üksused. Joonisel 15 on vasakul pool toodud nimisõna *konts* sõnavisandid ilma kontekstita ja paremal koos kontekstiga.

Joonis 15. Nimisõna *konts* sõnavisandid ilma kontekstita (vasakul) ja koos kontekstiga (paremal)

Jooniselt 15 selgub, et kollokatsiooni *madal konts* kasutatakse enamasti kaasaütlevas käändes (*madala kontsaga*) ning kollokatsiooni *kõrge konts* kasutatakse pigem lausetes, kus räägitakse kõrgete kontsade kandmisest.

4.4.3.3. Sarnased sõnad

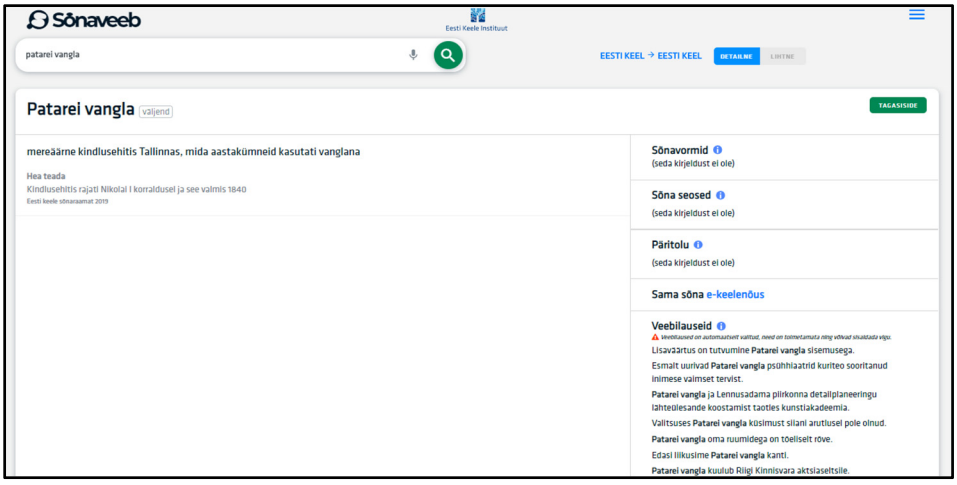
Sarnaste sõnade sakis kuvatakse automaatselt genereeritud pilv sõnadest, mis jagavad märksõnaga samu kollokaate – tavaliselt on nendel sõnadel sarnane tähendus. Sarnaste sõnade hulka võivad vahel sattuda ka sõnad, mis ei ole paradigmaatilises seoses või kuuluvad eri sõnaliikidesse. See tuleb sellest, et programm toetub sarnaste sõnade leidmisel üksnes eel- ja järelkontekstile. Sõnapilve keskel asuvad sõnad on otsitava sõnaga kõige sarnasemad ning kirja suurus näitab sõna sagedust (joonis 16).



Joonis 16. Nimisõna *ettekanne* sarnased sõnad

4.4.4. Keeleportaal Sõnaveeb

„Eesti keele õppekorpus 2018 (etSkELL)“ on integreeritud ka Eesti Keele Instituudi keeleportaali Sõnaveeb, kus kasutaja näeb leksikograafide koostatud info kõrval täiendavalt korpuslauseid. Eriti kasulikud on veebilaused sellistes sõnaartiklites, kus leksikograafi koostatud näitelaused puuduvad, nagu nt märksõnal *Patarei vangla*. Kuna mõiste *korpuslause* on sõnastiku tavakasutajatele võõras, kasutatakse kasutajaliideses mõistet *veebilause* (joonisel 17 all paremas nurgas). Veebilaused saadakse API (*application programming interface*) kaudu korpuspäringusüsteemist KORP.



Joonis 17. Märksõna *Patarei vangla* esitus Sõnaveebis

Sõnaveebis kuvatakse päringuvastuses vaikimisi kaks veebilauset, klikkides nupule „Näita rohkem“ kuvatakse maksimaalselt 26 lauset. Algselt kuvati veebilauseid Sõnaveebis GDEXi skoori alusel, kuid mõnel juhul olid esimesed kaks vaikimisi kuvatavat veebilauset vigased (nt sisaldasid lemmatiseerimise ja sõnaliigi märgenduse vigu, vt lähemalt ptk 5.2). Selle asemel, et kasutajale sama märksõna juures pidevalt samu vigaseid lauseid kuvada, otsustasid Sõnaveebi toimetajad lauseid kuvada skoori asemel juhuslikus järjekorras. Selline lähenemine tagab, et veebilausete esitus Sõnaveebis on sama dünaamiline nagu päring veebi otsinguportaalides, kus kasutaja saab iga kord pisut erineva tulemuse. See küll ei taga seda, et (esimesena kuvatavad) veebilaused ei sisalda vigu – vead sõltuvad endiselt väga palju märgendamise kvaliteedist. (Koppel, Kallas jt 2019)

Selliseid tänapäevaseid sõnastikke, mis pakuvad leksikograafi loodud või käsitsi valitud näitelausetele lisaks automaatselt tuvastatud korpuslauseid, on üsna vähe. Üks sellistest on näiteks inglise keele sõnaraamatu „Longman Dictionary of Contemporary English“ 5. trükk, kus kasutajatel on võimalus peale käsitsi valitud näitelausete lugeda kuni kümmet autentset lauset. Inglise veebisõnastikus Wordnik kuvatakse kasutajale juba suuremat hulka korpuslauseid ning ka Google'i automaattõlkel Google Translate on sarnane funktsioon olemas. (Cook jt 2014)

Eesti Keele Instituudi veebisõnastikes ei ole varem autentset korpusmaterjali kuvatud ning kasutajad on oma pikaajalisest sõnastike kasutamise kogemusest harjunud arvestama sellega, et kogu sõnastikus esitatava info on leksikograaf üle kontrollinud, toimetanud ning see on seega korrektne [P5]. Sõnaveeb on eesti leksikograafias esimene omataoline, kus kasutajale kuvatakse autentset ja toimetamata korpusmaterjali. Kuna Sõnaveebi kaudu saadetavas tagasisides on kasutajad osutanud sellele, et mõni veebilause nende meelest näitelauseks ei sobi, on veebilausete juurde lisatud hoiatus, et need on automaatselt valitud, toimetamata ning võivad sisaldada vigu (vt joonist 17). Sarnast hoiatust kasutavad ka Merriam-Websteri ja Collinsi veebisõnastikud. (Koppel, Kallas jt 2019)

5. PROBLEEMID JA EDASIARENDUSED

Korpuslausete automaatse valikuga kaasnevad mitmed kitsaskohad, mida on käsitletud ka artiklites [P5] ja Koppel, Kallas jt 2019. Järgnevalt kirjeldan lähemalt kõige sagedamini esile kerkinud problemaatilisi valdkondi, mis mõjutavad väljundi kvaliteeti: korpuse sisu ja maht, märgendamise kvaliteet ning grammatiline ja semantiline mitmesus. Pakun ka probleemidele võimalikke (reeglipõhiseid) lahendusi ning tutvustan uusi klassifikaatoreid, mida on plaanis järgmistes GDEXi eesti keele mooduli versioonides rakendada.

5.1. Korpuse sisu ja maht

John Sinclair (1991: 13) on öelnud, et korpuspäringu tulemused saavad olla ainult nii head, kui hea on korpus. GDEX 1.4 abil loodud „Eesti keele õppekorpuse 2018 (etSkELL)“ kvaliteet sõltub suurel määral selle aluseks olnud „Eesti keele ühendkorpuse 2017“ sisust. „Eesti keele ühendkorpuse 2017“ mahust umbes 80% moodustavad veebilehtedelt kogutud tekstid, mis tähendab, et olemuselt on ühendkorpus veebikorpus oma tüüpiliste probleemidega (tasakaalustamatus, allikate valik, masintõlkelised laused, vt lähemalt Gatto 2014).

Väitekirja kirjutamise ajal kogub tarkvarafirma Lexical Computing Ltd. Eesti Keele Instituudi tellimusel uut veebikorpust „Estonian Web 2019“. Korpuse sisu kvaliteedi parandamiseks alustatakse veebi kroolimist usaldusväärsetelt veebilehtedelt ning välditakse lehekülgi, millel on liiga pikad nimed (nt *johnsbestrealityestateandpaydayloansforstudents.com*).

Veebikorpuste sage probleemiallikas on masintõlkelised ja automaatselt genereeritud (*computer generated*) tekstid (Aharoni jt 2014, Nguyen-Son jt 2019). Nii on „Eesti keele ühendkorpuse 2017“ põhjal loodud „Eesti keele õppekorpusesse 2018 (etSkELL)“ sattunud kakskeelsetelt lehtedelt masintõlkelised laused, mida reeglipõhise lähenemisega ei ole võimalik tuvastada. Need laused vastavad küll GDEX 1.4 parameetritele (lauses esinevad sõnad on sagedad ja maksimaalselt 20 tähemärki pikad, lause on maksimaalselt 20 sõnet pikk, sisaldab tegusõna jmt), kuid ei ole grammatiliselt ega semantiliselt korrektsed (näide 16). [P5]

(16) Õpetused **roomakatoliku kirik** ei ole kaugeltki selge tõdesid Jumala sõna. (etSkELL)

Vältimaks automaatselt genereeritud tekstide sattumist väljundisse, eemaldatakse uuest veebikorpusest need dokumendid, mis üks kuu pärast veebi kroolimist enam ei eksisteeri, kuna artikli Koppel, Kallas jt 2019 kaasautori Vít Baisa kui veebikorpuste looja kogemus on näidanud, et automaatselt genereeritud tekstide eluiga on suhteliselt lühike. Lisaks kasutatakse musta nimekirja sellistest veebilehtedest, mille kohta on juba teada, et need sisaldavad ainult masintõlkelisi või automaatselt genereeritud tekste (nt *et.amazinghope.net* ja *ee.motion-free.website*) – st need jäetakse automaatselt kõrvale. Masintõlkeliste ja automaatselt genereeritud

lausete tuvastamist aitaks süntaksianalüsaatori väljundi arvestamine. Süntaksianalüsaatori abil saaks tuvastada vigase süntaktilise struktuuriga lauseid ja need kõrvaldada. Eesti keele jaoks on süntaksianalüsaator olemas, kuid veebikorpuste loomisel ei ole seda seni rakendatud. (Koppel, Kallas jt 2019)

Kui uus veebikorpus on valmis, siis saab selle sisu analüüsida võtmesõnade (*keywords*, vt Kilgarriff 2012) abil. Võtmesõnad annavad kompaktse pildi korpuses sisalduvate tekstide sisust. Näiteks saab võtmesõnana esinevate vulgarismidega tuvastada, missugustelt veebilehtedelt need tekstid on tulnud ning need dokumendid korpusest kõrvaldada.

Probleeme võib valmistada ka loodud korpuse maht. „Eesti keele õppekorpusest 2018 (etSkELL)“ ei leia kõikidele sõnadele näitelauseid, kuna õppekorpuse loomisel rakendati musta nimekirja (ptk 4.2.1). See tähendab, et kui emakeelne kasutaja otsib näiteks õppekorpusest näitelauseid sõnale *ajuinvaliid* või *rahvarämps*, siis ei saa sealt ühtegi vastet, kuna õppekorpuse loomisel kõrvaldati kõik laused, mis sisaldasid musta nimekirja kuuluvaid sõnu. Lahendus on luua eri konfiguratsioonid ja korpused eri sihtgruppidele – nii saaks emakeelsele kasutajale kuvada lauseid suuremast korpusest, mille loomisel pole leksikaalset filtrit rakendatud.

Lisaks on raske leida näitelauseid madala sagedusega sõnadele, näiteks esineb nimisõna *kalla* õppekorpuses väga harva. Madala sagedusega sõnadele näitelausete leidmiseks on üks võimalikke lahendusi kombineerida päringuid eri korpustest – kui sõna ei esine õppekorpuses, siis kuvatakse tulemusi ühendkorpusest.

Ideaalis oleks keeleõppijale vaja korpust, mis annab piisava ülevaate eri registrite keelekasutusest (Wilson 2013: 34, [P1]). Uue õppekorpuse loomisel tuleks lisaks arvesse võtta alliktekstide päritolu ja võimalusel ka allika aastat. See võimaldaks eri allikatest luua eraldi allkorpused ning teha lausepäring ainult konkreetsetest allikatest, näiteks eesti Vikipeediast ja perioodikaväljaannetest, ning jätta kõrvale blogi- ja foorumipostitused. Täiendav võimalus õppekorpuse kvaliteeti parandada on kokku koguda ja korpusesse lisada ka olemasolevate sõnastike näitelaused.

5.2. Märgendamise kvaliteet

„Eesti keele õppekorpus 2018 (etSkELL)“ on märgendatud EstNLTK versiooniga 1.4.1. Kuna märgendamine on automaatne protsess, esineb korpuses lemmatiseerimise, morfoloogilise analüüsi, lausepiiride ja mitmesõnaliste üksuste tuvastamise vigu, mille tõttu satuvad väljundisse mittesobivad või vigased laused.

5.2.1. Lemmatiseerimise ja morfoloogilise märgenduse vead

Üks sagedasemaid probleeme on vigane lemmatiseerimine. Näiteks annab päring *koha* õppekorpusest vastuseks lauseid, kus tegemist pole mitte kalaga, vaid nimisõna *koht* omastava käände vormiga (näide 17).[28]

(17) Meie klassi poisid võitsid II **KOHA**. (etSkELL)

Suur osa lemmatiseerimise vigadest on tulnud sellest, et õppekorpuse aluseks olnud „Eesti keele ühendkorpuse 2017" märgendamisel rakendati lokaalsele ehk lausesisesele kontekstile toetuvat ühestamist (*local context disambiguation*), ning juhul, kui mitmesused jäid lahendamata, jäi korpusesse ühestaja analüüsitud lemmade loendist ainult esimene. Ühestamise kvaliteeti oleks saanud oluliselt parandada, kui korpus oleks märgendatud dokumenditasemel ehk laiemat konteksti arvestades (*context-based disambiguation*). Nii oleks mitmetähenduslike sõnade puhul saanud kasutada oletust „üks lemma diskursuse kohta" (*one lemma per discourse*) ning sõna analüüsida dokumendis või teistes tekstiosades kõige sagedamini korduva lemmakandidaadi põhjal.

Üks reeglipõhine lahendus lemmatiseerimise ja vale morfoloogilise märgenduse vigasid vähendada oleks kontrollida lausete päringu käigus, kas korpuslauses esineva märksõna vorm on olemas märksõna grammatilises paradigmas. Täiendavalt saaks arvestada morfoloogiliste vormide sagedustega, nii et lausepäring väljastaks ainult sellised korpuslaused, kus märksõna on piisavalt sagedases vormis. Näiteks kui määr- ja tagasõna *kukil* oleks korpuses vigaselt märgendatud nimisõna *kukk* mitmuse alalütleva käände vormiks, siis ei satuks nimisõna *kukk* päringu korpuslausete väljundisse lauseid määr- ja tagasõnaks märgendatud sõnaga *kukil*, kuna *kukk* mitmuse alalütlevas käändes ei ole korpuses piisavalt sage vorm.

5.2.2. Lausestamine

Heiki-Jaan Kaalep ja Kadri Muischnek (2012) on välja töötanud parameetrid, millele toetudes saab ka süntaktilise analüüsita lausepiire tuvastada, toetudes kirjavahemärkidele, osalausepiiril olevatele üksiksõnadele ja tegusõna pöördelistele vormidele. „Eesti keele õppekorpuse 2018 (etSkELL)" lausestamisel on kasutatud EstNLTK 1.4.1 lausestajat (*sentence segmenter*) ja osalausestaja (*clause segmenter*) moodulit. Kuid teatud juhtudel, näiteks kui lauses esineb lühend, millele järgneb punkt, millele omakorda järgneb suure algustähega sõna (nt perekonnanimi) või number, esineb vigu lause piiride tuvastamisel (näited 18–19).

(18) Ühte eriala tutvustavasse loengusse tuli õppejõud **dr**. (etSkELL)

(19) Palun toetada meie ettepanekut **nr**. (etSkELL)

[28] Näites 17 on ka paradigmaväline ehk leksikaal-grammatiline vormisisene homonüümia (Puolakainen 2001: 13).

Näidetes 18 ja 19 ei ole lausepiiri üle otsustamine triviaalne ülesanne. Vigast märgendamist aitaks vähendada, kui võtta arvesse pärisnimede ja/või nimeüksuste märgendust, aga EstNLTK arendajatel praegu sellist lahendust pole.[29]

5.2.3. Mitmesõnalised üksused

Mitmesõnalise üksusena (*multi-word units*, *multi-word expressions*) käsitletakse üldjuhul püsiühendeid, perifrastilisi tegusõnu (väljend- ja ühendtegusõnad), idiomaatilisi ja kollokatiivseid ühendeid (Muischnek 2006: 12, Sag jt 2001). Mitmesõnalise üksuse moodustavad kaks või enam sõna(vormi), mida mingi tähenduse väljendamiseks koos kasutatakse (Muischnek 2006, Kaalep, Muischnek 2009: 157).

„Eesti keele sõnaraamatus 2019" esitatakse märksõnadena mitmesõnalisi üksusi, mida varasemas „Eesti keele seletavas sõnaraamatus" (2009) iseseisvana ei käsitletud, kuid mis keeles sellisena esinevad, nt *tähelepanu juhtima*, *löömaks minema*, *alkohoolne jook* ja *juua täis* (Langemets, Tiits jt 2018: 945). Need ei ole korpuses eraldi mitmesõnaliseks üksuseks märgendatud, ning seega võivad üksuse (nt *kokku leppima*) ühe komponendi (nt *leppima*) päringu väljundisse sattuda ka laused, kus märksõna on tegelikult mitmesõnalise üksuse osa (näide 20).

(20) Põhimõttelised asjad tuleb kokku **leppida**. (etSkELL)

Selleks, et selliseid lauseid väljundisse ei satuks, on oluline, et mitmesõnalised üksused oleksid korpuses märgendatud. See omakorda eeldab püsiühendite loendi olemasolu, mis võimaldaks automaatselt tuvastada sageli koos esinevaid sõnu. Korpuse märgendamisel tuleb igas lauses kontrollida, kas selles esineb püsiühendite loetelus olevaid sõnu, kusjuures märgendaja peab arvestama seda, et püsiühend ei esine alati selles vormis, nagu see on antud püsiühendite loendis; ning seda, et kõikide komponentide esinemine lauses ei tähenda automaatselt, et need moodustavad püsiühendi. (Kaalep, Muischnek 2009: 169–170) Samuti on osalausete piiride korrektne märgendamine ühendtegusõnade tuvastamisel oluline, kuna osalause pakub optimaalset konteksti kandidaatpaaride moodustamiseks (Kaalep, Muischnek 2012: 55, Uiboaed 2010: 324).

5.2.4. Leksikon

Märgendamise kvaliteedi parandamiseks on oluline pidevalt ajakohastada morfoanalüsaatori aluseks olevat leksikoni.[30] Eriti problemaatilised on grammatisatsiooni ja leksikalisatsiooni juhud, kus morfoanalüsaator märgendab lemmat ja sõnaliiki uuendamata leksikoni põhjal. Pealegi arvestab morfoanalüsaator „Morfoloogiliselt ühestatud korpuse" (*ca* 500 000 sõna) väljundit, mis ei pruugi katta

[29] Info pärineb isiklikust kirjavahetusest EstNLTK arendaja Siim Orasmaaga (07.08.2019).
[30] EstNLTK kasutab Vabamorfi leksikoni.

kõiki võimalikke lemma variante. Näiteks analüüsib morfoanalüsaator sõna *tasuta* läbivalt nimisõna *tasu* ilmaütleva käände vormiks[31], kuigi leksikonis on *tasuta* fikseeritud ka omadussõnalise lemmana. Üheks põhjuseks on see, et „Morfoloogiliselt ühestatud korpuses“ ei ole seda kordagi omadussõnalise lemmana märgendatud ning analüsaator järeldab, et sellist lemmat ei ole vaja variandina pakkuda. Sõnaveebi veebilausete päring käib API kaudu korpuspäringusüsteemist KORP lempose alusel, mis tähendab, et lisaks lemmale arvestatakse ka märksõna sõnaliiki. Kuna õppekorpuses sõna *tasuta* omadussõnana ei esine, siis leiab API üles ainult vigased veebilaused, kus sõna *tasuta* on tegelikult hoopiski vigaselt trükitud nimisõna *taust* käändeline vorm (joonis 17). (Koppel, Kallas jt 2019)

Joonis 17. Märksõna *tasuta* vigased veebilaused Sõnaveebis

5.2.5. Trükivead

Eraldi probleem on vigaselt trükitud sõnad, näiteks leiab nimisõnale *ahe* õppekorpusest lauseid, kus tegemist on valesti trükitud sõnadega *kaks* (näide 21) ja *vahe* (22).

(21) Ilm oli veidi jahe sprintide jaoks, sest külma õhku **ahe** suupoolega sisse ahmida pole just hea mõte. (etSkELL)

(22) Janne küsis: Mul on tütar 1,9 tal on ülahuule kida, hammastel on suur **v ahe**. (etSkELL)

[31] Nii käsitleti sõna *tasuta* ka varasemates sõnaraamatutes, näiteks „Eesti keele seletavas sõnaraamatus“ (2009), kuid 2019. aastal Sõnaveebis ilmunud sõnastikud („Eesti keele sõnaraamat 2019“ ja „Eesti keele naabersõnad 2019“) käsitlevad sõna *tasuta* aga juba iseseisva omadussõna ja määrsõnana.

5.3. Grammatiline mitmesus

Grammatilise mitmesuse all käsitletakse homonüümiat üldisemas mõttes ning sõnavormide põhjal eristatakse viit tüüpi mitmesust (Karlsson jt 1995, Vihma 1964, Viks 1977, 1984, Puolakainen 2001: 13):

1. lekseemide homonüümia,
2. paradigmasisene ehk grammatiline homonüümia,
3. paradigmaväline ehk leksikaal-grammatiline vormisisene homonüümia,
4. paradigmaväline vormidevaheline homonüümia,
5. kategoriaalne mitmesus.

Kuna eesti keeles on homonüümiat 45% ulatuses (Puolakainen 2001: 14), siis see avaldab mõju ka lausete päringule korpusest. Probleeme tekitab nii lekseemide homonüümia, paradigmaväline vormisisene homonüümia kui ka paradigmaväline vormidevaheline homonüümia.

Eesti keeles on leksikaalseid homonüüme umbes tuhat.[32] Näiteks sõnapäringu *tamm* vastuseks saab korpusest läbisegi lauseid, kus tamm on kas puu või vesiehitise tähenduses (näited 23–24) [P5].

(23) Meie kooli õuel kasvasid **tammed**. (etSkELL)

(24) Kopra loodud märgalade suurus sõltub **tammidest** ning nende asukohast. (etSkELL)

Homonüümide probleemi on väga keeruline automaatselt lahendada. Üks võimalikke reeglipõhiseid lahendusi oleks lausete päringus arvestada märksõna morfoloogilises paradigmas olevaid vorme. See ei aita lahendada erinevalt käänduvate homonüümide probleemi, kui märksõna on lauses nimetavas käändes (nt *tamm*), küll aga väheneks vigaste vastete arv alates omastavast käändest (*tamme* ja *tammi*). Teine võimalik reeglipõhine lahendus oleks kasutada lausete päringus märksõna kollokaate, nii et väljundis oleksid eelkõige sagedasemaid kollokatsioone sisaldavad näitelaused. Näiteks kui märksõnal *tamm* on kaks homonüümi ja kasutaja valib neist tähenduse ‘pais’, kuvatakse esmalt laused kollokatsioonidega *tamm puruneb* ja *tammi ehitama*. [P5]

Paradigmaväline vormisisene homonüümia esineb juhul, kui kattuvad eri sõnade samad vormid. Näiteks päringule *teod* saab korpusest vastuseks lauseid, kus see tähistab läbisegi nimisõnade *tigu* ja *tegu* mitmuse nimetava käände vorme (näited 25–26). Ka paradigmavälist vormisisest homonüümiat aitaks lahendada konteksti ehk kollokatsioonide arvestamine. Samuti oleks abiks korpuse semantiline märgendus (vt lähemalt ptk 5.4).

(25) **Teod** on peamiselt öised rändurid ja nende pahateod avastame hommikul. (etSkELL)

(26) Suured **teod** algavad suurelt unistamisega! (etSkELL)

[32] Allikas: https://sonaveeb.ee/about#sam (20.06.2019).

Paradigmavälise vormidevahelise homonüümia korral kattuvad eri sõnade eri vormid. Kuna korpuse märgendamisel ei kasutatud dokumenditasandil ühestamist, annab nimisõna päring *kalla* vastuseks lauseid, kus tegemist on hoopis tegusõna *kallama* käskiva kõneviisi 3. pöörde vormiga (näide 27). [P5]

(27) **Kalla** segu vormi ja kata vorm kilega. (etSkELL)

Vormihomonüümia probleemi lahendamise eeldus on arvestada ka märksõna morfoloogiliste tunnustega, eelkõige sõnaliigiga. See tähendab, et päringute koostamisel ei lähtuta mitte lemmast, vaid lemposest ehk lemmast koos sõnaliigiga. Seejuures võib toetuda leksikograafilistes andmebaasides juba sisalduvale märksõna morfoloogilisele ja grammatilisele infole.[33] Seda saab teostada näiteks API abil. Nii tehakse juba praegu keeleportaalis Sõnaveeb, kus API esitab veebilausete päringu KORPist lempose abil. See lahendab probleemi, kui homonüümid on eri sõnaliigist (nt *kerge-a* ja *kerge-s*), kuid annab vigaseid vastuseid, kui korpuses esineb lemmatiseerimise vigu ning sõnaliigi vale märgendust.

5.4. Semantiline mitmesus

Semantilise mitmesuse all käsitletakse mitmetähenduslikkust ehk polüseemiat (Karlsson jt 1995), näiteks saab päringule *leht* vastuseks lauseid, kus sõna esineb eri tähendustes (näited 28–30) [P5].

(28) Minu meelest on huvitav **lehti** lugeda. (etSkELL)

(29) **Lehti** ja õisi kogutakse õitsemise ajal. (etSkELL)

(30) Täpsemat infot leiad viisa nõuete **lehelt**. (etSkELL)

Polüseemsete sõnade eri tähendustele sobivate näitelausete leidmise üheks eelduseks on korpuse semantiline märgendamine (*sense annotation*). Seejuures on oluline, et leksikograafilise andmebaasi ja korpuse semantiliste märgendite nomenklatuur oleks sama. „Eesti keele põhisõnavara sõnastiku“ (2014) ja „Eesti keele sõnaraamatu“ (2018) koostamisel on kasutatud Margit Langemetsa (2010) välja töötatud semantilisi tüüpe, kuid neid ei ole kasutatud eesti korpuste märgendamisel.

Üks võimalikke reeglipõhiseid lahendusi polüseemse sõna eri tähendustele näitelausete leidmiseks oleks toetuda märksõna kollokaatidele, kui need on sõnastikus esitatud tähenduste kaupa, nagu on tehtud näiteks Sõnaveebis.

Täiendav võimalus oleks kasutada semantilist klasterdamist (*sense clustering*) (Martelli jt 2019, Boullosa jt 2017), mis vektorsemantika (*vector semantics*)

[33] Eesti Keele Instituudi sõnaraamatutes kasutatakse morfoloogilise info kuvamisel morfoloogilist andmebaasi (MAB).

algoritme kasutades jagab laused väiksematesse konteksti poolest sarnastesse alamhulkadesse, mis moodustavad semantilisi seotud plokke ehk klastreid.

5.5. Edasiarendused

5.5.1. Täiendavad klassifikaatorid

Nagu peatükis 4.2.2. mainitud, olen juba täiendanud lause alguses keelatud sõnade ja sõnapaaride nimekirja. Järgnevates GDEXi eesti keele mooduli versioonides täpsustan kindlasti ka tegusõna olemasolu lauses nõudvat klassifikaatorit, et see tuvastaks ainult sellised kandidaadid, mis sisaldavad tegusõna pöördelist vormi. Samuti on plaanis portugali mooduli eeskujul testida sagedusläve alammäära seadmist lauses esinevatele lemmadele ja sõnedele. Portugali keele GDEXi moodulis määrati lemma sagedusläve alammääraks 500, mis tähendab, et iga lauses esinev lemma peab korpuses esinema vähemalt 500 korda. Sõnede sagedusläve alammääraks seati 50, mis tähendab, et iga lauses esinev sõne peab korpuses esinema vähemalt 50 korda. (Kuhn 2017: 265) See, millised sagedusläved eesti lemmadele ja sõnedele seada, vajab testimist: suure tõenäosusega tuleb keeleõppijatele suunatud konfiguratsioonides seada kõrgem sageduslävi kui emakeelsetele kõnelejatele suunatud konfiguratsioonides, mis tähendab seda, et keeleõppijatele suunatud näitelaused sisaldaksid kõrgema sagedusega sõnu – mida madalam keeleoskustase, seda kõrgema sagedusega sõnad peaksid lauses esinema.

5.5.2. Eri sihtgruppidele kohandatud konfiguratsioonid ja uued õppekorpused

GDEXi mooduli arendamine on pidev protsess, eriti kuna on vajadus eri sihtrühmadele kohandatud konfiguratsioonide järele. Väitekirja kirjutamise ajal on päevakorral näiteks emakeelsele kõnelejale suunatud versioon ja kooliõpilastele[34] suunatud versioonid. Kuhn (2017: 256) õppis portugali moodulit arendades, et klassifikaatorite mõju kergemini märkamiseks on mõistlikum kasutada radikaalsemaid parameetreid. Ka siinkirjutaja kogemus on näidanud, et konfiguratsioonide testimine GDEX Editoris on seda aeganõudvam, mida rohkem on klassifikaatoreid, mis tähendab, et iga individuaalse parameetri mõju väljaselgitamine on väga ajamahukas töö. Plaanin edaspidi GDEXi konfiguratsioonides määrata olulisematele parameetritele vastamise eest suuremad lisapunktid ning mittevastamise eest suuremad karistused. Uute korpuste loomisel saaks kasutada korpusest automaatselt ekstraheeritud sõnavisandite grammatikaga ette määratud struktuuriga kollokatsioonide loendeid. See tähendab, et loend sisaldaks ainult sisusõnade

[34] „Eesti keele A1–C1 õpikute korpust“ täiendatakse 2019. aastal 1.–9. klassi eesti keele kui teise keele õpikute materjaliga, mis võimaldab omakorda analüüsida I (1.–3. klass), II (4.–6. klass) ja III (7.–9. klass) kooliastme sõnavara ja rakendada neid eri kooliastmetele suunatud GDEXi mooduli versiooni arendamisel.

sagedasemaid ja statistiliselt esilduvamaid kollokaate. Statistiliselt esilduvamate kollokaatide tuvastamisel töötab kõige paremini statistik logDice (Rychlý 2008, Kallas 2013). Kõige sagedasemaid kollokatsioone (nt esimest kümmet) saaks kasutada näitelausete ekstraheerimiseks korpusesse. Samuti tuleks mõelda võimalusele, et kasutaja saaks näitelausete kuvamisel filtreid (nt musta ja halli nimekirja) ise aktiveerida.

5.5.3. Leksikaalne filter

Sarnaselt jaapani mooduliga (Srdanović, Kosem 2016), kus keeleõppe eesmärgil loodud versioonid erinevad teineteisest peamiselt selle poolest, et need karistavad teatud sõnu ja lemmasid, mis konkreetsele keeleoskustasemele ei kuulu, ning annavad lisapunkte teatud hulga sõnade eest, mis sellele konkreetsele tasemele kuuluvad, plaanin leksikaalse filtri lisada ka olemasolevatesse üldistele keeleoskustasemetele suunatud GDEXi eesti keele mooduli versioonidesse. Leksikaalse filtrina saab kasutada 2018. aastal Eesti Keele Instituudis valminud sõnaloendeid A1-, A2- ja B1-keeleoskustasemetele[35] (Kallas, Koppel 2018a, 2018b, 2018c), B2- ja C1-keeleoskustaseme loendite avaldamine on planeeritud lähitulevikku. Üldistele keeleoskustasemetele loodud versioonide abil saab luua tasemekohased õppekorpused, mida saab tulevikus omakorda integreerida keeleportaali Sõnaveeb järgmistesse versioonidesse, kuhu on plaanis luua eri vaated eri sihtgruppidele. 2019. aasta seisuga on Sõnaveebis vaateid kaks: detailne ja lihtne. Detailne vaade on suunatud eelkõige emakeelsele kasutajale ning lihtne vaade eesti keele õppijale A2–B1-oskustasemel.

5.5.4. API sätted

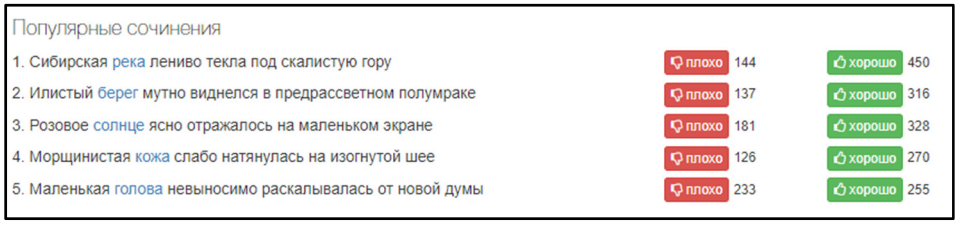
Õppekorpuse lausepäringutes võiks API toetuda kollokatsioonidele kui sõna tüüpilistele kasutusmustritele. Nagu peatükis 3.3. mainitud, loodi GDEX algselt sõnastike koostamiseks TBL-meetodiga, mis tähendab, et see tuvastab näitelauseid kõige edukamalt kollokatsioonidele toetudes. GDEX 1.4, mille abil „Eesti keele õppekorpus 2018 (etSkELL)“ loodi, sisaldab küll teise kollokaadi klassifikaatorit, mis annab iga lauses esineva kollokatsiooni kollokaadi eest lisapunkti, kuid see klassifikaator töötab vaid juhul, kui GDEXile on teada märksõna ja selle kollokaadid. GDEXi abil suurt korpust filtreerides rakenduvad klassifikaatorid, mille jaoks ei ole märksõna vaja teada, näiteks kontrollitakse lause ja sõnade pikkust, lauses esinevate sõnade sagedust korpuses, lause sõnaliigilist koosseisu, komade arvu lauses jmt. Kuna aga GDEXi abil suure korpuse lauseid filtreerides ei ole teada, missugune on märksõna, siis ei sisalda ka kõik GDEX 1.4 abil loodud õppekorpuse laused sagedasi kollokatsioone, kuna ilma märksõnata ei ole võimalik kollokaate tuvastada. Sõnaveebis saaks API aga veebilausete päringus arvestada

[35] Sõnavaraloendid on kättesaadavad aadressil http://www.eki.ee/keeletase/#/wordlistspdf (20.06.2019) ning alates märtsist 2020 ka keeleportaali Sõnaveeb sakil ”Õpetaja tööriistad”.

„Eesti keele naabersõnade 2019“ sõnastiku andmebaasis olevate otsisõna kollokaatidega, et väljundis oleksid eelkõige sagedasemaid kollokatsioone sisaldavad näitelaused. Teine võimalus olemasoleva õppekorpuse väljundit teise kollokaadi klassifikaatori abil parandada oleks lasta API-l korpuspäringus lauseid reaalajas GDEX 1.4 abil uuesti hinnata, kuid see teeks veebilausete päringu protsessi väga aeglaseks. Kolmas võimalus oleks lasta olemasoleva õppekorpuse lausetele GDEX 1.4 abil uued skoorid määrata, kuid seda tuleks teha iga Sõnaveebis oleva märksõna jaoks eraldi – nii saab GDEX märksõnade abil otsida nende sagedasemaid kollokaate (ning omakorda kollokatsioonide kollokaate) –, ning kuvada veebilauseid uuesti määratud GDEXi skoori alusel.

5.5.5. Õppekorpuse kvaliteedi parandamine kasutajate abil

Õppekorpuse puhastamisel on võimalik rakendada ka rahvahanget (*crowdsourcing*). Üks võimalikke viise on paluda kasutajatel hinnata korpusest automaatselt tuvastatud näitelausete sobivust. Keeleportaalis Sõnaveeb võiks see olla lahendatud sarnaselt vene keele assotsiatsiooni- ja sünonüümisõnastikuga Reright (joonis 18).



Joonis 18. Näitelausete sobivuse hindamine vene keele assotsiatsiooni- ja sünonüümisõnastikus Reright

Rahvahanke tulemusi arvestades saab luua kaks andmebaasi, millest ühte kuuluksid sobivaks hinnatud laused ehk head näitelaused ning teise sobimatuks hinnatud laused ehk halvad näitelaused. Neid andmebaase saab omakorda kasutada masinõppe algoritmi treenimiseks. Sarnast lähenemist on kasutanud näiteks Darja Fišer ja Jaka Čibej (2017) ning Tanara Z. Kuhn jt (2019).

Kokkuvõtteks võib öelda, et GDEX küll suudab tuvastada ja ekstraheerida häid näitelause kandidaate, kuid näitelausete automaatne valik ei sõltu üksnes GDEXiga määratud parameetritest. Tulemuse parandamise seisukohalt on oluline ka semantilise, süntaktilise ja morfoloogilise märgendamise täpsus. „Eesti keele ühendkorpuse 2019“ märgendamisel on plaanis kasutada uusimat EstNLTK versiooni 1.6, kuhu on võrreldes versiooniga 1.4.1 lisatud uusi reegleid ning värskendatud ka leksikoni, mis võib eelnevalt kirjeldatud probleeme osaliselt lahendada.

6. KOKKUVÕTE

Nii Eestis kui ka mujal Euroopas on viimase aastakümne jooksul hakatud sõnastikke koostama (pool)automaatselt. Sõnastike automaatne koostamine tähendab, et sõnaartikli üksused, sealhulgas näitelaused, ekstraheeritakse automaatselt korpuspäringusüsteemist sõnastikusüsteemi, kus leksikograaf neid vajadusel toimetab. Leksikograafidele on korpuslaused põhiline näitelausete allikas ning nende automaatseks tuvastamiseks kasutatakse masinõppemeetodit, reeglipõhist lähenemist või nende kahe kombinatsiooni ehk kombineeritud meetodit. Automaatselt tuvastatud näitelauseid on võimalik ekspertidele (nt lingvistidele, leksikograafidele) kuvada korpuspäringu- või sõnastikusüsteemi ning väliskasutajatele (nt keeleõppijatele) keele- või sõnastikuportaali osana (Kosem, Husák, McCarthy 2011).

Siinse väitekirja eesmärk oli välja töötada meetod, mida oleks võimalik rakendada näitelausete korpuspõhiseks automaattuvastamiseks. Eesti leksikograafias ei ole sellist uurimistööd varem tehtud. Väitekiri kuulub korpusleksikograafia ja automaatse leksikograafia valdkonda. Uurimismeetod on reeglipõhine lähenemine ning parameetrite häälestamiseks on osaliselt kasutatud ka masinõppe elemente.

Eesti Keele Instituudis on näitelausete automaatseks tuvastamiseks seni kasutatud korpuspäringusüsteemi Sketch Engine (Kilgarriff, Rychlý jt 2004) integreeritud tööriista Good Dictionary Examples ehk GDEX (Kilgarriff, Husák jt 2008, [P2]), mille keskmes on universaalne reeglipõhine valem, mis teatud eeldefineeritud lause tunnustele ehk parameetritele toetudes tuvastab kõige sobivamad näitelause kandidaadid. Kandidaadid reastatakse paremuse järjekorda, nii et kõige sobivamad kandidaadid on nimekirja eesotsas. Paremate tulemuste saavutamiseks tuleb GDEXi moodulit täiendada parameetritega, mis arvestavad keele spetsiifikat, näiteks sõnade ja lausete pikkust, märksõna asukohta lauses jmt. Siinse väitekirja eesmärk oli välja selgitada hea näitelause formaalsed parameetrid eesti keelele.

Väitekirja artiklites analüüsiti eesti keele (õppe)sõnastike („Eesti keele sõnaraamat 2019“, „Eesti keele põhisõnavara sõnaraamat“ (2014), „Eesti keele naabersõnad 2019“) näitelausete ja „Eesti keele A1–C1 õpikute korpuse“ lausete parameetreid ning loodi analüüsi tulemusi arvestades GDEXi eesti keele mooduli kuus erinevat versiooni: GDEX 1.2, GDEX 1.3, GDEX 1.4, etBasic-v1, etIndependent-v1 ja etProficient-v1. Esimesed kolm versiooni on teineteise edasiarendused ning loodud „Eesti keele naabersõnade 2019“ sõnastiku sihtgruppi ehk B2–C1-oskustasemel eesti keele valdajaid silmas pidades. etBasic-v1 on loodud A-tasemel, etIndependent-v1 B-tasemel ja etProficient-v1 C-tasemel eesti keele õppijat silmas pidades [P5]. GDEXi eesti keele mooduli eri versioonide testimiseks on kasutatud eesti keele korpusi „Eesti keele ühendkorpus 2013“ ja „Eesti keele ühendkorpus 2017“. Versiooni 1.2 kasutati „Eesti keele naabersõnade 2019“ sõnastiku andmebaasi täisautomaatsel genereerimisel (Kallas, Koppel, Tuulik 2015). Versiooni 1.3 abil loodi esimene autentseid lauseid sisaldav õppekorpus „EstonianNC GDEX“ [P1]. Kuna GDEX 1.4 [P3] abil loodi „Eesti keele õppe-

korpus 2018 (etSkELL)“, mis on keeleõppekeskkonna Sketch Engine for Estonian Language Learning ehk etSkELL ja keeleportaali Sõnaveeb veebilausete allikas, on see versioon väitekirja keskmes ning selle väljundit on ka evalveeritud [P4].

GDEX 1.4 abil korpusest leitavad laused on kõik täislaused ehk algavad suure tähega ja lõppevad lauselõpumärgiga. Lauses esinevate sõnade pikkus on maksimaalselt 20 tähemärki ning nende sagedus korpuses ei ole väiksem kui 5. Lause pikkus on minimaalselt 4 ja maksimaalselt 20 sõnet ning lause optimaalne pikkus on 6–12 sõnet. Lause sisaldab kindlasti tegusõna, aga ei sisalda teatud keelatud tähemärke. Lause ei alga teatud sõnade (nt *seejärel*, *järgnevalt*, *samuti*) ega sõnapaaridega (nt *samal põhjusel*, *sellest hoolimata*, *just sellepärast*), mis on sageli anafoorsed ehk viitavad tagasi eelnevale lausele või tekstilõigule; samuti mitte teatud sõnaliikidega (nt sidesõna, hüüdsõna, lühend). GDEX 1.4 parameetrite järgi saavad karistada korpuslaused, milles esinevad sõnad, mille sagedus korpuses on madalam kui 1000, ning laused, kus esineb rohkem kui kaks tegusõna, rohkem kui üks määrsõna, rohkem kui üks asesõna, rohkem kui üks sidesõna, rohkem kui üks pärisnimi, rohkem kui üks arvsõna ja rohkem kui üks koma. [P3]

Evalveerimiseks viidi läbi hindamisülesanne Eesti Keele Instituudis töötavate leksikograafide ja Tartu ja Tallinna Ülikoolis eesti keelt B2–C1-oskustasemel valdavate üliõpilaste seas. Hindamisülesande peamine eesmärk oli välja selgitada, kui suur hulk GDEX 1.4 parameetritele vastavatest korpuslausetest hinnatakse sobivaks ning kui suur hulk GDEX 1.4 parameetritele mitte vastavatest korpuslausetest hinnatakse sobimatuks näitelause kandidaadiks. Hindamisülesandele järgnes jätkuküsitlus, mille eesmärk oli välja selgitada, mis põhjusel üks või teine lause hindajate meelest sõnastiku näitelauseks ei sobinud. Tulemused näitasid, et 85% GDEX 1.4 abil valitud korpuslausetest hinnati sobivaks ning koguni 94% GDEX 1.4 parameetritele mitte vastavatest korpuslausetest hinnati sobimatuks näitelause kandidaadiks. See näitab, et GDEX suudab korpusest edukalt tuvastada sobivaid näitelauseid ning välja sortida sobimatud. Lausete sobimatuks hindamise põhjuseks toodi kõige sagedamini anafoorsust, konteksti puudumist, lause pikkust ja kõnekeelsust. [P4]

GDEX 1.4 abil loodud „Eesti keele õppekorpus 2018 (etSkELL)“ on allikaks keeleõppekeskkonnale etSkELL ja keeleportaali Sõnaveeb veebilausetele. etSkELL on olemuselt korpuspäringusüsteemi Sketch Engine lihtsustatud versioon, mille abil saab kasutada selle olulisemaid funktsioone: lugeda näiteid, vaadata naabersõnu (kollokatsioone) ja sarnaseid sõnu (tesaurust). etSkELLi eelis on veebiliidese tehniline lihtsus, läbipaistev metakeel ja väljastatud info piiratud maht. Keeleportaalis Sõnaveeb näeb kasutaja leksikograafide koostatud info kõrval täiendavalt ka korpuslauseid. Eriti kasulikud on need sellistes sõnaartiklites, kus leksikograafi koostatud näitelaused puuduvad. (Koppel, Kallas jt 2019)

GDEX 1.4 abil loodud õppekorpuse kvaliteet sõltub suurel määral aluseks olnud „Eesti keele ühendkorpuse 2017“ märgendamise kvaliteedist ja sisust. „Eesti keele ühendkorpus 2017“ koosneb umbes 80% ulatuses veebilehtedelt kroolitud tekstidest, mis tähendab, et ühendkorpus on suures osas just veebikorpus ning

sisaldab veebikorpusele tüüpilisi probleeme. Näiteks on veebist kroolitud korpustes sage probleem masintõlkelised ja automaatselt genereeritud tekstid. Samuti mõjutab väljundit loodud õppekorpuse maht. Kuna see on aluseks olevast korpusest tunduvalt väiksem, ei pruugi sealt leida näitelauseid kõikidele sõnadele, eriti kui õppekorpuse loomisel on rakendatud sõnavara piirangut ehk musta nimekirja. Lisaks on raske leida näitelauseid madala sagedusega sõnadele. Kuna õppekorpus on märgendatud automaatselt, esineb seal lemmatiseerimise, morfoloogilise analüüsi ning lausepiiride ja mitmesõnaliste üksuste tuvastamise vigu. Samuti mängib väljundi kvaliteedi juures rolli märgendamiseks kasutatav leksikon, mida on vaja pidevalt ajakohastada. Väljundit mõjutab ka grammatiline mitmesus, eriti lekseemide homonüümia, paradigmaväline vormisisene homonüümia ja paradigmaväline vormidevaheline homonüümia. Samuti mõjutab väljundit semantiline mitmesus ehk polüseemia. Neid kitsaskohti on võimalik osaliselt reeglipõhiselt lahendada, näiteks kontrollida korpuspäringus märksõna morfoloogilist paradigmat ja morfoloogilise vormi sagedust, märksõna sõnaliiki ja märksõna sagedasemaid kollokaate.

Edaspidi on plaanis „Eesti keele ühendkorpuse 2019“ põhjal luua uus õppekorpus. Samuti on plaanis arendada eri sihtgruppidele suunatud spetsiifilisemad GDEXi versioonid, kuhu on lisatud täiendavaid parameetreid, mis aitaksid mõningaid väitekirjas kirjeldatud kitsaskohti vältida. Uute konfiguratsioonide abil saab luua uued korpused, mida saab omakorda integreerida keeleportaali Sõnaveeb, kuhu on plaanis luua vaated eri sihtgruppidele.

Üks võimalik viis korpuse sisu puhastada on paluda sõnastiku kasutajatel näitelausete sobivust hinnata rahvahanke teel. Rahvahanke tulemusi rakendades saab välja töötada masinõppe algoritmi, mis ebasobiva (nt tundliku) sisuga laused ära tunneb ning korpusest kõrvaldab. Sarnast lähenemisviisi on plaanis edaspidi kasutada ka eesti keele korpuse kvaliteedi parandamiseks.

SUMMARY: CORPUS-BASED AUTOMATIC DETECTION OF EXAMPLE SENTENCES FOR DICTIONARIES FOR ESTONIAN LEARNERS

The aim of this dissertation is to develop a method for the automatic detection of authentic corpus sentences suitable for learners of Estonian L2. The dissertation topic belongs to the fields of corpus lexicography and automated lexicography.

Introduction

In Europe we are witnessing an increasing shift from traditional lexicography to e-lexicography. Almost half of the modern dictionaries in Europe are published only on the web (Kallas, Koeva et al. 2019). State-of-the-art lexicographic work requires the use of dictionary writing systems (DWS) and corpus query systems (CQS). Since 2011, the Institute of the Estonian Language has used Sketch Engine (Kilgarriff, Ruchlý et al. 2004) – a CQS widely used among lexicographers in Europe (Kallas, Koeva et al. 2019: 19) – and KORP. Until 2019, dictionaries were compiled in the web-based DWS EELex (Langemets, Loopmann, Viks 2006, Jürviste et al. 2011). Since 2019, dictionaries are compiled in a new DWS called Ekilex (Tavast et al. 2018).

Sketch Engine is used for automatic generation of dictionary databases, which is a growing practice in modern lexicography. The tool supports automatic generation of frequency lists (Kilgarriff 2010a), finding statistically significant collocates, generating Word Sketches (one-page summary of a word's grammatical and collocational behaviour) (Kilgarriff, Baisa et al. 2010), thesaurus (Rychlý & Kilgarriff 2007), extraction of definitions (Kovář et al. 2016), terms (Jakubíček et al. 2014), translational equivalents, and finding suitable example sentences (Kilgarriff, Husák et al. 2008). For automatic detection of good example sentences, the Good Dictionary Examples (GDEX) function (Kilgarriff, Husák et al. 2008) was implemented into Sketch Engine.

For Estonian, a module for Word Sketches and term extraction was developed by Kallas (2013) and Kallas, Suchomel et al. (2017). The different GDEX versions for Estonian were developed within the framework of this dissertation.

Outline

The dissertation is divided into six chapters. Chapter 1 gives an overview of the research objective and method and points out the goals of the dissertation. Chapter 2 describes the types of example sentences, points out the purpose of example sentences in a dictionary entry, discusses the features of a good dictionary example, and describes how authentic corpus sentences are used in learner lexico-

graphy and in language learning in general. Chapter 3 introduces different methods that are used to find good examples in large corpora, focuses on GDEX, and gives a short overview of GDEX configurations for seven different languages. Chapter 4 describes the six different configurations of GDEX for Estonian that were developed within the framework of this dissertation. The main focus is on version 1.4. The results of the evaluation of GDEX 1.4 are also presented. GDEX 1.4 was used to compile the Estonian Corpus for Learners 2018 (etSkELL), which is used in a corpus-based web tool etSkELL (Sketch Engine for Estonian Language Learning) and as a source of corpus sentences in the language portal Sõnaveeb. Chapter 5 discusses problems that have arisen from displaying authentic corpus sentences to end-users, proposes possible solutions, and envisages further developments. Chapter 6 summarizes the main conclusions. The appendices contain a summary table of parameters of all Estonian GDEX configurations, the developed Estonian GDEX configuration files, blacklist and greylist.

Goals of the Dissertation

The goals of this dissertation are as follows:

- to give an overview of what a good example is in terms of both traditional theoretical lexicography as well as in corpus lexicography;
- to analyse example sentences in Estonian dictionaries (Basic Estonian Dictionary (2014), Dictionary of Estonian (2019), Estonian Collocations Dictionary (2019)) and sentences in the Estonian Coursebook Corpus 2018 in order to ascertain the formal parameters of good dictionary examples for Estonian;
- to give an overview of methods being used for automatic detection of example sentences;
- to develop different Estonian GDEX configurations for learners at different CEFR levels;
- to evaluate the output of the GDEX 1.4 configuration;
- to use GDEX 1.4 to create a corpus that is targeted at learners of Estonian at a B2–C1 proficiency level and to implement GDEX 1.4 in a corpus-based web tool, etSkELL, and the language portal Sõnaveeb.

The Features and Purpose of Dictionary Examples

A good dictionary example is often described as natural or authentic, typical, informative and intelligible (Fox 1987, Harras 1989, Atkins & Rundell 2008, Kilgarriff, Husák et al. 2008). A typical example includes frequent and common syntactic and collocational patterns of the headword. A natural example features grammatical patterns (colligation) and includes only one register: e.g., if the sentence is informal, it should not include formal words. An informative sentence is self-contained, i.e., understandable without wider context (Harras 1989). An

intelligible example is not too long and does not contain complex syntax, rare or specialized vocabulary, or confusing proper names. (Atkins & Rundell 2008: 459–461)

Dictionaries are traditionally compiled with a specific target group in mind, e.g., native speakers or language learners; the whole contents of the dictionary entry should meet the needs of the target group. The features of the example sentences vary according to the type of dictionary and its target group. In dictionaries aimed at native speakers, example sentences illustrate the usage of the headword; they often complement the definition and help to distinguish between different senses of polysemous words. Example sentences in learner dictionaries help to place an unknown word in a learner's passive or active vocabulary. Example sentences in learner dictionaries should illustrate the usage and context of the headword, as well as its syntactic and collocational patterns. (Atkins & Rundell 2008) Example sentences aimed at learners should include information about grammar, collocations and phrases (Zöfgen 1986, Harras 1989, Laufer 1992, Byrne 2006).

In monolingual and bilingual learner dictionaries, example sentences are more important than in dictionaries aimed at native speakers. Sometimes, a definition without an example sentence can be incomprehensible. (Atkins & Rundell 2008) According to several studies (Frankenberg-Garcia 2012, 2014, Lew & Adamska-Sałaciak 2015, Burke 2003, Simpson 2003, Svensén 2009, Teral 2015), it is useful to add more than one example in the dictionary entry. Example sentences help to illustrate the definition, especially in the case of language learners, when definitions written in a foreign language can be difficult to understand (Lew & Adamska-Sałaciak 2015).

As previously described, each type of dictionary has its own rules when it comes to choosing example sentences. In passive dictionaries, example sentences are usually used to illustrate the headword and its subsenses. In active dictionaries, example sentences should provide grammatical support. When the target group is native speakers, example sentences can be grammatically and syntactically complex, and include rarer vocabulary. When the target group is language learners, the example sentences should be shorter, include frequently used vocabulary, and should not be grammatically and syntactically complex. Corpus sentences tend to either include too much context or, contrarily, include deictic or anaphoric references, and in many cases they need to be edited before adding them to a dictionary. Most typical editing strategies are to simplify the grammatical structure, replace proper names and abbreviations, delete irrelevant or disturbing clauses, etc. Even if the lexicographer has decided to edit the sentences, one has to find them in the corpus first. [P2]

Nowadays, the rapid development of e-lexicography has generated automatic tools to help detect various information units for dictionaries, including example sentences, which has resulted in automatic generation of dictionary databases. The automatic detection of example sentences requires the identification of language specific criteria that a good dictionary example should meet.

Automatic Detection of Example Sentences

Automatic detection of example sentences is a new practice in Estonian lexicography. There is a great need for this type of research due to the recent arrival of automatic lexicography in Estonia. The first project that implemented this approach was the Estonian Collocations Dictionary (ECD) (Kallas, Kilgarriff et al. 2015). The database of the ECD was automatically generated from the Estonian National Corpus 2013 and entailed automatic extraction of headwords, grammatical relations, statistics, collocates and example sentences.

There are three methods used to detect example sentences automatically: machine learning methods (Witten et al. 2016, Pilán 2018, Hastie et al. 2009, Søgaard 2013, Ljubešić & Peronja 2015, Boullosa et al. 2017, Tolmachev & Kurohashi 2017), rule-based methods (Didakowski et al. 2012, Baisa & Suchomel 2014), and a combination of the two (Lemnitzer et al. 2015, Pilán et al. 2013, Huang & Ku 2016).

The approach described in this dissertation is rule-based. The rules are set out in the GDEX configuration files. For the purpose of fine-tuning the parameters of the configuration, machine learning techniques have also been adopted.

Good Dictionary Examples (GDEX)

GDEX assigns a numerical score to each corpus sentence based on certain heuristics. Scores correspond to sentence values and vary from 0 (the worst) to 1 (the best), which helps to identify completely unsuitable candidates and rank all other (suitable) candidates according to their score so that the best candidates are presented at the top of the list. The computation of the GDEX score is based on a formula that deals with a variety of formal classifiers, paying attention to various features (parameters) of the sentence. The formula itself is described in the GDEX configuration files. Classifiers can be divided into hard classifiers and soft classifiers, both of which (in the case of the Estonian configuration) represent 50% of the score. Hard classifiers are mutually dependent (all conditions have to be met for a candidate sentence to receive 50% of the score), whereas soft classifiers are not (each of them are scored independently). Hard classifiers help to identify and sort truly inappropriate candidate sentences, whereas soft classifiers either reduce the overall score by penalizing the sentences or giving them bonus points. Classifiers contain a variety of predefined parameters that are interpreted as machine-measurable attributes. They can measure syntactic and lexical features of the sentence (e.g., sentence length, word frequency, keyword position, keyword repetition, etc.). The GDEX Editor can be used to simplify the development of configurations. This system evaluates sentences using two versions of the GDEX configuration and assigns two scores and ranks; thus, by comparing two configurations, the configuration developers can mark apt sentences and thus assess which set of parameters is more suitable for the task.

(Koppel, Kallas et al. 2019) The original purpose of creating GDEX was to help the computer do the 'groundwork' and to reduce the time spent by lexicographers in selecting example sentences from the corpus. Recently, GDEX has been used more widely, taking into account the needs of language learners as well as linguists and lexicographers (Baisa & Suchomel 2008, [P2], [P3]).

Results

As part of this dissertation, six different GDEX configurations for learners at different CEFR levels of Estonian were developed: GDEX 1.2. (Kallas, Kilgarriff et al. 2015), GDEX 1.3 [P1], GDEX 1.4 [P2, P3, P4], et-Basic-v1, etIndependent-v1, and etProficient-v1 [P5].

The parameters of these GDEX configurations are provided in Table 3. The character + indicates the existence and the character Ø the absence of the parameter, the character * indicates a penalty, and x indicates a high penalty. Hard classifiers are in bold, soft classifiers are in normal text.

GDEX 1.2 (Kallas, Kilgarriff et al. 2015) was developed in connection with the Estonian Collocations Dictionary (ECD) project (Kallas, Kilgarriff et al. 2015). ECD is a monolingual, online, corpus-driven, scholarly dictionary aimed at learners of Estonian as a foreign language or second language at the upper intermediate and advanced levels (CEFR levels B2–C1). In order to determine Estonian specific parameters, example sentences from the Basic Estonian Dictionary (2014) (BED) (Kallas, Tuulik et al. 2014) and the Dictionary of Estonian (2019) (DicEst) (Langemets, Tiits et al. 2018) were analysed. The main focus was on average sentence length, average word length, and the number of words in a sentence. By using GDEX 1.2, about 2.5 million sentences were extracted into the ECD database. The lexicographers' task in compiling the dictionary entry, among other things, was to choose one example out of the five extracted – the one that best illustrated the use of the collocation. In the process of postediting the ECD database, it became evident that extracted sentences included many anaphora, but did not include, for example, a verb. In order to obtain better results, it was decided to improve the parameters and thus develop a new version of the GDEX configuration.

GDEX 1.3 [P1] was used to compile a corpus (EstonianNC GDEX) of good corpus sentences meeting the criteria of GDEX 1.3. This was the first attempt to create a GDEX-based corpus targeted at learners of Estonian. The analysis of the corpus sentences revealed that the top candidate sentences still included many anaphora, proper names, numbers, and low frequency words. Anaphoric references cannot be understood without access to the wider context. Low frequency words tend to be difficult for language learners and sentences with proper names, especially people's names, tend to include too much personal information. To improve the output, a new version of the GDEX configuration was developed.

Table 3. The parameters of the Estonian GDEX configurations

parameter	**GDEX 1.2**	**GDEX 1.3**	**GDEX 1.4**	**etBasic-v1**	**etIndependent-v1**	**etProficient-v1**
sentence length (tokens)	5–20	5–20	4–20	3–14	3–18	4–23
optimal interval (tokens)	10–12	10–12	6–12	4–7	4–12	6–14
illegal characters (not allowed)	<|][>/\^@	<|\]\[>/\\^@	<|][>/\}{^@•·*#=_~	<|][>/\}{^@•·*#=_~	<|][>/\}{^@•·*#=_~	<|][>/\}{^@•·*#=_~
rare characters (are penalized)	A-Z0-9'.,!?)(;:-	A-Z0-9'.,!?)(;:“„”«»"…-]	A-Z0-9'.,!?)(;:“ „”«»"'×…§-	A-Z0-9'.,!?)(;:“ „”«»"'×…§-	A-Z0-9'.,!?)(;:“ „”«»"'×…§-	A-Z0-9'.,!?)(;:“ „”«»"'×…§-
sentence initial tags (not allowed)	J	J	J, I, Y, G, Z	J, I, Y, G, Z	J, I, Y, G, Z	J, I, Y, G, Z
lemma frequency < 1000 (are penalized)	Ø	Ø	+	+	+	+
verb forms (are penalized)	*ma*-infinitive forms (*-mata*, *-mast*, *-mas*, *-maks*), *des*-gerundive	*ma*-infinitive forms (*-mata*, *-mast*, *-mas*, *-maks*), *des*-gerundive	*ma*-infinitive forms (*-mata*, *-mast*, *-mas*, *-maks*), *des*-gerundive	*tud*-participle	impersonal forms (*-takse*, *-dakse*, *-akse*), impersonal forms (*-t*, *-d*, *-ta*, *-da*)	conditional mood forms (*-nuks*, *-taks*, *-tuks*), oblique mood forms (*-tavat*, *-tuvat*, *-nuvat*)
verb forms (not allowed)	Ø	Ø	Ø	*ma*-infinitive forms (*-mast*, *-maks*, *-mata*, *-tama*), *des*-form, conditional mood forms (*-nuks*, *-taks*, *-tuks*), oblique mood forms (*-tavat*, *-tuvat*, *-nuvat*), imperative mood forms (*-neg_gu*, *-tagu*), impersonal forms (*-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da*)	*ma*-infinitive forms (*-maks*, *-tama*), conditional mood forms (*-nuks*, *-taks*, *-tuks*), oblique mood forms (*-tavat*, *-tuvat*, *-nuvat*), imperative mood forms (*-neg_gu*, *-tagu*)	Ø

parameter	GDEX 1.2	GDEX 1.3	GDEX 1.4	etBasic-v1	etIndependent-v1	etProficient-v1
superlative form	Ø	Ø	Ø	*	Ø	Ø
verb complements	Ø	Ø	Ø	*	Ø	Ø
greylist	446 words + *mina* 'I/me', *sina* 'you', *tema* 'him/her', *see* 'this/it', *too* 'it', *siin* 'here', *seal* 'there'	446 words + *mina* 'I/me', *sina* 'you', *tema* 'him', *see* 'this/it', *too* 'it'	474 words + *see* 'this/it', *too* 'it', *siin* 'here', *siia* 'here', *siit* 'from here', *seal* 'there', *sinna* 'there', *sealt* 'from there'	451 words	451 words	451 words
blacklist		*siin* 'here'', *siia* 'here', *siit* 'from here', *seal* 'there', *sinna* 'there', *sealt* 'from there', *siis* 'then'	Ø	267 words	267 words	267 words
abbreviations	*	*	*	Ø	*	*
includes a verb	Ø	+	+	+	+	+
includes a noun	Ø	Ø	Ø	+	+	Ø
penalty for long words	Ø	Ø	Ø	>9 characters	>11 characters	Ø
keyword repetition	Ø	Ø	+	+	+	+
proper name	Ø	Ø	*	*	*	*
numeral + substantive *kroon* ('krona')	Ø	Ø	*	*	*	*

parameter	GDEX 1.2	GDEX 1.3	GDEX 1.4	etBasic-v1	etIndependent-v1	etProficient-v1
number of commas	Ø	Ø	* >1	* >1	* >1	* >1
number of conjunctions	Ø	Ø	* >1	* >1	* >1	* >1
number of pronouns	Ø	Ø	* >1	* >1	* >1	* >1
number of numerals	Ø	Ø	* >1	* >1	* >1	* >1
number of adverbs	Ø	Ø	* >1	* >1	* >1	* >1
number of proper names	Ø	Ø	* >1	* >1	* >1	* >1
number of verbs	Ø	Ø	* >2	* >2	* >2	* >2
weights	Ø	Ø	+	+	+	+
preferred subcorpus	Ø	+	+	Ø	Ø	Ø
sentence initial words (not allowed)	9 words	5 words	62 words	69 words	69 words	69 words
sentence initial word pairs (not allowed)	Ø	Ø	79 word pairs	85 word pairs	85 word pairs	85 word pairs
first word of the sentence written in all caps (not allowed)	Ø	Ø	+	+	+	+

In devising **GDEX 1.4** [P2**,** P3], machine learning techniques were adopted in which human-judged sets of examples were used as training datasets: one dataset contained the selected examples in the ECD database (so-called good examples) and the other one comprised of rejected or non-selected examples (so-called bad examples). The parameters obtained from the two datasets were used in the optimization of classifiers' values and weight attributions. The following features were analysed: sentence length; word length; keyword position; first tag; number of commas, adverbs, verbs, proper names, and pronouns in a sentence; etc. In addition to improving the parameters of existing classifiers, the analysis of the datasets helped to include completely new ones into the configuration, as well as to remove some of the non-relevant ones. In previous versions, each of the soft classifiers contributed the same share to 50% of the score. In GDEX 1.4, weights were assigned to each of the soft classifiers, except to those that were grouped together and shared the same weight due to shared characteristics (number of elements in the sentence). Optimal interval and word frequency proved to affect the output the most; hence, they were assigned the highest weights.

Evaluation of GDEX 1.4 [P4]. To evaluate the output of GDEX 1.4, an assessment was 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 assessment task was to determine whether, according to the above mentioned two types of annotators, authentic and unedited corpus sentences would be suitable as example sentences for learners' dictionaries at B2–C1 level. 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 inappropriate candidates; 3) example sentences compiled by lexicographers are suitable example sentences. Both types of annotators considered as many as 96% of the dictionary examples and 85% of corpus sentences chosen as good examples by GDEX 1.4 to be suitable. 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% were considered unsuitable. When the annotators were asked about the reasons for considering a sentence unsuitable, the most common arguments were that the sentences included anaphora and hence needed more context, or that the sentences were colloquial, too long, or too short.

Configurations for learners at different CEFR levels [P5]. GDEX 1.4 was used as a basis for the development of GDEX versions for three general language proficiency levels of Estonian L2. Version etBasic-v1 detects examples for learners at CEFR level A, version etIndependent-v1 detects examples for learners at CEFR level B, and version etProficient-v1 detects examples for learners at CEFR level C. In order to develop previously mentioned configurations, full sentences from the Estonian Coursebook Corpus 2018 were analysed. The coursebook corpus helped to identify which specific parameters characterize sentences in each proficiency level. Sentence and word length were analysed, as

well as the verb forms used, syntactic properties of the sentences, etc. etBasic-v1, etIndependent-v1, and etProficient-v1 can be refined by adding a lexical filter (so-called whitelist) that only detects sentences comprising of words at a certain level. For Estonian, vocabulary lists for A1, A2, and B1 levels can be used (Kallas & Koppel 2018a, 2018b, 2018c).

Problems and Perspectives

Various issues were observed when displaying authentic corpus sentences to end-users via the corpus-based web tool etSkELL and language portal Sõnaveeb: grammatical and semantic ambiguity; the inferior quality of the corpus; and mistakes in lemmatization, morphological analysis, sentence segmentation and multi-word unit detection.

As a **further development**, preparing GDEX-based corpora for different CEFR levels and evaluating their output is planned for the near future. Devising a GDEX configuration aimed at Estonian native speakers is also planned. New configurations will be supplemented with additional classifiers – e.g., lexical filter based on CEFR vocabulary lists (Kallas & Koppel 2018a, 2018b, 2018c) – and minimum frequencies for any tokens and lemmas (e.g., Kuhn 2017: 265). Implementing crowdsourcing is also intended as a way for users of the dictionary portal to assess the suitability by upvoting or downvoting authentic corpus sentences. A similar approach is also used by Fišer & Čibej (2017) and Kuhn et al. (2019).

KIRJANDUS

Adamska-Sałaciak, Arleta 2013. Issues in compiling bilingual dictionaries. – Howard Jackson (Ed), The Bloomsbury Companion to Lexicography. London: Bloomsbury, 213–231.

Aharoni, Roee, Moshe Koppel, Yoav Goldberg 2014. Automatic detection of machine translated text and translation quality estimation. – Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Vol. 2: Short Papers. Baltimore, Maryland: Association for Computational Linguistics, 289–295.

Apresjan, Valentina, Vít Baisa, Olga Buivolova, Olga Kultepina, Anna Maloletnjaja 2016. RuSkELL: Online Language Learning Tool for Russian Language. – Tinatin Margalitadze, George Meladze (Eds), Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity. Tbilisi: Ivane Javakhishvili Tbilisi State University, 292–299.

Aston, Guy 1997. Enriching the learning environment: Corpora in ELT. – Anne Wichmann, Steven Fligelstone, Tony McEnery, Gerry Knowles (Eds), Teaching and Language Corpora. Harlow: Longman, 51–64.

Atkins, Sue, Michael Rundell 2008. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Baicheng, Zhang 2009. Do example sentences work in direct vocabulary learning? – Issues in Educational Research, 19(2), 175–189.

Baisa, Vít, Vít Suchomel 2014. SkELL: Web interface for English language learning. – Aleš Horák, Pavel Rychlý (Eds), Proceedings of the Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014. Brno: Tribun EU, 63–70.

Boullosa Beto, Richard Eckart de Castilho, Alexander Geyken, Lothar Lemnitzer, Iryna Gurevych 2017. A tool for extracting sense-disambiguated example sentences through user feedback. – André Martins, Anselmo Peñas (Eds), Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia: Association for Computational Linguistics, 69–72.

Bowker, Lynne 2010. The contribution of corpus linguistics to the development of specialised dictionaries for learners. – Pedro A. Fuertes-Olivera (Ed), Specialised Dictionaries for Learners. (Lexicographica. Series Maior 136.) Berlin: de Gruyter, 155–170.

Burke, Sean Michael 2003. The design of online lexicons. – Piet van Sterkenburg (Ed), A Practical Guide to Lexicography. (Terminology and Lexicography Research and Practice 6.) Amsterdam: John Benjamins Publishing Company, 240–249.

Byrne, Jody 2006. Technical Translation: Usability Strategies for Translating Technical Documentation. Dordrecht: Springer.

Cobb, Tom 1997. Is there any measurable learning from hands-on concordancing? – System, 25, 301–315.

COBUILD = Collins COBUILD English Language Dictionary. London: HarperCollins Publishers, 1987.

Cook, Paul, Michael Rundell, Jay Han Lau, Timothy Baldwin 2014. Applying a word-sense induction system to the automatic extraction of diverse dictionary examples. – Andrea Abel, Chiara Vettori, Natascia Ralli (Eds), Proceedings of the XVI EURALEX International Congress: The User in Focus. Bolzano/Bozen: EURALEX, 319–328.

Cowie, Anthony P. 1978. The place of illustrative material and collocations in the design of a learners' dictionary. – Peter Strevens (Ed), In honour of A. S. Hornby. Oxford: Oxford University Press, 127–139.
DeKeyser, Robert 2008. Implicit and explicit learning. – Catherine J. Doughty, Michael H. Long (Eds), The Handbook of Second Language Acquisition. Blackwell Publishing Ltd, 313–348.
Didakowski, Jörg, Lothar Lemnitzer, Alexander Geyken 2012. Automatic example sentence extraction for a contemporary German dictionary. – Ruth Vatvedt Fjeld, Julie Matilde Torjusen (Eds), Proceedings of the XV EURALEX International Congress, 7–11 August. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 343–349.
Dodd, Bill 1997. Exploring texts through the concordancer: Guiding the learner. – Anne Wichmann, Steven Fligelstone, Tony McEnery, Gerry Knowles (Eds), Teaching and Language Corpora. Harlow: Longman, 131–145.
Eesti keele naabersõnad 2019. Jelena Kallas, Kristina Koppel, Maria Tuulik, Geda Paulsen (Toim). Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee/ (20.06.2019).
Eesti keele põhisõnavara sõnastik. Jelena Kallas, Mai Tiits, Maria Tuulik (Toim). Madis Jürviste, Kristina Koppel, Maria Tuulik (Koost). Tallinn: Eesti Keele Sihtasutus, 2014.
Eesti keele sõnaraamat 2019. Margit Langemets, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll (Toim). Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee/ (20.06.2019).
EKSS = Eesti keele seletav sõnaraamat 1–6. „Eesti kirjakeele seletussõnaraamatu" (1988–2007) 2., täiendatud ja parandatud trükk. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll (Toim). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus, 2009.
Erelt, Mati, Helle Metslang 2017. Eesti keele süntaks. (Eesti keele varamu 3.) Tartu: Tartu Ülikooli Kirjastus.
Fišer, Darja, Jaka Čibej 2017. The potential of crowdsourcing in modern lexicography. – Vojko Gorjanc, Polona Gantar, Iztok Kosem, Simon Krek (Eds), Dictionary of Modern Slovene: Problems and Solutions. Ljubljana: Ljubljana University Press, Faculty of Arts, 212–228.
Fox, Gwyneth 1987. The case of examples. – John Sinclair (Ed), Looking Up. An Account of the COBUILD Project in Lexical Computing. London: Collins ELT, 137–149.
Frankenberg-Garcia, Ana 2012. Learners' use of corpus examples. – International Journal of Lexicography, 25(3), 273–296.
Frankenberg-Garcia, Ana 2014. The use of corpus examples for language comprehension and production. – ReCall, 26(2), 128–146.
Gantar, Polona, Iztok Kosem, Simon Krek 2016. Discovering automated lexicography: The case of the Slovene lexical database. – International Journal of Lexicography, 29(2), 200–225.
Gantar, Polona, Simon Krek 2011. Slovene lexical database. – Daniela Majchráková, Radovan Garabík (Eds), Natural Language Processing, Multilinguality: Sixth International Conference, 20–21 October, Modra, Slovakia. Brno: Tribun EU, 72–80.
Gatto, Maristella 2014. Web as Corpus: Theory and Practice. London: A&C Black.
Gavioli, Laura 2005. Exploring Corpora for ESP Learners. (Studies in Corpus Linguistics 21.) Amsterdam: John Benjamins Publishing.
Halling, Anneliis 2016. Eesti keele keeleressursse kasutav õppeprogramm käänete õppimiseks. Bakalaureusetöö. Tartu: Tartu Ülikool. http://hdl.handle.net/10062/56227 (24.09.2019).

Hannay, Mike 2003. Types of bilingual dictionaries. – Piet van Sterkenburg (Ed), A Practical Guide to Lexicography. (Terminology and Lexicography Research and Practice 6.) Amsterdam: John Benjamins Publishing Company, 145–153.

Harras, Gisela 1989. Zu einer Theorie des lexikographischen Beispiels. – Wörterbücher: Ein internationales Handbuch zur Lexikographie. 1. Teilband. (HSK 5/1.) Berlin: Walter de Gruyter, 607–614.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman 2009. Unsupervised learning. – The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag, 485–585.

Hirata, Yoko, Yoshihiro Hirata 2018. Students' evaluation of SkELL: The 'Sketch Engine for Language Learning'. – Simon K. S. Cheung, Lam-for Kwok, Kenichi Kubota, Lap-Kei Lee, Jumpei Tokito (Eds), Blended Learning. Enhancing Learning Success. ICBL 2018. (Lecture Notes in Computer Science 10949). Cham: Springer, 368–377.

Holdt, Špela Arhar, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski, Marko Robnik-Šikonja 2018. Thesaurus of Modern Slovene: By the community for the community. – Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (Eds), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018. Ljubljana University Press, Faculty of Arts, 401–410.

Hubbard, Philip, James Coady, John Graney, Kouider Mokhtari, Jeff Magoto 1986. Report on a pilot study of the relationship of high-frequency vocabulary knowledge and reading proficiency in ESL readers. – Ohio University Papers in Linguistics and Language Teaching, 8, 48–57.

Huang, Chieh-Yang, Lun-Wei Ku 2016. GiveMeExample: Learning confusing words by example sentences. – IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA. Piscataway, NJ: IEEE Press, 1414–1417.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, Vít Suchomel 2014. Finding terms in corpora for many languages with the Sketch Engine. – Shuly Wintner, Marko Tadić, Bogdan Babych (Eds), Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg: EACL, 53–56.

Johns, Tim 1991. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. – Tim Johns, Philip King (Eds), Classroom Concordancing. ELR Journal, 4, 27–45.

Jürviste, Madis, Jelena Kallas, Margit Langemets, Maria Tuulik, Ülle Viks 2011. Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. – Iztok Kosem, Karmen Kosem (Eds), Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of the eLex 2011, Bled, 10–12 November. Ljubljana: Trojina, Institute for Applied Slovenian Studies, 106–112.

Kaalep, Heiki-Jaan 1998. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. – Keel ja Kirjandus, 1, 22–29.

Kaalep, Heiki-Jaan, Kadri Muischnek 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. – Eesti Rakenduslingvistika Ühingu aastaraamat, 5, 157–172.

Kaalep, Heiki-Jaan, Kadri Muischnek 2012. Osalausete tuvastamine eestikeelses tekstis kui iseseisev ülesanne. – Eesti Rakenduslingvistika Ühingu aastaraamat, 8, 55–68.

Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. (Humanitaarteaduste dissertatsioonid 32.) Tallinn: Tallinna Ülikool.

Kallas, Jelena, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Margit Langemets, Jan Michelfeit, Maria Tuulik, Ülle Viks 2015. Automatic generation of the Estonian Collocations Dictionary database. – Iztok Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek (Eds), *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age.* Proceedings of the eLex 2015 Conference, 11–13 August, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 1–20.

Kallas, Jelena, Svetla Koeva, Margit Langemets, Carole Tiberius, Iztok Kosem 2019. Lexicographic practices in Europe: Results of the ELEX survey on user needs. – 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 the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 519–536.

Kallas, Jelena, Kristina Koppel 2018a. Eesti keele B1-taseme sõnavara. Tallinn: Eesti Keele Instituut. http://www.eki.ee/keeletase/lists/B1.pdf (14.02.2019).

Kallas, Jelena, Kristina Koppel 2018b. Eesti keele A2-taseme sõnavara. Tallinn: Eesti Keele Instituut. http://www.eki.ee/keeletase/lists/A2.pdf (14.02.2019).

Kallas, Jelena, Kristina Koppel 2018c. Eesti keele A1-taseme sõnavara. Tallinn: Eesti Keele Instituut. http://www.eki.ee/keeletase/lists/A1.pdf (14.02.2019).

Kallas, Jelena, Kristina Koppel, Maria Tuulik 2014. Eesti keele põhisõnavara sõnastik. – Oma Keel, 2, 87–89.

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

Kallas, Jelena, Margit Langemets, Kristina Koppel, Maria Tuulik 2019. State-of-the-art on monolingual lexicography for Estonia. – Slovenščina 2.0, 7(1), 25–38. https://doi.org/10.4312/slo2.0.2019.1.25-38

Kallas, Jelena, Vit Suchomel, Maria Khokhlova 2017. Automated identification of domain preferences of collocations. – Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa (Eds), Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference. Leiden, Netherlands 19–21 September. Brno: Lexical Computing Ltd., 309–320.

Kallas, Jelena, Maria Tuulik 2011. Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 59–75.

Kallas, Jelena, Maria Tuulik, Margit Langemets 2014. The Basic Estonian Dictionary: The first monolingual L2 learner's dictionary of Estonian. – Andrea Abel, Chiara Vettori, Natascia Ralli (Eds), Proceedings of the XVI EURALEX International Congress: The User in Focus. Bolzano/Bozen: EURALEX, 1109–1119.

Kaneta, Taku 2011. Folded or unfolded: Eye-tracking analysis of L2 learners' reference behavior with different types of dictionary. – Asialex2011 Proceedings, Kyoto Terrsa, Japan, August 22–24. Lexicography: Theoretical and Practical Perspectives. Asian Association for Lexicography, 219–224.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, Arto Anttila 1995. Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text. Berlin/New York: Mouton de Gruyter.

Kaufmann, Nicolas, Thimo Schulze, Daniel Veit 2011. More than fun and money. Worker motivation in crowdsourcing – A Study on Mechanical Turk. – Proceedings of the

Seventeenth Americas Conference on Information Systems (AMCIS 2011). Red Hook, NY, Curran Associates, Inc., 3012–3022.
Kilgarriff, Adam 2009. Corpora in the classroom without scaring the students. – Proceedings of the 18th International Symposium on English Teaching, Taipei.
Kilgarriff, Adam 2010a. Comparable corpora within and across languages, word frequency lists and the KELLY project. – Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, 1–5. http://www.fb06.uni-mainz.de/lk/bucc2010/ (07.09.2019).
Kilgarriff, Adam 2010b. A detailed, accurate, extensive, available English lexical database. – Proceedings of the NAACL HLT 2010 Demonstration Session. Association for Computational Linguistics.
Kilgarriff, Adam 2012. Getting to know your corpus. – International Conference on Text, Speech and Dialogue. Proceedings of the 15th International Conference, TSD 2012, Brno, Czech Republic, September 3–7, 2012. (Lecture Notes in Computer Science 7499.) Berlin/Heidelberg: Springer, 3–15.
Kilgarriff, Adam 2013. Using corpora as data source for dictionaries. – Howard Jackson (Ed), The Bloomsbury Companion to Lexicography. London: Bloomsbury, 77–96.
Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel 2014. The Sketch Engine: ten years on. – Lexicography, 1(1), 7–36.
Kilgarriff, Adam, Vít Baisa, Pavel Rychlý, Miloš Jakubíček 2015. Longest-commonest match. – Iztok Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek (Eds), *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age.* Proceedings of the eLex 2015 Conference, 11–13 August, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 397–404.
Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý 2008. GDEX: Automatically finding good dictionary examples in a corpus. – Elisenda Bernal, Janet DeCesaris (Eds), Proceedings of the XIII EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425–432.
Kilgarriff, Adam, Vojtěch Kovář, Simon Krek, Irena Srdanović, Carole Tiberius 2010. A quantitative evaluation of word sketches. – Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6–10 July). Ljouwert: Fryske Academy, 372–379.
Kilgarriff, Adam, Vojtěch Kovář, Pavel Rychlý 2010. Tickbox lexicography. – Sylviane Granger, Magali Paquot (Eds), New Challenges, New Applications. Proceedings of the eLex 2009: eLexicography in the 21st Century. Louvain-la-Neuve: Presses Universitaires de Louvain, 411–418.
Kilgarriff, Adam, Fredrik Marcowitz, Simon Smith, James Thomas 2015. Corpora and language learning with the Sketch Engine and SKELL. – Revue française de linguistique appliquée, 20(1), 61–80.
Kilgarriff, Adam, Pavel Rychlý, Pavel Smr, David Tugwell 2004. The Sketch Engine. – Geoffrey Williams, Sandra Vessier (Eds), Proceedings of the XI EURALEX International Congress. Lorient, France: Université de Bretagne Sud, 105–115.
Koppel, Kristina, Jelena Kallas, Maria Khokhlova, Vít Baisa, Vít Suchomel, Jan Michelfeit 2019. 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 the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 763–782.

Koppel, Kristina, Arvi Tavast, Margit Langemets, Jelena Kallas 2019. 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 the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 434–452.

Kosem, Iztok, Polona Gantar, Simon Krek 2013. Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. – Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets, Maria Tuulik (Eds), Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, 17–19.

Kosem, Iztok, Milos Husák, Diana McCarthy 2011. GDEX for Slovene. – Iztok Kosem, Karmen Kosem (Eds), Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of the eLex 2011, Bled, 10–12 November. Ljubljana: Trojina, Institute for Applied Slovenian Studies, 151–159.

Kosem, Iztok, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Cyprian Laskowski 2018. Collocations Dictionary of Modern Slovene. – Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (Eds), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018. Ljubljana University Press, Faculty of Arts, 989–997.

Kovář, Vojtěch, Monika Močiariková, Pavel Rychlý 2016. Finding definitions in large corpora with Sketch Engine. – Nicoletta Calzolari et al. (Eds), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association, 391–394.

Krek, Simon 2019. Natural Language Processing and automatic knowledge extraction for lexicography. – International Journal of Lexicography, 32(2), 115–118.

Krek, Simon, Cyprian Laskowski, Marko Robnik-Šikonja 2017. From translation equivalents to synonyms: Creation of a Slovene thesaurus using word co-occurrence network analysis. – Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa (Eds), Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference. Leiden, Netherlands 19–21 September. Brno: Lexical Computing Ltd., 93–109.

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. European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect), 14–16 March, Lisbon, Portugal. https://www.researchgate.net/publication/331813170_Crowdsourcing_corpus_cleaning_for_language_learning_-_an_approach_proposal (20.06.2019).

Kuhn, Tanara Zingano, José Pedro Ferreira 2016. Building a corpus of written academic texts in Portuguese. – Teaching and Language Corpora Conference (TaLC12), 20–23 July, Giessen, Germany. Book of Abstracts, 103.

Langemets, Margit 2010. Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras. Tallinn: Eesti Keele Sihtasutus.

Langemets, Margit, Tarja Heinonen, Kristina Koppel, Ülle Viks, Indrek Hein 2017. From monolingual to bilingual dictionary: The case of semi-automated lexicography on the example of Estonian-Finnish dictionary. – Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa (Eds), Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference. Leiden, Netherlands 19–21 September. Brno: Lexical Computing Ltd., 155–171.

Langemets, Margit, Andres Loopmann, Ülle Viks 2006. The IEL dictionary management system of Estonian. – DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems: Pre-EURALEX workshop, Turin, 5th September 2006. Turin: University of Turin, 11–16.

Langemets, Margit, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll 2018. Eesti keel uues kuues: Eesti keele sõnaraamat 2018. – Keel ja Kirjandus, 12, 942–958.

Lázaro, Jorge, Juan-Manuel Torres-Moreno, Gerardo Sierra, M. Teresa Cabré, Andrés Torres 2017. Genex+, a semantic-based automatic extractor of examples applied to bilingual terms. – Research in Computing Science, 145, 51–67.

Laufer, Batia 1992. Corpus-based versus lexicographer examples in comprehension and production of new words. – Hannu Tommola et al. (Eds), Proceedings of the V EURALEX International Congress. Tampere: University of Tampere, 71–76.

Leech, Geoffrey 1997. Teaching and language corpora: A convergence. – Anne Wichmann, Steven Fligelstone, Tony McEnery, Gerry Knowles (Eds), Teaching and Language Corpora. Harlow: Longman, 1–23.

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.

Lemnitzer, Lothar, Christian Pölitz, Jörg Didakowski, Geyken Alexander 2015. Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German. – Iztok Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek (Eds), *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age.* Proceedings of the eLex 2015 Conference, 11–13 August, Herstmonceux Castle, United Kingdom. Ljubljana/ Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 21–31.

Lew, Robert 2004. Which Dictionary for Whom? Receptive Use of Bilingual Monolingual and Semi-Bilingual Dictionaries by Polish Learners of English. Poznań: Motivex.

Lew, Robert, Arleta Adamska-Sałaciak 2015. A case for bilingual learners' dictionaries. – ELT Journal, 69(1), 47–57.

Ljubešić, Nikola, Peronja Mario 2015. Predicting corpus example quality via supervised machine learning. – Iztok Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek (Eds), *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age.* Proceedings of the eLex 2015 Conference, 11–13 August, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 477–485.

Läti-eesti sõnaraamat. Merle Madisson, Aive Mandel, Tauno Nõulik, Anita Tannenberg, Arvi Tavast. Tallinn: Eesti Keele Sihtasutus, 2015. https://doi.org/10.15155/3-00-0000-0000-0000-05B49L (29.01.2020).

Martelli, Federico, Roberto Navigli, Paolo Spadoni, Giovanni Stilo, Paola Velardi 2019. Lexical-Semantic Analytics for NLP: Sense Clustering. https://elex.is/wp-content/uploads/2019/08/ELEXIS_D3_1_Lexical_semantic_analytics_for_NLP_sense_clustering_Final.pdf (20.06.2019).

MCD 2010 = Macmillan Collocations Dictionary for Learners of English. Australia: Macmillan Education, 2010.

McEnery, Tony, Andrew Hardie 2012. Corpus Linguistics: Method, Theory and Practice. (CambridgeTextbooks in Linguistics.) Cambridge: Cambridge University Press.

Měchura, Michal 2017. Introducing Lexonomy: An open-source dictionary writing and publishing system. – Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa (Eds), Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference. Leiden, Netherlands 19–21 September. Brno: Lexical Computing Ltd., 662–679.

Muischnek, Kadri 2006. Verbi ja noomeni püsiühendid eesti keeles. (Dissertationes philologiae Estonicae Universitatis Tartuensis 17.) Tartu: Tartu Ülikooli Kirjastus.

Nguyen-Son, Hoang-Quoc, Tran Phuong Thao, Seira Hidano, Shinsaku Kiyomoto 2019. Detecting Machine-Translated Paragraphs by Matching Similar Words. – arXiv preprint. arXiv:1904.10641.

Pilán, Ildikó 2018. Automatic Proficiency Level Prediction for Intelligent Computer-Assisted Language Learning. Doctoral thesis. (Data linguistica 29.) University of Gothenburg.

Pilán, Ildikó, Elena Volodina, Richard Johansson 2013. Automatic selection of suitable sentences for language learning exercises. – Linda Bradley, Sylvie Thouësny (Eds), 20 Years of EUROCALL: Learning from the Past, Looking to the Future: 2013 EUROCALL Conference Proceedings. Dublin, Voillans: Research-publishing.net, 218–225.

Pomikalek, Jan, Vít Suchomel 2012. Efficient web crawling for large text corpora. – Adam Kilgarriff, Serge Sharoff (Eds), Proceedings of the 7th Web-as-Corpus Workshop, Lyon, France, 39–43.

Puolakainen, Tiina 2001. Eesti keele arvutigrammatika: morfoloogiline ühestamine. (Dissertationes mathematicae Universitatis Tartuensis 27.) Tartu: Tartu Ülikooli kirjastus.

Raadik, Lydia 2016. Sugude kujutamine „Eesti õigekeelsussõnaraamatu ÕS 2013“ näidetes. Magistritöö. Tallinna Ülikool, Humanitaarteaduste instituut.

Raamdokument 2007 = Euroopa keeleõppe raamdokument: õppimine, õpetamine, hindamine. Tartu: Haridus- ja Teadusministeerium, 2007.

Richards, Jack C., Richard W. Schmidt 2013. Longman Dictionary of Language Teaching and Applied Linguistics. New York: Routledge.

Rundell, Michael 1998. Recent trends in pedagogical lexicography. – Bernard Caron (Ed), Proceedings of the 16th International Congress of Linguistics, Vol. 2. New York: Pergamon, 159–162.

Rychlý, Pavel 2008. A lexicographer-friendly association score. – Petr Sojka, Aleš Horák (Eds), Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008. Brno: Masaryk University, 6–9.

Rychlý, Pavel, Adam Kilgarriff 2007. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). – Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 41–44.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger 2001. Multiword Expressions: A Pain in the Neck for NLP. – LinGO Working Paper No. 2001-03.
Simpson, John 2003. The production and use of occurrence examples. – Piet van Sterkenburg (Ed), A Practical Guide to Lexicography. (Terminology and Lexicography Research and Practice 6.) Amsterdam: John Benjamins Publishing Company, 260–272.
Sinclair, John 1987. Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. London: Collins ELT.
Sinclair, John 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.
Srdanović, Irena, Iztok Kosem 2016. GDEX for Japanese: automatic extraction of good dictionary example candidates. – Lexicographic Resources for Human Language Technology, GLOBALEX 2016, Portorož, 24 May 2016. Paris: European Language Resources Association, 57–64.
Svensén, Bo 2009. A Handbook of Lexicography. The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.
Søgaard, Anders 2013. Semi-supervised learning and domain adaptation in natural language processing. – Synthesis Lectures on Human Language Technologies, 6(2), 1–103.
Zöfgen, Ekkehard 1986. Kollokation, Kontextualisierung, (Beleg-)Satz. Anmerkungen zur Theorie und Praxis des lexikographischen Beispiels. – Französische Sprachlehre und bon usage. Festschrift für Hans Wilhelm Klein zum 75. Geburtstag. München: Max Hueber, 219–238.
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 (Eds), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018. Ljubljana University Press, Faculty of Arts, 749–761.
Tavast, Arvi, Marju Taukar 2013. Mitmekeelne oskussuhtlus. http://tavast.ee/public/opik/opik.pdf (20.06.2019).
Teral, Maarika 2015. Arvutipõhine eesti keele õpe: vahendid ja hinnangud nende efektiivsusele Tartu ülikooli keelekursuste näitel. (Dissertationes philologiae Estonicae Universitatis Tartuensis 37.) Tartu: Tartu Ülikooli Kirjastus.
Tiberius, Carole, Dirk Kinable 2015. Using and configuring GDEX for Dutch. – Slaidiesitlus. http://www.elexicography.eu/wp-content/uploads/2015/04/ENeLWG3_GDEX4Dutch.pdf (04.02.2016).
Tiberius, Carole, Tanneke Schoonheim 2016. The *Algemeen Nederlands Woordenboek (ANW)* and its lexicographical process. – Hardarik Blühdorn, Mechthild Elstermann, Annette Klosa (Hg.), Lexikographische Prozesse bei Intemetwörterbüchern. (OPAL. Online publizierte Arbeiten zur Linguistik 1.) Mannheim: Institut für Deutsche Sprache, 20–28.
Tolmachev, Arseny, Sadao Kurohashi 2017. Automatic extraction of high-quality example sentences for word learning using a determinantal point process. – Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. Stroudburg: Association for Computational Linguistics, 133–142.
Uiboaed, Kristel 2010. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel. – Eesti Rakenduslingvistika Ühingu aastaraamat, 6, 307–326.
Vare, Silvi 2012. Eesti keele sõnapered. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.
Vihma, Helgi 1964. Homonüümi mõiste. – Emakeele Seltsi aastaraamat, X, 45–56.

Viks, Ülle 1977. Sõnaliik kui niisugune. – Keel ja Kirjandus, 9, 521–524.
Viks, Ülle 1984. Sõnavormide homonüümia eesti keeles. – Keel ja Kirjandus, 2, 97–105.
Wilson, James 2013. Technology, pedagogy and promotion: How can we make the most of corpora and Data-Driven Learning (DDL) in language learning and teaching? – Higher Education Academy research report (July 2013). https://www.heacademy.ac.uk/system/files/corpus_technology_pedagogy_promotion2.pdf (07.09.2019).
Witten, Ian H., Eibe Frank, Mark A. Hall, Christopher J. Pal 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
ÕS 2018 = Õigekeelsussõnaraamat ÕS 2018. Tiiu Erelt, Tiina Leemets, Sirje Mäearu, Maire Raadik (Toim). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.

Võrguviited

ANW: Algemeen Nederlands Woordenboek. http://anw.inl.nl (20.06.2019).
Bing Microsoft Translation. https://www.bing.com/translator (20.06.2019).
Collins Dictionary. https://collinsdictionary.com (20.06.2019).
DANTE: A lexical database for English. http://www.webdante.com/ (20.06.2019).
EELex: leksikograafi töökeskkond. https://eelex.eki.ee (20.06.2019).
Eesti keele koondkorpus. https://www.cl.ut.ee/korpused/kasutajaliides/index.php?lang=et (20.06.2019).
Eesti keele süntaksianalüsaator. https://korpused.keeleressursid.ee/syntaks/ (20.06.2019).
Eesti murrete sõnaraamat. http://www.eki.ee/dict/ems/ (20.06.2019).
Ekilex: Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteem. https://doi.org/10.15155/3-00-0000-0000-0000-0823CL (29.01.2020).
Elexis: European lexicographic infrastructure. https://elex.is/ (20.06.2019).
enetCollect: Keeleõppe ja rahvahanke ühendamise Euroopa võrgustik. https://enetcollect.eurac.edu/ (20.06.2019).
ESTMORF: eesti keele morfoloogiline analüsaator. http://www.filosoft.ee/html_morf_et/morfoutinfo.html (20.06.2019).
estNLTK: Open source tools for Estonian natural language processing. https://estnltk.github.io/estnltk/1.2/index.html# (20.06.2019).
estNLTK 1.4.1. https://estnltk.github.io/estnltk/1.4.1/ (20.06.2019).
estNLTK 1.4.1. lausestaja (*sentence segmenter*). https://estnltk.github.io/estnltk/1.4.1/tutorials/text.html#tagging-clauses (20.06.2019).
estNLTK 1.4.1. osalausestaja (*clause segmenter*). https://estnltk.github.io/estnltk/1.4.1/api/clausesegmenter.html?highlight=sentence (20.06.2019).
estNLTK 1.4.1. ühestaja (*disambiguation*). https://estnltk.github.io/estnltk/1.4.1/tutorials/disambiguation.html (20.06.2019).
EstonianNC GDEX. https://www.sketchengine.eu/ (20.06.2019).
etSkELL: Sketch Engine for Estonian Language Learning. https://etskell.sketchengine.co.uk/run.cgi/ (20.06.2019).
GDEX Editor. https://gdexed.sketchengine.eu/ (20.06.2019).
GDEXi juhend. https://www.sketchengine.eu/documentation/manual-for-gdex/ (20.06.2019).

GDEXi konfiguratsioonifaili süntaks. https://www.sketchengine.eu/syntax-of-gdex-configuration-files/ (20.06.2019).
GiveMeExample. http://givemeexample.com/GiveMeExample (20.06.2019).
Google Forms. https://docs.google.com/forms/ (20.06.2019).
Google Translate. https://translate.google.com/ (20.06.2019).
Jaccardi sarnasuse indeks. https://en.wikipedia.org/wiki/Jaccard_index (20.06.2019).
Kallas, Jelena, Kristina Koppel 26.03.2018. Eesti keele ühendkorpus 2017. [Estonian National Corpus 2017.] Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-071e7l (20.06.2019).
Kallas, Jelena, Kristina Koppel 26.03.2018. Eesti keele A1–C1 õpikute korpus 2018. [Estonian Coursebook Corpus 2018.] Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-071e9l (20.06.2019).
Kallas, Jelena, Kristina Koppel 26.03.2018. Eesti keele õppekorpus 2018 (etSkELL). [The Estonian Corpus for Learners 2018 (etSkELL).] Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-07335l (20.06.2019).
KORP: korpuspäringusüsteem. https://korp.keeleressursid.ee/ (20.06.2019).
Laur, Sven 2018. EstNLTK teek Pythoni jaoks (ver.1.6.2b). Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-0015EL (24.01.2020).
Lause parameetrite analüsaator. http://www.eki.ee/keeletase/statistics/ (20.06.2019).
Lexonomy: A cloud-based, open-source platform for writing and publishing dictionaries. https://www.lexonomy.eu/ (20.06.2019).
Longman Dictionary of Contemporary English. http://ldoce.longmandictionariesonline.com/main/Home.html (20.06.2019).
MAB = Eesti Keele Instituudi eesti keele morfoloogiline andmebaas 2019. Ülle Viks, Indrek Hein, Katrin Tsepelina (Koost). Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee (14.02.2019).
Merriam-Webster Dictionary. https://www.merriam-webster.com/ (20.06.2019).
Morfoloogiliselt ühestatud korpus. https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/clutee-lehed/143-morfkorpus (20.06.2019).
OneClick Dictionary. https://www.sketchengine.eu/user-guide/lexicographers/ (06.02.2020).
Pybossa. https://pybossa.com/ (20.06.2019).
Reright: vene keele assotsiatsiooni- ja sünonüümisõnastik. http://www.reright.ru/ (20.06.2019).
ruSkELL: Sketch Engine for Russian Language Learning. https://ruskell.sketchengine.co.uk/run.cgi/skell (20.06.2019).
SkELL: Sketch Engine for Language Learning. http://skell.sketchengine.co.uk/run.cgi/skell (20.06.2019).
SkELLi keeleõppekeskkondade sari. https://www.sketchengine.eu/skell/?highlight=SkELL (20.06.2019).
Sketch Engine. https://www.sketchengine.eu/ (20.06.2019).
Sopomenke 1.0: Thesaurus of Modern Slovene. https://viri.cjvt.si/sopomenke/eng/ (20.06.2019).
Sõnaveeb 2019. [Language portal.] https://doi.org/10.15155/3-00-0000-0000-0000-0823EL (29.01.2020).
Tasakaalus korpus. https://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=et (20.06.2019).
Wordnik. https://www.wordnik.com/ (20.06.2019).
Õppeprogramm käänete õppimiseks. http://prog.keeleressursid.ee/opimekaandeid/ (20.06.2019).

LISA 1. Koondtabel GDEXi eesti keele mooduli eri versioonide parameetritest

Tabelis märgib märk + parameetri olemasolu, Ø parameetri puudumist, x keelamist ja * karistamist. Tugevate klassifikaatorite alla liigituvad parameetrid on esitatud paksus kirjas.

	versioon GDEX 1.2	**versioon GDEX 1.3**	**versioon GDEX 1.4**	**versioon etBasic-v1**	**versioon etIndependent-v1**	**versioon etProficient-v1**
lause pikkus (sõnedes)	5–20	5–20	4–20	3–14	3–18	4–23
lause optimaalne pikkus	10–12	10–12	6–12	4–7	4–12	6–14
keelatud märgid	<|][>/\^@	<|\]\[>/\\^@	<|][>/\}{^@•·*#=_~	<|][>/\}{^@•·*#=_~	<|][>/\}{^@•·*#=_~	<|][>/\}{^@•·*#=_~
karistada saavad märgid	A-Z0-9'.,!?)(;:-	A-Z0-9'.,!?)(;:“„”«»"…-]	A-Z0-9'.,!?)(;:“ „”«»"’×…§-	A-Z0-9'.,!?)(;:“ „”«»"’×…§-	A-Z0-9'.,!?)(;:“ „”«»"’×…§-	A-Z0-9'.,!?)(;:“ „”«»"’×…§-
lause alguses keelatud sõnaliigid	J	J	J, I, Y, G, Z	J, I, Y, Z	J, I, Y, Z	J, I, Y, Z
karistus sõnadele, mille sagedus korpuses on madalam kui 1000	Ø	Ø	+	+	+	+
karistada saavad tegusõnavormid	*-mata*, *-mast*, *-mas*, *-maks*, *-des*	*-mata*, *-mast*, *-mas*, *-maks*, *-des*	*-mata*, *-mast*, *-mas*, *-maks*, *-des*	*-tud*	*-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da*	*-nuks*, *-taks*, *-tuks*, *-tavat*, *-tuvat*, *-nuvat*

	versioon GDEX 1.2	**versioon GDEX 1.3**	**versioon GDEX 1.4**	**versioon etBasic-v1**	**versioon etIndependent-v1**	**versioon etProficient-v1**
keelatud tegusõnavormid	Ø	Ø	Ø	*-mast*, *-maks*, *-mata*, *-tama*, *-des*, *-nuks*, *-taks*, *-tuks*, *-tavat*, *-tuvat*, *-nuvat*, *-neg_gu*, *-tagu*, *-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da*	*-maks*, *-tama*, *-nuks*, *-taks*, *-tuks*, *-tavat*, *-tuvat*, *-nuvat*, *-neg_gu*, *-tagu*	Ø
omadussõna ülivõrde vorm	Ø	Ø	Ø	*	Ø	Ø
tegusõnavormi juurde kuuluv sõna (nt *plehku*)	Ø	Ø	Ø	*	Ø	Ø
hall nimekiri	446 sõna + *mina*, *sina*, *tema*, *see*, *too*, *siin*, *seal*	446 sõna + *mina*, *sina*, *tema*, *see*, *too*	474 sõna + *see*, *too*, *siin*, *siia*, *siit*, *seal*, *sinna*, *sealt*	451 sõna	451 sõna	451 sõna
must nimekiri	Ø	*siin*, *siia*, *siit*, *seal*, *sinna*, *sealt*, *siis*	Ø	267 sõna	267 sõna	267 sõna
lühendid	*	*	*	Ø	*	*
sisaldab tegusõna	Ø	+	+	+	+	+

	versioon GDEX 1.2	versioon GDEX 1.3	versioon GDEX 1.4	versioon etBasic-v1	versioon etIndependent-v1	versioon etProficient-v1
sisaldab nimisõna	Ø	Ø	Ø	+	+	Ø
karistus pikkadele sõnadele	Ø	Ø	Ø	>9 tähemärki	>11 tähemärki	Ø
märksõna kordus	Ø	Ø	x	x	x	x
pärisnime esinemine lauses	Ø	Ø	*	*	*	*
arvsõna + nimisõna *kroon*	Ø	Ø	*	*	*	*
komade arv lauses	Ø	Ø	* >1	* >1	* >1	* >1
sidesõnade arv lauses	Ø	Ø	* >1	* >1	* >1	* >1
asesõnade arv lauses	Ø	Ø	* >1	* >1	* >1	* >1
numbrite arv lauses	Ø	Ø	* >1	* >1	* >1	* >1
määrsõnade arv lauses	Ø	Ø	* >1	* >1	* >1	* >1

	versioon GDEX 1.2	**versioon GDEX 1.3**	**versioon GDEX 1.4**	**versioon etBasic-v1**	**versioon etIndependent-v1**	**versioon etProficient-v1**
pärisnimede arv lauses	Ø	Ø	* >1	* >1	* >1	* >1
tegusõnade arv lauses	Ø	Ø	* >2	* >2	* >2	* >2
nõrkade klassifikaatorite kaalud	Ø	Ø	+	+	+	+
eelistatud allkorpus	Ø	+	+	Ø	Ø	Ø
lause alguses keelatud sõnad	9 sõna	5 sõna	62 sõna	69 sõna	69 sõna	69 sõna
lause alguses keelatud sõnapaarid	Ø	Ø	79 sõnapaari	85 sõnapaari	85 sõnapaari	85 sõnapaari
lause alguses keelatud suurtähtedega kirjutatud sõnad	Ø	Ø	+	+	+	+

Sõnaliikide eristamise aluseks on eesti keele morfoloogilise analüsaatori ESTMORFi (vt Kaalep 1998) sõnaliikide eristus:

S substantiiv ehk nimisõna
V verb ehk tegusõna
A adjektiiv ehk omadussõna
G genitiivatribuut ehk käändumatu omadussõna
P pronoomen ehk asesõna
D adverb ehk määrsõna
K kaassõna
J konjunktsioon ehk sidesõna
N numeraal ehk arvsõna
I interjektsioon ehk hüüdsõna
Y lühend
X tegusõna juurde kuuluv sõna, millel ei ole eraldi sõnaliiki (nt *plehku*)
Z kirjavahemärk

Tegusõnavormide eristamise aluseks on eesti keele tekstianalüsaatori EstNLTK märgendus:

- *ma*-tegevusnime käändelised vormid (*-mast*, *-maks*, *-mata*, *-tama*)
- *des*-gerundiiv
- *tud*-kesksõna
- tingiva kõneviisi vormid (*-nuks*, *-taks*, *-tuks*)
- kaudse kõneviisi vormid (*tavat-*, *tuvat-*, *nuvat-*, *tuks*)
- käskiva kõneviisi vormid (*neg_gu*, *-tagu*)
- umbisikulise tegumoe vormid (*-takse*, *-dakse*, *-akse*, *-t*, *-d*, *-ta*, *-da*)

LISA 2. GDEX 1.2 konfiguratsioonifail

```
formula: >
     (50 * all(is_whole_sentence(), length > 5, length < 20, max([len(w) for w in words])
< 20, blacklist(words, illegal_chars), 1-match(lemmas[0], adverbs_bad_start),
min([word_frequency(w, 250000000) for w in words]) > 5)
  + 50 * optimal_interval(length, 10, 12)
  * greylist(words, rare_chars, 0.05) * 1.09
  * greylist(lemposs, anaphors, 0.1)
  * greylist(lemmas, bad_words, 0.25)
  * greylist(tags, abbreviation, 0.5)
  * (0.5 + 0.5 * (tags[0] != conjunction))
  * (1 - 0.5 * (tags[0]==verb) * match(featuress[0], verb_nonfinite_suffix))
  ) / 100

variables:
  illegal_chars: ([<|\]\[>/\\^@])
  rare_chars: ([A-Z0-9'.,!?)(;:-])
  conjunction: J
  abbreviation: Y
  anaphors: ^(mina-p|sina-p|tema-p|see-p|too-p|siin-d|seal-d)$
  adverbs_bad_start: ^(nagu|siin|siia|siit|seal|sinna|sealt|siis|seejärel)$
  verb: V
  verb_nonfinite_suffix: ^(mata|mast|mas|maks|des)$
  bad_words:
^(loll|jama|kurat|kapo|kanep|tegelt|sitt|pätt|in|mats|homo|pagan|joodik|idioot|nats|point|
kesik|aa|neeger|veits|jurama|narkomaan|jobu|siuke|õps|perse|tibi|riist|aint|tiss|pask|raisk|
raisk|värdjas|prostituut|pedofiil|mupo|gei|suli|porno|kaabakas|pepu|peldik|kaka|piss|tibla|
möla|lollakas|luuser|lits|tatt|pissima|vets|lesbi|ment|pede|inf|ee|pervert|narkar|okse|bemm|
penskar|kusema|kakama|rullnokk|tola|ruulima|junn|tglt|pubekas|peer|out|pedofiilia|muti|
tõbras|sittuma|kaak|totakas|pee|kuramus|debiilik|tutt|diiler|ila|kommar|pilu|raibe|kusi|
fašist|paganama|keppima|tra|moll|tips|kuradima|pohhui|pederast|pandav|kommu|jõmm|
vänt|beib|friik|nolk|tegelinski|totu|möga|oss|mölakas|lurjus|mõrd|fašistlik|kaif|noku|
argpüks|tatikas|mate|ajukääbik|liputama|vibraator|lollpea|sitane|memmekas|lõust|somm|
idikas|bordell|kärvama|kärakas|kemps|hoor|iiling|pederastia|narkots|vant|hui|hui|venku|
sitasti|nodi|soperdis|tõusik|puuks|äbarik|vitt|libu|hulkur|enivei|looder|peeretama|peda|
tolgus|lontrus|pohh|hängima|sunnik|jätis|türa|jura|laiskvorst|drive-in|kiim|matslik|sittama|
debiil|rops|mimm|kurivaim|sitapea|jota|nahhui|tšikk|veitsa|bitch|dire|linnavurle|russ|
prost|emps|tumba|burks|shoppama|pabul|keska|tohman|peldik|bemar|kretiin|liputaja|
tainapea|varganägu|litakas|värdjalik|haip|litsakas|molu|kaltsakas|vanka|lojus|kähkukas|
sovett|närakas|töllmokk|prükkar|häbe|hoorama|kagebiit|tipsi|kantpea|skinhead|hullar|
keelekas|lasteporno|ruts|nikkuma|chillima|šoppama|komnoor|lausloll|logard|tuhvlialune|
piff|mata|matslus|lipakas|pasandama|ropsima|memmepoeg|tattnina|puuksutama|skiso|
natsistlik|molkus|pohui|tillu|pissipott|kusik|jobukakk|niuke|litsimaja|närukael|pohuism|
haipima|pilusilm|juhmakas|puts|julk|vurle|pursui|blatnoi|komu|kuram|tuss|baaba|näss|
hoorus|kakane|sitakott|lita|sopakas|pohhuist|grupiseks|nuss|ponks|joomar|skinn|samakas|
trulla|frits|eniveis|sitahunnik|kurask|jokkis|huinjaa|sitahais|sakuska|tots|amf|morda|
nussima|sakumm|pissitama|sitavares|sitaratas|kili|jeestlane|gümna|platnoi|gigolo|mamps|
sekspomm|kürb|tšillima|kaki|pilukas|ladna|duubeldama|jobukari|tšau|krõhva|haisukott|
```

perseli|kehka|klassijuss|pohhuistlik|kabistama|plää|linnusitt|nahui|sitajunn|toksikomaan|
pordumaja|labrakas|narkomuul|joobar|masuurikas|nässakas|kräu|ciao|lirva|persevest|
lirva|koinima|sitaauk|tsillima|samagonn|töpa|sopajoodik|tšuhnaa|larhv|ajukääbus|
kiimlema|jobi|porduelu|sitahäda|tõprakari|kirvenägu|odratolgus|kakima|kiimakott|kräkk|
saksmann|bomž|kusev|plebei|pasahunnik|sakusment|pasakott|kabajantsik|kiimalus|milf|
pano|litapoeg|jobutama|pohhuilt|grupiks|topakas|hooramaja|türapea|küberseks|pepuvahe|
kusene|kusija|hoorapoeg|pizdets|hoholl|pasanteeria|bitš|kekats|kakanoku|panomees|
nadikael|pärakas|tolbajoob|kusetama|bljät|bizdets|pleiboi|pasahais|kagebist|praagamagu|
bljat|kiimlus|pedetsema|nihhuijaa|nehhui|häbedus|häbemepilu|jobama|kuselema|
kagebeelane|munapiiks|oolrait|beibe|jobutus|sigarijunn|sitavedaja|dolbajoob|jobisema|
pipravitt|türahiinlane|perseklile|tindinikkuja)$

LISA 3. GDEX 1.3 konfiguratsioonifail

```
formula: >
    (50 * all(is_whole_sentence(), length > 5, length < 20, max([len(w) for w in words])
< 20, count_matches(tags, verb) > 0, blacklist(words, illegal_chars), blacklist(lemmas,
bad_adverbs_any), not match(lemmas[0], bad_adverbs_first), min([word_frequency(w)
for w in words]) > 5)
  + 50 * optimal_interval(length, 10, 12)
  * greylist(words, rare_chars, 0.05) * 1.09
  * greylist(lemposs, anaphors, 0.1)
  * greylist(lemma_lcs, bad_words, 0.25)
  * greylist(tags, abbreviation, 0.5)
  * (0.5 + 0.5 * (tags[0] != conjunction))
  * max(0, 1 - 0.5 * len([t for t in tokens if t.tag==verb and match(t.features,
verb_nonfinite_suffix)]))
  ) / 100

frequency_reference_corpus: estonianRC
variables:
  illegal_chars: ([<|\]\[>/\\^@])
  rare_chars: ([A-Z0-9'.,!?)(;:“„„«»„…-])
  conjunction: J
  abbreviation: Y
  anaphors: ^(mina-p|sina-p|tema-p|see-p|too-p|siin-d|seal-d)$
  bad_adverbs_any: ^(siin|siia|siit|seal|sinna|sealt|siis)$
  bad_adverbs_first: ^(näiteks|kui|ühesõnaga|seejärel|nagu)$
  verb: V
  verb_nonfinite_suffix: ^(mata|mast|mas|maks|des)$
  bad_words:
^(loll|jama|kurat|kapo|kanep|tegelt|sitt|pätt|in|mats|homo|pagan|joodik|idioot|nats|point|
kesik|aa|neeger|veits|jurama|narkomaan|jobu|siuke|õps|perse|tibi|riist|aint|tiss|pask|raisk|
raisk|värdjas|prostituut|pedofiil|mupo|gei|suli|porno|kaabakas|pepu|peldik|kaka|piss|tibla|
möla|lollakas|luuser|lits|tatt|pissima|vets|lesbi|ment|pede|inf|ee|pervert|narkar|okse|bemm|
penskar|kusema|kakama|rullnokk|tola|ruulima|junn|tglt|pubekas|peer|out|pedofiilia|muti|
tõbras|sittuma|kaak|totakas|pee|kuramus|debiilik|tutt|diiler|ila|kommar|pilu|raibe|kusi|
fašist|paganama|keppima|tra|moll|tips|kuradima|pohhui|pederast|pandav|kommu|jõmm|
vänt|beib|friik|nolk|tegelinski|totu|möga|oss|mölakas|lurjus|mõrd|fašistlik|kaif|noku|
argpüks|tatikas|mate|ajukääbik|liputama|vibraator|lollpea|sitane|memmekas|lõust|somm|
idikas|bordell|kärvama|kärakas|kemps|hoor|iiling|pederastia|narkots|vant|hui|hui|venku|
sitasti|nodi|soperdis|tõusik|puuks|äbarik|vitt|libu|hulkur|enivei|looder|peeretama|peda|
tolgus|lontrus|pohh|hängima|sunnik|jätis|türa|jura|laiskvorst|drive-in|kiim|matslik|
sittama|debiil|rops|mimm|kurivaim|sitapea|jota|nahhui|tšikk|veitsa|bitch|dire|linnavurle|
russ|prost|emps|tumba|burks|shoppama|pabul|keska|tohman|peldik|bemar|kretiin|liputaja|
tainapea|varganägu|litakas|värdjalik|haip|litsakas|molu|kaltsakas|vanka|lojus|kähkukas|
sovett|närakas|töllmokk|prükkar|häbe|hoorama|kagebiit|tipsi|kantpea|skinhead|hullar|
keelekas|lasteporno|ruts|nikkuma|chillima|šoppama|komnoor|lausloll|logard|tuhvlialune|
piff|mata|matslus|lipakas|pasandama|ropsima|memmepoeg|tattnina|puuksutama|skiso|
natsistlik|molkus|pohui|tillu|pissipott|kusik|jobukakk|niuke|litsimaja|närukael|pohuism|
haipima|pilusilm|juhmakas|puts|julk|vurle|pursui|blatnoi|komu|kuram|tuss|baaba|näss|
```

hoorus|kakane|sitakott|lita|sopakas|pohhuist|grupiseks|nuss|ponks|joomar|skinn|samakas|
trulla|frits|eniveis|sitahunnik|kurask|jokkis|huinjaa|sitahais|sakuska|tots|amf|morda|
nussima|sakumm|pissitama|sitavares|sitaratas|kili|jeestlane|gümna|platnoi|gigolo|mamps|
sekspomm|kürb|tšillima|kaki|pilukas|ladna|duubeldama|jobukari|tšau|krõhva|haisukott|
perseli|kehka|klassijuss|pohhuistlik|kabistama|plää|linnusitt|nahui|sitajunn|toksikomaan|
pordumaja|labrakas|narkomuul|joobar|masuurikas|nässakas|kräu|ciao|lirva|persevest|
lirva|koinima|sitaauk|tsillima|samagonn|töpa|sopajoodik|tšuhnaa|larhv|ajukääbus|
kiimlema|jobi|porduelu|sitahäda|tõprakari|kirvenägu|odratolgus|kakima|kiimakott|kräkk|
saksmann|bomž|kusev|plebei|pasahunnik|sakusment|pasakott|kabajantsik|kiimalus|milf|
pano|litapoeg|jobutama|pohhuilt|grupiks|topakas|hooramaja|türapea|küberseks|pepuvahe|
kusene|kusija|hoorapoeg|pizdets|hoholl|pasanteeria|bitš|kekats|kakanoku|panomees|
nadikael|pärakas|tolbajoob|kusetama|bljät|bizdets|pleiboi|pasahais|kagebist|praagamagu|
bljat|kiimlus|pedetsema|nihhuijaa|nehhui|häbedus|häbemepilu|jobama|kuselema|
kagebeelane|munapiiks|oolrait|beibe|jobutus|sigarijunn|sitavedaja|dolbajoob|jobisema|
pipravitt|türahiinlane|perseklile|tindinikkuja)$

LISA 4. GDEX 1.4 konfiguratsioonifail

```
formula: >
  (50 * all(
  is_whole_sentence(),
  length > 4,
  length < 20,
  max([len(w) for w in words]) < 20,
  count_matches(tags, verb) > 0,
  blacklist(words, illegal_chars),
  blacklist(lemma_lcs, bad_words),
  not match(lemmas[0], bad_first_word),
  not match(space_separated(words), bad_first_two),
  not match(tags[0], bad_first_tag),
  match(words[0], lowercase),
  min([word_frequency(w) for w in words]) > 5,
  keyword_repetition(lemmas) == 1
  )
  + 9 * max(0, 1 - sum([0.5 for lemma in lemmas if lemma_frequency(lemma) < 1000]))
  + 9 * optimal_interval(length, 6, 12)
  + 5 * greylist(words, rare_chars, 0.05) * 1.09
  + 7 * greylist(lemposs, anaphors, 0.5)
  + 5 * greylist(lemma_lcs, taboo_words, 0.5)
  + 2 * greylist(tags, abbreviation, 0.5)
  + 2 * greylist(tags, proper_name, 0.1)
  + 2 * (1 - 0.4 * (count_matches(lemmas, 'kroon') and count_matches(tags, 'N')))
  + 2 * max(0, 1 - 0.5 * len([t for t in tokens if t.tag==verb and match(t.features,
verb_nonfinite_suffix)]))
  + 2 * min(1, sum([0.2 for score in lemma_collocation_scores(fromw=-5, tow=5,
minfreq=5, mincnt=3, maxitems=10, colfunc='PROD_MI.LOG_F')[max(0,kw_start-
5):kw_end+5] if score > 0]))
  + 5 * (1 - 0.2 * max(0, count_matches(words, comma) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, pronoun) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, verb) - 2))
  * (1 - 0.2 * max(0, count_matches(tags, conjunction) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, proper_name) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, number) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, adverb) - 1))
  ) / 100

frequency_reference_corpus: estonianRC_filosoft
variables:
  illegal_chars: ([<|\]\[>/\\} {^@•·*#=_~])
  rare_chars: ([A-Z0-9'.,!?)(;:“„„«»„’×…§-])
  lowercase: ([a-z])
  verb: V
  abbreviation: Y
  proper_name: H
  pronoun: P
```

adverb: D
number: N
conjunction: J
comma: (,)
anaphors: ^(see-p|too-p|siin-d|siia-d|siit-d|seal-d|sinna-d|sealt-d)$
verb_nonfinite_suffix: ^(mata|mast|mas|maks|des)$
bad_first_tag: (I|Y|G|Z|J)
bad_first_word:
^(*aga*|*ega*|*ehk*|*esiteks*|*hoolimata*|*ikka*|*iseasi*|*jah*|*ju*|*just*|*järelikult*|*järgnevalt*|*ka*|*lihtsalt*|*muidu*|*nad*|*nagu*|*nemad*|*niisiis*|*niisugune*|*nimelt*|*no*|*noh*|*nõnda*|*näiteks*|*ometi*|*pealegi*|*pigem*|*põhjuseks*|*samamoodi*|*samas*|*samuti*|*seal*|*sealjuures*|*see*|*see-eest*|*seega*|*seejuures*|*seejärel*|*seepeale*|*seepärast*|*seetõttu*|*seevastu*|*sellegipoolest*|*sellekohaselt*|*sellepärast*|*selletõttu*|*seniks*|*sestap*|*siin*|*siis*|*säärane*|*tagajärjeks*|*teiseks*|*teisisõnu*|*tere*|*too*|*vastupidi*|*või*|*võrdluseks*|*ühesõnaga*|*ülejäänud*)$
bad_first_two: ^((Ainult et)|(Ainult nii)|(Ehk siis)|(Ehk teisisõnu)|(Eriti kui)|(Eriti juhul)|(Eriti just)|(Eriti siis)|(Eriti veel)|(Isegi siis)|(Just need)|(Just nii)|(Just niikaua)|(Just nimelt)|(Just see)|(Just seepärast)|(Just seetõttu)|(Just sellega)|(Just sellepärast)|(Just selletõttu)|(Kõige selle)|(Küll aga)|(Lisaks sellele)|(Muidugi eeldusel)|(Muidugi ka)|(Nii et)|(Nüüd aga)|(Peale seda)|(Peale selle)|(Sama asi)|(Sama kehtib)|(Samal aastal)|(Samal ajal)|(Samal hommikul)|(Samal põhjusel)|(Samal päeval)|(Samal viisil)|(Samal õhtul)|(Samal ööl)|(Samal öösel)|(Seda enam)|(Seda eriti)|(Seda kõike)|(Sellisel juhul)|(Sellisel moel)|(Sellisel puhul)|(Teisel juhul)|(Teisel korral)|(Teiselt poolt)|(Teisest küljest)|(Teisiti öeldes)|(Teiste sõnadega)|(Välja arvatud)|(Vastasel juhul)|(Vastasel korral)|(Veel enam)|(Viimasel juhul)|(See omakorda)|(Selleks ajaks)|(Selleks on)|(Selleks peab)|(Selleks pead)|(Selleks peaks)|(Selleks peame)|(Sellele vaatamata)|(Selles mõttes)|(Ses mõttes)|(Selles osas)|(Selles valguses)|(Sellest hoolimata)|(Sellest johtuvalt)|(Sellest lähtudes)|(Sellest lähtuvalt)|(Sellest omakorda)|(Sellest tulenevalt)|(See tähendab)|(Vaatamata sellele)|(Veelgi enam)|(Ühelt poolt))
bad_words:
^(aare|mugu|leid|peidukoht|kellegil|kellegile|kah|zhest|tatt|öigus|öigustus|öigustama|läits|loll|lollakas|lollpea|lollike|lausloll|sitt|shantazheerima|sittuma|sittama|sitane|sitasti|sitahais|sitapea|sitaauk|sitajunn|sitavares|sitaratas|sitakott|sitamaitse|sitavedaja|sitahäda|sitahunnik|linnusitt|junn|perse|perseli|perseklile|persevest|perseauk|pepu|pee|pask|pasandama|pasane|pasahais|pasakott|pasahunnik|pasapea|pasanteeria|peer|peeretama|peerukott|puuks|puuksutama|puuksutamine|kakapuuks|vitt|pipravitt|vitupea|tutt|tuss|puts|munn|munnikari|türa|tra|türapea|riist|vänt|kürb|prostituut|prost|hoor|hoorama|hooramaja|hoorapoeg|hoorus|pordumaja|bordell|porduelu|lits|litsakas|litsimaja|litsinahk|litsitama|lita|litakas|lipakas|litapoeg|libu|lirva|porno|lasteporno|pornograafia|peldik|kaka|kakama|kakima|kakane|kakanoku|kaki|kakine|piss|pissima|pissine|pissitama|pissipott|kusi|kusema|kusev|kuselema|kusene|kusija|kusik|kusetama|kiim|kiimlema|kiimlus|käsikiimlus|kiimalus|kiimakott|keppima|keppimine|keppija|nikkuma|nikkumine|nikkuja|koinima|grupikas|grupiseks|suuseks|anaalseks|sekspomm|küberseks|nuss|nussima|tindinikkuja|kähkukas|keelekas|pedofiilia|pedofiil|pede|pedekas|pedene|pedendus|pederast|pedetsema|pederastia|pederastiapropaganda|autopede|pedepropaganda|pervert|homo|gei|lesbi|lesbiline|narkar|narkomaan|tibi|tips|beib|beibe|tšikk|piff|tots|tipsi|idioot|lausidioot|täisidioot|idioodikari|idikas|pätt|mats|mupo|kapo|suli|pagan|joodik|sopajoodik|joobar|jama|kurat|tegelt|nats|point|värdjas|kaabakas|mõlakas|debiilik|lurjus|tolgus|lontrus|topakas|tola|jobu|tohman|kesik|tibla|ajukääbik|penskar|rullnokk|memmekas|

platnoi|gigolo|mamps|argpüks|tatikas|soperdis|tõusik|äbarik|hulkur|kommu|kommar|
komnoor|logard|tuhvlialune|baaba|näss|jõmm|oss|mõrd|friik|nolk|närakas|töllmokk|
prükkar|kagebiit|tegelinski|totu|pubekas|muti|tõbras|luuser|kaak|totakas|vant|vanka|
venku|sovett|pilukas|russ|somm|frits|kili|jeestlane|hoholl|vurle|pursui|blatnoi|bomž|
plebei|komu|neeger|diiler|ment|toksikomaan|ajukääbus|jobi|tõprakari|kirvenägu|kretiin|
liputaja|tainapea|varganägu|odratolgus|narkomuul|memmepoeg|närukael|tattnina|joomar|
skinn|skiso|masuurikas|nässakas|krõhva|haisukott|pleiboi|kagebist|praagamagu|tšuhnaa|
kabajantsik|milf|jurama|siuke|veits|möla|vets|okse|bemm|ruulima|tglt|kuramus|aint|tiss|
ila|pilu|raibe|raisk|fašist|fašistlik|fashist|pilusilm|juhmakas|kekats|panomees|nadikael|
dolbajoob|paganama|kuradima|kuram|pandav|dildo|möga|kaif|noku|noks|mate|liputama|
vibraator|lõust|kärvama|kärakas|kemps|fiiling|narkots|nodi|enivei|looder|peda|hängima|
sunnik|jätis|jura|laiskvorst|matslik|debiil|rops|mimm|kurivaim|jota|veitsa|bitch|bitš|
linnavurle|emps|tumba|burks|pabul|peldik|peller|bemar|värdjalik|kagebeelane|haip|molu|
kaltsakas|lojus|kantpea|skinhead|hullar|ruts|chillima|šoppama|shoppama|tšillima|tsillima|
julk|mata|matslus|ropsima|natsistlik|molkus|tillu|jobukakk|niuke|moll|haipima|sopakas|
ponks|samakas|trulla|eniveis|kurask|jokkis|sakuska|sakumm|samagonn|amf|morda|ladna|
duubeldama|jobukari|tšau|ciao|kabistama|labrakas|kräu|larhv|kräkk|saksmann|sakusment|
pano|jobutama|pepuvahe|pärakas|tolbajoob|hui|nahhui|nahui|pohhui|pohui|pohh|
pohuism|pohhuist|pohhuistlik|pohhuilt|bljat|nihhuijaa|nehhui|huinjaa|bljät|bljääd|bljäd|
bljää|bizdets|pizdets|plää|häbe|häbedus|häbemepilu|häbememokk|häbemekarv|
häbemekink|jobama|munapiiks|oolrait|jobutus|sigarijunn|jobisema|türahiinlane|töpa|õps|
kehka|keska|dire|gümna|klassijuss|aganoh|kanep|türnüffel|sis|drive-in|aa|in|out|inf|
ee|no|noh|ete|bla|blabla|blablabla|blablablabla|pääle|pääl|päält|päälegi|awesome|crash|
story)$

LISA 5. etBasic-v1 konfiguratsioonifail

```
formula: >
  (50 * all(
  is_whole_sentence(),
  length > 3,
  length < 14,
  max([len(w) for w in words]) < 20,
  count_matches(tags, verb) > 0,
  count_matches(tags, substantive) > 0,
  not [t for t in tokens if t.tag==verb and match(t.features,
'^maks|mast|tama|mata|des|neg_gu|nuks|nuksin|taks|tuks|tagu|tavat|vat|neg_vat|gem|
neg_gem|tuvat|nuvat|dakse|takse|akse|t|d|ta|da$')],
  blacklist(words, illegal_chars),
  blacklist(lemma_lcs, bad_words),
  not match(lemmas[0], bad_first_word),
  not match(space_separated(words), bad_first_two),
  not match(tags[0], bad_first_tag),
  match(words[0], lowercase),
  min([word_frequency(w) for w in words]) > 25,
  keyword_repetition(lemmas) == 1
  )
  + 8 * max(0, 1 - sum([0.5 for lemma in lemmas if lemma_frequency(lemma) < 5000]))
  + 8 * optimal_interval(length, 4, 7)
  + 5 * greylist(words, longer_than_nine, 0.5)
  + 7 * greylist(words, rare_chars, 0.05) * 1.09
  + 5 * greylist(lemposs, anaphors, 0.5)
  + 2 * greylist(lemma_lcs, taboo_words, 0.5)
  + 2 * greylist(tags, abbreviation, 0.5)
  + 3 * greylist(tags, proper_name, 0.5)
  + 2 * (1 - 0.4 * (count_matches(lemmas, 'kroon') and count_matches(tags, 'N')))
  + 4 * max(0, 1 - 0.7 * len([t for t in tokens if t.tag==verb and match(t.features,
verb_nonfinite_suffix)]))
  + 4 * (1 - 0.2 * max(0, count_matches(words, comma) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, pronoun) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, verb) - 2))
  * (1 - 0.2 * max(0, count_matches(tags, conjunction) - 1))
  * (1 - 0.5 * max(0, count_matches(tags, proper_name) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, number) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, adverb) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, abbreviation) - 0))
  * (1 - 0.2 * max(0, count_matches(tags, superlative) - 0))
  * (1 - 0.2 * max(0, count_matches(tags, plehku) - 0))
  ) / 100

variables:
  illegal_chars: ([<|\]\[>/\\}{^@•·*#=_~])
  rare_chars: ([A-Z0-9'.,!?)(;:“„„«»„’×…§-])
  lowercase: ([a-z])
```

longer_than_nine: (.....)
verb: V
substantive: S
abbreviation: Y
proper_name: H
pronoun: P
adverb: D
number: N, O
conjunction: J
superlative: U
plehku: X
comma: (,)
anaphors: ^(see-p|too-p|siin-d|siia-d|siit-d|seal-d|sinna-d|sealt-d)$
verb_nonfinite_suffix: ^(mas|tama|tud|dud|taks|tuks)$
bad_first_tag: (I|Y|Z|J|C|U)
bad_first_word:
^(nagu|samamoodi|ühesõnaga|niisugune|niisiis|nad|nemad|see|ka|samuti|ja|samas|aga|
näiteks|kuid|nüüd|seega|et|seetõttu|sellepärast|siis|just|nimelt|ehk|või|ometi|eks|ent|
teiseks|seepärast|ning|esiteks|seal|pigem|siin|ülejäänud|sest|seejärel|ega|jah|muidu|tere|
seejuures|hoolimata|lihtsalt|ikka|järelikult|nõnda|põhjuseks|seevastu|järgnevalt|
see-eest|too|säärane|vastupidi|siit|sellegipoolest|sestap|seepeale|ju|teisisõnu|
võrdluseks|iseasi|sealjuures|tagajärjeks|sellekohaselt|selletõttu|pealegi|seniks)$
bad_first_two: ^((Küll aga)|(Umbes nagu)|(Nii et)|(Nüüd aga)|(Kuid juhtumisi)|
(Samal ajal)|(Ühelt poolt)|(Ainult et)|(Vaatamata sellele)|(Ainult nii)|(Ehk siis)|
(Ehk teisisõnu)|(Eriti kui)|(Eriti juhul)|(Eriti just)|(Eriti siis)|(Eriti veel)|(Isegi siis)|
(Just need)|(Just nii)|(Just see)|(Just seetõttu)|(Muidugi eeldusel)|(Muidugi ka)|
(Nii nagu)|(Peale seda)|(Peale selle)|(Sama asi)|(Sama kehtib)|(Seda enam)|(Seda eriti)|
(Seda kõike)|(Sellisel juhul)|(Sellisel moel)|(Sellisel puhul)|(Teisel juhul)|
(Teisel korral)|(Teiselt poolt)|(Teisest küljest)|(Teisiti öeldes)|(Teiste sõnadega)|
(Välja arvatud)|(Vastasel juhul)|(Vastasel korral)|(Lisaks sellele)|(Veel enam)|
(Veelgi enam)|(Viimasel juhul)|(See omakorda)|(Selleks ajaks)|(Selleks on)|
(Selleks peab)|(Selleks pead)|(Selleks peaks)|(Selleks peame)|(Selleks peavad)|
(Sellele vaatamata)|(Selles mõttes)|(Ses mõttes)|(Selles osas)|(Selles valguses)|
(Sellest hoolimata)|(Sellest johtuvalt)|(Sellest lähtudes)|(Sellest lähtuvalt)|
(Sellest omakorda)|(Sellest tulenevalt)|(See tähendab)|(Samal põhjusel)|(Samal viisil))
bad_words:
^(sitt|sittuma|sittagi|sittama|sitane|sitasti|sitahais|sitapea|sitaauk|sitajunn|sitavares|
sitaratas|sitakott|sitamaitse|sitavedaja|sitahäda|sitahunnik|sitamaja|linnusitt|junn|perse|
perseli|perseklile|persevest|perseauk|pask|paskagi|pasahais|pasakott|pasahunnik|pasapea|
pasapeeter|pasapüks|pasarahe|pasanteeria|vitt|pipravitt|vitupea|tutt|tuss|puts|munn|
munnikari|türa|tra|türapea|kürb|keppima|keppimine|keppija|nikkuma|nikkumine|nikkuja|
tindinikkuja|nikk|koinima|larhv|lipakas|litapoeg|morda|mordu|nuss|nussima|pano|
panomees|praagamagu|hoorama|hoorapoeg|hoorajääger|libu|lirva|lita|lits|kiimakott|
pede|pedekas|pedene|pedendus|pederast|pedetsema|piider|daun|taun|raibe|raip|raisk|
pasandama|pasane|prost|onama|pervo|pirtsperse|sitaks|sitanikerdis|sitaseen|sitavanka|
sitavesi|sitavikat|kirbusitt|litsimaja|litsinahk|lohh|bitch|debiilik|debiil|ajudoonor|
ajugeenius|ajuinvaliid|ambaal|bimbo|imbetsill|jobi|jorss|kärbsesitt|ohmoon|ajukääbik|
ajukäabus|hoor|hoorus|hui|nahhui|nahui|pohhui|pohui|pohh|pohuism|pohhuist|
pohhuistlik|pohhuilt|bljat|nihhuijaa|nehhui|huinjaa|bljät|bljääd|bljäd|bljää|bizdets|pizdets|

plää|hooramaja|idikas|idioot|lausidioot|täisidioot|idioodikari|jota|karuperse|kanaaju|
kanapea|ahvinägu|kusik|homo|frits|kili|rüssä|rüssa|tibla|pajuvenelane|uusvenelane|
moskoviit|murjan|murjam|negru|niger|tšuhnaa|bosniakk|eurovenelane|hoholl|jeestlane|
neeger|vant|vanka|venku|sovett|pilukas|russ|somm|pilusilm|kaltsupea|alfons|baaba|
haisukott|hampelmann|hilpharakas|ilamokk|töllmokk|tõpranägu|inimjätis|inimrämps|
jobukakk|kabajantsik|kaebupunn|kitupunn|kekutis|kirjasolkija|koeraraibe|kopikakoi|
kretiin|krõhva|krõnks|krõõp|krõõt|kõlkapüks|lakard|lakkekrants|lakekrants|lakkekauss|
lakekauss|lapsevänts|lojus|lollikari|lontrus|luhva|lupard|lõhverdis|lödipüks|vedelpüks|
memmepoeg|memmetütar|mammapoja|mammatütar|mats|matsikeel|matslik|matslus|
molkus|molu|mudakoon|munajoodik|mõrd|mölakas|nolk|paksmagu|penskar|plebei|prole|
pulgajunkur|punaparun|rahvarämps|rasvamagu|rasvarull|ristikoer|seapeet|tallalakkuja|
talumats|tattnokk|tolgus|tšinovnik|tuulenuusutaja|tõusik|täikrae|täinahk|töll|töllakas|
untsantsakas|untsakas|vurle|värdjas|väärakas|fuck|fakk|fck)$
  taboo_words:
^(kaabakas|kaak|aare|adrekas|aftekas|mugu|leid|peidukoht|kellegil|kellegile|kah|zhest|
tatt|öigus|öigustus|öigustama|läits|alkašš|atu|autopede|elueit|ess|kirvenägu|kännuämblik|
loll|lollakas|lollpea|lollike|lausloll|shantazheerima|pepu|pee|peer|peeretama|peerukott|
puuks|puuksutama|puuksutamine|kakapuuks|riist|vänt|prostituut|pordumaja|bordell|
porduelu|litsakas|litsitama|litakas|porno|lasteporno|pornograafia|peldik|kaka|kakama|
kakima|kakane|kakanoku|kaki|kakine|piss|pissima|pissine|pissitama|pissipott|kusi|
kusema|kusev|kuselema|kusene|kusija|kusetama|kiim|kiimlema|kiimlus|käsikiimlus|
kiimalus|grupikas|grupiseks|suuseks|anaalseks|sekspomm|küberseks|kähkukas|keelekas|
pedofiilia|pedofiil|pederastia|pederastiapropaganda|pedepropaganda|pervert|gei|lesbi|
lesbiline|narkar|narkomaan|tibi|tibin|tips|beib|beibe|tšikk|piff|tots|tipsi|tiinekas|tiiner|pätt|
mupo|kapo|suli|pagan|joodik|sopajoodik|joobar|jama|kurat|tegelt|nats|point|lurjus|
topakas|tola|jobu|tohman|kesik|rullnokk|memmekas|platnoi|gigolo|mamps|argpüks|
tatikas|soperdis|äbarik|hulkur|kommu|komnoor|kommar|logard|tuhvlialune|näss|jõmm|
oss|friik|närakas|prükkar|kagebiit|tegelinski|totu|pubekas|muti|tõbras|luuser|totakas|
pursui|blatnoi|bomž|komu|diiler|ment|toksikomaan|tõprakari|liputaja|tainapea|taignapea|
varganägu|odratolgus|narkomuul|närukael|tattnina|joomar|skinn|skiso|masuurikas|
nässakas|pleiboi|kagebist|kabajantsik|milf|jurama|siuke|veits|möla|vets|bemm|benss|
bensukas|ruulima|tglt|kuramus|aint|tiss|ila|pilu|fašist|fašistlik|fashist|juhmakas|kekats|
nadikael|dolbajoob|paganama|kuradima|kuram|pandav|dildo|möga|kaif|noku|noks|
liputama|lõust|kärvama|kärakas|kemps|fiiling|narkots|nodi|enivei|looder|peda|hängima|
sunnik|jätis|jura|laiskvorst|rops|mimm|kurivaim|veitsa|bitš|linnavurle|emps|tumba|burks|
pabul|peldik|peller|bemar|värdjalik|kagebeelane|haip|kaltsakas|lojus|kantpea|skinhead|
hullar|ruts|chillima|šoppama|shoppama|tšillima|tsillima|julk|mata|ropsima|natsistlik|tillu|
niuke|moll|haipima|sopakas|ponks|samakas|trulla|eniveis|kurask|jokkis|sakuska|sakumm|
samagonn|amf|ladna|duubeldama|jobukari|tšau|ciao|kabistama|labrakas|kräu|kräkk|
saksmann|sakusment|jobutama|pepuvahe|pärakas|tolbajoob|jobama|munapiiks|oolrait|
jobutus|sigarijunn|jobisema|türahiinlane|töpa|õps|kehka|keska|dire|gümna|klassijuss|
aganoh|türnüffel|sis|drive-in|aa|in|out|inf|ee|no|noh|ete|bla|blabla|blablabla|blablablabla|
pääle|pääl|päält|päälegi|awesome|crash|story|blackout|bläkk|crack|davai|eur|leheneeger|
nannipunn|nilbik|pehmo|transa|pigi|liivaneeger|kirbukott|blufivend|lambakari|lambapea|
loodrinahk)$

LISA 6. etIndependent-v1 konfiguratsioonifail

```
formula: >
  (50 * all(
  is_whole_sentence(),
  length > 3,
  length < 18,
  max([len(w) for w in words]) < 20,
  count_matches(tags, verb) > 0,
  count_matches(tags, substantive) > 0,
  not [t for t in tokens if t.tag==verb and match(t.features,
'^maks|tama|neg_gu|tagu|nuks|nuksin|taks|tuks|tavat|vat|neg_vat|gem|neg_gem|nuvat|
tuvat$')],
  count_matches(tags, substantive) > 0,
  blacklist(words, illegal_chars),
  blacklist(lemma_lcs, bad_words),
  not match(lemmas[0], bad_first_word),
  not match(space_separated(words), bad_first_two),
  not match(tags[0], bad_first_tag),
  match(words[0], lowercase),
  min([word_frequency(w) for w in words]) > 25,
  keyword_repetition(lemmas) == 1
  )
  + 50 * max(0, 1 - sum([0.5 for lemma in lemmas if lemma_frequency(lemma)
< 5000]))
  * optimal_interval(length, 4, 12)
  * greylist(words, longer_than_eleven, 0.1)
  * greylist(words, rare_chars, 0.05) * 1.09
  * greylist(lemposs, anaphors, 0.5)
  * greylist(lemma_lcs, taboo_words, 0.5)
  * greylist(tags, abbreviation, 0.5)
  * greylist(tags, proper_name, 0.1)
  * (1 - 0.4 * (count_matches(lemmas, 'kroon') and count_matches(tags, 'N')))
  * max(0, 1 - 0.5 * len([t for t in tokens if t.tag==verb and match(t.features,
verb_nonfinite_suffix)]))
  * (1 - 0.2 * max(0, count_matches(words, comma) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, pronoun) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, verb) - 2))
  * (1 - 0.2 * max(0, count_matches(tags, conjunction) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, proper_name) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, number) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, adverb) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, abbreviation) - 0))
  * (1 - 0.2 * max(0, count_matches(tags, plehku) - 0))
  ) / 100

frequency_reference_corpus: estonianRC_filosoft
variables:
  illegal_chars: ([<|\]\[>/\\} {^@•·*#=_~])
```

rare_chars: ([A-Z0-9'.,!?)(;:“„„«»„’×…§-])
lowercase: ([a-z])
longer_than_eleven: (.....)
verb: V
substantive: S
abbreviation: Y
proper_name: H
pronoun: P
adverb: D
number: N, O
conjunction: J
superlative: U
plehku: X
comma: (,)
anaphors: ^(see-p|too-p|siin-d|siia-d|siit-d|seal-d|sinna-d|sealt-d)$
verb_nonfinite_suffix:
^(mata|mast|mas|des|tama|nuks|taks|takse|dakse|akse|t|d|ta|da|tuks)$
bad_first_tag: (I|Y|Z|J|C|U)
bad_first_word:
^(nagu|samamoodi|ühesõnaga|niisugune|niisiis|nad|nemad|see|ka|samuti|ja|samas|aga|
näiteks|kuid|nüüd|seega|et|seetõttu|sellepärast|siis|just|nimelt|ehk|või|ometi|eks|ent|
teiseks|seepärast|ning|esiteks|seal|pigem|siin|ülejäänud|sest|seejärel|ega|jah|muidu|tere|
seejuures|hoolimata|lihtsalt|ikka|järelikult|nõnda|põhjuseks|seevastu|järgnevalt|see-eest|
too|säärane|vastupidi|siit|sellegipoolest|sestap|seepeale|ju|teisisõnu|võrdluseks|iseasi|
sealjuures|tagajärjeks|sellekohaselt|selletõttu|pealegi|seniks)$
bad_first_two: ^((Küll aga)|(Umbes nagu)|(Nii et)|(Nüüd aga)|(Kuid juhtumisi)|
(Samal ajal)|(Ühelt poolt)|(Ainult et)|(Vaatamata sellele)|(Ainult nii)|(Ehk siis)|
(Ehk teisisõnu)|(Eriti kui)|(Eriti juhul)|(Eriti just)|(Eriti siis)|(Eriti veel)|(Isegi siis)|
(Just need)|(Just nii)|(Just see)|(Just seetõttu)|(Muidugi eeldusel)|(Muidugi ka)|
(Nii nagu)|(Peale seda)|(Peale selle)|(Sama asi)|(Sama kehtib)|(Seda enam)|(Seda eriti)|
(Seda kõike)|(Sellisel juhul)|(Sellisel moel)|(Sellisel puhul)|(Teisel juhul)|
(Teisel korral)|(Teiselt poolt)|(Teisest küljest)|(Teisiti öeldes)|(Teiste sõnadega)|
(Välja arvatud)|(Vastasel juhul)|(Vastasel korral)|(Lisaks sellele)|(Veel enam)|
(Veelgi enam)|(Viimasel juhul)|(See omakorda)|(Selleks ajaks)|(Selleks on)|
(Selleks peab)|(Selleks pead)|(Selleks peaks)|(Selleks peame)|(Selleks peavad)|
(Sellele vaatamata)|(Selles mõttes)|(Ses mõttes)|(Selles osas)|(Selles valguses)|
(Sellest hoolimata)|(Sellest johtuvalt)|(Sellest lähtudes)|(Sellest lähtuvalt)|
(Sellest omakorda)|(Sellest tulenevalt)|(See tähendab)|(Samal põhjusel)|(Samal viisil))
bad_words:
^(sitt|sittuma|sittagi|sittama|sitane|sitasti|sitahais|sitapea|sitaauk|sitajunn|sitavares|
sitaratas|sitakott|sitamaitse|sitavedaja|sitahäda|sitahunnik|sitamaja|linnusitt|junn|perse|
perseli|perseklile|persevest|perseauk|pask|paskagi|pasahais|pasakott|pasahunnik|pasapea|
pasapeeter|pasapüks|pasarahe|pasanteeria|vitt|pipravitt|vitupea|tutt|tuss|puts|munn|
munnikari|türa|tra|türapea|kürb|keppima|keppimine|keppija|nikkuma|nikkumine|nikkuja|
tindinikkuja|nikk|koinima|larhv|lipakas|litapoeg|morda|mordu|nuss|nussima|pano|
panomees|praagamagu|hoorama|hoorapoeg|hoorajääger|libu|lirva|lita|lits|kiimakott|pede|
pedekas|pedene|pedendus|pederast|pedetsema|piider|daun|taun|raibe|raip|raisk|
pasandama|pasane|prost|onama|pervo|pirtsperse|sitaks|sitanikerdis|sitaseen|sitavanka|
sitavesi|sitavikat|kirbusitt|litsimaja|litsinahk|lohh|bitch|debiilik|debiil|ajudoonor|

ajugeenius|ajuinvaliid|ambaal|bimbo|imbetsill|jobi|jorss|kärbsesitt|ohmoon|ajukääbik|
ajukäabus|hoor|hoorus|hui|nahhui|nahui|pohhui|pohui|pohh|pohuism|pohhuist|
pohhuistlik|pohhuilt|bljat|nihhuijaa|nehhui|huinjaa|bljät|bljääd|bljäd|bljää|bizdets|pizdets|
plää|hooramaja|idikas|idioot|lausidioot|täisidioot|idioodikari|jota|karuperse|kanaaju|
kanapea|ahvinägu|kusik|homo|frits|kili|rüssä|rüssa|tibla|pajuvenelane|uusvenelane|
moskoviit|murjan|murjam|negru|niger|tšuhnaa|bosniakk|eurovenelane|hoholl|jeestlane|
neeger|vant|vanka|venku|sovett|pilukas|russ|somm|pilusilm|kaltsupea|alfons|baaba|
haisukott|hampelmann|hilpharakas|ilamokk|töllmokk|tõpranägu|inimjätis|inimrämps|
jobukakk|kabajantsik|kaebupunn|kitupunn|kekutis|kirjasolkija|koeraraibe|kopikakoi|
kretiin|krõhva|krõnks|krõõp|krõõt|kõlkapüks|lakard|lakkekrants|lakekrants|lakkekauss|
lakekauss|lapsevänts|lojus|lollikari|lontrus|luhva|lupard|lõhverdis|lödipüks|vedelpüks|
memmepoeg|memmetütar|mammapoja|mammatütar|mats|matsikeel|matslik|matslus|
molkus|molu|mudakoon|munajoodik|mõrd|mölakas|nolk|paksmagu|penskar|plebei|prole|
pulgajunkur|punaparun|rahvarämps|rasvamagu|rasvarull|ristikoer|seapeet|tallalakkuja|
talumats|tattnokk|tolgus|tšinovnik|tuulenuusutaja|tõusik|täikrae|täinahk|töll|töllakas|
untsantsakas|untsakas|vurle|värdjas|väärakas|fuck|fakk|fck)$
  taboo_words:
^(kaabakas|kaak|aare|adrekas|aftekas|mugu|leid|peidukoht|kellegil|kellegile|kah|zhest|
tatt|öigus|öigustus|öigustama|läits|alkašš|atu|autopede|elueit|ess|kirvenägu|kännuämblik|
loll|lollakas|lollpea|lollike|lausloll|shantazheerima|pepu|pee|peer|peeretama|peerukott|
puuks|puuksutama|puuksutamine|kakapuuks|riist|vänt|prostituut|pordumaja|bordell|
porduelu|litsakas|litsitama|litakas|porno|lasteporno|pornograafia|peldik|kaka|kakama|
kakima|kakane|kakanoku|kaki|kakine|piss|pissima|pissine|pissitama|pissipott|kusi|
kusema|kusev|kuselema|kusene|kusija|kusetama|kiim|kiimlema|kiimlus|käsikiimlus|
kiimalus|grupikas|grupiseks|suuseks|anaalseks|sekspomm|küberseks|kähkukas|keelekas|
pedofiilia|pedofiil|pederastia|pederastiapropaganda|pedepropaganda|pervert|gei|lesbi|
lesbiline|narkar|narkomaan|tibi|tibin|tips|beib|beibe|tšikk|piff|tots|tipsi|tiinekas|tiiner|pätt|
mupo|kapo|suli|pagan|joodik|sopajoodik|joobar|jama|kurat|tegelt|nats|point|lurjus|
topakas|tola|jobu|tohman|kesik|rullnokk|memmekas|platnoi|gigolo|mamps|argpüks|
tatikas|soperdis|äbarik|hulkur|kommu|komnoor|kommar|logard|tuhvlialune|näss|jõmm|
oss|friik|närakas|prükkar|kagebiit|tegelinski|totu|pubekas|muti|tõbras|luuser|totakas|
pursui|blatnoi|bomž|komu|diiler|ment|toksikomaan|tõprakari|liputaja|tainapea|taignapea|
varganägu|odratolgus|narkomuul|närukael|tattnina|joomar|skinn|skiso|masuurikas|
nässakas|pleiboi|kagebist|kabajantsik|milf|jurama|siuke|veits|möla|vets|bemm|benss|
bensukas|ruulima|tglt|kuramus|aint|tiss|ila|pilu|fašist|fašistlik|fashist|juhmakas|kekats|
nadikael|dolbajoob|paganama|kuradima|kuram|pandav|dildo|möga|kaif|noku|noks|
liputama|lõust|kärvama|kärakas|kemps|fiiling|narkots|nodi|enivei|looder|peda|hängima|
sunnik|jätis|jura|laiskvorst|rops|mimm|kurivaim|veitsa|bitš|linnavurle|emps|tumba|burks|
pabul|peldik|peller|bemar|värdjalik|kagebeelane|haip|kaltsakas|lojus|kantpea|skinhead|
hullar|ruts|chillima|šoppama|shoppama|tšillima|tsillima|julk|mata|ropsima|natsistlik|tillu|
niuke|moll|haipima|sopakas|ponks|samakas|trulla|eniveis|kurask|jokkis|sakuska|sakumm|
samagonn|amf|ladna|duubeldama|jobukari|tšau|ciao|kabistama|labrakas|kräu|kräkk|
saksmann|sakusment|jobutama|pepuvahe|pärakas|tolbajoob|jobama|munapiiks|oolrait|
jobutus|sigarijunn|jobisema|türahiinlane|töpa|õps|kehka|keska|dire|gümna|klassijuss|
aganoh|türnüffel|sis|drive-in|aa|in|out|inf|ee|no|noh|ete|bla|blabla|blablabla|blablablabla|
pääle|pääl|päält|päälegi|awesome|crash|story|blackout|bläkk|crack|davai|eur|leheneeger|
nannipunn|nilbik|pehmo|transa|pigi|liivaneeger|kirbukott|blufivend|lambakari|lambapea|
loodrinahk)$

LISA 7. etProficient-v1 konfiguratsioonifail

```
formula: >
  (50 * all(
  is_whole_sentence(),
  length > 4,
  length < 23,
  max([len(w) for w in words]) < 20,
  count_matches(tags, verb) > 0,
  blacklist(words, illegal_chars),
  blacklist(lemma_lcs, bad_words),
  not match(lemmas[0], bad_first_word),
  not match(space_separated(words), bad_first_two),
  not match(tags[0], bad_first_tag),
  match(words[0], lowercase),
  min([word_frequency(w) for w in words]) > 5,
  keyword_repetition(lemmas) == 1
  )
  + 9 * max(0, 1 - sum([0.5 for lemma in lemmas if lemma_frequency(lemma) < 5000]))
  + 9 * optimal_interval(length, 6, 14)
  + 5 * greylist(words, rare_chars, 0.05) * 1.09
  + 7 * greylist(lemposs, anaphors, 0.5)
  + 5 * greylist(lemma_lcs, taboo_words, 0.5)
  + 2 * greylist(tags, abbreviation, 0.5)
  + 2 * greylist(tags, proper_name, 0.1)
  + 2 * (1 - 0.4 * (count_matches(lemmas, 'kroon') and count_matches(tags, 'N')))
  + 2 * max(0, 1 - 0.5 * len([t for t in tokens if t.tag==verb and match(t.features,
verb_nonfinite_suffix)]))
  + 2 * min(1, sum([0.2 for score in lemma_collocation_scores(fromw=-5, tow=5,
minfreq=5, mincnt=3, maxitems=10, colfunc='PROD_MI.LOG_F')[max(0,kw_
start-5):kw_end+5] if score > 0]))
  + 5 * (1 - 0.2 * max(0, count_matches(words, comma) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, pronoun) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, verb) - 2))
  * (1 - 0.2 * max(0, count_matches(tags, conjunction) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, proper_name) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, number) - 1))
  * (1 - 0.2 * max(0, count_matches(tags, adverb) - 1))
  ) / 100

frequency_reference_corpus: estonianRC_filosoft
variables:
  illegal_chars: ([<|\]\[>/\\} {^@•·*#=_~])
  rare_chars: ([A-Z0-9'.,!?)(;:“,„„«»„’×…§-])
  lowercase: ([a-z])
  verb: V
  abbreviation: Y
  proper_name: H
  pronoun: P
```

adverb: D
number: N, O
conjunction: J
comma: (,)
anaphors: ^(see-p|too-p|siin-d|siia-d|siit-d|seal-d|sinna-d|sealt-d)$
verb_nonfinite_suffix: ^(nuks|taks|tuks|tavat|tuvat|nuvat)$
bad_first_tag: (I|Y|Z|J)
bad_first_word:
^(nagu|samamoodi|ühesõnaga|niisugune|niisiis|nad|nemad|see|ka|samuti|ja|samas|aga|
näiteks|kuid|nüüd|seega|et|seetõttu|sellepärast|siis|just|nimelt|ehk|või|ometi|eks|ent|
teiseks|seepärast|ning|esiteks|seal|pigem|siin|ülejäänud|sest|seejärel|ega|jah|muidu|tere|
seejuures|hoolimata|lihtsalt|ikka|järelikult|nõnda|põhjuseks|seevastu|järgnevalt|see-eest|
too|säärane|vastupidi|siit|sellegipoolest|sestap|seepeale|ju|teisisõnu|võrdluseks|iseasi|
sealjuures|tagajärjeks|sellekohaselt|selletõttu|pealegi|seniks)$
bad_first_two: ^((Küll aga)|(Umbes nagu)|(Nii et)|(Nüüd aga)|(Kuid juhtumisi)|
(Samal ajal)|(Ühelt poolt)|(Ainult et)|(Vaatamata sellele)|(Ainult nii)|(Ehk siis)|
(Ehk teisisõnu)|(Eriti kui)|(Eriti juhul)|(Eriti just)|(Eriti siis)|(Eriti veel)|(Isegi siis)|
(Just need)|(Just nii)|(Just see)|(Just seetõttu)|(Muidugi eeldusel)|(Muidugi ka)|
(Nii nagu)|(Peale seda)|(Peale selle)|(Sama asi)|(Sama kehtib)|(Seda enam)|(Seda eriti)|
(Seda kõike)|(Sellisel juhul)|(Sellisel moel)|(Sellisel puhul)|(Teisel juhul)|
(Teisel korral)|(Teiselt poolt)|(Teisest küljest)|(Teisiti öeldes)|(Teiste sõnadega)|
(Välja arvatud)|(Vastasel juhul)|(Vastasel korral)|(Lisaks sellele)|(Veel enam)|
(Veelgi enam)|(Viimasel juhul)|(See omakorda)|(Selleks ajaks)|(Selleks on)|
(Selleks peab)|(Selleks pead)|(Selleks peaks)|(Selleks peame)|(Selleks peavad)|
(Sellele vaatamata)|(Selles mõttes)|(Ses mõttes)|(Selles osas)|(Selles valguses)|
(Sellest hoolimata)|(Sellest johtuvalt)|(Sellest lähtudes)|(Sellest lähtuvalt)|
(Sellest omakorda)|(Sellest tulenevalt)|(See tähendab)|(Samal põhjusel)|(Samal viisil))
bad_words:
^(sitt|sittuma|sittagi|sittama|sitane|sitasti|sitahais|sitapea|sitaauk|sitajunn|sitavares|
sitaratas|sitakott|sitamaitse|sitavedaja|sitahäda|sitahunnik|sitamaja|linnusitt|junn|perse|
perseli|perseklile|persevest|perseauk|pask|paskagi|pasahais|pasakott|pasahunnik|pasapea|
pasapeeter|pasapüks|pasarahe|pasanteeria|vitt|pipravitt|vitupea|tutt|tuss|puts|munn|
munnikari|türa|tra|türapea|kürb|keppima|keppimine|keppija|nikkuma|nikkumine|nikkuja|
tindinikkuja|nikk|koinima|larhv|lipakas|litapoeg|morda|mordu|nuss|nussima|pano|
panomees|praagamagu|hoorama|hoorapoeg|hoorajääger|libu|lirva|lita|lits|kiimakott|pede|
pedekas|pedene|pedendus|pederast|pedetsema|piider|daun|taun|raibe|raip|raisk|
pasandama|pasane|prost|onama|pervo|pirtsperse|sitaks|sitanikerdis|sitaseen|sitavanka|
sitavesi|sitavikat|kirbusitt|litsimaja|litsinahk|lohh|bitch|debiilik|debiil|ajudoonor|
ajugeenius|ajuinvaliid|ambaal|bimbo|imbetsill|jobi|jorss|kärbsesitt|ohmoon|ajukääbik|
ajukäabus|hoor|hoorus|hui|nahhui|nahui|pohhui|pohui|pohh|pohuism|pohhuist|
pohhuistlik|pohhuilt|bljat|nihhuijaa|nehhui|huinjaa|bljät|bljääd|bljäd|bljää|bizdets|pizdets|
plää|hooramaja|idikas|idioot|lausidioot|täisidioot|idioodikari|jota|karuperse|kanaaju|
kanapea|ahvinägu|kusik|homo|frits|kili|rüssä|rüssa|tibla|pajuvenelane|uusvenelane|
moskoviit|murjan|murjam|negru|niger|tšuhnaa|bosniakk|eurovenelane|hoholl|jeestlane|
neeger|vant|vanka|venku|sovett|pilukas|russ|somm|pilusilm|kaltsupea|alfons|baaba|
haisukott|hampelmann|hilpharakas|ilamokk|töllmokk|tõpranägu|inimjätis|inimrämps|
jobukakk|kabajantsik|kaebupunn|kitupunn|kekutis|kirjasolkija|koeraraibe|kopikakoi|
kretiin|krõhva|krõnks|krõõp|krõõt|kõlkapüks|lakard|lakkekrants|lakekrants|lakkekauss|
lakekauss|lapsevänts|lojus|lollikari|lontrus|luhva|lupard|lõhverdis|lödipüks|vedelpüks|

memmepoeg|memmetütar|mammapoja|mammatütar|mats|matsikeel|matslik|matslus|
molkus|molu|mudakoon|munajoodik|mõrd|mölakas|nolk|paksmagu|penskar|plebei|prole|
pulgajunkur|punaparun|rahvarämps|rasvamagu|rasvarull|ristikoer|seapeet|tallalakkuja|
talumats|tattnokk|tolgus|tšinovnik|tuulenuusutaja|tõusik|täikrae|täinahk|töll|töllakas|
untsantsakas|untsakas|vurle|värdjas|väärakas|fuck|fakk|fck)$
  taboo_words:
^(kaabakas|kaak|aare|adrekas|aftekas|mugu|leid|peidukoht|kellegil|kellegile|kah|zhest|
tatt|öigus|öigustus|öigustama|läits|alkašš|atu|autopede|elueit|ess|kirvenägu|kännuämblik|
loll|lollakas|lollpea|lollike|lausloll|shantazheerima|pepu|pee|peer|peeretama|peerukott|
puuks|puuksutama|puuksutamine|kakapuuks|riist|vänt|prostituut|pordumaja|bordell|
porduelu|litsakas|litsitama|litakas|porno|lasteporno|pornograafia|peldik|kaka|kakama|
kakima|kakane|kakanoku|kaki|kakine|piss|pissima|pissine|pissitama|pissipott|kusi|
kusema|kusev|kuselema|kusene|kusija|kusetama|kiim|kiimlema|kiimlus|käsikiimlus|
kiimalus|grupikas|grupiseks|suuseks|anaalseks|sekspomm|küberseks|kähkukas|keelekas|
pedofiilia|pedofiil|pederastia|pederastiapropaganda|pedepropaganda|pervert|gei|lesbi|
lesbiline|narkar|narkomaan|tibi|tibin|tips|beib|beibe|tšikk|piff|tots|tipsi|tiinekas|tiiner|pätt|
mupo|kapo|suli|pagan|joodik|sopajoodik|joobar|jama|kurat|tegelt|nats|point|lurjus|
topakas|tola|jobu|tohman|kesik|rullnokk|memmekas|platnoi|gigolo|mamps|argpüks|
tatikas|soperdis|äbarik|hulkur|kommu|komnoor|kommar|logard|tuhvlialune|näss|jõmm|
oss|friik|närakas|prükkar|kagebiit|tegelinski|totu|pubekas|muti|tõbras|luuser|totakas|
pursui|blatnoi|bomž|komu|diiler|ment|toksikomaan|tõprakari|liputaja|tainapea|taignapea|
varganägu|odratolgus|narkomuul|närukael|tattnina|joomar|skinn|skiso|masuurikas|
nässakas|pleiboi|kagebist|kabajantsik|milf|jurama|siuke|veits|möla|vets|bemm|benss|
bensukas|ruulima|tglt|kuramus|aint|tiss|ila|pilu|fašist|fašistlik|fashist|juhmakas|kekats|
nadikael|dolbajoob|paganama|kuradima|kuram|pandav|dildo|möga|kaif|noku|noks|
liputama|lõust|kärvama|kärakas|kemps|fiiling|narkots|nodi|enivei|looder|peda|hängima|
sunnik|jätis|jura|laiskvorst|rops|mimm|kurivaim|veitsa|bitš|linnavurle|emps|tumba|burks|
pabul|peldik|peller|bemar|värdjalik|kagebeelane|haip|kaltsakas|lojus|kantpea|skinhead|
hullar|ruts|chillima|šoppama|shoppama|tšillima|tsillima|julk|mata|ropsima|natsistlik|tillu|
niuke|moll|haipima|sopakas|ponks|samakas|trulla|eniveis|kurask|jokkis|sakuska|sakumm|
samagonn|amf|ladna|duubeldama|jobukari|tšau|ciao|kabistama|labrakas|kräu|kräkk|
saksmann|sakusment|jobutama|pepuvahe|pärakas|tolbajoob|jobama|munapiiks|oolrait|
jobutus|sigarijunn|jobisema|türahiinlane|töpa|õps|kehka|keska|dire|gümna|klassijuss|
aganoh|türnüffel|sis|drive-in|aa|in|out|inf|ee|no|noh|ete|bla|blabla|blablabla|blablablabla|
pääle|pääl|päält|päälegi|awesome|crash|story|blackout|bläkk|crack|davai|eur|leheneeger|
nannipunn|nilbik|pehmo|transa|pigi|liivaneeger|kirbukott|blufivend|lambakari|lambapea|
loodrinahk)$

LISA 8. Must nimekiri

aa, *aganoh*, *ahvinägu*, *aind*, *ajudoonor*, *ajugeenius*, *ajuinvaliid*, *ajukääbik*, *ajukäämbus*, *alfons*, *ambaal*, *awesome*, *baaba*, *bizdets*, *bitch*, *blackout*, *bljat*, *bljäd*, *bljät*, *bljää*, *bljääd*, *bosniakk*, *crash*, *daun*, *debiil*, *debiilik*, *ee*, *eeslinägu*, *ete*, *eurovenelane*, *fakk*, *fck*, *frits*, *fuck*, *hobusenägu*, *hoholl*, *homo*, *hoor*, *hooraelu*, *hooraja*, *hoorajääger*, *hoorama*, *hooramaja*, *hooramine*, *hoorapoeg*, *hooratöö*, *hoorjääger*, *hoorus*, *hui*, *huinjaa*, *hädaperse*, *idikas*, *idioodikari*, *idioot*, *ilamokk*, *imbetsill*, *inimjätis*, *inimrämps*, *jeestlane*, *jobi*, *jobukakk*, *jorss*, *junn*, *juudasitt*, *kabajantsik*, *kaltsupea*, *kanaaju*, *kanapea*, *kanaperse*, *karuperse*, *kellegiga*, *kellegil*, *kellegile*, *kellegina*, *kellegisse*, *kellegist*, *kellegita*, *keppija*, *keppima*, *keppimine*, *kiimakott*, *kili*, *kirbusitt*, *koinima*, *konnasööja*, *konnaõgija*, *kretiin*, *krõhva*, *kusik*, *kärbsesitt*, *kürb*, *larhv*, *lausidioot*, *linnusitt*, *lipakas*, *lirva*, *lita*, *litapoeg*, *lits*, *litsilööja*, *litsilöömine*, *litsimaja*, *litsinahk*, *lohh*, *millegiga*, *millegil*, *millegile*, *millegina*, *millegisse*, *millegist*, *millegita*, *molu*, *morda*, *mordu*, *moskoviit*, *mudakoon*, *munn*, *munnikari*, *murjam*, *murjan*, *mõrd*, *mõtetu*, *nahhui*, *nahui*, *neeger*, *negru*, *nehhui*, *niger*, *nihhuijaa*, *nikk*, *nikkuja*, *nikkuma*, *nikkumine*, *nuss*, *nussima*, *ohmoon*, *onama*, *oosom*, *oossom*, *pajuvenelane*, *paksmagu*, *pano*, *panomees*, *pasahais*, *pasahunnik*, *pasakott*, *pasandaja*, *pasandama*, *pasandamine*, *pasandus*, *pasane*, *pasanteeria*, *pasapea*, *pasapeeter*, *pasaperse*, *pasapüks*, *pasarahe*, *pask*, *paskagi*, *pede*, *pedekas*, *pedendus*, *pedene*, *pederast*, *pedetsema*, *perse*, *perseauk*, *persekil*, *perseklile*, *perseli*, *persepugeja*, *persepugemine*, *persetama*, *persetamine*, *persevest*, *perssepugeja*, *perssepugemine*, *pervo*, *piider*, *pilukas*, *pilusilm*, *pipravitt*, *pirtsperse*, *pizdets*, *plää*, *pohh*, *pohhui*, *pohhuilt*, *pohui*, *praagamagu*, *libu*, *prost*, *puts*, *püksiperse*, *rahvarämps*, *raibe*, *raip*, *raisk*, *rasvamagu*, *rotinägu*, *rsk*, *russ*, *rätipea*, *rüssa*, *rüssä*, *sis*, *sitaauk*, *sitahais*, *sitahunnik*, *sitahäda*, *sitajunn*, *sitakott*, *sitaks*, *sitakäi*, *sitamaitse*, *sitamaja*, *sitane*, *sitanikerdis*, *sitapea*, *sitaratas*, *sitaseen*, *sitasti*, *sitavanka*, *sitavares*, *sitavedaja*, *sitavesi*, *sitavikat*, *sitt*, *sittagi*, *sittaja*, *sittama*, *sittamine*, *sittuja*, *sittuma*, *sittumine*, *slut*, *somm*, *sorry*, *sovett*, *story*, *tahmanägu*, *taibohh*, *tainanägu*, *taipohh*, *taun*, *tibla*, *tindinikkuja*, *tra*, *tšuhnaa*, *tuss*, *tõpranägu*, *täikrae*, *täinahk*, *täisidioot*, *tänavalibu*, *töllmokk*, *türa*, *türahiinlane*, *türaimeja*, *türapea*, *uusvenelane*, *vanka*, *vant*, *venku*, *vitt*, *vitupea*, *värdjas*.

LISA 9. Hall nimekiri

aare, *adrekas*, *aftekas*, *aint*, *ajuhiiglane*, *alkašš*, *amf*, *anaalseks*, *argpüks*, *atu*, *autopede*, *beib*, *beibe*, *bemar*, *bemm*, *benss*, *bensukas*, *bimbo*, *bitš*, *bla*, *blabla*, *blablabla*, *blablablabla*, *blatnoi*, *blufivend*, *bläkk*, *bomž*, *bordell*, *burks*, *chillima*, *ciao*, *crack*, *davai*, *diiler*, *dildo*, *dire*, *dolbajoob*, *drive-in*, *dzhäss*, *duubeldama*, *elueit*, *emps*, enivei, *eniveis*, *ess*, *eur*, *fashist*, *fašist*, *fašistlik*, *fiiling*, *friik*, *gei*, *gigolo*, *grupikas*, *grupiseks*, *gümna*, *haip*, *haipima*, *hampelmann*, *homopropaganda*, *hulkur*, *hullar*, *hängima*, *ila*, *ilalõug*, *in*, *inf*, *jama*, *jobama*, *jobisema*, *jobu*, *jobukari*, *jobutama*, *jobutus*, *jokkis*, *joobar*, *joodik*, *joomapunker*, *joomar*, *joomaurgas*, *jota*, *juhmakas*, *julk*, *jura*, *jurama*, *jõmm*, *jätis*, *kaabakas*, *kaak*, *kabajantsik*, *kabistama*, *kadakasaks*, *kadakasakslane*, *kaebupunn*, *kagebeelane*, *kagebiit*, *kagebist*, *kah*, *kaif*, *kaka*, *kakama*, *kakane*, *kakanoku*, *kakapuuks*, *kaki*, *kakima*, *kakine*, *kaltsakas*, *kantpea*, *kapo*, *keelekas*, *kehka*, *kekats*, *kekspüks*, *kekutis*, *kemps*, *kesik*, *keska*, *kiim*, *kiimaline*, *kiimalus*, *kiimalus*, *kiimane*, *kiimas*, *kiimlema*, *kiimlus*, *kirbukott*, *kirjasolkija*, *kirvenägu*, *kitupunn*, *klassijuss*, *kommar*, *kommu*, *komnoor*, *komu*, *kopikakoi*, *krõnks*, *krõõp*, *krõõt*, *kräkk*, *kräu*, *kuda*, *kuradi*, *kuradima*, *kuram*, *kuramus*, *kurask*, *kurat*, *kurivaim*, *kuselema*, *kusema*, *kusene*, *kusetama*, *kusev*, *kusi*, *kusija*, *kõlkapüks*, *kõlupea*, *kõverkael*, *kähkukas*, *kännuämblik*, *kärakas*, *kärvama*, *käsikiimlus*, *küberseks*, *laadna*, *labrakas*, *ladna*, *laiskvorst*, *lakard*, *lakekauss*, *lakekrants*, *lakkekauss*, *lakkekrants*, *lambajutt*, *lambakari*, *lambapea*, *lapsevänts*, *lapsprostituut*, *lasteporno*, *lausloll*, *leheneeger*, *leid*, *lesbi*, *lesbiline*, *liivaneeger*, *linnauntsakas*, *linnavurle*, *liputaja*, *liputama*, *litakas*, *litsakas*, *litsitama*, *logard*, *lojus*, *lojus*, *loll*, *lollakas*, *lollikari*, *lollike*, *lollpea*, *lontkõrv*, *lontrus*, *looder*, *loodrinahk*, *luhva*, *lumpen*, *lupard*, *lurjus*, *luuser*, *lõhverdis*, *lõust*, *läits*, *lödipüks*, *maadam*, *madaam*, *mammapoja*, *mammatütar*, *mamps*, *masuurikas*, *mata*, *mats*, *matsikeel*, *matslik,* *matslus*, *meh*, *memmekas*, *memmepoeg*, *memmetütar*, *ment*, *milf*, *mimm*, *molkus*, *moll*, *mugu*, *munajoodik*, *munapiiks*, *mupo*, *muti*, *mäh*, *möga*, *möh*, *möla*, *mölakas*, *mölama*, *mölin*, *mölisema*, *nadikael*, *nannipunn*, *narkar*, *narkomaan*, *narkomuul*, *narkots*, *nats*, *natsistlik*, *neh*, *netu*, *nilbik*, *niuke*, *njaa*, *njah*, *njeetu*, *njetu*, *no*, *nodi*, *noh*, *nojaa*, *nojah*, *noks*, *noku*, *nolk*, *noneh*, *nonoh*, *nonoo*, *noojah*, *närakas*, *närukael*, *näss*, *nässakas*, *odratolgus*, *oki*, *okk*, *oolrait*, *oss*, *out*, *pabul*, *pagan*, *paganama*, *pandav*, *peda*, *pedepropaganda*, *pederastia*, *pederastiapropaganda*, *pedofiil*, *pedofiilia*, *pee*, *peer*, *peeretama*, *peerukott*, *pehmo*, *peidukoht*, *pekkis*, *peldik*, *peller*, *penskar*, *pepu*, *pepuvahe*, *pervert*, *piff*, *pigi*, *pilu*, *piss*, *pissima*, *pissine*, *pissipott*, *pissitama*, *platnoi*, *plebei*, *pleiboi*, *pohhuist*, *pohhuistlik*, *pohuism*, *point*, *ponks*, *porduelu*, *pordumaja*, *porno*, *pornograafia*, *prole*, *prostituut*, *prükkar*, *pubekas*, *pudulojus*, *pulgajunkur*, *punaparun*, *pursui*, *puuks*, *puuksutama*, *puuksutamine*, *pägalik*, *pärakas*, *päss*, *pätt*, *pääl*, *pääle*, *päälegi*, *päält*, *rasvarull*, *riist*, *ristikoer*, *rops*, *ropsima*, *rullnokk*, *ruts*, *ruulima*, *saksmann*, *sakumm*, *sakuska*, *sakusment*, *samagonn*, *samakas*, *seapeet*, *sekspomm*, *shantazheerima*, *shoppama*, *sigarijunn*, *siuke*, *skinhead*, *skinn*, *skiso*, *sopajoodik*, *sopakas*, *soperdis*, *sorri*, *suli*, *sunnik*, *suuseks*, *svensson*, *šoppama*, *zhest*, *tahmanina*, *tahmapea*, *taignapea*, *tainapea*, *tallalakkuja*, *talumats*, *tatikas*, *tatt*, *tattnina*, *tattnokk*, *tavai*, *tegelinski*, *tegelt*, *tglt*, *tibi*, *tibin*, *tiinekas*, *tiiner*, *tiirane*, *tillu*, *tips*, *tipsi*, *tiss*, *tohman*, *toksikomaan*, *tola*, *tolapea*, *tolbajoob*, *tolgus*, *tolmuahv*, *topakas*, *totakas*, *tots*, *totu*, *transa*, *trulla*, *tsillima*, *tšau*, *tšikk*, *tšillima*, *tšinovnik*, *tuhapea*, *tuhvlialune*, *tumba*, *tutt*, *tuulenuusutaja*, *tõbras*, *tõprakari*, *tõusik*, *töll*, *töllakas*, *töpa*, *türnüffel*, *untsakas*, *untsantsakas*, *ussisugu*, *varganägu*, *vedelpüks*, *veits*, *vets*, *vitsa*, *vuhva*, *vurle*, *vänt*, *värdjalik*, *väärakas*, *õps*, *äbarik*, *öigus*, *öigustama*, *öigustus*.

PUBLIKATSIOONID

ELULOOKIRJELDUS

Nimi: Kristina Koppel
Sünniaeg: 02.05.1985
Kodakondsus: Eesti
e-post: kristina.koppel@eki.ee

Haridus:

2015–2020	Tartu Ülikool, eesti ja soome-ugri keeleteadus (üldkeeleteadus), PhD
2007–2009	Tartu Ülikool, eesti ja soome-ugri keeleteadus (eesti keel), MA
2004–2007	Tartu Ülikool, eesti ja soome-ugri keeleteadus (eesti ja soome keel), BA
1994–2004	Kohtla-Järve Järve Gümnaasium
1992–1994	Risti Põhikool

Teenistuskäik:

2016–…	Eesti Keele Instituut, leksikograaf-nooremteadur
2012–2016	Eesti Keele Instituut, leksikograaf
2009–2012	Eesti Keele Instituut, assistent

Teadusorganisatsiooniline tegevus:

2019–...	Eesti Rakenduslingvistika Ühingu juhatuse liige
2017–...	keeleõppe ja rahvahanke ühendamise Euroopa võrgustiku (enetCollect) juhtkomitee liige
2015–...	Eesti Rakenduslingvistika Ühingu liige
2011–...	Emakeele Seltsi liige
2014–2017	Euroopa elektroonilise leksikograafia (ENeL) võrgustiku liige

CURRICULUM VITAE

Name: Kristina Koppel
Date of birth: 02.05.1985
Citizenship: Estonian
E-mail: kristina.koppel@eki.ee

Education:

2015–2020	University of Tartu, Estonian and Finno-Ugric Linguistics (general linguistics), PhD
2007–2009	University of Tartu, Estonian and Finno-Ugric Linguistics (Estonian), MA
2004–2007	University of Tartu, Estonian and Finno-Ugric Linguistics (Estonian and Finnish), BA
1994–2004	Kohtla-Järve Järve High School
1992–1994	Risti Primary School

Professional employment:

2016–…	Institute of the Estonian Language, lexicographer / junior researcher
2012–2016	Institute of the Estonian Language, lexicographer
2009–2012	Institute of the Estonian Language, assistant

Membership in professional organisations:

2019–…	Member of the board of the Estonian Association for Applied Linguistics
2017–...	Member of the management committee of the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect)
2015–...	Member of the Estonian Association for Applied Linguistics
2011–...	Member of the Mother Tongue Society
2014–2017	Member of the European Network of e-Lexicography (ENeL)

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS

1. **Anna Verschik.** Estonian Yiddish and its contacts with coterritorial languages. Tartu, 2000, 196 p.
2. **Silvi Tenjes.** Nonverbal means as regulators in communication: socio-cultural perspectives. Tartu, 2001, 214 p.
3. **Ilona Tragel.** Eesti keele tuumverbid. Tartu, 2003, 196 lk.
4. **Einar Meister.** Promoting Estonian speech technology: from resources to prototypes. Tartu, 2003, 217 p.
5. **Ene Vainik.** Lexical knowledge of emotions: the structure, variability and semantics of the Estonian emotion vocabulary. Tartu, 2004, 166 p.
6. **Heili Orav.** Isiksuseomaduste sõnavara semantika eesti keeles. Tartu, 2006, 175 lk.
7. **Larissa Degel.** Intellektuaalsfäär intellektuaalseid võimeid tähistavate sõnade kasutuse põhjal eesti ja vene keeles. Tartu, 2007, 225 lk.
8. **Meelis Mihkla.** Kõne ajalise struktuuri modelleerimine eestikeelsele tekst-kõne sünteesile. Modelling the temporal stucture of speech for the Estonian text-to-speech synthesis. Tartu, 2007, 176 lk.
9. **Mari Uusküla.** Basic colour terms in Finno-Ugric and Slavonic languages: myths and facts. Tartu, 2008, 207 p.
10. **Petar Kehayov.** An Areal-Typological Perspective to Evidentiality: the Cases of the Balkan and Baltic Linguistic Areas. Tartu, 2008, 201 p.
11. **Ann Veismann.** Eesti keele kaas- ja määrsõnade semantika võimalusi. Tartu, 2009, 145 lk.
12. **Erki Luuk.** The noun/verb and predicate/argument structures. Tartu, 2009, 99 p.
13. **Andriela Rääbis.** Eesti telefonivestluste sissejuhatus: struktuur ja suhtlus-funktsioonid. Tartu, 2009, 196 lk.
14. **Liivi Hollman.** Basic color terms in Estonian Sign Language. Tartu, 2010, 144 p.
15. **Jane Klavan.** Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy. Tartu, 2012, 285 p.
16. **Krista Mihkels.** Keel, keha ja kaardikepp: õpetaja algatatud parandus-sekventside multimodaalne analüüs. Tartu, 2013, 242 lk.
17. **Sirli Parm.** Eesti keele ajasõnade omandamine. Tartu, 2013, 190 lk.
18. **Rene Altrov.** The Creation of the Estonian Emotional Speech Corpus and the Perception of Emotions. Tartu, 2014, 145 p.
19. **Jingyi Gao.** Basic Color Terms in Chinese: Studies after the Evolutionary Theory of Basic Color Terms. Tartu, 2014, 248 p.
20. **Diana Maisla.** Eesti keele mineviku ajavormid vene emakeelega üliõpilaste kasutuses. Tartu, 2014, 149 lk.

21. **Kersten Lehismets.** Suomen kielen väylää ilmaisevien adpositioiden *yli*, *läpi*, *kautta* ja *pitkin* kognitiivista semantiikkaa. Tartu, 2014, 200 lk.
22. **Ingrid Rummo.** A Case Study of the Communicative Abilities of a Subject with Mosaic Patau Syndrome. Tartu, 2015, 270 p.
23. **Liisi Piits.** Sagedamate inimest tähistavate sõnade kollokatsioonid eesti keeles. Tartu, 2015, 164 lk.
24. **Marri Amon.** Initial and final detachments in spoken Estonian: a study in the framework of Information Structuring. Tartu, 2015, 216 p.
25. **Miina Norvik.** Future time reference devices in Livonian in a Finnic context. Tartu, 2015, 228 p.
26. **Reeli Torn-Leesik**. An investigation of voice constructions in Estonian. Tartu, 2015, 240 p.
27. **Siiri Pärkson.** Dialoogist dialoogsüsteemini: partneri algatatud parandused. Tartu, 2016, 314 lk.
28. **Djuddah A. J. Leijen.** Advancing writing research: an investigation of the effects of web-based peer review on second language writing. Tartu, 2016, 172 p.
29. **Piia Taremaa.** Attention meets language: a corpus study on the expression of motion in Estonian. Tartu, 2017, 333 p.
30. **Liina Tammekänd**. Narratological analysis of Võru-Estonian bilingualism. Tartu, 2017, 217 p.
31. **Eva Ingerpuu-Rümmel**. Teachers and learners constructing meaning in the foreign language classrooms: A study of multimodal communication in Estonian and French classes. Tartu, 2018, 218 p.
32. **Kaidi Rätsep.** Colour terms in Turkish, Estonian and Russian: How many basic blue terms are there? Tartu, 2018, 181 p.
33. **Kirsi Laanesoo**. Polüfunktsionaalsed küsilaused eesti argivestluses. Tartu, 2018, 176 lk.
34. **Maria Reile**. Estonian demonstratives in exophoric use: an experimental approach. Tartu, 2019, 240 p.
35. **Helen Türk**. Consonantal quantity systems in Estonian and Inari Saami. Tartu, 2019, 149 p.
36. **Andra Rumm**. Avatud küsimused ja nende vastused eesti suulises argivestluses. Tartu, 2019, 217 lk.
37. **Eleri Aedmaa**. Detecting Compositionality of Estonian Particle Verbs with Statistical and Linguistic Methods. Tartu, 2019, 271 p.